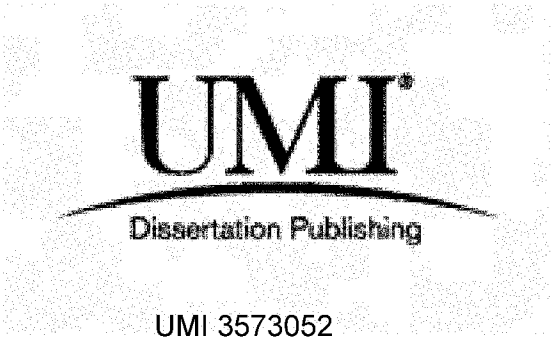
# PREDICTING MAMMALIAN ENHANCERS

# FROM DNA SEQUENCE

by
Dongwon Lee

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
March, 2013







ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

# Abstract

It is widely appreciated that transcriptional regulation plays a central role in most biological processes. However, our current understanding of the molecular mechanisms by which the activity of regulatory elements is modulated in diverse biological contexts still remains incomplete. With the advent of massively parallel sequencing technologies, genome-wide identification of regulatory elements has become increasingly accessible and provides unique opportunities to develop sequence-based models to predict the function and activity of the regulatory elements.

My dissertation directly addresses this issue with the development of a new support vector machine based framework, referred to as kmer-SVM. I demonstrated that kmer-SVM can accurately predict EP300 bound enhancers using only genomic sequence and an unbiased set of general sequence features. I further showed that the kmer-SVM method reveals both enriched and depleted predictive sequence features that are critical for specifying these enhancer activities. Moreover, kmer-SVM can be used to identify novel enhancers, which have been further validated by multiple independent experiments both *in vivo* and *in vitro*.

In the second part of my dissertation, I significantly improved my original kmer-SVM by using as features robustly estimated longer *k*-mer frequencies. With this new method, called gkm-SVM, I showed that cell-specific genomic c-Myc binding can be accurately predicted from local DNA sequence. I further applied the gkm-SVM to predict fine scale structure of the regulatory regions to identify functional transcription factor binding sites

(TFBSs) *within* the regulatory regions at a single base pair resolution. To summarize the fine scale structure predictions, I also developed a systematic approach to build *de novo* position weight matrices (PWMs) from the identified TFBSs, and discovered that they are almost perfectly matched to known PWMs. Moreover, in each case, their cognate factors are known to play a role in the specification of the cell type in which they are identified.

These approaches are applicable to any dataset that can be framed as sequence classification and quickly generates testable hypotheses for further experiments to dissect the underlying molecular mechanisms which control these regulatory elements. I believe that these efforts will significantly contribute to our understanding of transcriptional regulatory networks.

**Advisor:**
Michael A. Beer, Ph.D.

**Thesis Readers:**
Michael A. Beer, Ph.D.
Andrew S. McCallion, Ph.D.

**Thesis Committee:**
Michael A. Beer, Ph.D.
Andrew S. McCallion, Ph.D.
Aravinda Chakravarti, Ph.D.
Joel Bader, Ph.D.

# Acknowledgments

First of all, I would like to express my sincere appreciation to my advisor Mike Beer for all of his support and advice during my graduate studies. His genuine enthusiasm for his research continually inspired me and encouraged me to continue to pursue a research career.

I would also like to acknowledge my thesis committee members, Dr. Aravinda Chakravarti, Dr. Andrew S. McCallion, and Dr. Joel Bader for their comments, suggestions and feedback. In addition, I would like to specifically thank Dr. Rachel Karchin for introducing me to ideas in her FCBBII class that were critical in developing the methods which ultimately became a basis of this work.

This work would not have been possible without my several collaborators. I would first like to thank Dr. Andrew S. McCallion for providing me unique opportunities to work on exciting projects. I specially thank David Gorkin for his suggestions and scientific discussions, and for his enthusiasm in pushing validation of kmer-SVM model, especially testing the impact of SNPs on the SVM score. I also thank Christopher Fletez-Brant for the idea of the kmer-SVM web-server, and for developing the Galaxy modules. I want to specifically acknowledge my lab member Mahmould Ghandi for developing a significantly improved feature generating algorithm by using gapped k-mer distribution.

I thank all of my lab members, Donavan Cheng, Mahmoud Ghandi, Rahul Karnik, Jun Kyu Rhee, and Navneeta Bansal for their help and discussion. I also thank the undergraduate students who worked with me. In particular, I thank Alessandro Asoni for

helping me analyze a large amount of ENCODE ChIP-seq data sets, and John Lee for helping me develop a new ChIP-seq peak boundary refining algorithm. I also want to specially thank my friends, Changhee Lee and Hawon Lee, who supported me and gave me many helpful comments and suggestions on several manuscripts.

Finally, I want to thank my wife, Sanghee, for her unconditional love and support.

# Dedication

This thesis is dedicated to my parents, my parents-in-law, my wife, and my daughter.

# Table of Contents

# List of Figures

# List of Tables

# 1 Introduction

## 1.1 Overview

Uncovering the function of regulatory DNA elements in the genome is an essential step in the development of a more complete understanding of complex biological processes. These elements, by regulating the expression of their associated genes, are widely believed to play a key role in human development and disease. As highlighted by a catalog of published genome-wide association studies (GWAS) by The National Human Genome Research Institute (NHGRI) [1], almost 90% of common human single nucleotide polymorphisms (SNPs), significantly associated with a phenotypic trait or a disease ($p$-value$<10^{-8}$), are located in intergenic or intronic regions. Thus, changes in regulatory elements are implicated in most human diseases and traits of clinical significance. Much effort in current genomic studies is, therefore, devoted to understanding regulatory systems on genome-wide levels, i.e. identifying the networks of regulatory factors and their interactions, and specifying how they modulate their target genes.

Why does a specific genomic locus act as an enhancer in one cell type, but not in another? What mutations in that locus would affect its function, and how? In mathematical terms, building a sequence-based discriminative model that can separate a set of regulatory regions with shared functions from others, such as cell-type specific

enhancers, is a critical step in being able to address these questions. In addition, it would be particularly useful for inferring the underlying molecular mechanisms by which enhancers function. Such models not only provide a means to detect additional regulatory regions that have similar sequence properties and, thereby, presumably similar biological functions, but also give clues about transcription factors (TFs) that may play crucial roles in specifying the regulatory regions by exploiting predictive sequence features. These approaches were successful especially in simple model organisms [2]. However, until recently, development of sequence-based predictive models for mammalian regulatory system has been limited by their extremely large genome sizes, making it difficult to identify even a small set of regulatory regions of a common function.

Now, with the advent of new massively parallel sequencing technologies, experimental identification of regulatory regions directly or indirectly bound by a protein in a specific cell type have become routine tasks in many biomedical investigations. Furthermore, large scale collaborative efforts have generated broad maps of regulatory regions associated with many key TFs, coactivators and epigenomic histone markers in a wide spectrum of human cell-lines [3] and mouse tissues [4]. These publically available large-scale genomic datasets provide us with unique opportunities to systematically study the underlying mechanisms of a wide range of regulatory elements.

In my dissertation, I determined to address this fundamental problem, predicting mammalian enhancers from DNA sequence, by combining state-of-the-art supervised machine learning techniques with systematic integration of diverse collections of large-scale genomic datasets. I first successfully developed a new sequence-based

discriminative method which can accurately discriminate mammalian enhancers from random genomic DNA regions [5]. In collaboration with the McCallion lab, we further experimentally validated our computationally predicted enhancers both *in vivo* and *in vitro* [6]. To make the method more easily accessible to broader community, we also developed and launched a public web server that provides its full functionality [7]. In the second part of my dissertation, I have significantly improved the initial method using more robust set of features [8], [9] in collaboration with my lab colleague. Not only did we increase the overall accuracy of the original method, but also made it possible to predict transcription factor binding sites at single base-pair resolution. Taken together, I believe that these new approaches will ultimately broaden our understanding of mammalian transcriptional regulatory system.

## 1.2 Thesis Organization

To provide background information on transcriptional enhancers in Chapter 2, I start with a brief introduction to the biology of transcriptional regulation by focusing on the function of enhancers in the transcriptional regulation process. Then, I provide a short review of enhancer's roles in human diseases, and introduce current experimental methods for enhancer detection. Since my dissertation mainly discusses the development of new computational methods for enhancer prediction, I also provide a brief review of the previous methods and discuss their limitations.

In Chapter 3, I introduce a new sequence-based discriminative framework, kmer-SVM, and demonstrate how the kmer-SVM methods can accurately predict various types

of EP300 bound enhancers. I provide a thorough analysis and discussion about predictive sequence features from trained kmer-SVMs and their biological relevance. Another main feature of kmer-SVM is the ability to predict additional enhancers from the genome. I discuss how the kmer-SVM method can be used to predict novel enhancers, and provide multiple evidences of the validity of our predictions.

In Chapter 4, I introduce a couple of additional applications of the kmer-SVM method. Since the initial development of kmer-SVM, it has been successfully applied to several other biological problems. The first part of Chapter 4 focuses on the genome-wide prediction of EP300 bound enhancers in melanocytes and experimental validations of the predicted enhancers in collaboration with the McCallion lab. In this study, I introduce an advanced ChIP-seq data processing procedure and discuss how the advanced procedure can improve the classification performance of the kmer-SVM. In an effort to make our kmer-SVM method accessible to broader community, the second application of kmer-SVM is about the development of the public web server that provides the full function of the kmer-SVM method.

The second part of my dissertation discusses the development of a much improved method and its application. In Chapter 5, I introduce a significantly improved feature set and demonstrate how the new features can be incorporated into the original kmer-SVM framework and enhance the classification performance. In Chapter 6, I adopt the new method, referred to as gkm-SVM, developed in Chapter 5 to predict differential genomic binding patterns of the c-Myc TF in diverse cell types. I further develop a new algorithm to summarize the predictive sequence features from trained gkm-SVMs, and demonstrate

that cell-type specific c-Myc binding is essentially determined by the bindings of other TFs near the c-Myc binding sites. Finally, Chapter 7 discusses general issues and future directions.

# 2 Background

## 2.1 Transcriptional Enhancers

Transcriptional enhancers are a key component of the information encoded in the human genome.[1] As a primary step in the regulation of protein quantities, transcriptional regulation significantly contributes to virtually all biological processes, especially those of embryonic development and cell fate determination. In this section, I first introduce a brief history of mammalian transcription studies followed by discussion of the role of enhancers as a key element in the transcriptional regulation process (2.1.1). I then review previous efforts to identify disease-causing mutations, and introduce some representative examples of human disorders caused by mutations in enhancers (2.1.2). Current experimental methods to identify enhancers are then discussed in detail (2.1.3).

### 2.1.1 Mammalian Transcription and Enhancers

Since the pioneering work of Jacob and Monod [11], it was recognized that the rate of transcription was a regulated process which had the potential to impact all downstream processes. Early work on the lysis-lysogeny switch in phage lambda [12] identified DNA binding transcription factors as the key molecular mediators of this process. Subsequent

[1] An earlier version of Chapter 2 was published in the book, titled "*Genome Analysis: Current Procedures and Applications.* [10]"

studies on the lac-operon in *E. coli* and additional pathways in yeast showed how transcription factors could interact with small molecules or surface signaling receptors to modulate the expression of sets of genes in order to respond to external signals and conditions, by activating or repressing immediately downstream genes. These same basic molecular events are essential steps in the processing of information regarding the extracellular or developmental state of multicellular organisms via cell-cell or diffuse morphogen interactions with cell-surface receptors in embryonic development.

Early studies in mammalian systems uncovered a significantly different regulatory element structure. Mammalian transcription typically involves sets of regulatory elements, which act at a distance, in contrast to the local regulation in single cellular organisms. These elements can broadly be classified as promoters, enhancers, and insulators. Promoters are regulatory elements immediately upstream from transcriptional start sites and are locations of RNA polymerase II complex initiation. Enhancers are distal regulatory elements that can enhance transcriptional activities of specific genes roughly independent of their orientation and distance from the site of action [13]. Insulators block interactions between distal enhancers and promoters.

The scale of these DNA regulatory elements are on the order of several hundred base pairs (up to a few kilobases) of DNA sequence, comparable to the intergenic space in the prokaryotic or yeast genomes. In contrast, however, these elements can act over a much longer intervening distance in linear sequence; many examples of enhancer-promoter interactions are known to act over several hundred kb, and up to a megabase, via DNA looping [14], [15]. While insulators block promoter-enhancer interactions in some cases

and contribute to the establishment of specific promoter-enhancer connections, the general mechanisms which determine specific promoter-enhancer association are not yet clear.

It is likely that this modular structure of mammalian transcriptional regulation and the larger mammalian and vertebrate genome size have evolved concomitantly. It has been convincingly argued that intronic recombination can facilitate modular protein evolution by allowing exon shuffling [16], and this process was associated with a burst of evolutionary diversity at the time of metazoan radiation. Similarly, the dissociation of enhancers and promoters may have allowed more modular regulatory evolution. Typically, a developmentally regulated gene will have several distal enhancers that can interact with the gene's promoter in different cell types or at different times during development. The increased genome size at a roughly fixed gene number greatly increases the available regulatory space relative to single-celled organisms, and would have allowed this modular structure to evolve more easily than in a very compact genome. Intergenic recombination events that swap enhancers to new locations can allow the rapid evolution of novel combinations of enhancers and regulatory circuits. It seems likely that this was crucial in the evolution of the increased cell diversity and developmental regulatory complexity of multicellularity.

Since the discovery of the interferon-beta enhancer [13], much progress has been made toward elucidating the fundamental molecular mechanisms by which enhancers regulate the transcription of their target genes. The transcriptional consequences of enhancers are mediated through direct binding of sequence specific transcription factors

(TFs) and their interactions with complexes, which modulate chromatin accessibility. It is now fairly well established that all of three main classes of regulatory elements (promoters, enhancers, and insulators) are comprised of binding sites of several cooperating TFs. While all of these elements share some common features, recent studies indicate that a key difference between enhancers and promoters and insulators is that while most enhancers are activated in a temporal or cell-type specific manner, promoters and insulators tend to have a broader function, i.e., they are active across most cell types in human and mouse [17], [18].

### 2.1.2 Enhancers and Human Disease

Most human traits have a familial hereditary component, but demonstrate complex patterns of inheritance. Genome wide association studies (GWAS) have been widely used to identify complex trait loci, and now more than 7,000 SNPs are known to be associated with variation in more than 700 traits and diseases [1]. Although GWAS can still only explain a small fraction (5-10%) of the phenotypic variance [19], the list of regulatory mutations responsible for heritable susceptibility to specific disease is growing at a steady rate. The significant role of regulatory variation in complex trait heritability is underscored by the finding that the majority (>80%) of the trait associated SNPs are non-exonic [20]. When a regulatory variant is identified, it is often hypothesized that that the variant disrupts a TF binding site, or creates a new binding site, or both.

Although it can be extremely difficult to identify causal regulatory variants associated with human disorders and their underlying molecular mechanisms, there are some well-

studied cases. The classic example of how disruption of a regulatory element can directly cause human disease is in the β-globin locus, where removal of the locus control region can disrupt the high expression levels required in erythroid cells and lead to thalassemia [21], [22]. The first vertebrate insulator to be systematically characterized was also in the β–globin locus [23]. Preaxial polydactyly (PPD), a limb malformation, is another dramatic example. PPD in humans was genetically mapped to a 450 kb chromosomal locus on 7q36, but it was not until a fortuitous mutant mouse was generated by a transgene insertion that these mutations were associated with misexpression of *Sonic hedgehog* (*Shh*), approximately 1 megabase (Mb) away from the gene [24]. This transgene insertion disrupted a limb distal enhancer of *Shh,* in which several different single-nucleotide variations, segregating with the PPD phenotype in four unrelated human families, have been identified [25].

Other important examples include cases in which disruption of a PAX6 enhancer are associated with aniridia [26]; deletion of sequences approximately 1 Mb upstream of *POU3F4* causing X-linked deafness [27]; and a common SNP in RET intron 1 associated with a high risk of Hirschsprung's disease [28], [29]. In a more recent example, two SNPs associated with prostate cancer are both found in an enhancer which regulates the 1 Mb distant *SOX9* gene, and these variants modulate the enhancer's activity by strengthening or weakening binding sites for FoxA1, AP1, and AR within the enhancer [30].

Since the early observation of the extreme similarity of proteins across different vertebrate species, both in sequence and biochemical activity, it has been postulated that much evolutionary diversity is generated by changes in gene regulation mediated by

mutation of regulatory elements [31], [32]. It is thus reasonable to expect that human evolution is also operating largely at the level of regulatory variation, and that much of the existing variation in the human population is contributing to variable susceptibility to common disease.

### 2.1.3 Experimental Methods for Genome-wide Enhancer Detection

Early experimental exploration of the genomic regulatory landscape in *Drosophila* used a method known as enhancer trapping, employing a mobile GAL4 gene inserted randomly into the genome, driving GAL4 expression from flanking genomic enhancers [33]. When crossed with lines carrying a GAL4 responsive reporter gene, the patterns of the flanking enhancer's activity can be observed. The genome sequencing projects ushered in a new generation of systematic approaches which aim to map all genomic regulatory elements. The initial sequencing of the human genome revealed that the gene number was surprisingly low: 20,000-25,000 genes, comparable to other model organisms (*C. elegans*, *Drosophila*). This was an initial indication that the differences between humans and these model organisms is not due to a dramatic increase in the repertoire of tissue specific genes, but instead results from a dramatic evolution in the structure and repertoire of regulatory elements. Subsequent sequencing of the mouse genome [34] showed that roughly 5% of the human genome was under selection, and since only 1.5% of the human genome was accounted for by protein coding genes, it was strongly suggested that at least 3.5% of the genome encoded regulatory function.

Genome sequencing allowed the construction of the first generation of microarrays,

which used hybridization to detect DNA bound by key components of regulatory complexes. This method, known as ChIP-chip, was used successfully to map TF-DNA interactions of hundreds of TFs in yeast [35], and similar techniques (ChIP-PET) were applied to detect genome wide binding of MYC in human cells [36].

Soon after the Human Genome Project was completed, the ENCODE pilot project was then initiated with the ambitious goal of cataloguing all possible functional elements in the human genome [37]. Although only 1% of the human genome was initially evaluated with various techniques, several biological insights were already gained from this pilot project. One striking finding was that a surprisingly large fraction of potential regulatory DNA elements did not appear to be evolutionarily constrained (as assessed by sequence alignment). These initial findings encouraged a more complete approach applying similar analysis to the whole genome [3].

At about the same time, the parallel development of next-generation sequencing technologies and the accompanying dramatic drop in the cost of deep sequencing enabled completely different approaches to the identification of regulatory elements. For example, chromatin immunoprecipitation followed by sequencing (ChIP-seq) using antibodies specific to the protein of interest is now a routine process when genomic occupancies of a certain DNA interacting factor are in question. Specifically, enhancer specific protein markers, such as EP300/CREBBP coactivators and covalent modifications of histone proteins (H3K4me1 and H3K27ac) have further facilitated the identifications of genome-wide enhancers in mammalian genomes [17], [38], [39]. DNase-seq is another example of using next-generation sequencing technologies to detect regulatory elements [40], [41].

The accessibility of the genome to DNA binding factors is not invariant, but rather extremely variable under different conditions, and it has long been known that highly accessible DNA is associated with various kinds of regulatory elements. To experimentally detect such regions, biologists have taken advantage of the fact that deoxyribonuclease I (DNase I), an endonuclease, exhibits varying cleavage efficiency depending on the DNA accessibility. Now, combined with new sequencing technology, DNase-seq has become a standard technique to find genome-wide open chromatin regions.

The initial successful ENCODE pilot project has now been significantly expanded to the main human and mouse ENCODE projects fully equipped with new sequencing technologies. The ENCODE project [3] has produced maps of chromatin accessibility via DNaseI hypersensitivity [42], genomic binding of many key TFs via ChIP-seq [43], and specific chromatin marks in a broad array of human cell lines [44] and mouse tissues [4]. Together, these maps have identified a large set of previously undocumented regulatory regions and either directly or indirectly reflect cell-specific TF occupancy. Each of these experimental methods is subject to its own limitations and they are under continuing development. In particular, the set of genomic regions defined as being positive in any of these assays will depend sensitively on the signal threshold chosen. However, as I will demonstrate, these data provide a rich substrate which can be used to develop a predictive computational model of enhancer activity.

## 2.2 Computational Prediction of Enhancers

While consensus on the general mechanisms is rapidly emerging, we do not yet understand how enhancers work at the level of a predictive model of regulatory element activity, which can specify the set of cell types and environmental conditions under which the enhancer would stimulate the expression of its target gene(s). Further, a predictive enhancer model should describe how specific mutations to that enhancer sequence would affect its activity. My philosophy here is akin to protein coding gene prediction: although direct experimental validation of each individual gene's transcription is ultimately required to verify the predictions, current gene prediction algorithms which integrate incomplete and perhaps noisy experimental data (e.g. ESTs) and genomic sequence features (e.g. ORFs and models of splice donors and acceptors) are able to provide a highly accurate picture of the set of protein coding genes in many organisms [45]. In the case of enhancer prediction, we know that the key features are transcription factor binding sites, but we have incomplete knowledge of their binding specificities and function, especially which TF binding events are able to modulate chromatin structure and vice versa.

In this section, I first review early computational approaches to enhancer prediction (2.1), I then introduce methods which use primary DNA sequence and frame the problem as a discriminative classification problem (2.2), and present a recent successful SVM-based discriminative approach (2.3).

### 2.2.1 Early Approaches Based on TF binding sites clustering and conservation

Attempts to predict regulatory elements from DNA sequence in mammalian genomes still face major challenges in computational biology. As we have gained more knowledge about regulatory elements, various strategies to computationally identify regulatory regions have been developed. Until recently, however, none of the previous methods have shown success rates in predicting mammalian enhancers that would encourage their use as a general tool in biological or medical investigation, as assessed in a benchmark study [46].

Several early approaches took advantage of the observation that TFBSs tend to cluster together within relatively short DNA stretches ranging from several hundreds to thousands base pairs. These approaches showed some success, especially in the *Drosophila* genome [47]–[50]; for review, see [51], but application to mammalian genomes has been much less promising. These methods essentially identify regions that harbor TFBSs more than expected by chance within a given window of DNA sequence by using relatively simple counting methods or more sophisticated probabilistic methods such as Hidden Markov Models. However, this strategy always relies on prior knowledge about TF binding specificities, which is still thought to be far from complete. Also, some of these methods only identify regions where TFs are densely clustered without regard to the identity of the TFs in the combinations, a biologically implausible assumption that might lead to large number of false positives in their predictions.

There are other strategies that utilize sequence conservation information in

combination with aforementioned methods. Since regulatory function is under evolutionary constraint, it is a widely accepted idea that significant fraction of conserved non-coding DNA is likely to function as regulatory elements [34], although the converse is not necessarily true [37], [52]–[54]. Sequence conservation information can be used to detect putative regulatory elements under purifying selection by comparing different species, as well as to detect individual TFBSs within regulatory elements, a technique known as phylogenetic footprinting. Several methods have been developed based on this idea, mostly focusing on the *Drosophila* genome [55]–[57], and some for mammalian genomes [58]–[60]. However, subsequent experiments are always required since sequence conservation gives essentially no information about the element's specific biological function. Moreover, these validation experiments are typically labor intensive and time consuming. One notable study set out to systematically assay conserved non-coding regions in the human genome *in vivo* using a LacZ reporter system in transgenic mice to discover developmental tissue-specific enhancers. Among over 2,000 regions tested so far, they discovered that at least 40~50% can act as tissue specific enhancers at a single developmental time in the early mouse embryo [61].

Unfortunately, all the efforts discussed so far have achieved only limited success and systematic computational approaches fell far short of desired predictive accuracy [46], especially in mammalian genomes. These results strongly suggest that current knowledge about TF binding specificities and overall sequence conservation information is not sufficient to describe the function of regulatory elements from primary DNA sequence.

### 2.2.2 Early Sequence Based Discriminative Approaches

The demonstrated limitations of sequence conservation and current knowledge about TF binding specificities as predictors of regulatory function have led many computational biologists to develop more sophisticated approaches. Models which integrate limited experimental evidence of chromatin state or cofactor binding and sequence features enriched in these regions can provide a more accurate description of the genomic enhancer landscape than either approach used in isolation. Such sequence-based models can be framed as classifiers which can discriminate between regulatory elements and non-regulatory DNA after training on a suitable but incomplete set of sequence regions with the function of interest.

One of the first successful studies using a sequence based discriminative method, also known as "Regulatory Potential", showed that simple Markov models can distinguish regulatory regions from non-regulatory regions with reasonable accuracy [62]–[64]. In this approach, two 2nd-order Markov models separately trained on a set of aligned known regulatory regions and a set of neutral DNA regions (ancestral repeats) were used to calculate the log-odds ratio of a given DNA sequence. This remarkably simple method demonstrated for the first time that regulatory elements in mammalian genomes can be predicted from primary DNA sequence without prior knowledge of TF binding specificities, although the overall accuracy was not high enough to be useful for genome-wide prediction. More recently, several studies have achieved notable successes in predicting different classes of regulatory elements in mammalian genome using various techniques: transcription start site prediction using SVMs [65]; promoter prediction using

logistic regression [66]; enhancer prediction using LASSO regression [67]; and enhancer prediction using SVMs [5], [6]. These recent successes are mostly due to (1) state-of-the-art supervised machine learning algorithms and (2) appropriate experimental data sets for model training. Especially for enhancer prediction, one of the most accurate methods to date is kmer-SVM [5]. Since the development of the original kmer-SVM, this model has been applied to other problems, and some of computationally predicted regions have been experimentally validated both *in vitro* and *in vivo* [6]. In the next section, we will briefly discuss the general properties of SVM methods.

### 2.2.3 Support Vector Machines

Since the development of support vector machines (SVMs) in the early 1990s [68], [69], the SVM has become one of the most popular machine learning techniques and has been successfully applied to almost every problem in computational biology, for reviews, see [70] and [71]. A SVM is a general binary classifier that learns a decision boundary, called a hyperplane, by maximizing margins between the two sets in the feature vector space, formalized as follows. Suppose you have $N$ number of $n$-dimensional real valued vectors $\mathbf{x}_i \in R^n$ with associated class labels $y_i \in \{+1, -1\}$ for $i = 1, \ldots, N$. Then, a hyperplane is found by minimizing $\|\mathbf{w}\|^2$ such that $y_i(\mathbf{x}_i \cdot \mathbf{w} + b) \geq 1$ for any $i = 1, \ldots, N$. In practice, however, the optimal solution is obtained by maximizing the Wolfe's dual form:

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j),$$

subject to $\alpha_i \geq 0$ for any $i = 1, \ldots, N$, and $\sum_{i=1}^{N} \alpha_i y_i = 0$. In the dual form, the inner product $(\mathbf{x}_i \cdot \mathbf{x}_j)$ can be considered as a measure of the similarity between two data points $i$ and $j$ in the feature space. Moreover, since this is the only term that has feature vectors in the object function, it can be replaced by a more general function, called a kernel function $K(\mathbf{x}_i \cdot \mathbf{x}_j)$. This generalization makes SVMs very powerful methods because it relaxes the requirement of an explicit feature space as long as a kernel function between any two data points is defined. A very simple yet powerful measure of sequence similarity is the $k$-spectrum kernel [72], which calculates the inner product of frequencies of all possible $k$-mers of two sequences. This kernel was first introduced to classify functional domains from protein sequence, and have been successfully used in several different contexts such as nucleosome positioning [73], as well as enhancer predictions [5].

# 3 Kmer-SVM: A New Discriminative Framework for Enhancer Prediction

In this chapter, I describe a new support vector machine (SVM) based framework, referred to as kmer-SVM, in great detail.[2] Kmer-SVM is originally developed to predict EP300-bound enhancers using only genomic sequence and an unbiased set of general sequence features. Here, I show that my kmer-SVM method can accurately predict diverse sets of EP300-bound enhancers. I also demonstrate that the predictive sequence features identified by the kmer-SVM reveal biologically relevant sequence elements enriched in the enhancers, but I 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. I finally show that kmer-SVM can be used to predict novel enhancers with reasonable precision, which provides a high confidence list of enhancer targets for subsequent experimental investigation.

## 3.1 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 [13].

[2] An earlier version of Chapter 3 was published in the journal, *Genome Research* [5].

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 [74]. Factors bound to these sequences often interact with common coactivators, which in turn recruit the basal transcription machinery [14], [75].

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 [76]).

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 [61], [77], [78]. 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 [37], [52], [53].

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 [79], [80], chromatin signatures [17], [38], or coactivators [39], [81]. 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 [17], [38]. The transcriptional coactivators EP300 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 [82]. Several studies have reported genome-wide mapping of EP300/CREBBP bound enhancers in different contexts, for example, tissue-specific activity in dissected

mouse tissue [39], and environment dependent activity in neurons [81]. 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 [83] or JASPAR [84], or a combination of the two. Many of these existing approaches were assessed by Su *et al.* [46], 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 [64]. In more recent work, Leung and Eisen [85] 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 [67].

Here I present a discriminative computational framework, referred to as kmer-SVM, to detect enhancers from DNA sequence alone that does not rely on conservation or known TF binding specificities. I use a support vector machine (SVM) [68], [69] to differentiate

enhancers from non-functional regions, using DNA sequence elements as features. SVMs have been successfully applied in many biological contexts (reviewed in [70] and [71]): cancer tissue classification [86]; protein domain classification [72], [87], [88]; splice site prediction [89], [90]; and nucleosome positioning [73]. In my case, because of the potentially diverse mechanisms which direct EP300 and CREBBP binding, I 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, I investigate EP300/CREBBP binding in mouse embryos [39], activated cultured neurons [81], and embryonic stem (ES) cells [91]. My 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. I evaluate the kmer-SVM 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.

I show that sequence features in the experimentally identified enhancer set are sufficient to accurately discriminate enhancers from random genomic regions. I also show that the most predictive sequence elements are related to biologically relevant transcription factor binding sites. Notably, my kmer-SVM method also finds that some sequence elements are significantly *absent* in the enhancers (those with large negative SVM weights). For example, I 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 [92]. In addition, I 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.

I further apply my kmer-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 my approach.

## 3.2 Methods

### 3.2.1 Positive Data Sets

As positive data sets, I initially used the genome-wide *in vivo* EP300 binding sites identified by ChIP-seq [39], 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 I directly use the entire sequences without modification. I also analyzed two other data sets [81], [91]. Chen *et al.* reported 524 EP300 binding sites in mouse embryonic stem cells, and Kim *et al.* reported about 12,000 neural activity dependent CREBBP binding sites in stimulated cultured mouse cortical

neurons. Since both CREBBP data sets report only summits of the ChIP-seq signals, 100 bp or 400 bp in both directions from these summits were extended to obtain sequences for further analysis.

### 3.2.2 Generating Negative Data Sets

I generated negative sequence sets to match the distribution of sequence length and repeat element fraction of the corresponding positive sets (Figure 3-1). Repeat fractions were calculated using the repeat masked sequence data from the UCSC genome browser [93]. I 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, $\alpha$ 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.

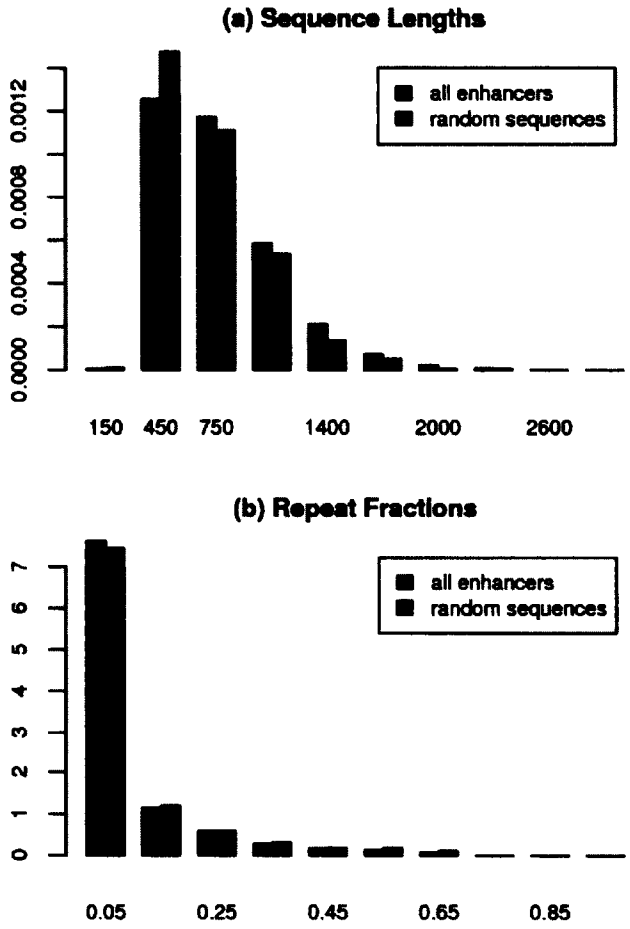



**Figure 3-1: Comparison of the sequence properties between enhancers and random genomic sequences**
For the null sequence model, I selected random sequences from the genome to match the repeat fraction and length distribution of the sequences in the positive data set. The combined set of all Visel's EP300 bound regions is shown in red, and my null sequence set is shown in blue.

### 3.2.3 Sequence Features

I use a standard SVM method [68], [69] with a full set of *k*-mers as features, also known as the *k*-spectrum kernel [72]. I have found that this kernel produces one of the best results, is easy to interpret, and can easily represent a combination of TF binding

sites. To implement the *k*-spectrum kernel, I generate a *k*-mer count vector for the full set of distinct *k*-mers, for each sequence. Then I normalize the count vector so that $||\mathbf{x}||=1$, to reduce the effect of the variable length of different enhancers, which is referred to 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 original spectrum kernel function is 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.

### 3.2.4 SVM Function and Feature Selection

After a SVM training, I construct the SVM weight vector $\mathbf{w}$ from the $\alpha_i$ of support vectors as follows.

$$\mathbf{w} = \sum_{i=1}^{N} y_i \alpha_i \mathbf{x}_i$$

The SVM function, or "SVM score", can also be defined by $\alpha_i$ , which represents the distance of any vector $\mathbf{x}$ from the decision boundary, and determines the predicted label of the vector $\mathbf{x}$.

$$f_{\text{SVM}}(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b = \sum_{i=1}^{N} y_i \alpha_i (\mathbf{x}_i \cdot \mathbf{x}) + b$$

To select predictive features, I sort the *k*-mers by the SVM weights, and choose a desired number of *k*-mers from the top and the bottom of the sorted list.

### 3.2.5 Other Kernel Methods

In addition to *k*-spectrum kernel, I tried various kernel functions to compare the performances between different methods. As a non-linear kernel function, I tested the Gaussian kernel, which uses the same feature vectors as the *k*-spectrum kernel, but uses a different similarity measure via the following kernel function.

$$K_{\text{Gaussian}}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2})$$

I also tested more sophisticated string kernel methods such as (*k*,*m*)-mismatch kernel [88] and KIRMES [94]. 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, reflecting the fact that some TFs bind degenerate sites. My implementation utilizes the Shogun machine learning toolbox [95], [96] and SVM light [97].

### 3.2.6 Naïve Bayes Classifier

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

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

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

### 3.2.7 PhastCons Score

To measure conservation, I use PhastCons [98], based on a two-state phylogenetic hidden Markov model (phylo-HMM). PhastCons outputs the probability that each aligned column was generated by the "conserved" state given the model parameters and the multiple alignments. I used this PhastCons conservation score for alignment of 29 vertebrate genomes available at the UCSC Genome Browser [93] to assess how well the predictive sequence elements in the enhancers are evolutionarily conserved. I calculated the average PhastCons score over all bases of each *k*-mer in each sequence, and obtained one score for each *k*-mer ranging from 0 to 1, reflecting overall conservation.

## 3.3 Results

### 3.3.1 Enhancers Can Be Accurately Predicted From DNA Sequence

My primary concern in this study 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. I initially focus on recent genome-wide experiments that identified EP300 binding sites by

ChIP-seq [39] 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 generally unknown. While EP300 ChIP may not detect all the enhancers active under these conditions, I initially analyze this dataset to identify sequence features responsible for EP300 binding in these tissues.

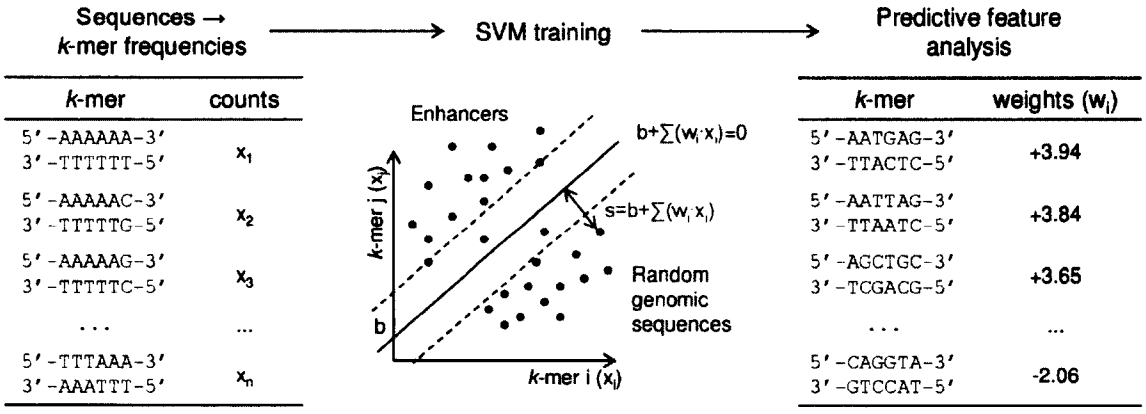



**Figure 3-2: Overview of my methodology**
*k*-mer frequencies are calculated for each of the EP300-bound and negative genomic training sequences. These feature vectors ($x_1,\ldots,x_n$) are used to find SVM weights, **w**, which describe a hyperplane that separates the positive and negative training sets.

To model DNA sequence features, I use a support vector machine (SVM) framework. In brief, a 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 3-2 and more 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. I 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 [99], because the binding specificity of many transcription factors (TFs) has yet to be determined, I prefer *k*-mers (oligomers of length *k*) because they are an unbiased, general, and complete set of sequence features. To evaluate classification performance, I use a standard five-fold cross validation method. Briefly, 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), I measure the quality of the classifier by calculating the area under the ROC curve (auROC). I 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 the following ROC figures.

To further test sensitivity to various assumptions in SVM construction, I repeated these cross-validation experiments on Visel's each tissue-specific enhancer set using SVM classifiers with different types of kernels: Spectrum kernels [72], Mismatch spectrum kernels [88] and Gaussian kernels as shown in Figure 3-3. 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, I 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.

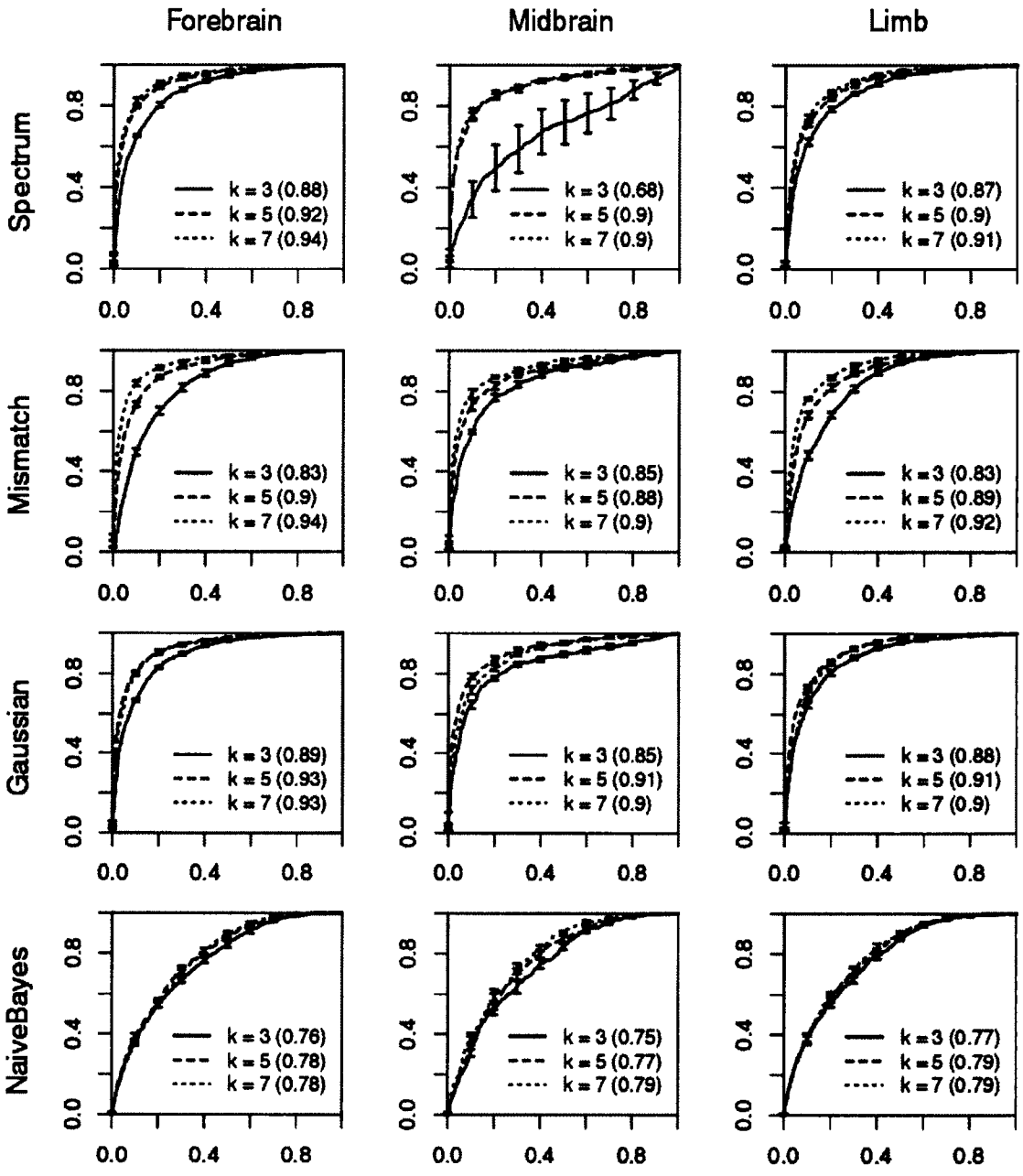



**Figure 3-3: Classification results with different k-mers and methods**
Using the full set of k-mers, SVM Classification results on Visel's data sets with three different kernels (Spectrum, Mismatch, and Gaussian) and Naïve Bayes classification results are shown. SVMs outperform Naïve Bayes classifiers in every case but one which failed to converge (SVM with 3-spectrum kernel on Midbrain). Three different lengths of k-mers, k=3, 5, 7, are tested. Generally, larger k exhibits better performance in terms of auROCs with some exceptions caused by over-fitting.

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 3-3 and Figure 3-4). In general, larger *k*-mers achieved superior performance (Figure 3-3), 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 3-4A, 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, I adopt the 6-mer spectrum kernel as my standard model for the remainder of the study.

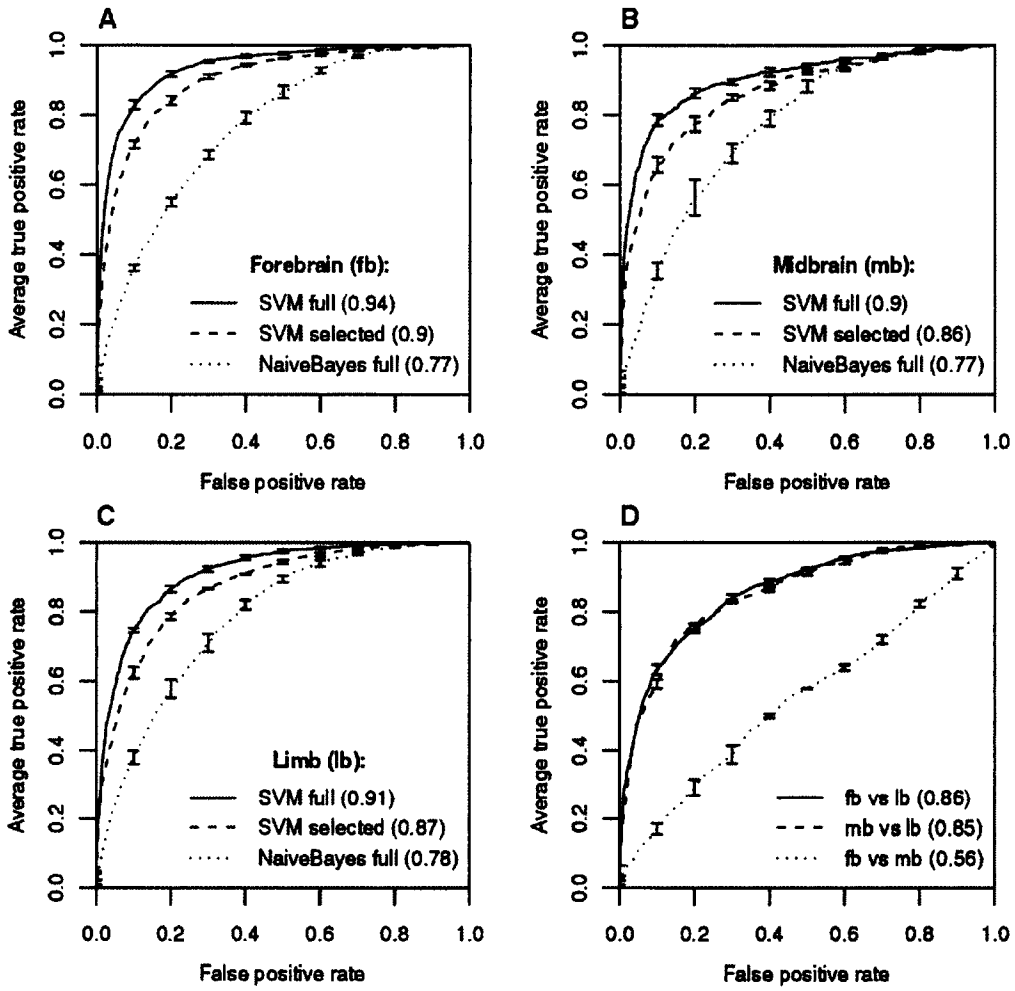



**Figure 3-4: Classification results on each tissue-specific enhancer set**
(A) Classification of forebrain enhancers vs. random genomic sequences. (B) Classification of midbrain enhancers vs. random genomic sequences. (C) Classification of limb enhancers vs. random genomic sequences. Each graph in A, 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 Naive Bayes classifier (dotted). Each curve is an average of five cross-fold validations; error bars denote one standard deviation over the five cross-fold validation sets. Numbers in parentheses indicate the area under each ROC curve (auROC). Both the full SVM and SVM with selected features perform very well and significantly better than Naive Bayes. Individually, each tissue-specific set can be accurately discriminated from nonenhancer 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.

Besides distinguishing individual enhancer sets from random genomic sequences, I next tested whether my kmer-SVM method could also distinguish between enhancers in different tissues (forebrain, midbrain, and 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 3-4D). 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, I have freedom to choose the size of the negative sequence set. The genomic ratio of enhancers to non-enhancer sequence is very large (I estimate that enhancers comprise 1-2% of the genome in a given cell-type), and ideally I would compare alternative prediction methods using a very large negative set. However, some of the computational methods I compared could not handle such large amounts of sequence due to memory constraints. To compare between datasets, I used the same ratio between positives and negatives. To test the scaling with negative set size, I 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, but auPRC drops as $n$ becomes large and the high scoring tail of the negative sequences becomes competitive with the true positive sequences (Figure 3-5). However, the trends of auROC and auPRC are usually consistent.

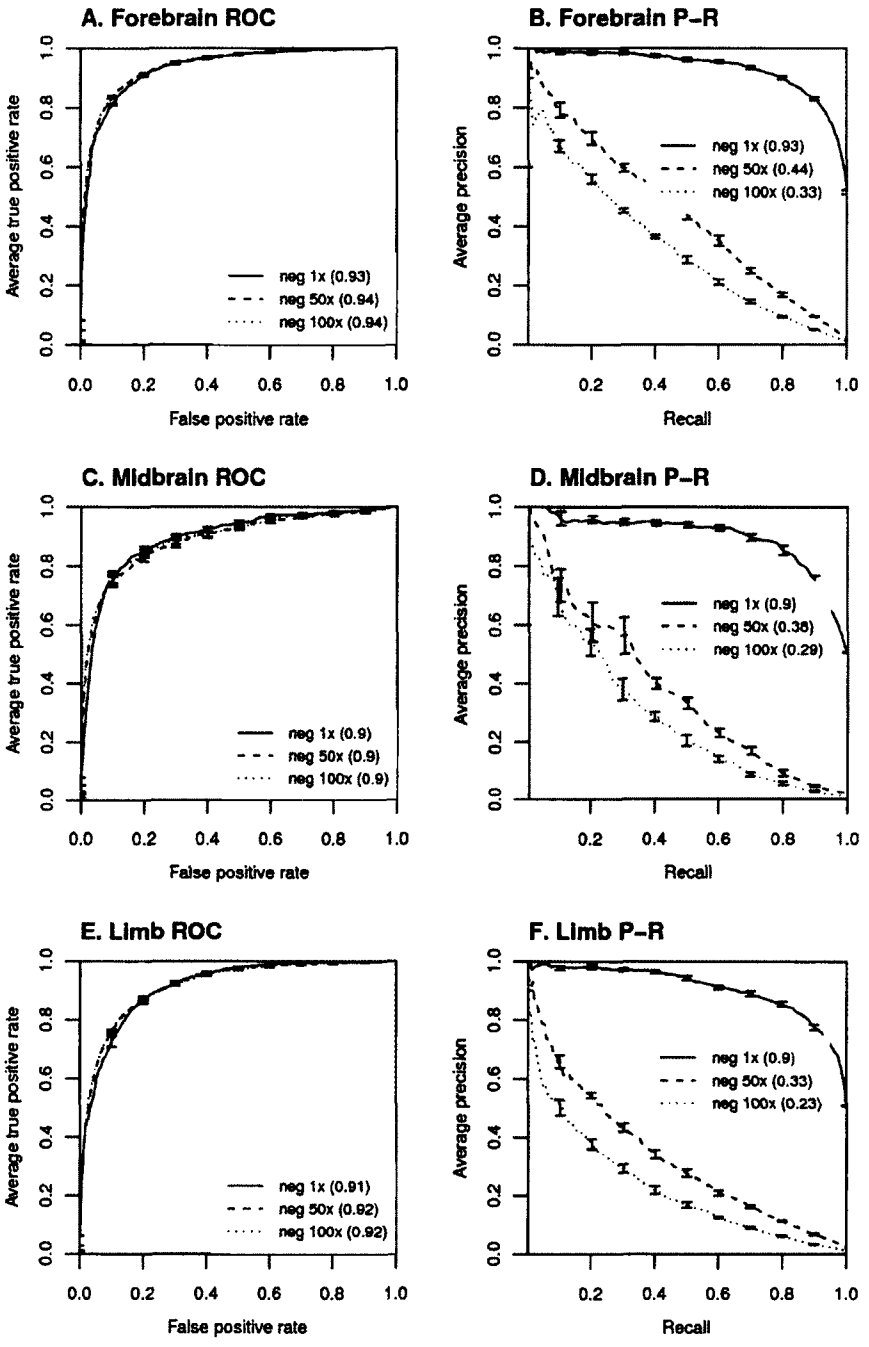



**Figure 3-5: Comparison between ROC curves and Precision-Recall curves with larger negative sets**

ROC curves and Precision-recall curves for tissue-specific enhancers vs. random genomic sequences with larger negative set sizes are shown. Since the genomic ratio of enhancers to non-enhancer sequence is very large, I tested three negative set sizes (1x; 50x larger; and 100x larger than the positive enhancer set) for each case. For large negative set size, auPRC is a more reliable measure of performance than the auROC curve, which is independent of negative set size, as expected.

### 3.3.2 Most Predictive Sequence Elements Are Known TFBSs

I next investigated which subsets of sequence features that allowed the SVM to successfully discriminate enhancers from random sequence. The SVM 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 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.

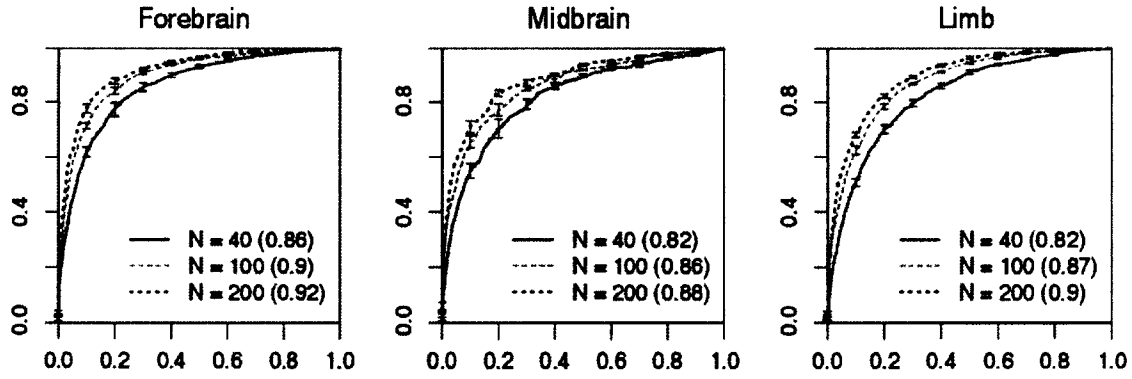



**Figure 3-6: SVM classifications with selected 6-mers**
Using only selected 6-mers, results of SVMs with spectrum kernels are presented. For each classification, a half of N 6-mers with the largest positive SVM weights and a half of N 6-mers with the largest negative SVM weights were selected (N=40, 100 and 200)

I conducted the SVM classification again, using only the subset of *k*-mers with largest positive and negative SVM weights (Figure 3-6). 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 3-4A, B, and C). Interestingly, the most frequently observed *k*-mers do not always have the largest SVM weights or vice versa. I find only a weak correlation between SVM weights and *k*-mer frequencies (Figure 3-7). 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.

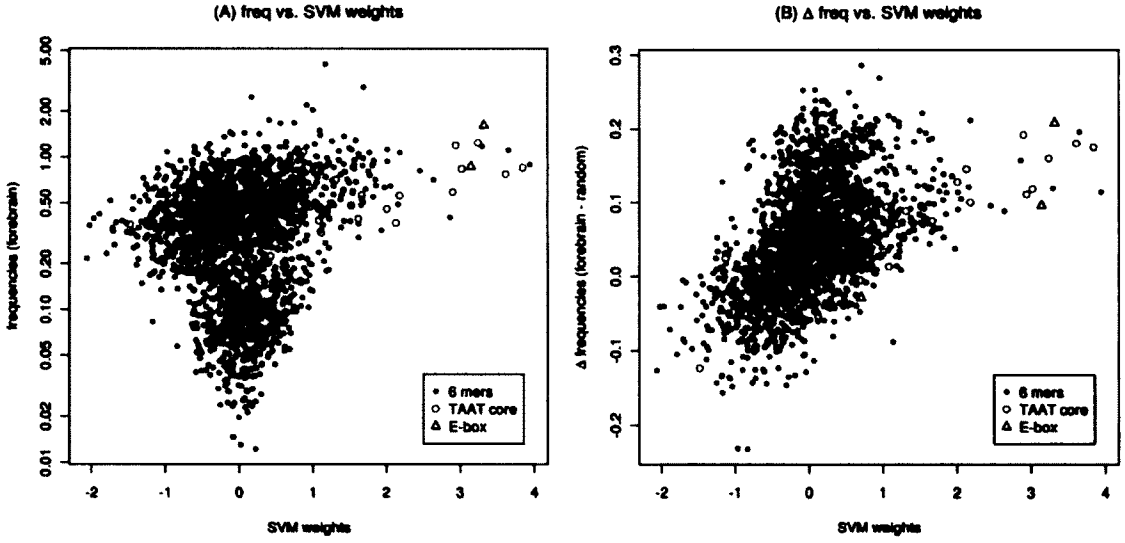



**Figure 3-7: Comparison between frequencies and SVM weights of 6-mers**
While the SVM features which are assigned large positive weights are generally over-represented in the EP300 bound regions relative to background genomic sequence, there is not a strictly direct correlation between SVM weights and k-mer frequencies. (A) 6-mer frequency in forebrain vs. SVM weights. (B) Normalized frequency difference between forebrain and random sequences, $\Delta f = (fb-rand)/(fb+rand)/2$

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. I systematically scored each of the predictive *k*-mers with PWMs for known motifs available in public databases (JASPAR [84], TRANSFAC [83], and UniPROBE [100]) using the TOMTOM package [101]. Because the databases contain many PWMs from families of TFs with similar specificity, many PWMs often score highly for a given *k*-mer, so I report for each *k*-mer the family of matched TFs with *q*-value <0.1 [102], 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 3-1 shows the fifteen 6-mers with the largest positive SVM weights. The elements that positively contribute to EP300 binding include many *k*-mers with TAAT or ATTA cores, which are bound by the homeodomain family [103]. Several homeodomain proteins have restricted expression in the embryonic mouse forebrain, and are required for proper forebrain development, such as *Otx* and *Dlx* [104]–[106]. 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 [107]–[109], and are also known to interact with the coactivator EP300/CREBBP [82].

| 6-mers | Reverse Complement | SVM weight | Database Family Match | Top Matched TFs (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 |

**Table 3-1: Fifteen 6-mers with the largest positive SVM weights**

One of the distinguishing features of my 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. I consistently observe that ZEB1-related *k*-mers have the largest negative weights in forebrain enhancers (Table 3-2). 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 [92]. 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.

| 6-mers | Reverse Complement | SVM weight | Database Family Match | Top Matched TFs (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 3-2: Five 6-mers with the largest negative SVM weights**

Although some negative weight *k*-mers are predictive (*e.g.* ZEB1), on average the positive weights in Table 3-1 are more predictive than the negative weights (Table 3-2) 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 3-8 (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).

### 3.3.3 Predictive Sequence Elements Are Conserved And Clustered within Enhancers

In their previous analysis, Visel *et al.* showed that most EP300-bound regions are enriched in evolutionarily constrained non-coding regions [39]. However, not all

sequences in the EP300-bound regions (average length 750-800 bp) are conserved, rather, several more localized peaks of conservation (10-100 bp) 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. I 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, I measured the degree to which individual *k*-mers were present in conserved regions by averaging the PhastCons conservation score [98] over each instance of the *k*-mer (see Methods), and examined its correlation with SVM weight. Figure 3-8 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.

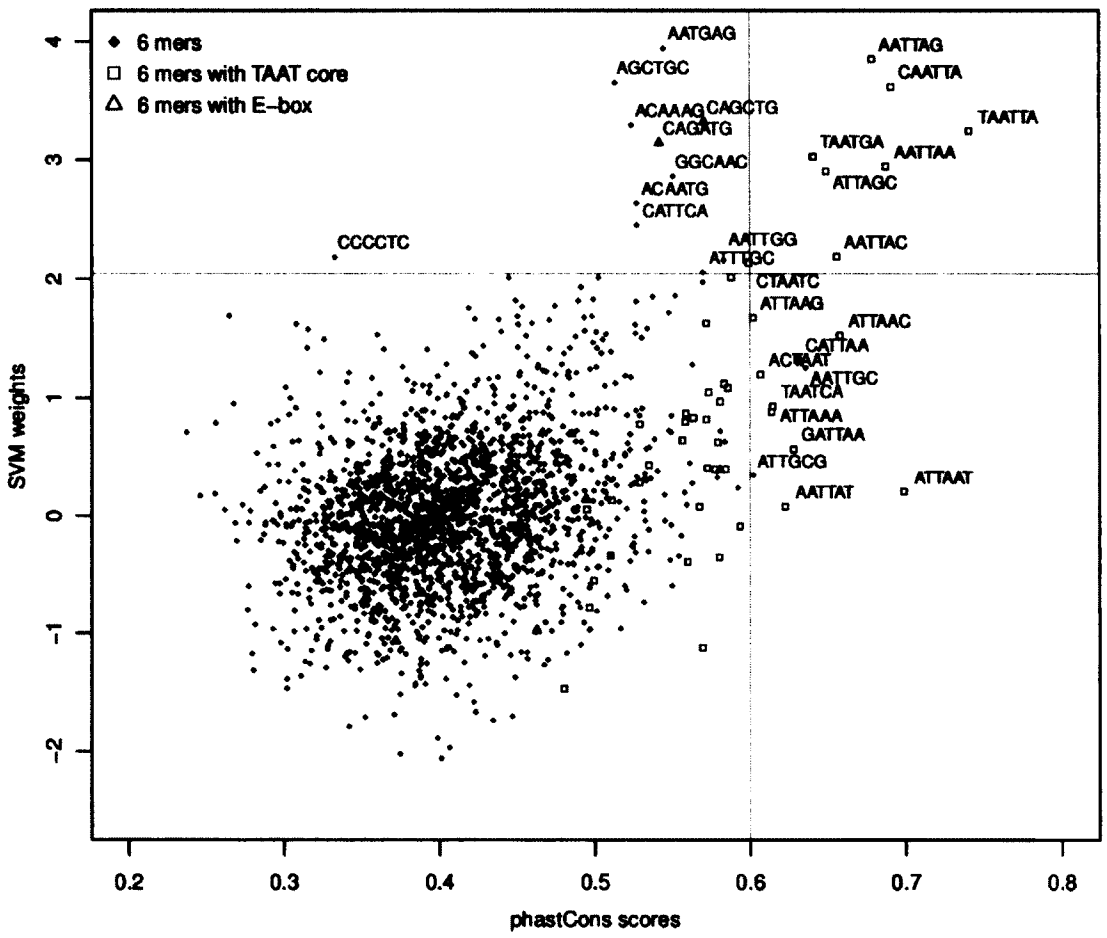



**Figure 3-8: SVM weights vs. PhastCons scores**
Scatter plot between SVM weights and PhastCons conservation scores for 6-mers in forebrain enhancers are presented. Predictive SVM sequence features are more conserved. Two well-known TFBS, TAAT cores (red rectangles), and E-box elements (blue triangles) 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 is displayed.

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 [2]. In Figure

3-9, I 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, I 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 3-9, and the forebrain distribution often deviates from the null distribution by several standard deviations, especially for small spacing. I 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 3-9), we observe periodic enrichments with characteristic spacing of 10–11 bp. The highest peak is around 11 bp, almost two times higher than the null distribution. These positional correlations suggest cooperative binding interactions in phase with the 10.5 bp DNA helix periodicity, consistent with previous observations [59], [110], and local physical interactions between the factors that bind these DNA sequence elements.

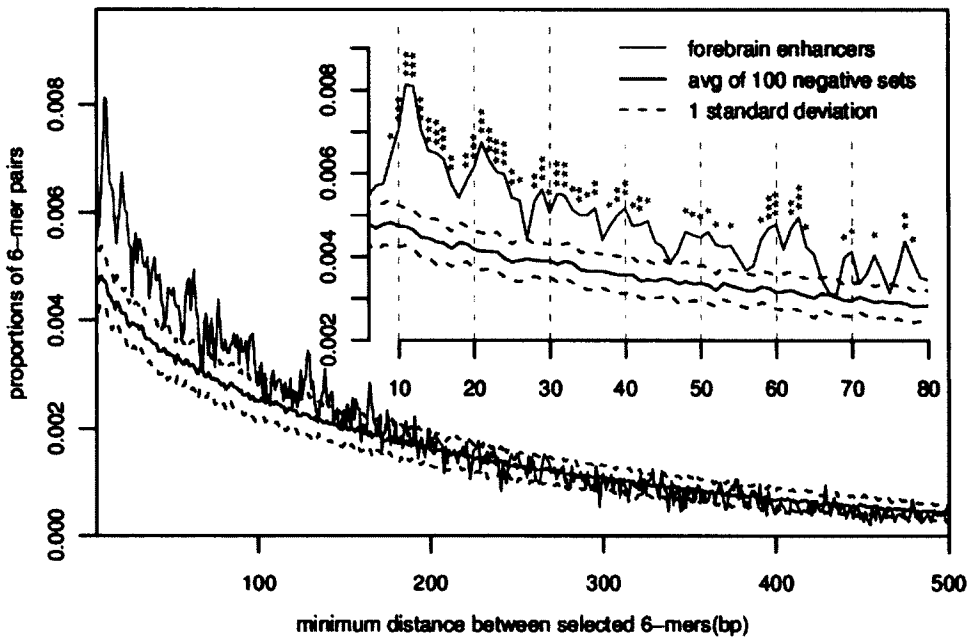



**Figure 3-9: Distributions of minimum pairwise distances**
Distributions of minimum pairwise distances between the most predictive sequence features (Ten 6-mers with the largest positive SVM weights) in forebrain enhancers vs. random genomic sequences are shown. To measure the significance of these differences, I generated 100 distinct full negative genomic sequence sets (using my null model, see section 3.2.2). 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 11 bp, 22 bp, etc., are observed, suggesting positioning in phase with the periodicity of the DNA helix. *P*-values are indicated: * <0.01, ** <0.001, *** <0.0001.

### 3.3.4 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, I scanned the entire genome systematically with my kmer-SVM trained on forebrain enhancers. I segmented the mouse genome sequence into 1,000 bp regions with 500 bp overlap, resulting in about 5.2 million overlapping sequence regions. To compare with the 2453 forebrain region "EP300 training set", I followed Visel and

removed centromeric regions, telomeric regions, and regions containing at least 70% repeats, (however, this filter had minimal impact on my predictions). I then scored all these 1,000 bp regions using the SVM with the $k$=6 spectrum kernel for forebrain enhancers.

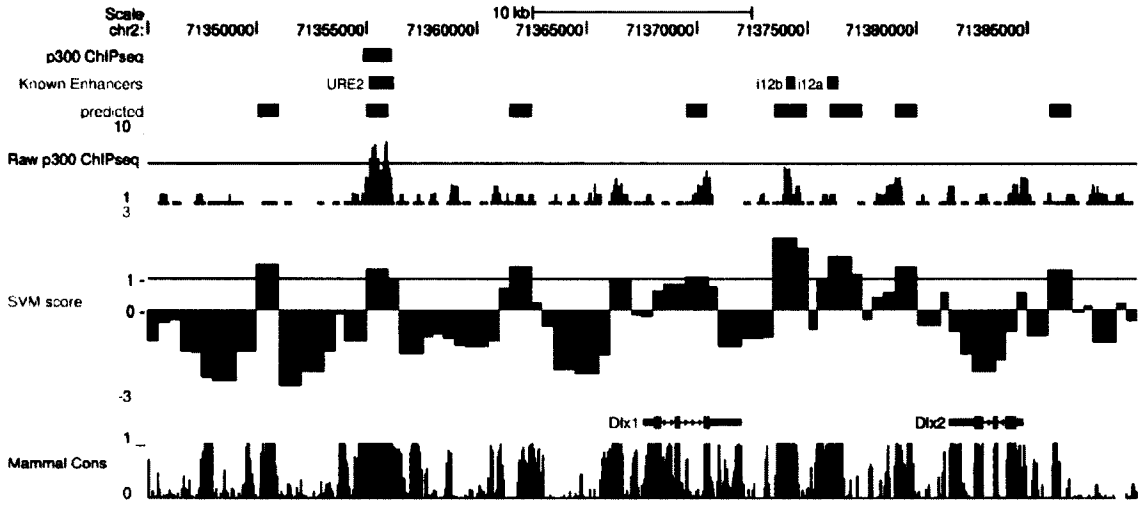



**Figure 3-10: An example of genome-wide SVM enhancer prediction**
A well studied region around *Dlx1* and *Dlx2* is shown here, both known to be expressed in forebrain. While the predicted enhancers often overlap the training EP300 set (blue), novel enhancers are also predicted, and often identify previously 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 (blue).

An example of the continuous SVM score along the *Dlx1/2* locus is shown in Figure 3-10 ("Raw SVM Score"). *Dlx1* and *2* are expressed in the mouse forebrain [104], [111], [112]. Besides the sole EP300 training set element in this region (URE2, labeled "EP300 ChiPseq" in Figure 3-10), two other enhancers within this locus have been experimentally validated ("Known Enhancers", labeled i12a and i12b) [111]. These

enhancers (i12a and i12b) were detected by my kmer-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.* [39]. Comparing the "Raw EP300 ChIPseq" to "Raw SVM score" in Figure 3-10 shows striking correlation: most of my 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, I evaluated the genome-wide correlation between my kmer-SVM predicted regions and EP300 read density. In Figure 3-11 I plot the EP300 ChIP-seq read density as a function of distance from the center of each of the top 1% SVM scoring regions. I 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.

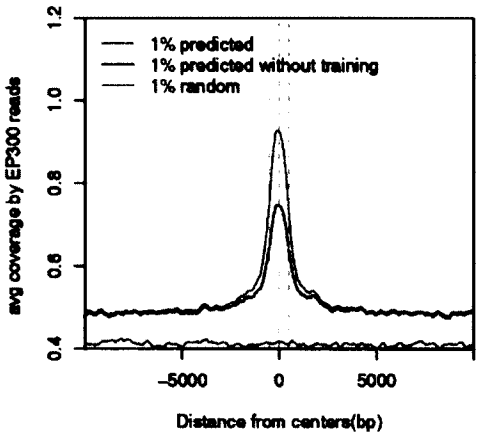



**Figure 3-11: Average EP300 ChIP-seq read coverage in the SVM predicted regions** EP300 reads are significantly enriched in the SVM predicted regions: The middle point of the top 1% SVM predicted regions in forebrain were aligned at 0 bp, the sequence around each peak was extended +-10kb in each direction, and the average coverage of EP300 reads in the surrounding regions is shown. Significant enrichments compared to random genomic sequence (by about two fold) is observed even after those regions which overlap with the original training set are excluded.

To define a high confidence set of enhancer predictions, I 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. I can estimate my false discovery rate (the expected fraction of predicted positives which are false positives, FP/(FP+TP)) from the P-R curves in Figure 3-5B. 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. To trade off precision and recall, I choose a cutoff which corresponds to 50% recall, which at 1x is a SVM score of 1.0. For the large negative sets, precision is about 50% when recall is 50%, and I therefore estimate the false discovery rate to be about 50%. In other words, at this cutoff (SVM>1.0), on the training set, I capture 50% of the EP300 training set regions, and an equal number of negative regions.

In what follows I will be comparing the properties of my kmer-SVM predicted enhancer regions (SVM scores greater than 1.0), the EP300 training set regions, and non-enhancer genomic regions (SVM scores less than 1.0). These three sets are all distinct, i.e. each genomic 1,000 bp region can only belong in one class. Any 1,000 bp region which overlaps a region in the training set by as little as 1 bp is excluded from the SVM sets and included in the EP300 training set. I will show that the EP300 training set and SVM predicted regions have similar properties, much different than the non-enhancer regions.

At a SVM score threshold of 1.0, I predict 33,232 1,000 bp regions in the genome (outside of the EP300 training set), or 26,920 enhancers after merging overlapping

regions, and I 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 can be found in Supplementary Material of Lee *et al.* [5]. I 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 3-3), and the correlation of individual SVM scores between two different SVMs is high (Pearson correlation coefficient=91.5%). 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.

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

**Table 3-3: Overlap between top SVM scoring regions predicted by two separately trained SVMs**

To assess the validity of my genome-wide predictions with independent experimentation, I quantified the DNaseI hypersensitivity of the high scoring forebrain SVM regions with experiments in embryonic mouse whole brain provided by the mouse ENCODE project [4]. 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 my predictions as tissue specific enhancers. In Figure 3-12, I split the predicted 1,000 bp 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). I plot the distributions of average intensity of DNaseI hypersensitivity of the different SVM scoring classes in Figure 3-12A, 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 3-12B 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. With this understanding, I can evaluate the precision with which the kmer-SVM detects DNaseI sensitive enhancers. Because the SVM score and DNaseI signals are continuous, I consider DNaseI signal

greater than 10 to be positive (open chromatin), and DNaseI less than 2 to be negative (not open) for purposes of quantification, consistent with the distributions in Figure 3-12. Then, regions with DNaseI signal greater than 10 and SVM score greater than 1.0 are true positive predictions, and regions with DNaseI signal less than 2 and SVM score greater than 1.0 are false positive predictions. The precision is calculated as TP/(TP+FP), or the accuracy of the predicted positives. As shown in Table 3-4, predictions with SVM score greater than 1.0 have a 56.3% precision, and more stringent predictions with SVM score greater than 1.5 have a 74.5% precision. These results are consistent with my previous estimate that 50% of my novel predictions are true enhancers functioning in mouse brain.

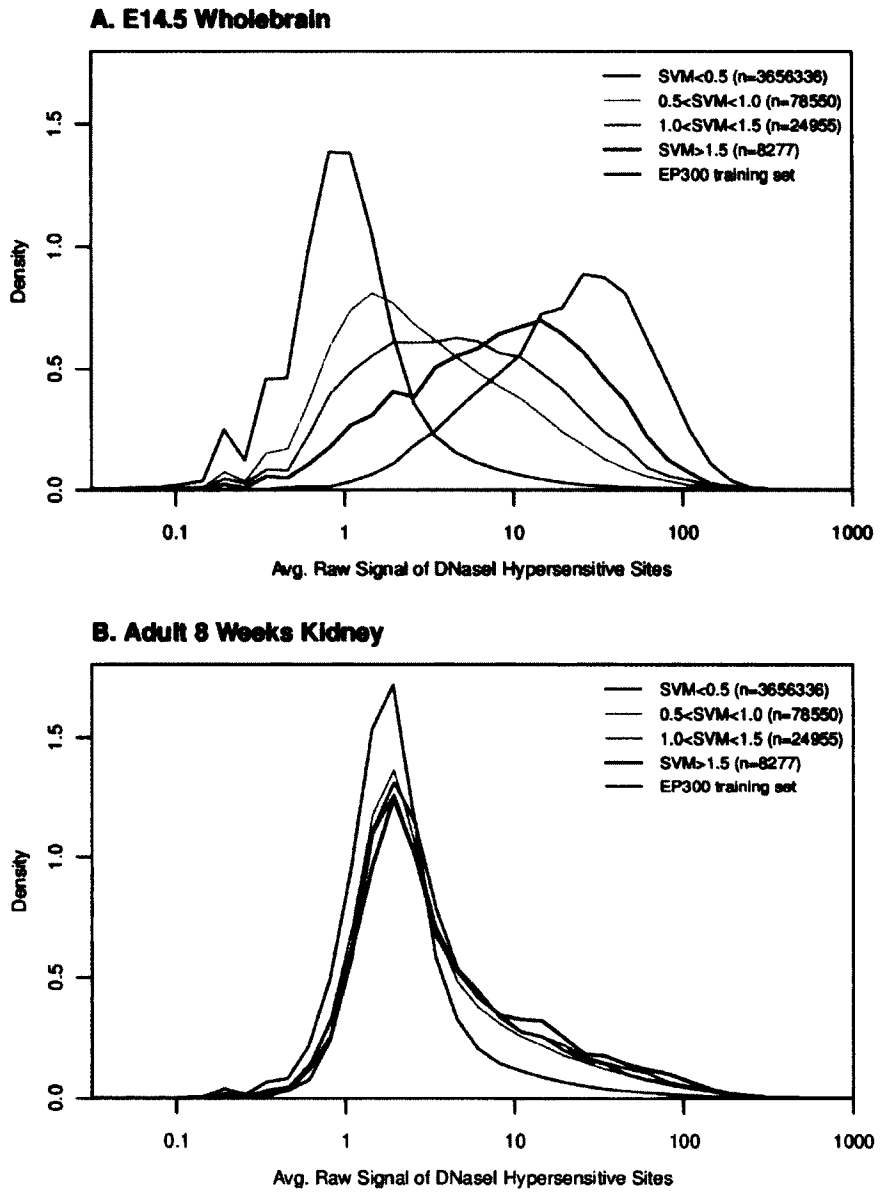



**Figure 3-12: Distributions of average intensity of the DNaseI hypersensitivity**
To independently confirm my predictions with DNaseI measurements in embryonic mouse brain, distributions of average intensity of DNaseI hypersensitivity of different forebrain SVM scoring regions are plotted. (A) DNaseI Hypersensitivity measured in E14.5 Wholebrain (B) DNaseI Hypersensitivity measured in adult 8 weeks kidney, as a negative control. I observe significant enrichments only in high scoring SVM predicted regions in brain.

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

**Table 3-4: Precision of detecting DNase I hypersensitive enhancers**

To further support the biological significance of these novel SVM predicted enhancers, I examined their proximity to forebrain expressed genes. Microarray experiments [39] identified 885 (495) genes over-expressed (under-expressed) in the forebrain at E11.5. I examined the intergenic distance between the EP300 training set regions and the transcription start site (TSS) of the nearest over-expressed genes. I also found the distance between my kmer-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 3-13, both the EP300 training set and my predicted enhancer regions are significantly enriched near (within 10k bp) 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 10k bp 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, my kmer-SVM predicted regions are significantly enriched within 10k bps of forebrain under-expressed genes (Figure 3-13). 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.

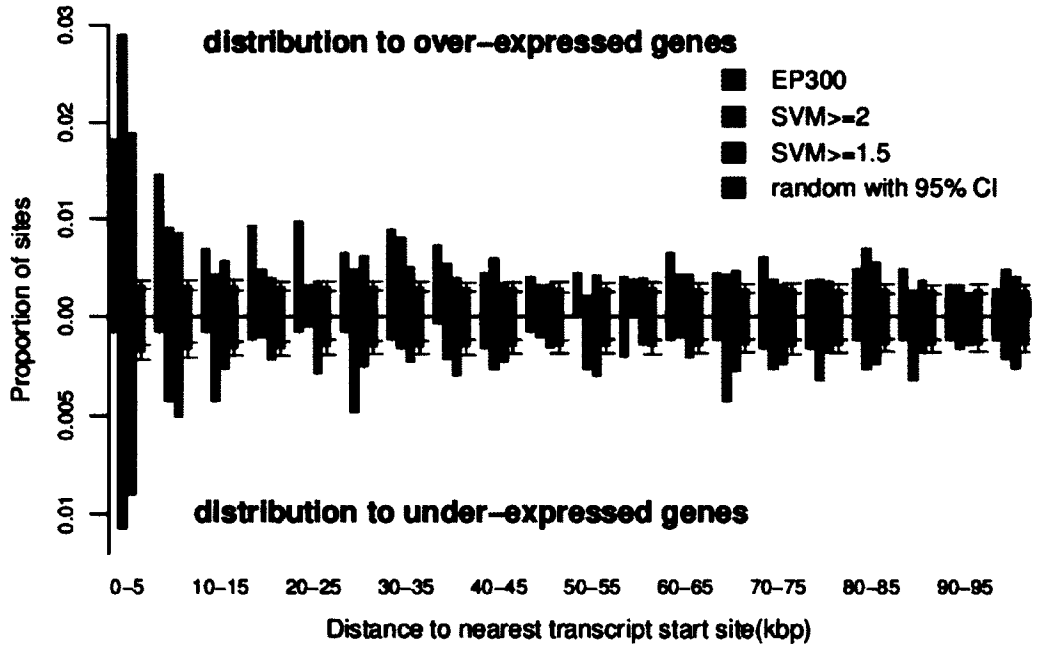



**Figure 3-13: Genome-wide distributions of SVM predicted regions**
The distribution of the distance between the EP300 and SVM predicted regions and the nearest forebrain expressed gene are shown. Any region which overlapped a training set region was excluded from the analysis. Both the EP300 (red) and SVM predicted regions are preferentially located within 10k bp of the TSS of a forebrain *over-expressed* gene (above the axis). Only the SVM predicted regions show significant clustering toward TSS of a forebrain *under-expressed* gene (below the axis). As a null set, I compare to the average of 100 randomized genomic positions, with 95% confidence interval shown (grey)

### 3.3.5 SVM Also Predicts Human Enhancers

I next assessed the ability of my kmer-SVM to predict human enhancers. I found human orthologous regions (hg18) of the mouse EP300 training set with the liftOver utility from the UCSC genome browser [93]. With 70% or greater identity, 2205 of the 2453 forebrain enhancers were successfully mapped onto the human genome. I discarded 13 mapped sequences longer than 3k bps. I then trained SVMs to discriminate this positive human training set from an equal number of human random sequences generated by my null model, and achieved reasonably high auROC=0.87 (Figure 3-14). I also tested more stringent orthology cutoffs (requiring 90% and 95% identity instead of 70%), and found that the overall performance was very similar (Figure 3-14). This result also demonstrates the robustness of my kmer-SVM predicted enhancers. Thus a 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.

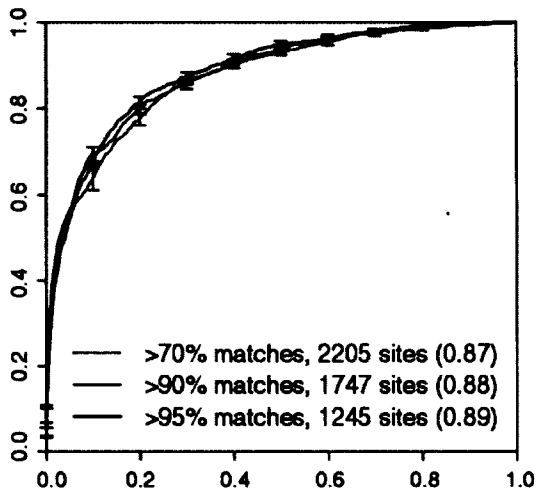



**Figure 3-14: Classification of human orthologous regions of the EP300 mouse forebrain set**
A positive human test set was generated by sequence alignment of the mouse EP300 training set regions to the human genome, varying the stringency for assigning homologous regions (70% identical, 90% identical, and 95% identical) all three of these sets can be classified with high accuracy (auROC=0.87, 0.88, 0.89), and classification power is relatively unaffected by the cut-off for determining homologous regions.

In addition, I 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. I 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 [113] is captured by both SVMs (Figure 3-15). I then systematically analyzed the entire genome to assess how many top SVM scoring regions overlap each other (Table 3-5). Although the overlaps are not as significant as scores using only different negative sets (Table 3-3), a large fraction of top SVM scoring regions

are still shared between the two SVMs, so to a large degree, a SVM trained on mouse can be used to successfully predict human enhancers. This result is in general agreement with *in vivo* experimental results [114] 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.

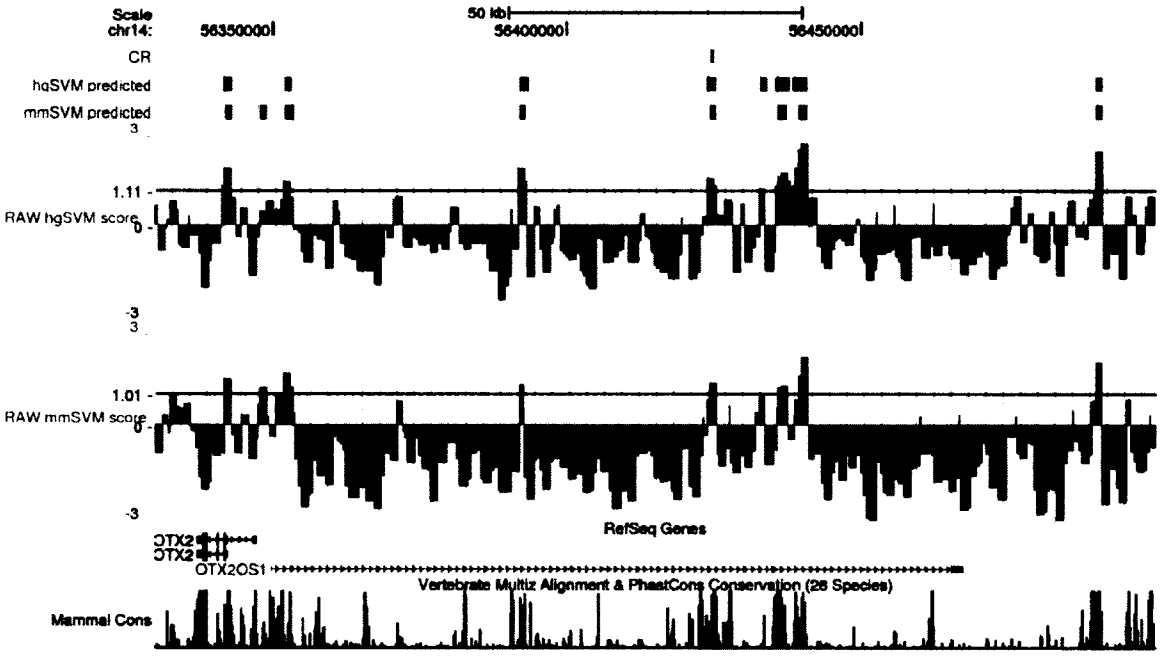



**Figure 3-15: SVM predictions at the human Otx2 locus**
The human genome *Otx2*, which is known to play a role in forebrain development, is scored by two SVMs (mmSVM and hgSVM). The raw hgSVM and mmSVM scores are quite similar, and most of the predicted enhancers above the 1% threshold overlap. One of these enhancers has been experimentally verified to have enhancer activity (CR).

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

**Table 3-5: Overlap between human putative enhancers predicted by two SVMs trained on mouse or human**

### 3.3.6 Comparison Between Different EP300/CREBBP ChIP-seq Datasets Reveals Sequence Elements Important for Pluripotency

The success of my kmer-SVMs in predicting EP300 binding in mouse embryonic brain and limb motivated a comparison with other EP300/CREBBP ChIP-seq data sets. I 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 [81], and EP300-bound regions in cultured mouse embryonic stem cells [91]. I will refer to these as "CREBBP neuron" and "EP300 ES" in the following discussion. I was 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). EP300 ES enhancers do not significantly overlap with any other set (fb, mb, lb, or CREBBP neuron). This indicates that EP300 mediated embryonic neuronal development is linked to CREBBP mediated neural activity dependent transcription via extensively shared

common regulatory regions. I 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 3-1 and Table 3-6), which further indicates common modes of regulation.

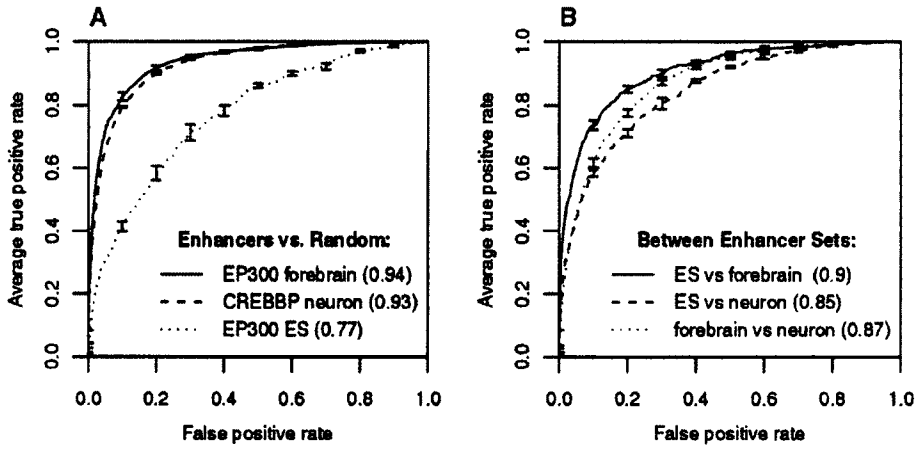



**Figure 3-16: Classifications of other EP300 enhancer sets**
(A) Classification of EP300 forebrain enhancers, neuronal stimulus dependent enhancers (CREBBP neuron), and mouse embryonic stem cell enhancers (EP300 ES) vs. random genomic sequence are shown. Although the embryonic stem cell dataset is somewhat less accurately classified, my kmer-SVMs successfully discriminate EP300 or CREBBP bound regions from random sequences. (B) Classification of EP300 forebrain, CREBBP neuron, and EP300 ES datasets vs. each other is also robust.

Figure 3-16A 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, I can also successfully classify these sets (EP300 forebrain, CREBBP neuron, EP300 ES) against each other, as show in Figure 3-16B. 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 3-1 and Table 3-6). However, when classified against each other, I 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 3-8). Therefore, homeodomain, SOX, and bHLH binding sites may play more prominent roles in neural developmental processes than in neural activity dependent transcription.

| 6-mers | Reverse Complement | SVM weight | Database Family Match | Top Matched TFs (q-val <0.1) |
|---|---|---|---|---|
| GACTCA | TGAGTC | 5.68 | Leucine Zipper | BACH1, NFE2, BACH2, JUNDM2, AP1 |
| CAGATG | CATCTG | 5.25 | HLH | ZNF238, TAL1:TCF3, TAL1:TCF4, TCF3 |
| GAGTCA | TGACTC | 5.19 | Leucine Zipper | BACH1, BACH2, NFE2, JUNDM2, AP1 |
| ACGTCA | TGACGT | 5.05 | Leucine Zipper | ATF6, CREB, ATF1 |
| TGCCAA | TTGGCA | 4.80 | Nuclear Factor I | HIC1, NFIC, NF1 |
| AGATGG | CCATCT | 4.29 | HLH | TAL1:TCF3, TAL1:TCF4, YY1, TCF3 |
| CATATG | CATATG | 4.13 | - | - |
| AATTAG | CTAATT | 4.02 | Homeodomain | VSX2, PRRX2, EVX2, PDX1, GBX2 |
| GGCAAC | GTTGCC | 3.63 | - | - |
| CTGGCA | TGCCAG | 3.47 | Nuclear Factor I | HAND1::TCF3, HIC1, NFIC, TGIF2 |
| TAATTA | TAATTA | 3.42 | Homeodomain | OTP, PROP1, HOXA, ALX1, LHX3 |
| GATTCA | TGAATC | 3.30 | - | - |
| CGTCAC | GTGACG | 3.18 | Leucine Zipper | PAX3, CREB, ATF1, JUNDM2 |
| GCGTCA | TGACGC | 3.12 | Leucine Zipper | CREB, ATF6, BACH1, BACH2 |
| AATTAC | GTAATT | 3.10 | Homeodomain | PRRX2, HOXA6, HOXA1, HOXC8, DLX1 |
| GGTCAA | TTGACC | -3.58 | Nuclear Receptor | PPARG, RORA2, HNF4A, RORA1, ESRRA |
| CTGACC | GGTCAG | -3.98 | Nuclear Receptor | RORA2, RORA1, NR2F2, ESRRA, PPARG |
| AGGTGA | TCACCT | -3.99 | Zinc-finger | ZEB1 |
| CAGGTA | TACCTG | -4.08 | Zinc-finger | ZEB1 |
| GGGTCA | TGACCC | -4.64 | Nuclear Receptor | NR2F2, ESRRA, HNF4A, RXRA, PPARG |

**Table 3-6: Predictive 6-mers of CREBBP Neuron**

| 6-mers | Reverse Complement | SVM weight | Database Family Match | Top Matched TFs (q-val <0.1) |
|---|---|---|---|---|
| ACAATG | CATTGT | 1.45 | SOX | SOX17, SOX9, SOX5, SOX10, SOX30 |
| ATTGTC | GACAAT | 1.19 | SOX | SOX17 |
| ACAAAG | CTTTGT | 1.06 | SOX | SOX4, SOX11, SOX10, HNF4 |
| TATGCA | TGCATA | 1.00 | Homeodomain | POU2F1, POU3F3, POU2F3, POU2F2 |
| GAGCTA | TAGCTC | 0.95 | - | - |
| CAAAAG | CTTTTG | 0.90 | - | - |
| AGGTCA | TGACCT | 0.89 | Nuclear Receptor | RORA1, PPARG, RORA2, ESRRB, RAR |
| AAAGCC | GGCTTT | 0.89 | - | - |
| AATTCC | GGAATT | 0.88 | - | - |
| AAGGTC | GACCTT | 0.88 | Nuclear Receptor | PPARG, ESRRB, ESRRA, RAR, NR2F2 |
| TCTACA | TGTAGA | 0.85 | - | - |
| TAACAA | TTGTTA | 0.85 | SOX | SOX5 |
| CCGGAA | TTCCGG | 0.84 | ETS | ELK4, GABPA, NRF2, STAT3, ELK1 |
| GGTGAC | GTCACC | 0.82 | Leucine Zipper | SREBP, CREB, AP1, ATF, RAR |
| CATTCA | TGAATG | 0.79 | SOX | HBP1 |
| AACATG | CATGTT | -0.75 | - | - |
| GCTAGA | TCTAGC | -0.76 | - | - |
| CTGATA | TATCAG | -0.82 | - | - |
| AATAAA | TTTATT | -0.83 | Homeodomain | HOXD13, FOXC1, HOXB13 |
| ACAAAT | ATTTGT | -0.85 | - | - |

**Table 3-7: Predictive 6-mers of embryonic stem cells**

| ES (+) vs forebrain (-) | | ES (+) vs neuron (-) | | forebrain (+) vs. neuron (-) | |
|---|---|---|---|---|---|
| AGGTCA | 2.83 | AGGTCA | 1.38 | ACAAAG | 3.20 |
| ACCTTG | 2.34 | CAATAG | 1.08 | AGCTGC | 2.63 |
| AGGTGA | 1.97 | ACCTTG | 0.954 | TAATGA | 2.47 |
| CCTTGA | 1.77 | AAGGTC | 0.853 | CAGCTG | 2.47 |
| AAGGTG | 1.76 | CCTTGA | 0.760 | GAACAA | 2.46 |
| CACACC | 1.64 | GGTCAC | 0.730 | AAAGGG | 2.33 |
| GGTGGA | 1.52 | CCGGAG | 0.709 | GGATTA | 2.23 |
| CAGGTA | 1.51 | ACCTGA | 0.671 | ACAATG | 2.22 |
| CACCTG | 1.48 | CACCTG | 0.666 | CAATTA | 2.22 |
| CTGACC | 1.45 | CAGGTA | 0.639 | AATTAG | 2.10 |
| ACCTGG | 1.44 | TCTACA | 0.628 | CAATGG | 2.08 |
| AGGTAA | 1.42 | GGTCAA | 0.627 | ATTAGC | 2.05 |
| GAGTCA | 1.38 | ACACCC | 0.624 | GGCCCC | 2.01 |
| CTAGAA | 1.32 | GAACCC | 0.624 | ACAATA | 1.89 |
| AGGAAG | 1.30 | AGGTGA | 0.619 | GAGGCC | 1.88 |
| | | | | | |
| CTGGCA | -1.36 | ACAGAT | -0.585 | ATTTCA | -1.49 |
| AGGGGG | -1.38 | ATGACG | -0.593 | GTGCCA | -1.52 |
| AATTAG | -1.39 | AACATG | -0.605 | CTCATC | -1.53 |
| GCTGCC | -1.41 | ATGCCA | -0.629 | ACTCAT | -1.60 |
| CAGCTG | -1.49 | AATATG | -0.637 | ACGTCA | -1.67 |
| CAGATG | -1.53 | CTAAAA | -0.647 | AACATG | -1.69 |
| ACAAAG | -1.54 | TAATTA | -0.660 | AAATTA | -1.71 |
| TAATTA | -1.64 | GGCAAC | -0.660 | GTTCCA | -1.73 |
| AGCTGC | -1.65 | CAGATG | -0.670 | GACTCA | -1.77 |
| GGCAAC | -1.66 | ACGTCA | -0.726 | CTAAAA | -1.90 |
| CAATTA | -1.68 | TGCCAA | -0.773 | GAATCA | -1.90 |
| AATGAG | -1.68 | CTGGCA | -0.776 | CATCTC | -1.94 |
| AATTAA | -1.76 | ATGTCA | -0.817 | GATTCA | -2.34 |
| CACAAA | -1.80 | GAAATA | -0.824 | GAGTCA | -2.46 |
| TAATGA | -1.88 | AAATAG | -0.833 | TGCCAA | -3.12 |

**Table 3-8: Comparison of predictive 6-mers from the different data sets**

I also assessed the biological significance of the predictive *k*-mers in these new data sets. I find that most of the predictive *k*-mers can be related to known TFBSs (Table 3-6 and Table 3-7), 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 3-6), are the target of heterodimers of the regulators Fos and Jun,

which play critical roles in neural activity dependent transcription regulation [115]. 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 3-6 [81], [115]. 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 3-6. The SRF binding site is also known as a CArG box, whose consensus sequence is CCWTATAWGG [84]. 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 3-6. Therefore, all well characterized TFBSs known to play a role in neuronal activation are successfully captured by my kmer-SVM. Interestingly, I 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 [116], [117]. I found only one relevant example of neural activity dependent expression of a homeobox protein, LMX1B [118]. There may be still unknown mechanisms involving NFI and homeodomain proteins in the context of neural activity dependent transcriptional regulation, but broadly speaking, my results indicate significant pleiotropy between neuronal developmental pathways and neural activity dependent signaling pathways.

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

**Figure 3-17: 6-mer SVM weights across the SOX2-OCT4-NANOG binding site**
Many large weight *k*-mers from the SVM trained on the EP300 ES dataset are subsequences that tile across the SOX2-OCT4 consensus oligomer.

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, my method identifies factors known to be crucial for maintaining ES identity: I find high scoring binding sites for NANOG-POU5F1(also known as OCT4)-SOX2 SOX-family factors (Table 3-7), essentially the same binding sites found in previous studies [91], [119]. I 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 3-7 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 [91],

and the 6-mer subsequences within the combined binding site (CATTGTYATGCAAAT) have high SVM weights. Figure 3-17 shows how large weight *k*-mers tile across this extended known binding site. In addition, I 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 [91]. More interestingly, I 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 3-6), 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.

### 3.3.7 SVM Can Predict Other ChIP-seq Data Sets

Until this point I have applied my kmer-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. It should be noted that popular *de novo* motif finding methods such as AlignACE [120] or MEME [121] 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.* [91] did successfully identify Sox2, Oct4, and Nanog binding sites in the EP300 ES data with Weeder [119], the EP300 ES data set was the smallest and least diverse of the data sets I analyzed.

To directly assess the ability of my kmer-SVM to predict binding of individual transcription factors, I analyzed ChIP-seq results on the TF ZNF263. I chose ZNF263, a 9-finger C2H2 zinc finger which is predicted to have a binding site of ~24 bp, to assess how well *k*-mers can represent extended degenerate binding sites. I used ChIP-seq data on ZNF263 in a K562b cell line [122] which identified 1418 strongly bound regions. Predicting against a 50x random negative set yielded auROC=0.938 and auPRC=0.51 (Figure 3-18B,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 [122], 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 my approach should have significant predictive value for a wide range of binding data.

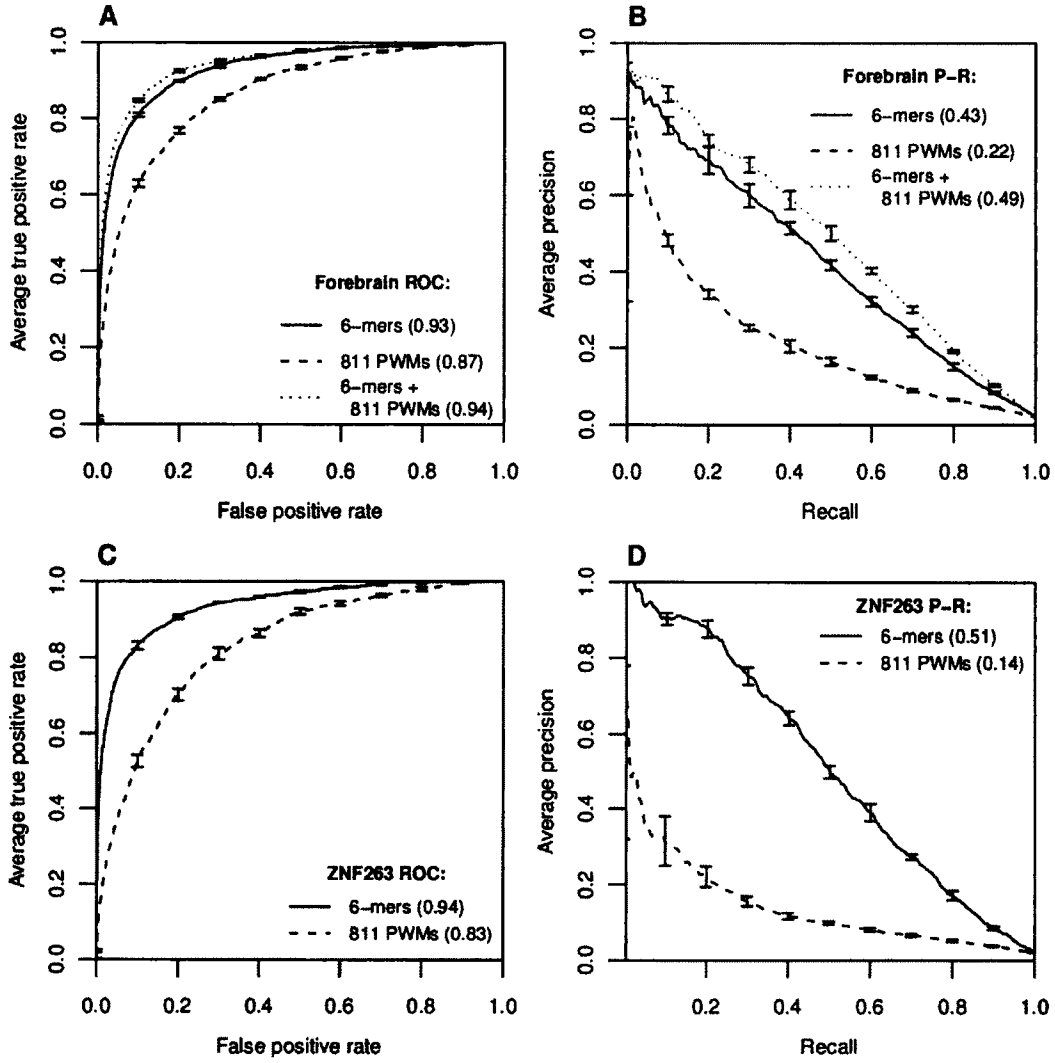



**Figure 3-18: PWM vs. k-mers as feature sets on forebrain and ZNF263**
(A) On the forebrain enhancers, my kmer-SVM is more accurate than known PWMs alone, but a combination of *k*-mers and PWMs performs slightly better. (B) These differences in auROC translate to a dramatic reduction in auPRC for PWMs relative to *k*-mers only or combined *k*-mers and PWMs. (C) my kmer-SVM predicts ZNF263 bound regions from ChIP-seq with high accuracy (auROC=0.94), but the 811 PWM SVM is less accurate (auROC=0.83). (D) Again the lower auROC for PWMs corresponds to a significant decrease in auPRC for PWMs on the ZNF263 data (0.14 vs. 0.51), and a much higher false discovery rate.

### 3.3.8 Comparison to Alternative Approaches

As an alternative to *k*-mers, I also tried using known PWMs as features in a SVM. I

used 811 PWMs from existing databases of known TF specificities [84], TRANSFAC [83], and UniPROBE [100]. When using these features, I 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 my *k*-mer approach (Figure 3-18A), 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 3-18B). 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. I 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 3-18B,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 my kmer-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 [123], might allow clearer identification of informative sequence features from within the *k*-mer SVM, but would not affect predictive performance.

I also compare my approach to alternative kernel methods. I applied the weighted degree kernel with shifts (WDS) [89] to the CREBBP neuron data set (as WDS requires input sequences of equal length), and found auROC=0.83, compared to auROC=.93 for

my kmer-SVM. A notable SVM based approach which incorporates positional information between general *k*-mer features (KIRMES) has been recently described [94], [124]. I 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 my kmer-SVM result. Additionally, the periodic spatial distribution in Figure 3-9 suggests that a model based on difference in angle (similar to Hallikas et al. [59]) would be more appropriate than the Gaussian spatial dependence used in KIRMES. Another approach to predict promoters [66] used PWMs and L1-logistic regression. I found little difference between logistic regression and SVM: using the same *k*-mer feature vectors in L1-logistic regression yielded auROC=0.92 on the EP300 forebrain dataset, using publicly available software [125].

## 3.4 Discussion

In this chapter, I 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 my 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, I predict that 50% of our 26,920 predicted non-overlapping enhancers with forebrain SVM scores above 1.0 are true positives. This is a conservative estimate of the ability to detect novel enhancers, since when scanning the genome I have scored 1,000 bp 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. I have shown that I can predict human enhancers based on these mouse enhancer experiments by measuring the overlap between human enhancers predicted by a SVM trained on the mouse sequence, and comparing these predictions to a SVM trained on human sequence orthologous to the mouse enhancer sequences. Finally, by comparing between other

EP300/CREBBP ChIP-seq data sets, I 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, my 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.

Although I have provided evidence that my kmer-SVM predicted regions are likely functional, to what degree I are predicting these enhancers accurately based on sequence features which are tissue-specific? Alternatively, I 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 [126], [127], whose activity could be modulated by tissue-specific factors, much as Pol II operates generally. Ultimately this should be determined by individual experiments, but I here address this problem computationally by investigating overlaps between forebrain and limb-specific predicted regions, which I then compare with the overlaps between EP300-enriched regions in forebrain and limb. For this comparison, I independently

determined EP300-enriched regions from the raw data set using the same threshold criteria as the previous study [39] except that I have used fixed-length 1,000 bp regions, rather than the ChIP-seq determined peak regions. With a 1% false discovery rate (FDR), I 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 my kmer-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 my hypothesis that kmer-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.

# 4 Applications of kmer-SVM

## 4.1 Melanocyte Enhancer Prediction

### 4.1.1 Summary

Since the original development of kmer-SVM [5], it has been successfully applied to several different biological problems. One of the first applications of my kmer-SVM method was to predict EP300-bound enhancers in melanocytes.[3] Collaborating with the McCallion lab, we built a kmer-SVM model using experimentally identified melanocyte enhancers and showed that our kmer-SVM can successfully discern the melanocyte enhancers and learn predictive sequence features known to play major roles in melanocytes development and differentiation. We then used the trained kmer-SVM to computationally predict several thousand putative melanocyte enhancers in addition to the experimentally identified ones, and validated a subset of these predictions both *in vivo* and *in vitro*. We also predicted human melanocyte enhancers using the same kmer-SVM model, and showed that the predicted enhancers exhibit several biologically relevant properties. In summary, we further established the validity of our newly developed method by providing multiple lines of experimental evidence through collaborative efforts.

[3] An earlier version of Chapter 4.1 was published in the journal, *Genome Research* [6].

### 4.1.2 Positive Training Set And Peak Refinement Algorithm

EP300 ChIP-seq and H3K4me1 ChIP-seq experiments in melanocytes were used to define the 2489 melanocyte enhancer training set. Multiple criteria have been applied to determine a high confidence set of melanocyte enhancers. In order to be considered as a putative melanocyte enhancer, it first should be observed as a peak (determined by MACS [128]) in both biological replicates of the EP300 ChIP-seq experiments. Additionally, H3K4me1 peaks should also be present in both flanking sites of a candidate region. These preprocessing steps including ChIP-seq experiments were all done by collaborators in the McCallion lab, and more detailed explanations can be found in Gorkin *et al.* [6].

To increase the quality of our kmer-SVM model, we additionally refined the original melanocyte enhancer set before training. We frequently observe that many peak calling algorithms tend to augment peaks with much longer flanking regions that exhibit only moderate ChIP-seq signal intensity. We found that this was also the case for the original melanocyte enhancer set and these extra flanking regions could limit the overall classification results. We reasoned that most of these flanking regions would provide little information, which then could significantly contribute to the noise in the positive set. To resolve this issue, I developed a simple peak refinement algorithm to find a common length sub-region that maximizes the ChIP-seq signal intensities within each original peak. Briefly, I first built a whole gnome track that count the number of extended mapped reads (200 bp) at every 20 bp segment of the genome using MACS software [128]. I refer to this set of numbers as "ChIP-seq signal intensity". I then found a common length sub-

region which maximizes the sum of the ChIP-seq signal intensities of the sub-region. To account for multiple biological replicates, I used as the objective function $F(x)$ the sum of logarithm of the ChIP-seq signal of each replicate as shown below:

$$F(x) = \underset{x}{\operatorname{argmax}}\left(\sum_{rep}\log\left(\sum_{p=x}^{x+\mathrm{LENGTH}}\left(\mathrm{Intensity}_{rep}(p)+1\right)\right)\right)$$

$\mathrm{Intensity}_{rep}(p)$ is the ChIP-seq signal intensity at the position $p$ in the *rep* data set, and LENGTH is a common length of sub-regions. To determine the optimal length, I generated three training sets with different common lengths; 200, 400 and 600 bp and compared between their classification performances. I finally chose 400 bp as the optimal common length for the subsequent analysis as determined by the best auROC (Figure 4-1). Comparison of classification results between the original set and the optimal common length set further reveals that this simple peak refining step can effectively improve the classification accuracy (Figure 4-1).

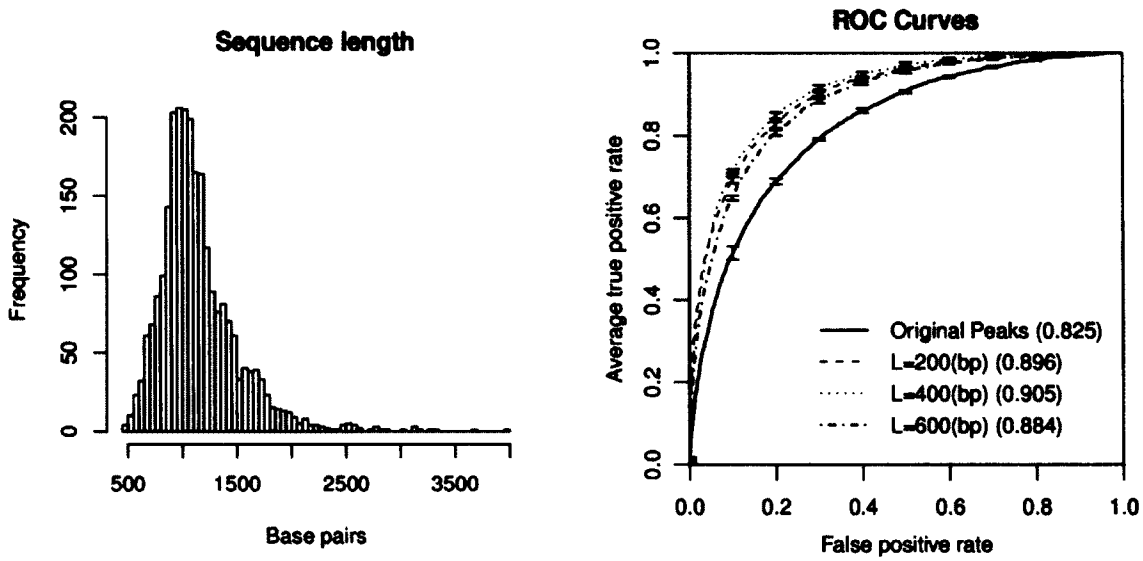



**Figure 4-1: Comparison between different peak lengths**
(*Left*) Length distribution of the original melanocyte enhancer set is shown. Median length is 1,077 bp, which is much longer than the optimal common length determined by the best auROC. (*Right*) ROC curves for three different common peak lengths (200, 400 and 600 bp) are presented. The best auROC was achieved when a common peak length was 400 bp (dotted green). A simple peak refinement algorithm can significantly improve the accuracy of our kmer-SVM model.

### 4.1.3 Negative Training Set With Improved Null Model

For negative sequences, I found a 50X larger set of random genomic 400 bp sequences to account for the imbalance between potential regulatory DNA and non-regulatory DNA in the genome. Here I improved my original null sequence sampling algorithm by matching GC contents as well as repeat fraction of the positive set (Figure 4-2). I additionally excluded from sampling any potential EP300-bound regions with Poisson test *P*-value less than 0.1 (equivalent to 10 ChIP-seq reads). At each sampling step, I randomly selected a region from the positive set, calculated its length, GC content, and repeat fraction, and sampled an identical length genomic sequence that closely matched other two properties (the GC content and the repeat fraction). I repeated the sampling

process until I obtained 50X sequences. It should be noted that I used only 10X larger negative sets for the preliminary analysis shown in Figure 4-1.

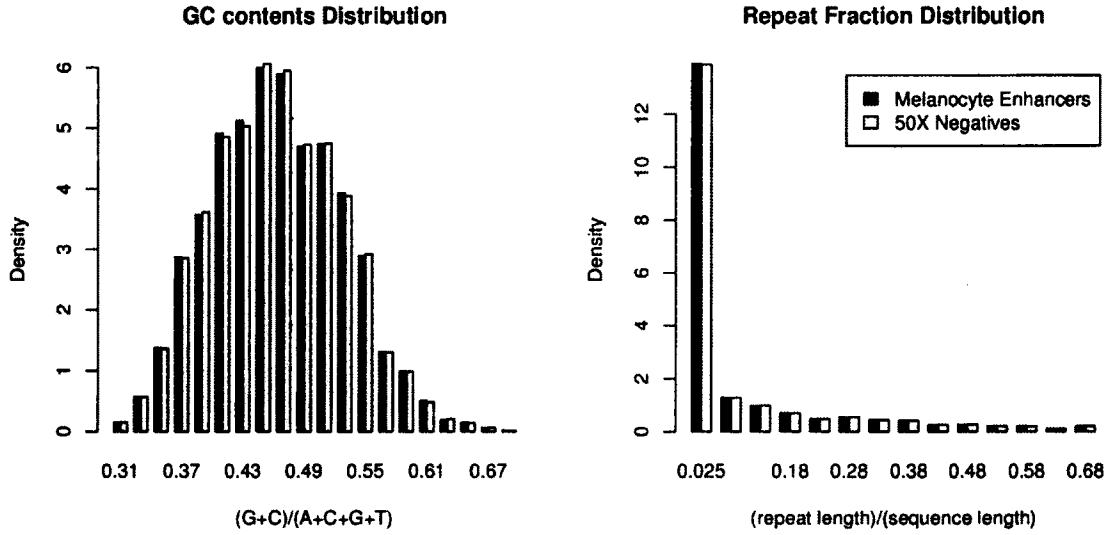



**Figure 4-2: Comparison of the sequence properties of the melanocyte enhancer training datasets**
(*Left*) GC content distributions of enhancers and random genomic regions are shown. (*Right*) Repeat fraction distributions are shown.

### 4.1.4 Kmer-SVM Can Accurately Discriminate Melanocyte Enhancers From Genomic DNA

To identify the specific sequence signals which may in turn suggest mechanisms of melanocyte enhancer activity, we built a kmer-SVM model on the melanocyte enhancer set after refining peaks to a common length regions and then removing any enhancers which were >70% repeats (See section 4.1.2). Here, we used the refined melanocyte enhancer set as positive sequences, a 50X larger random genomic sequence set as negative sequences, and the full set of 2,080 distinct 6-mers as features.

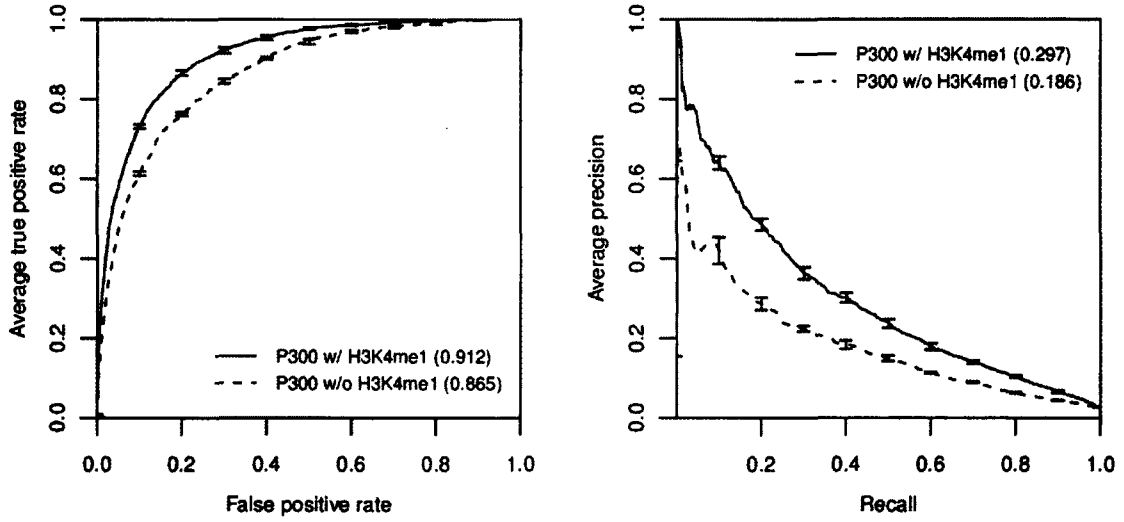



**Figure 4-3: Classification results of melanocyte enhancers using kmer-SVM**
(*Left*) ROC curves of five-fold cross validation results are shown. The EP300 bound regions flanked by H3K4me1 peaks (solid line) are classified better than the regions without the histone marks (dashed line), although both sets are successfully classified. (*Right*) Precision-recall (PR) curves of the same classification results are presented. The difference in PR curves is more evident than in ROC curves.

After training, we assessed our kmer-SVM classifier by its ability to accurately predict the class of reserved test sets via standard five-fold cross validation, as shown by the area under the receiver operating characteristic curves (auROC) and precision-recall curves (auPRC) in Figure 4-3. The melanocyte enhancer trained kmer-SVM achieved auROC of 0.912 and auPRC of 0.297, comparable with our results on other enhancer sets, providing independent verification of the quality of the experimental EP300 bound enhancer identification. In addition to the original melanocyte enhancer set, we also trained a kmer-SVM model on melanocyte EP300 bound regions without neighboring H3K4me1 signals. This relaxed criterion identifies 1134 additional putative melanocyte enhancers, and also shows reasonable auROC and auPRC although it is less accurate (Figure 4-3).

| 6-mer | Reverse complement | SVM weight | Top matched TF(s) |
|---|---|---|---|
| GAGTCA | TGACTC | 5.13 | JUN, FOS |
| GACTCA | TGAGTC | 4.43 | JUN, FOS |
| CACGAG | CTCGTG | 4.09 | MYCN (MITF*) |
| CGTGAC | GTCACG | 4.00 | PAX2 (PAX3*) |
| ACGTCA | TGACGT | 3.89 | ATF1, JDP2, CREB1 |
| ACAAAG | CTTTGT | 3.72 | SOX10 |
| CACATG | CATGTG | 3.54 | MYCN (MITF*) |
| ACCACA | TGTGGT | 3.48 | RUNX1 |
| AAGAAT | ATTCTT | 3.45 | GATA6, SOX5 |
| ATTCCA | TGGAAT | 3.40 | TEAD1 |
| ACCTGG | CCAGGT | -3.55 | ZEB1, TCF3 |
| ACAGGT | ACCTGT | -3.84 | ZEB1 |
| CACCTG | CAGGTG | -4.20 | ZEB1 |
| ACACCT | AGGTGT | -4.38 | ZEB1 |
| CAGGTA | TACCTG | -4.84 | ZEB1 |

**Table 4-1: Predictive 6-mers of melanocytes**

We then determined to evaluate predictive sequence features identified by the kmer-SVM. The most predictive (large SVM weight) features provide insight into the mechanisms which specify enhancer function. As shown in Table 4-1, the most positively predictive 6-mers correspond to binding sites for TFs known to be directly involved in melanocyte biology, including MITF, SOX10, FOS/JUN, and TEAD1. Interestingly, we also identified ZEB1 related 6-mers as the most negatively predictive 6-mers. These TF binding sites have also been predicted as negative sequence features in several other EP300 data sets. This again indicates that EP300 coactivator binding requires depletion of ZEB1 binding sites, in concordance with our earlier results on forebrain enhancers [5].

### 4.1.5 Kmer-SVM Predicts Additional Melanocyte Enhancers

We further determined to investigate regions which the kmer-SVM classifier predicts

as additional melanocyte enhancers based on their sequence. To predict additional enhancers, we segmented the mouse genome into 400 bp regions with 300 bp overlap and scored all of these segments with the kmer-SVM. The top 10,000 SVM scoring regions were chosen for further analysis, corresponding to an SVM cut-off score of 1.0, yielding a precision of 0.74 and recall of 0.05 estimated from the PR curve in Figure 4-3. We then eliminated any predicted regions overlapping the original 2,489 melanocyte enhancers (508 regions overlapping 348 enhancers from the original training set), finally obtaining a set of 7,361 SVM predicted melanocyte enhancers after merging any overlaps. These predicted enhancers show evolutionary sequence conservation, but at a reduced level relative to the original set of 2,489 putative enhancers measured by average PhastCons scores [98] (Figure 4-4A and B). ChIP-seq signals of EP300 and H3K4me1 of these SVM predicted loci also show similar patterns to the original melanocyte enhancer set, although the ChIP-seq signal at these loci is much lower than at regions of the original training set (Figure 4-4C, D, E, and F).

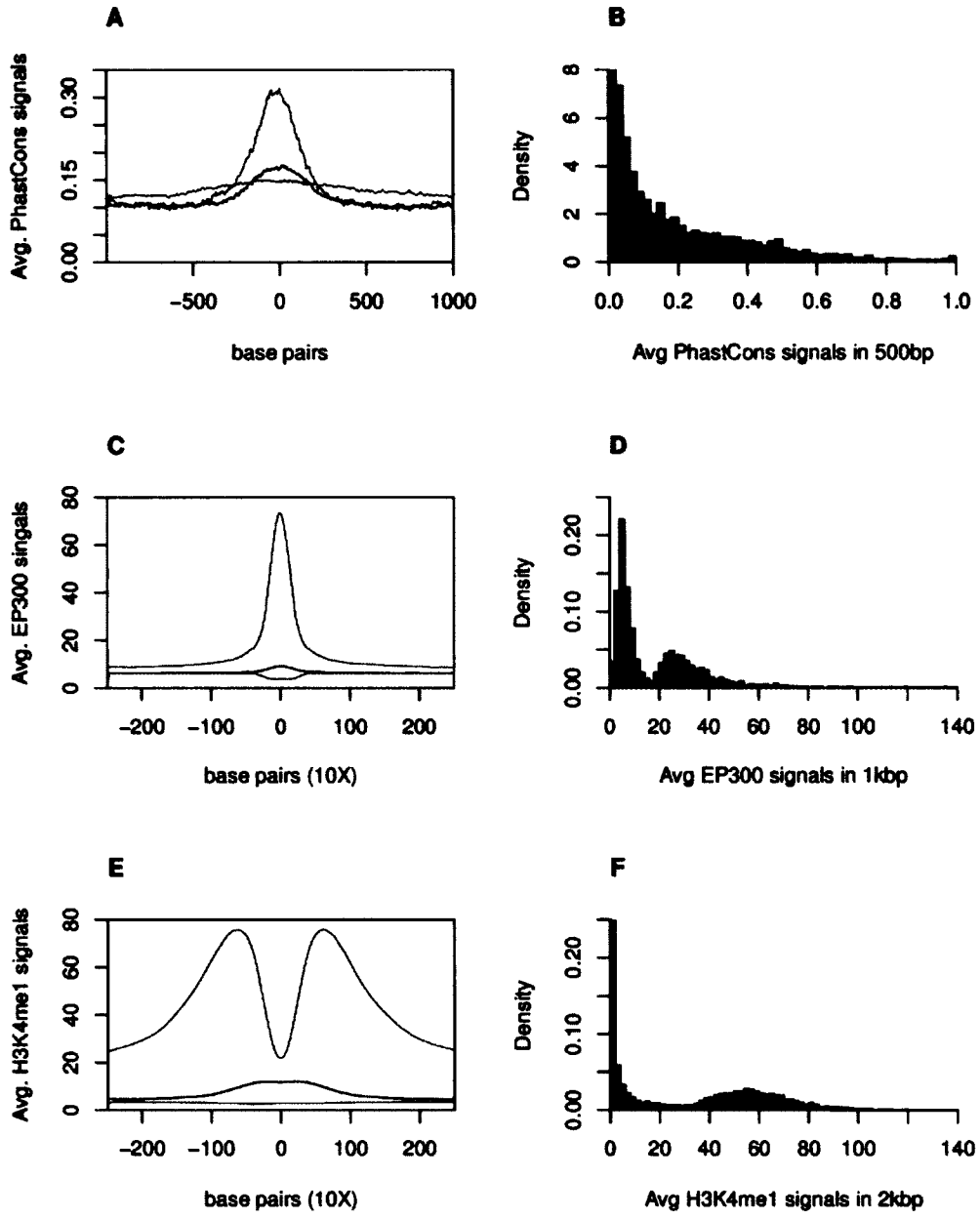



**Figure 4-4: Comparison of genomic signals in three different sets**
Average PhastCons scores (A and B), EP300 ChIP-seq signals (C and D), and H3K4me1 signals (E and F) are shown for three different sets; 2489 melanocyte enhancers (red), 7361 SVM predicted enhancers (blue), and 10000 random genomic regions (green). (A, C, and E) For each set, an average signal at each position across regions aligned at centers is shown. (B, D, and F) distributions of average signals of regions are shown. In all cases, significantly elevated signals are observed for the training set. SVM predicted regions also exhibited elevated signals to some extent analogous to the original training set, although the overall intensity is significantly low.

To further validate the predicted enhancers, our collaborators in the McCallion lab tested our top scoring kmer-SVM predictions (n=11) in luciferase assays [6]. The majority of the enhancers tested direct luciferase expression *in vitro* more than threefold higher than the minimal promoter alone (8/11; 73%). Even more significantly, several constructs direct expression more than fivefold (6/11; 55%) and even 10-fold higher (3/11; 27%), comparable to the original EP300 bound melanocyte enhancers (Figure 4-5). They also tested the enhancer activity of the top three predicted enhancers *in vivo* in transgenic zebrafish assays [6], and found that at least two of the three constructs assayed drove expression of GFP in the melanocytes of transgenic zebrafish (Figure 4-6). These multiple experimental results further validate our computational predictions as melanocyte enhancers.

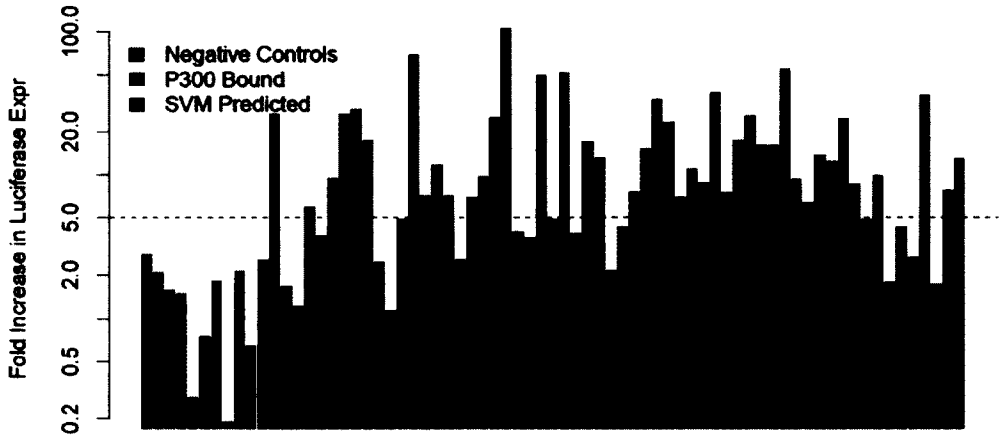



**Figure 4-5: Validation of SVM predicted enhancers *in vitro* with luciferase assay**
Luciferase assays verify that majority of the top scoring SVM predicted enhancers tested (n=11) can significantly drive expression (green) compared to negative regions (red). The expression pattern is similar to the randomly selected fifty EP300 bound regions shown as blue (Data was generated by the McCallion lab.)

We also used the SVM classifier to make predictions in the human genome using the

same approach as described above for mouse, predicting 7,788 human melanocyte enhancers. The predicted human enhancers show sequence constraint even though sequence conservation was not explicitly used in making predictions (Figure 4-7). In addition, the predicted human enhancers display elevated DNaseI hypersensitivity (DHS) in human primary melanocytes, but not in an unrelated cell type (Gm12878, a lymphoblastoid cell line) as shown in Figure 4-8, which is a feature of active enhancers [129]. Cell-type dependant DHS of our SVM predicted human melanocyte enhancers strongly suggests that the activity of the predicted enhancers is largely specific to the melanocyte lineage.

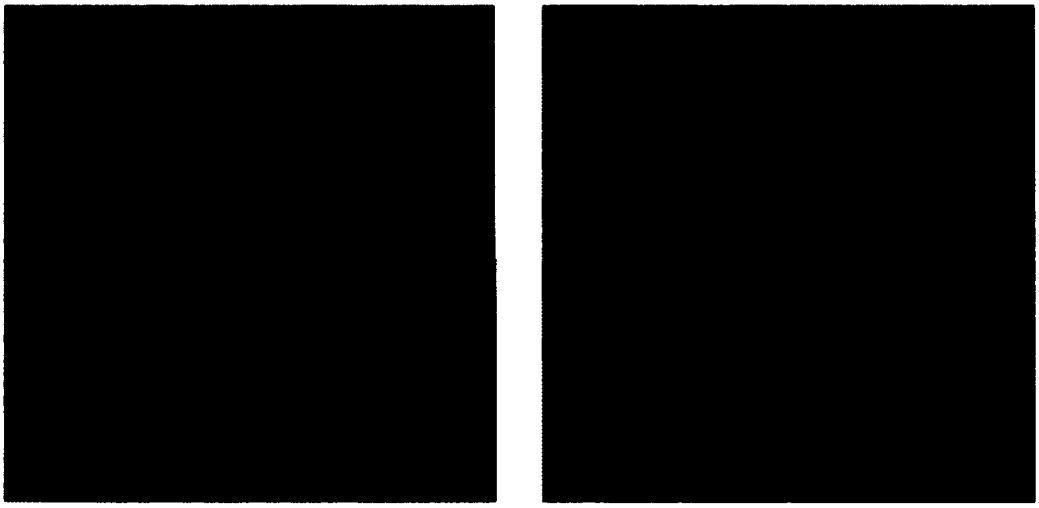


**Figure 4-6: Validation of predicted enhancers *in vivo* with transgenic zebrafish assay**
Two SVM predicted enhancers drove GFP expression in melanocytes in transgenic zebrafish assay (images were generated by the McCallion lab.)

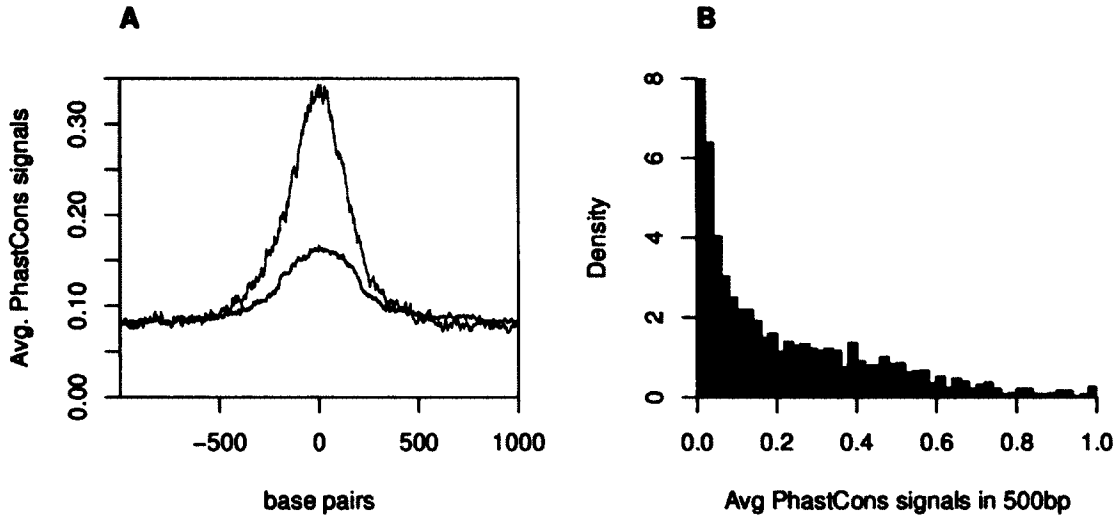



**Figure 4-7: Comparison between PhastCons scores of predicted human melanocyte enhancers**

Average PhastCons scores are shown for two different sets; 1,963 human orthologous regions of mouse melanocyte enhancers (red), and 7,788 SVM predicted human melanocyte enhancers (blue). (A) For each set, an average signal at each position across regions aligned at centers is displayed. (B) Distributions of average signals of regions are displayed. Similar to the analysis for mouse melanocyte enhancers, SVM predicted regions show elevated signals to some extent, although the overall intensity is significantly lower than the orthologous regions.

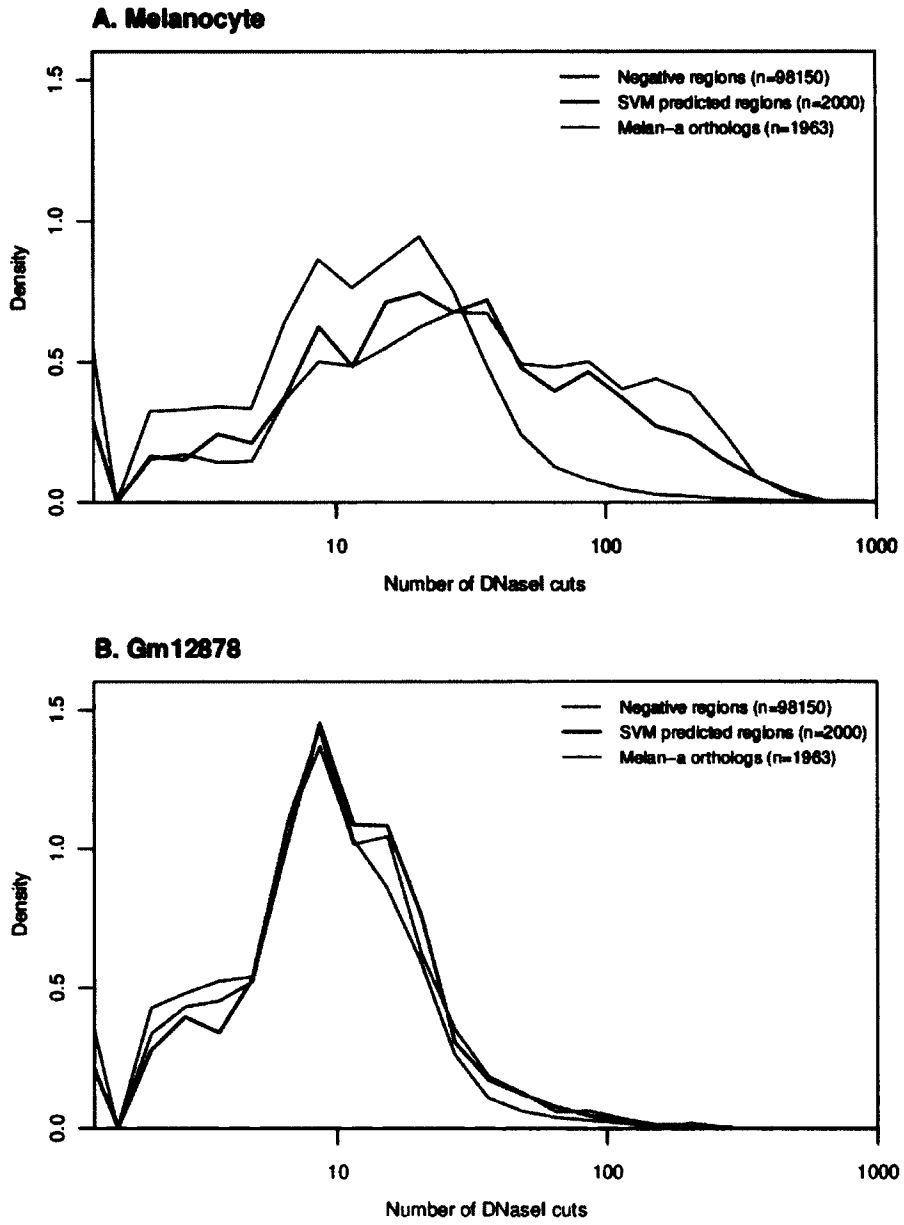



**Figure 4-8: Distributions of the number of the DNaseI cuts**
Distributions of the number of DNaseI cut of three different sets (2000 top scoring regions (blue), human orthologs of mouse melanocyte enhancers (green), and random genomic regions (red)) are plotted. (A) DNaseI Hypersensitivity measured in human primary melanocyte (B) DNaseI Hypersensitivity measured in lymphoblastoid cell line (Gm12878), as a negative control. We observe significant enrichments only in SVM predicted regions (blue) and orthologous regions (green) in the melanocyte data set, but not in the other cell type.

### 4.1.6 Discussion

In this collaborative effort, we demonstrate the validity of our kmer-SVM predictions in many different ways. One remarkable accomplishment is the successful experimental validation of the kmer-SVM predictions both *in vitro* and *in vivo*. Although we chose to test a subset of the top scoring regions by the kmer-SVM, the success rate of the validation experiments is noteworthy considering the fact that our predictions have been minimally optimized. For example, lengths and positions of predicted enhancers could be optimized by developing more sophisticated strategies for scanning the genome. Moreover, other genomic information, such as conservation scores and repeat masked regions, could also be used to reduce false positives of our predictions. Alternatively, further improvement in our predictions could be made by the development of better models. For example, incorporating synergistic cooperative binding mechanisms in our model is one of promising research areas. In conclusion, although the kmer-SVM method still has plenty room for improvement, we convincingly show via multiple lines of experimental and computational analysis that our method can accurately predict enhancers from primary DNA sequences.

## 4.2 Kmer-SVM Web Server

### 4.2.1 Summary

To make my kmer-SVM method more easily accessible to the public, I also developed a web server that provides full functionality of the kmer-SVM method in collaboration with the McCallion lab [7]. I expect that this kmer-SVM web server will aid many experimentalists in the analysis of their own genomic datasets in several ways. First, kmer-SVM methods can provide an independent measure of the quality of their genomic datasets, and can also be used to optimize thresholds and the treatment of biological replicates, as genomic datasets with reduced noise can be predicted with higher accuracy. Second, identifying recognizable TF binding sites among the most significant positive kmer-SVM features gives further confidence in the positive set of genomic regions. Third, the identification of unexpected TF binding sites among the most predictive features frequently generates novel hypotheses for subsequent experiments. Fourth, the highest scoring *k*-mers provide focused targets for mutations predicted to modulate enhancer activity in validation experiments. Finally, predictive kmer-SVM features can be used to prioritize targets among disease associated SNPs within larger haplotypes in linkage disequilibrium.

We have chosen to use the Galaxy platform [130] as a framework for our web server, but our tool can also be used as a standalone set of programs. Here we describe its use, and demonstrate how to use the kmer-SVM to extract useful information from datasets

generated by high-throughput sequencing-based experiments.

### 4.2.2 Overview of kmer-SVM Galaxy Module

Our proposed analysis pipeline to identify regulatory DNA sequence features consists of three main components: 1) Generating the positive and negative sequence sets, 2) Training the SVM classifier, and 3) Analyzing its performance and predictive sequence features. While the positive training sequence set is provided by the experimenter in the form of a BED file of coordinates or sequence data in FASTA format, including genomic coordinates, the negative set is generated by our module "Generate Null Sequence". SVM training is fairly transparent, takes the positive and negative sequence sets as input, and produces a set of kmer weights and predicted class labels as output using cross-validation. Finally the performance of the SVM classifier is summarized by ROC and Precision-Recall curves, and features are ranked by their significance. Figure 4-9 shows the general workflow generated by Galaxy's "Workflow Editing Tool." This Figure uses the actual Galaxy module names and data files, and this Workflow can also be used as a template for a typical analysis pipeline.

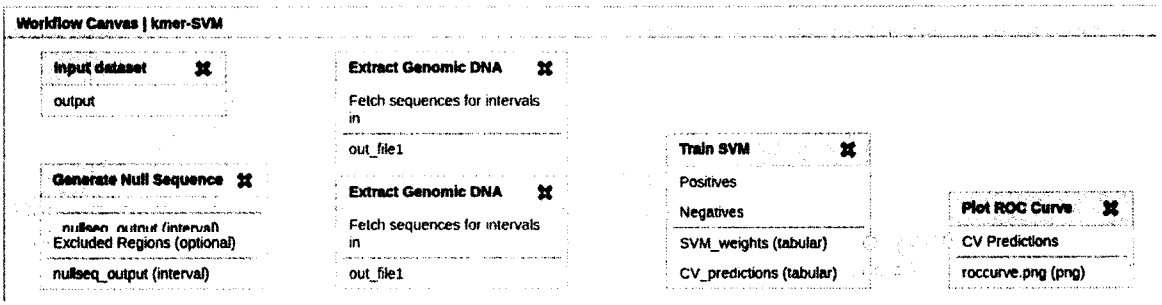



**Figure 4-9: An example workflow for the kmer-SVM module**
This workflow consists of three different components from the kmer-SVM module, "Generate Null Sequence", "Train SVM," and "Plot ROC Curve," and one built-in Galaxy module, "Extract Genomic DNA."

## 4.2.3 Details of Core Modules

### 4.2.3.1 Generation of sequence sets

My kmer-SVM classifier takes as training data a FASTA file of positive sequences obtained through ChIP-seq, DNase-seq or another experimental assay, and a negative sequence set. To ensure that the SVM identifies sequence features specific to the positive regions, it is necessary to match the GC content, length, and repeat fraction when constructing the negative set, otherwise sequence features could be predictive simply by their enrichment or absence in the biased negative set. We refer to the set of the three distributions of GC, length, and repeats in the positive set as its "sequence profile," and the "Generate Null Sequence" module matches this sequence profile for the negative set by using the following random sampling procedure. First, a positive sequence is randomly selected, and "Generate Null Sequence" samples a region from the genome for a match in terms of length, GC content and repeat fraction which does not overlap any positive sequence or existing negative sample by even one base pair. This random

selection process is then repeated until the negative set has reached the requested size. This random selection process uses a precomputed table of genomic indices that are currently provided for the mouse and human genome. The full negative sequence set then by construction closely approximates the sequence profile of the positive set. In special cases, users can exclude regions other than the input positive sequences from consideration for negative sequence generation through the "Excluded Regions" option. We recommend using a negative set which is larger than the positive set, as doing so generally improves the statistical robustness of the classifier. We allow the user to specify the size of the negative set as an integral multiple of the number of positive sequences (say 10X) in the "# of Fold-Increase field". As some positive sequences may not have exact matches in terms of GC content or repeat fraction, users can specify the percentage of GC content or repeat fractions by which a generated null sequence may differ from its corresponding positive sequence. This additional flexibility speeds the generation of the negative set, and affects how precisely the negative set sequence profile matches the positive set sequence profile. Also, distinct realizations of null sequence sets may be generated by varying the "Random Number Seed" parameter. The output of the Generate Null Sequence tool is a BED file describing the coordinates of the negative genomic intervals.

After the coordinates are specified, the actual sequences needed for SVM training are generated from the positive and negative BED file coordinates by the built-in Galaxy tool: "Fetch Sequences/Extract Genomic DNA", whose output is FASTA format DNA sequence files.

#### 4.2.3.2 SVM Training

A support vector machine (SVM) [68], [69] is a classifier which attempts to find a hyper-plane boundary in feature space that separates elements of the positive and negative sequence sets. SVMs use techniques known as 'kernels' to efficiently find this hyper-plane. A set of kernels called "string kernels" have been developed for analyses of sequence data sets and have achieved great success in computational biology [71]. "Train SVM" uses one of these string kernels, specifically, the spectrum kernel [72]. In my model, the features are the complete set of *k*-mers, and their frequencies are calculated from the input FASTA files. The training module "Train SVM" generates the normalized *k*-mer count vector for each sequence, and then finds the SVM internal parameters (support vectors) that most accurately distinguish the positive and negative sets. Currently, "Train SVM" supports two kernels: the spectrum kernel (using a single length kmer) and the weighted spectrum kernel (using a user specified range of k's, with equal weighting). In both cases reverse complement *k*-mers are treated as separate instances of the same feature. This module was implemented by using SVM shogun toolbox [96].

"Train SVM" performs two tasks: it generates a set of ranked kmer-SVM weights, and it generates a set of class predictions using cross-validation. A given *k*-mer's score can be thought of as a measure of the degree to which that *k*-mer contributes to the discriminatory power of the classifier. The weights are output to the table labeled "Weights."

#### 4.2.3.3 Cross-validation

As is standard in machine-learning, cross-validation is used to assess classifier performance. The initial positive and negative sets are randomly partitioned into $n$ distinct sets (for $n$-fold CV), and the ROC and PR performance of each test set is generated using a classifier trained on the other $n$-1 sets. The number of CV sets is a parameter which can be specified by the user. This is repeated for all $n$ partitions such that in the end each partition is used for both training and test-set scoring. The result of this process is the set of scores for test-set sequences in each round of cross-validation, output in the table labeled "Predictions."

Three parameters for SVM learning are adjustable ($k$, $C$, and $E$). If the spectrum kernel is used, $k$ specifies a single $k$-mer length, while if the weighted spectrum kernel is used minimum and maximum values for $k$ must be set. Using a single $k$ is somewhat easier to interpret in the beginning, as the vocabulary is simpler. Using a range of $k$ values does have the advantage that similar $k$-mers of slightly varying length and composition should all receive significant weights, increasing confidence in interpretation. Also using a range (e.g. 5-8) usually performs incrementally better than a single $k$ in terms of overall classification accuracy.

The SVM maximizes the margin between the positive and negative sequences while simultaneously minimizing errors (sequences on the wrong side of the boundary). The relative importance of misclassification error is weighted by the regularization parameter, $C$. In practice this affects over-fitting. A small $C$ will result in less over-fitting of the SVM at the expense of slightly greater training classification error. With unbalanced

positive and negative set sizes it is often recommended to use a separate regularization parameter for positive and negative sequences, reflecting the relative importance of errors. We specify this using an additional parameter *Positive_Set_Weight* or *PSW*. The regularization parameter for the positive set is $C * PSW$, while for the negative set it is $C$. The default setting is $PSW = 1+\log(N/P)$, which weighs positives more heavily when the negative set is very large. In practice, results are insensitive to $C$ and *PSW* unless there is a significant imbalance between the positive and negative set sizes. Finally, the precision parameter $E$ constrains the precision of the SVM classifier. Increasing $E$ results in a reduced number of support vectors, and can lead to a more robust classifier by reducing the requirements on the accuracy of the classifier on the training set [131]. In practice, the results should be insensitive to the choice of $E$, and the default value is recommended.

#### 4.2.3.4 Interpretation of kmer SVM weights

The output of SVM training is a list of $k$-mer weights, and it is the weighted sum of normalized $k$-mer counts in a sequence which determines the predicted class. In biological terms, the presence of $k$-mer with large positive weights significantly increases a sequence's likelihood of being positive (e.g., being an enhancer, or being bound by a TF in a specific cell type). Large negative weights are equally informative, as their absence significantly increases the probability of being positive (e.g. a binding site for a transcriptional repressor). The weights file output by "Train SVM" lists all $k$-mer and their corresponding scores. The SVM weight is a continuous valued quantity and large absolute value is a direct measure of significance. It is the scores with large absolute

values that will be of particular value to the biologist. The transcription factors binding the highest and lowest scoring *k*-mer, if previously studied, can be found using database matching programs such as TOMTOM [101], using the UniPROBE TRANSFAC, and JASPAR databases [83], [84], [100]. We have found that large positive scoring *k*-mer are usually recognizable as transcription factor binding sites known to be important in the cell type of interest, while large negative scoring *k*-mer have identified an important role for repressors in previously unknown contexts [5].

#### 4.2.3.5 ROC Analysis

The area under the curve (AUC) of the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves are measures of the accuracy of the classifier. The area under the ROC curve corresponds to the probability that a randomly selected positive sequence will score higher than a randomly selected negative sequence. For each possible SVM score threshold, we calculate the true positive rate ($TPR=TP/(TP+FN)$, or sensitivity) and false positive rate ($FPR=FP/(FP+TN)$, or 1-specificity) at this threshold. The ROC, or Receiver Operating Characteristic curve, plots *TPR* vs. *FPR*. The PR, or Precision-Recall curve, plots Precision vs. Recall, where, $Precision= TP/(TP+FP)$ and $Recall=TPR$.

The ROC and PR curves are slightly different measures of the classification performance of the trained SVM: the ROC emphasizes true and false positive rates, while the PR curve emphasizes true positive predictions. This difference results in the ROC possibly overestimating the accuracy of a classifier for data sets with large imbalances in

the positive and negative class sizes, as is typical of genomic predictions with large negative sets. The PR curve is more appropriate in the case of large negative sets, yielding more accurate evaluations of classifier performance, because it directly assess the accuracy of positive predictions.

### 4.2.4 Details of Auxiliary Modules

#### 4.2.4.1 Score Sequences of Interest

Once the SVM is trained, in addition to classifying the cross-validation test-sets, it can be used to score any sequence of interest. An additional detail is that while the rank of the SVM scores is significant, the scale of the SVM scores is not. We therefore turn this SVM score into a probability that the element is positive, by reporting the posterior probability that each sequence is in the positive class, using the algorithm described in [132], [133]. "Score Sequences of Interest", takes as input a set of sequences in FASTA format and outputs the SVM score and posterior probability. Parameters to produce this posterior probability are included in the weight table output by "Train SVM." "Score Sequences of Interest" can also be used to make genome wide predictions using the module "Split Genome," which splits a genome into chunks of a length $c$ bp that overlap each other by $v$ bp. The results of "Split Genome" can then be used as input for "Score Sequences of Interest."

#### 4.2.4.2 Sequence Profiles

As discussed above, the sequence profiles, or distributions of length, GC content and repeat fraction content in the positive and negative sequences are matched by "Generate Null Sequence." It may be useful to compare the sequence profiles of other sets of genomic intervals, so we have provided an additional module to perform this analysis. For a given BED file, this module calculates and reports the sequence profile of the regions specified by these coordinates.

### 4.2.5 Examples

#### 4.2.5.1 Prediction of ESRRB bound regions in mouse ES cells

To take a specific example, we first consider the ChIP-seq dataset of Chen *et al.* [91], who identified binding loci of transcription factors in mouse embryonic stem (ES) cells. As an example, we analyze their ChIP-seq data for ESRRB (estrogen-related-receptor beta) known to play a role in maintaining the pluripotency of ES cells [134]. Because the ESRRB bound regions reported by Chen *et al.* [91] were short (10-30 bp), we extended from the midpoint of these regions and used 100 bp elements as the positive sequence set. Following the workflow in Figure 4-9, we then used "Generate Null Sequence" to produce a 10x negative set. The top five positive and negative *k*-mers reported by "Train SVM" are shown in Table 4-2. For the purpose of direct comparison, the ESRRB PWM found and reported by Chen *et al.* [91] is also shown in Figure 4-10. As expected the top *k*-mers span the core motif of the ESRRB binding site, but interestingly, several SVM predicted *k*-mers contribute to the specificity of the ESRRB. For example AAGGTCA

(1st), AGGTCA (2nd) and CAAGGT (3rd) , etc., have large positive weights, but AGGTCC and AGGTCT have large negative weights, showing that A or G is allowed in the binding site at the 11th position of the PWM, but that C and T are not. This subtlety is not reflected in the PWM found by Weeder [135], the motif discovery algorithm used in Chen *et al.* [91].

| 6-mers | Revcomp | SVM Scores |
|---|---|---|
| **Top Positive 6-mers** | | |
| AAGGTC | GACCTT | 10.05 |
| AGGTCA | TGACCT | 8.47 |
| ACCTTG | CAAGGT | 5.33 |
| AGGTCG | CGACCT | 5.17 |
| GGTCAA | TTGACC | 4.01 |
| **Top Negative 6-mers** | | |
| GCAATA | TATTGC | -2.05 |
| TGACCA | TGGTCA | -3.33 |
| AAGGTA | TACCTT | -4.23 |
| AGACCT | AGGTCT | -4.55 |
| AGGTCC | GGACCT | -4.98 |

**Table 4-2: Predictive 6-mers of ESRRB binding sites in ES cells**


**Figure 4-10: the ESRRB PWM model**
This PWM model is previously found and reported by Chen *et al.* [91] using Weeder algorithm [135].

Following the workflow in Figure 4-9, we generated the ROC and PR curves for Chen's ESRRB dataset as shown in Figure 4-11. These curves are typical of an accurate classifier, and we obtained summary statistics of AUC ROC = 0.921 and AUC PR = 0.74 for this dataset.

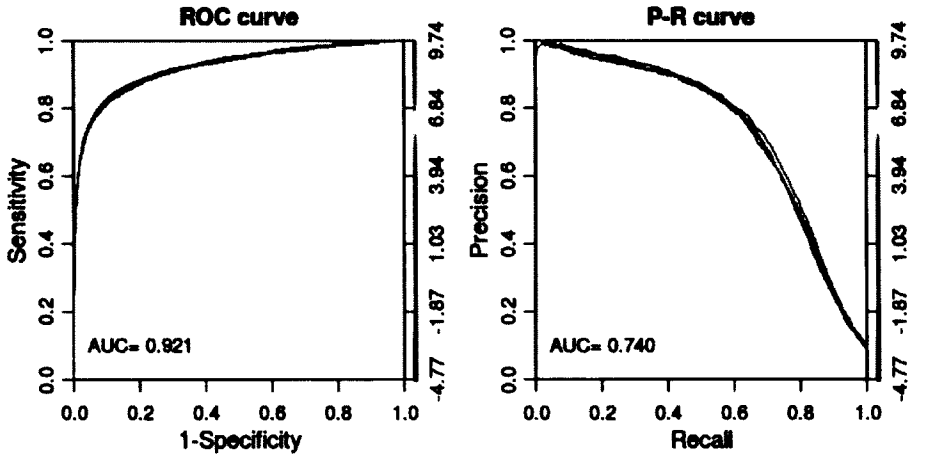



**Figure 4-11: Classification result of ESRRB binding sites in ES cells**
ROC (*Left*) and Precision-Recall (*Right*) curves for a kmer-SVM trained on ESRRB bound genomic loci in ES cells vs. 10X random genomic sequence are presented.

#### 4.2.5.2 Prediction of distinct Glucocorticoid Receptor bound regions in 3134 and AtT20 cells

We next show how our kmer-SVM can be applied to identify sequence features responsible for directing the binding of a single TF to different genomic locations in distinct tissues, developmental states, or cell lines. As an example, John *et al.* [136] investigated the genomic binding of the Glucocorticoid Receptor (GR) TF in response to hormone stimulation in two divergent cell lines. Specifically, glucocorticoid receptor (GR) binding was profiled via ChIP-seq on a mouse mammary adenocarcinoma derived cell line (3134) and mouse pituitary (AtT20) cells. The binding of GR in these two cell lines were largely at non-overlapping genomic loci. John *et al.* [136] showed that the consensus GR binding element (GRBE) was present in both 3134 and AtT20 bound regions, but that distinct sets of accessory sequence motifs were detected in the two cell lines, including binding sites for AP1, AML1, HNF3, TAL1 and NF1 (Figure 4-12).

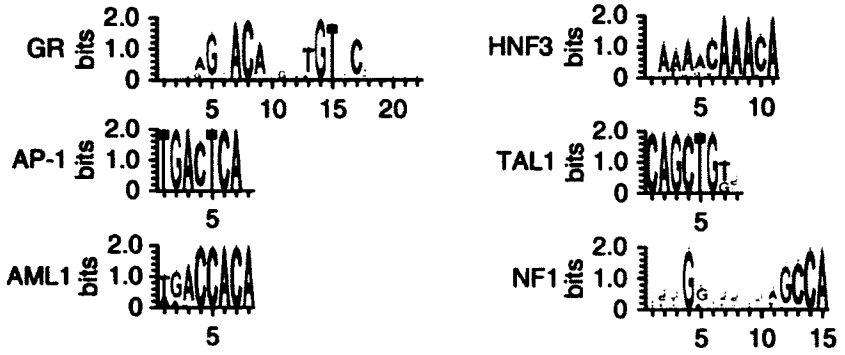



**Figure 4-12: PWM models of accessory TFBSs for GR binding**
The accessory factor binding sites of GR which have been previously found and reported by John *et al.* [136] are shown; AP1 and AML1 are specific for 3134 cell type, whereas HNF3, AML1, and AP1, are specific for atT20 cell type.

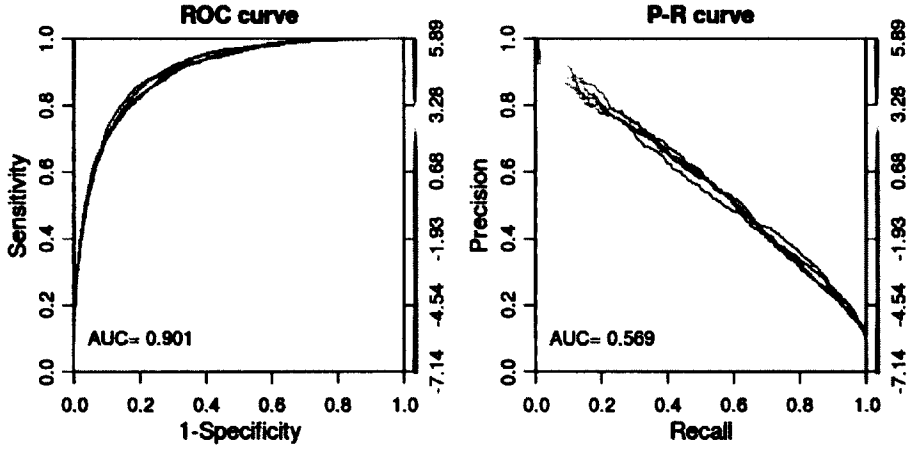



**Figure 4-13: Classification results of GR bound loci in 3134 cells**
ROC (*Left*) and Precision-Recall (*Right*) curves for a kmer-SVM trained on GR bound genomic loci in 3134cells vs. 10X random genomic sequence are presented.

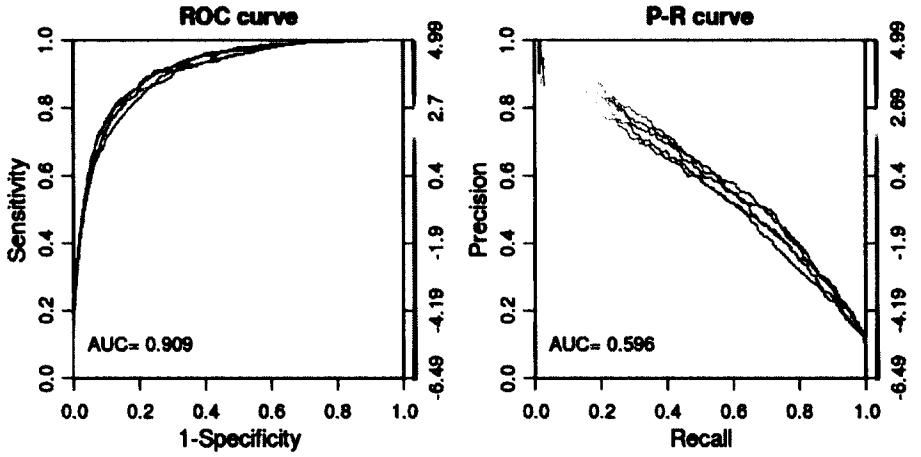



**Figure 4-14: Classification results of GR bound loci in atT20 cells**
ROC (*Left*) and Precision-Recall (*Right*) curves for a kmer-SVM trained on GR bound genomic loci in 3 atT20 cells vs. 10X random genomic sequence are presented.

We followed the Galaxy pipeline described above to train a kmer-SVM on the ChIP-seq GR bound loci in 3134 cells vs. 10X random genomic sequence, and separately on GR bound loci in AtT20 cells vs. 10X random genomic sequence, using the coordinates in John *et al.* [136] as positive set input. Our kmer-SVM classifier achieved an AUROC of 0.901 and AUPRC of 0.569 (Figure 4-13) in 3134 cells, and AUROC of 0.909 and AUPRC of 0.596 in the AtT20 cell line (Figure 4-14), indicating that GR binding in both cell lines is highly predictable based on sequence. The top then positive and negative weight *k*-mers for each cell line are shown in Table 4-3, recovering *k*-mers that span the GRBE and binding sites for accessory factors reported in John *et al.* [136] (Figure 4-12). While high scoring *k*-mers matching the GRBE consensus were found in both cell lines, the accessory factors are specific to each cell line. In 3134 cells the top two ranking *k*-mers both match AP-1, and the 8$^{th}$ and 9$^{th}$ highest *k*-mers in 3134 cells matched AML1. Our kmer-SVM also identified TEAD1 as the 5$^{th}$ most important *k*-mers (ACATTC), a

binding site not found in John *et al.* [136]. In addition, four of the most negative *k*-mers match the binding site for ZEB1 or Snail, a common negative sequence feature in our analysis [5], indicating that the absence of ACCT or AGGT is predictive for GR bound regions. Thus we hypothesize that either the presence of a ZEB1 binding site would directly inhibit the binding of GR, presumably through the binding of ZEB1 or another factor that binds specifically to this site. In other cases, this binding site could otherwise disrupt the normal function of the enhancer elements, and is thus required to be absent.

| 3134 Cells | | | | AtT20 Cells | | | |
|---|---|---|---|---|---|---|---|
| **6-mers** | **Revcomp** | **SVM Score** | **TF Match** | **6-mers** | **Revcomp** | **SVM Score** | **TF Match** |
| **Top Positive 6-mers** | | | | **Top Positive 6-mers** | | | |
| GACTCA | TGAGTC | 7.23 | AP1 | GGAACA | TGTTCC | 5.36 | GRBE |
| GAGTCA | TGACTC | 6.93 | AP1 | CAGATG | CATCTG | 5.33 | TAL1 |
| GGAACA | TGTTCC | 6.10 | GRBE | CAGCTG | CAGCTG | 4.79 | TAL1 |
| AGAACA | TGTTCT | 4.99 | GRBE | GTAAAC | GTTTAC | 4.56 | HNF3 |
| ACATTC | GAATGT | 4.80 | TEAD1 | CAAACA | TGTTTG | 4.43 | HNF3 |
| GGTACA | TGTACC | 4.71 | GRBE | TGCCAA | TTGGCA | 4.20 | NF1 |
| ATGTTC | GAACAT | 4.67 | GRBE | GCAAAC | GTTTGC | 4.09 | HNF3 |
| ACCACA | TGTGGT | 4.64 | AML1 | ACGTCA | TGACGT | 3.99 | CREB |
| AACCAC | GTGGTT | 4.61 | AML1 | AGAACA | TGTTCT | 3.97 | GRBE |
| AGTACA | TGTACT | 4.58 | GRBE | ATGACG | CGTCAT | 3.97 | CREB |
| **Top Negative 6-mers** | | | | **Top Negative 6-mers** | | | |
| AGGTAA | TTACCT | -2.34 | ZEB1 | ATGTAG | CTACAT | -2.20 | |
| CCTATA | TATAGG | -2.39 | | TTATAA | TTATAA | -2.20 | |
| GTGCAC | GTGCAC | -2.43 | | ATATAT | ATATAT | -2.24 | |
| AGACCC | GGGTCT | -2.46 | | CTTATA | TATAAG | -2.26 | |
| AACTCA | TGAGTT | -2.62 | AP1-var | AAAGTT | AACTTT | -2.34 | |
| CACTCA | TGAGTG | -2.80 | AP1-var | GCCCAC | GTGGGC | -2.36 | |
| CGAGAC | GTCTCG | -2.95 | | CATATA | TATATG | -2.39 | |
| ACAGGT | ACCTGT | -3.29 | ZEB1 | GGACCC | GGGTCC | -2.42 | |
| CACCTG | CAGGTG | -3.88 | ZEB1 | AGGGTC | GACCCT | -2.48 | |
| CAGGTA | TACCTG | -4.29 | ZEB1 | CAGGTA | TACCTG | -2.88 | ZEB1 |

**Table 4-3: Predictive 6-mers of GR binding in 3134 cells and AtT20 cells**

In the AtT20 cells, a separate set of accessory sites is found: the 4$^{th}$, 5$^{th}$ and 7$^{th}$ most positive *k*-mers match HNF3, while the 2$^{nd}$ and 3$^{rd}$ match TAL1. The 6$^{th}$ ranked *k*-mers

matched NF1. The 8th and 10th ranked *k*-mers match CREB, not reported in John *et al.* [136]. In summary, our analysis uncovered most of the accessory factors described in the original study, but also identifies novel positive and novel negative binding sites. Further, we demonstrate that these features are predictive, in the sense that these features can be used to accurately classify the positive and negative regions, and are not simply over- (or under-) represented in one of the sets.

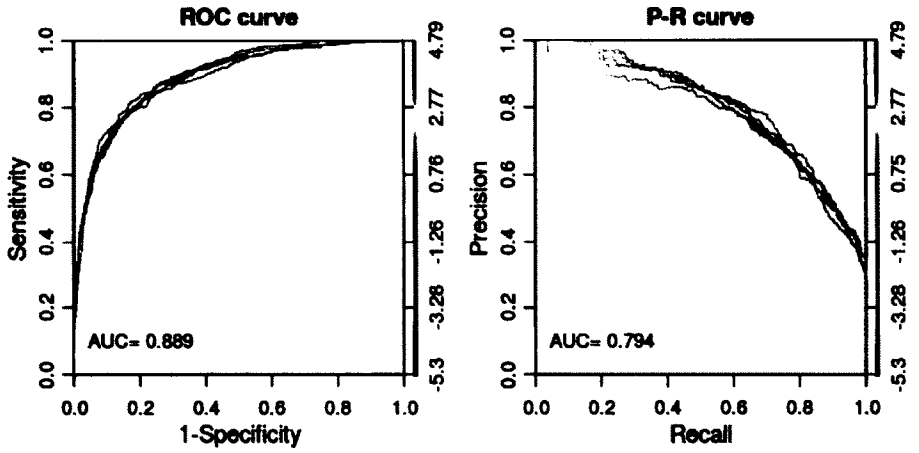



**Figure 4-15: Classification results of GR bound regions in AtT20 cells vs. GR bound regions in 3134 Cells**
ROC (*Left*) and Precision-Recall (*Right*) curves for a kmer-SVM trained on GR bound regions in AtT20 cells vs. GR bound regions in 3134 Cells.

We next demonstrate that our kmer-SVM is able to directly distinguish the GR bound regions in 3134 cells from the GR bound regions in AtT20 cells from DNA sequence. In this case we do not use random genomic sequence as the negative set, but instead train a kmer-SVM using the AtT20 regions as the positive sequence set, and the 3134 regions as the negative sequence set. The ROC and PR curves are shown in Figure 4-15, yielding AUROC of 0.889 and AUPRC of 0.794. Thus DNA sequence is sufficient to distinguish

the cell specific binding of GR. Now, since both sets are bound by GR, the *k*-mers weights shown in Table 4-4 do not include the GRBE, since it is present in both sets. The distinguishing features are now binding sites for the GR accessory factors. The *k*-mers CAGGTG (ZEB1) which was negative for 3134 vs. random is now the most positive *k*-mers for AtT20 vs. 3134. The other positive *k*-mers match the AtT20 specific accessory factors TAL1 and HNF3. The negative weight *k*-mers are the 3134 specific accessory factors AML1 and AP1. This demonstrates that these accessory sequence elements are predictive of the tissue specific binding of GR, because the sequence information in the accessory factor binding sites is sufficient to distinguish GR binding in these two contexts. We emphasize that this is a much stronger statement than simply observing the enrichment of distinct sequence features in the two cases: we claim further that these sequence features are sufficient to specify which GR binding sites will be occupied in each tissue. This differential occupancy is determined by the presence of binding sites for accessory factors which can be identified from the *k*-mers weights.

| 6-mers | Revcomp | SVM Scores | TF Match |
|---|---|---|---|
| **Top atT20 specific 6-mers** | | | |
| CACCTG | CAGGTG | 5.95 | ZEB1 |
| CAGATG | CATCTG | 4.24 | TAL1 |
| CAAACA | TGTTTG | 4.15 | HNF3 |
| GTAAAC | GTTTAC | 4.13 | HNF3 |
| ATTTAC | GTAAAT | 3.39 | HNF3 |
| CAGCTG | CAGCTG | 3.36 | TAL1 |
| **Top 3134 specific 6-mers** | | | |
| CCACAA | TTGTGG | -2.62 | AML1 |
| AACCAC | GTGGTT | -2.89 | AML1 |
| AGTCAT | ATGACT | -3.05 | AP1 |
| GACTCA | TGAGTC | -3.75 | AP1 |
| GAGTCA | TGACTC | -3.77 | AP1 |
| ACCACA | TGTGGT | -4.39 | AML1 |

**Table 4-4: Predictive 6-mers of distinguishing GR bound loci in atT20 from 3134**

#### 4.2.5.3 Prediction of distinct EWS-FLI bound regions in EWS502 and HUVEC cells

While the previous example showed that binding of a sequence specific TF to different loci in different tissues was predictable from DNA sequence, we now turn to an example where a wild-type and mutant TF were shown to bind distinct regions, and that this differential binding is also predictable from DNA sequence. Most Ewing-Sarcoma tumors harbor a mutation which creates an oncogenic chimerical EWS-FLI transcription factor by fusing the transactivation domain of EWS to the DNA-binding domain of FLI. Patel *et al.* [137] showed that this chimeric EWS-FLI TF targets different genomic regions in tumor cells and in non-tumor cells, and that additionally the wild-type protein FLI1 binds to largely the same regions as the fusion protein in non-tumor cells. Specifically, the authors assayed binding in the EWS502 cell line (derived from a Ewing Sarcoma tumor) and primary human endothelial cells (HUVEC). They reported a preferential binding for regions containing repeats of the tetranucleotide GGAA by EWS-FLI in both EWS502 and HUVEC cells (although the tumor cell line showed a greater enrichment). Additionally, binding of EWS-FLI in HUVEC cells was shown to be enriched in ETS, AP1, and GATA motifs, but that these accessory motifs were largely absent from the EWS-FLI bound regions in EWS502 cells.

To analyze these datasets, we used as positive sets the ChIP-seq regions in Patel *et al.* [137] bound by EWS-FLI in EWS502 cells and HUVEC cells, and we generated separate 10X negative sets for each cell line. After training the kmer-SVM, in EWS502 cells the AUROC was 0.965 and AUPRC was 0.972 [[figure]], and in HUVEC cells the AUROC for this dataset was 0.962 and AUPRC was 0.961 [[figure]], again showing that the cell

line specific binding of the EWS-FLI TF is predictable from primary DNA sequence features. In this case, the training data was optimized for length by the peak-calling algorithm ZINBA [138] which may account for the extremely high AUC. Another possible factor is that the repeat fraction in these positive sets is relatively high.

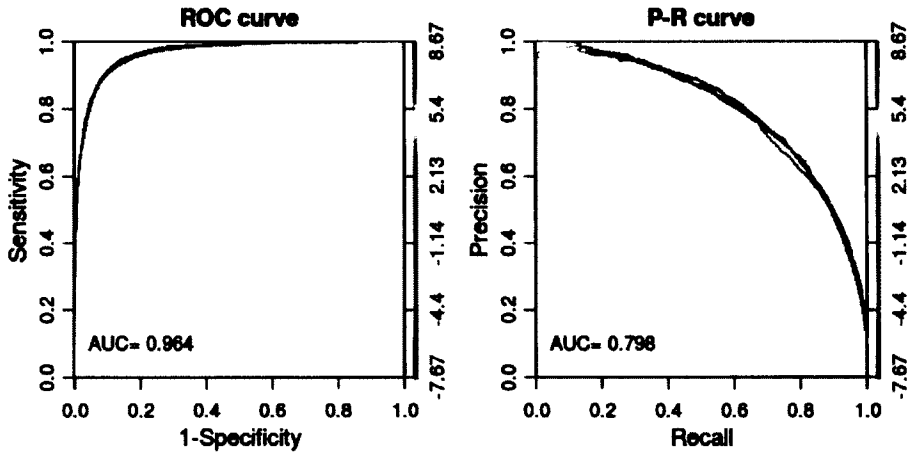



**Figure 4-16: Classification results of EWS-FLI bound regions in EWS502**
ROC (*Left*) and Precision-Recall (*Right*) curves for a kmer-SVM trained on EWS-FLI bound loci in EWS502 cells vs. 10X random genomic regions are presented.

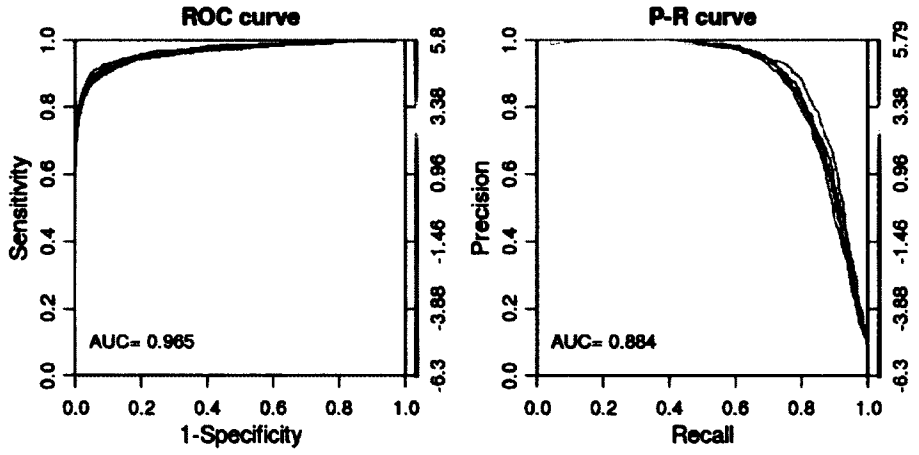



**Figure 4-17: Classification results of EWS-FLI bound regions in HUVEC**
ROC (*Left*) and Precision-Recall (*Right*) curves for a kmer-SVM trained on EWS-FLI bound loci in HUVEC cells vs. 10X random genomic regions are presented.

Our method finds some motifs common to both cell lines. Positive sequence features reflect both the ETS motif recognized by FLI1 and the repetitive structure reported by Patel *et al.* [137], with the ETS motif GGAA as part of the highest ranked *k*-mers in both cell lines, as shown in Table 4-5. Negative weight *k*-mers are again found to be significant. *k*-mers which disrupt the repetitive GGAA structure (e.g. TGGAAG) score negatively in both cell lines, but more negatively in EWS502 cells. Notably, many of the most negative *k*-mers for both cell lines contain AGGT, again emphasizing the importance of the absence of ZEB1 or Snail repressor family binding sites for EWS-FLI binding or function.

| EWS502 | | | | HUVEC | | | |
|---|---|---|---|---|---|---|---|
| **6-mers** | **Revcomp** | **SVM Score** | **TF Match** | **6-mers** | **Revcomp** | **SVM Score** | **TF Match** |
| **Top Positive 6-mers** | | | | **Top Positive 6-mers** | | | |
| CCGGAA | TTCCGG | 8.97 | Ets | CCGGAA | TTCCGG | 9.83 | Ets |
| CAGGAA | TTCCTG | 7.49 | Ets | CAGGAA | TTCCTG | 9.70 | Ets |
| ATTTCC | GGAAAT | 6.93 | Ets | ATTTCC | GGAAAT | 9.58 | Ets |
| CGGAAA | TTTCCG | 6.56 | Ets | ACTTCC | GGAAGT | 9.45 | Ets |
| AGGAAG | CTTCCT | 6.23 | Ets | CGGAAA | TTTCCG | 8.92 | Ets |
| CGGAAG | CTTCCG | 5.72 | Ets | AGGAAG | CTTCCT | 8.82 | Ets |
| ACTTCC | GGAAGT | 5.18 | Ets | GACTCA | TGAGTC | 7.53 | AP1 |
| ACAGGA | TCCTGT | 5.17 | | CGGAAG | CTTCCG | 7.09 | Ets |
| ACCGGA | TCCGGT | 4.26 | Ets | AGGAAA | TTTCCT | 6.70 | |
| AGGAAA | TTTCCT | 4.14 | | ACAGGA | TCCTGT | 6.56 | |
| **Top Negative 6-mers** | | | | **Top Negative 6-mers** | | | |
| CAAGGA | TCCTTG | -1.84 | TEAD1 | ACGTAT | ATACGT | -2.26 | |
| AGTCCA | TGGACT | -1.92 | | ACCTGC | GCAGGT | -2.27 | ZEB1 |
| AGGTGA | TCACCT | -1.97 | ZEB1 | AGACTC | GAGTCT | -2.28 | |
| CTGGAC | GTCCAG | -1.99 | | AGGTGA | TCACCT | -2.34 | ZEB1 |
| TGGAAA | TTTCCA | -2.29 | Ets-var | CTTCCA | TGGAAG | -2.58 | Ets-var |
| ACAGGT | ACCTGT | -2.33 | ZEB1 | AGGAAT | ATTCCT | -2.82 | |
| AAGGTG | CACCTT | -2.60 | | AGGTAA | TTACCT | -2.90 | ZEB1 |
| AGGAAT | ATTCCT | -2.61 | TEAD1 | ACAGGT | ACCTGT | -3.10 | ZEB1 |
| CTTCCA | TGGAAG | -2.67 | Ets-var | CACCTG | CAGGTG | -3.24 | ZEB1 |
| CAGGTA | TACCTG | -3.28 | ZEB1 | CAGGTA | TACCTG | -4.45 | ZEB1 |

**Table 4-5: Predictive 6-mers of EWS-FLI binding in EWS502 and HUVEC**

Cell line specific *k*-mers recover the AP1 motif reported in Patel *et al.* [137], and a potentially novel role for TEAD1. The HUVEC specific accessory factor AP1 is found as a high scoring motif in HUVEC cells but not EWS502 cells. Two highly negative *k*-mers in EWS502 cells correspond to the binding site for TEAD1. TEAD1 has been implicated in tumor suppression and growth control, and because the absence of TEAD1 binding sites is predictive of EWS-FLI binding in EWS502 cells but not HUVEC cells, it is tempting to speculate that TEAD1 binding would disrupt EWS-FLI binding in EWS502 cells but not in HUVEC cells.

### 4.2.6 Discussion

We have shown that our kmer-SVM model as offered in this web server is able to find predictive sets of DNA sequence features in several different genomic datasets, and can be used to assess and explore the genomic data and generate testable hypotheses for subsequent biological analysis. Using the existing sequence tools and pipeline flow of the Galaxy platform has greatly facilitated the ease of distribution of this tool. The examples we have highlighted above, in addition to our previous results on mouse EP300 bound enhancers [5] and melanocyte enhancers [6], emphasize several key benefits of our kmer-SVM analysis. Using our web server, users can find the essential sequence features which distinguish a set of experimentally determined genomic regions from random sequence, and identify key accessory factors and repressive elements for biological interpretation and follow up investigations. In addition, users can use the kmer-SVM to score alternative sequence sets or entire genomes to make predictions of the activity of these

regions in the relevant context.

To the best of my knowledge, there is only one web server available that provides tools [96] with some similarity to the kmer-SVM. The web server, whose URL address is http://galaxy.raetschlab.org, offers simple SVM functions including several string kernels as well as common kernels, such as linear and Gaussian. It also provides means to evaluate prediction performance using ROC and PR curves. This server, however, is mainly intended for general use of SVMs by users with a certain level of computational experience. In contrast, our kmer-SVM is specifically designed to allow biologists with no prior machine learning expertise to quickly and rigorously analyze regulatory sequence data sets. To do so, our tool incorporates modules with functionality required for regulatory sequence analyses and takes into account the specific properties of regulatory elements. First, we modified the spectrum kernel function to account for the fact that TFs bind to double stranded DNA. We not only count an exact *k*-mer, but also count its reverse complement *k*-mer. Redundant *k*-mers are then eliminated from the final feature set to remove the possible bias caused by double counting. Second, we offer a module that generates negative sequence sets to match the distribution of sequence length, GC content, and repeat fraction of the corresponding positive sets. Third, we provide a means to explain results of SVMs by calculating SVM weights of *k*-mers from a list of support vectors, the primary output of SVM training. We hope that release of our kmer-SVM method in this form will facilitate open availability and ease of access to the broader research community.

# 5 Gkm-SVM: An Improved Framework For Enhancer Prediction

## 5.1 Introduction

In the second part of my dissertation, I further improved the original kmer-SVM method in collaboration with my lab colleagues. A TFBS is usually described by a set of sequences with some gaps (as non-informative positions) and a typical length of a TFBS is ranging between 8 and 20. Although the original kmer-SVM method can capture TFBSs longer than the *k*-mer length by tiling across TFBSs, this loses some spatial information in the binding site, and overall classification accuracy is also significantly impaired when long TFBSs are important predictive features. One major issue of the kmer-SVM using longer *k*-mer frequencies as features is that they generate extremely sparse feature vectors (i.e. most *k*-mers simply do not appear in a training sequence and they thus receive zero frequencies), which has the potential to cause a severe overfitting problem. Therefore, the original kmer-SVM method was limited in practice to *k*-mer lengths of 8~10.

In this study, we successfully developed more robust classifiers with higher accuracy by using estimated frequencies of long *k*-mers (typically between 10 and 20) as features in the framework of our SVM models [8], [9]. Here, we refer to this new method as gkm-SVM. We will briefly introduce the *k*-mer frequency estimation method, and then show

that our new gkm-SVM method significantly outperforms the original kmer-SVM on CTCF ChIP-seq dataset and our previously analyzed EP300 forebrain ChIP-seq dataset. This much improved performance of gkm-SVM will serve as the basis of the development of more sophisticated fine scale feature detection methods discussed in later chapters.

## 5.2 Methods

### 5.2.1 Robust *k*-mer Frequency Estimation Using Gapped *k*-mers

To obtain longer *k*-mer frequencies with sufficient statistical power, we determined to estimate them from the gapped *k*-mer frequency distribution of a training sequence. A key assumption in this approach is that the full set of gapped *k*-mers provides a sufficient description of the relevant set of TFBSs. It is also motivated by the structural molecular observation that most TFs make direct contacts with the DNA binding site in a relatively small set of base pair positions, but potentially spread across a wider (15-20 bp) stretch of DNA. Based on these observations, we hypothesize that estimated *k*-mer frequencies from gapped *k*-mers might describe TFBSs significantly better than exact *k*-mer frequencies. Here, we consider the mapping from *k*-mers to gapped *k*-mers. Among the all possible sets of *k*-mer frequencies that could be produced from the same gapped *k*-mer frequency distribution, we developed a method to estimate the "most likely" *k*-mer frequency set using a maximum likelihood estimation algorithm with a Gaussian prior. A mapping from gapped *k*-mer frequencies to estimated *k*-mer frequencies can be

formalized as follows.

Let $x_j$ be an estimated $j^{th}$ $k$-mer frequency, and let $\mathbf{y} = (y_1, \ldots, y_M)$ be a gapped $k$-mer frequency vector directly observed from a sequence, where $M$ is the number of all gapped $k$-mers with $u$ informative positions ($M = 4^u\binom{k}{u}$). We then consider a mapping $\mathbf{w}^{(j)} = (w_1, \ldots, w_M)$ that determines $x_j$ from $\mathbf{y}$.

$$x_j = \mathbf{w}^{(j)} \cdot \mathbf{y} = \sum_{i=1}^{M} w_i^{(j)} y_i$$

An $i^{th}$ element of $\mathbf{w}^{(j)}$ is proven to take the following form.

$$w_i^{(j)}(m) = w(m) = \frac{(-1)^m}{4^k\binom{k}{u-m}} \frac{k-u}{k-u+m} \sum_{t=0}^{u-m} \binom{k}{t} 3^t$$

$m$ is the number of mismatches between the $j^{th}$ $k$-mer and $i^{th}$ gapped $k$-mer. Since $w_i^{(j)}(m)$ only depends on $m$, it can be simply denoted as $w(m)$ [8]. This method and the proof of the above formula for $w(m)$ by my lab colleagues is described in Ghandi *et al.* [8]. Here I test the use of these estimated $k$-mer frequencies as features in our SVM prediction of regulatory elements.

### 5.2.2 Gkm-SVM and gkm-Kernel

Direct calculations of estimated $k$-mer frequencies using the formula derived from the previous section 5.2.1 are computationally impractical when $k$ becomes large. To resolve this issue for our SVM framework, we develop a new kernel method, referred to as gkm-kernel, which can efficiently calculate the inner product of two estimated $k$-mer frequency vectors.

We first consider a general linear kernel function as follows. Let $\mathbf{x}^{(1)} = (x_1^{(1)}, \ldots, x_N^{(1)})$ and $\mathbf{x}^{(2)} = (x_1^{(2)}, \ldots, x_N^{(2)})$ be the two estimated *k*-mer frequency vectors of a sequence 1 ($S_1$) and a sequence 2 ($S_2$), respectively. Then, the linear kernel function with a L2-norm normalization method can be written as follows.

$$K(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) = \frac{\langle \mathbf{x}^{(1)}, \mathbf{x}^{(2)} \rangle}{\|\mathbf{x}^{(1)}\| \|\mathbf{x}^{(2)}\|}$$

Since the inner product in the kernel function is only affected by the *k*-mers that appear in the sequences, the full enumeration of *k*-mer frequency vectors can be avoided, via a method developed by my lab colleague M. Ghandi, and fully described in Ghandi *et al.* [9]. In short, the inner product can be rewritten in terms of *k*-mers only appearing in the two sequences:

$$\langle \mathbf{x}^{(1)}, \mathbf{x}^{(2)} \rangle = \sum_{i=0}^{n_1} \sum_{j=0}^{n_2} \mathrm{c}(u_i^{(1)}, u_j^{(2)})$$

where $n_1$ and $n_2$ are the number of *k*-mers in the two sequences, $S_1$ and $S_2$, respectively, and $\mathrm{c}(u_i^{(1)}, u_j^{(2)})$ is the corresponding coefficient, which is proven to be a function of a number of mismatches between the two *k*-mers. We can further rewrite the equation by using the fact that the only parameter that affects the inner product is the number of mismatches between all possible *k*-mers pairs from the two sequences.

$$\langle \mathbf{x}^{(1)}, \mathbf{x}^{(2)} \rangle = \sum_{m=0}^{l} \mathrm{N}_m(S_1, S_2)\, \mathrm{c}(m)$$

where $\mathrm{N}_m(S_1, S_2)$ is the number of *k*-mer pairs with *m* mismatches, and $\mathrm{c}(m)$ is given by the following equation.

$$c(m) = \sum_{m_1}\sum_{m_2}\sum_{t}\binom{k-m}{t}\binom{m}{m_1-t}\binom{m_1-t}{r}3^t 2^r w(m_1)w(m_2)$$

$$r = m_2 - (t + m - (m_1 - t))$$

The function $w(m_1)$ and $w(m_2)$ are both $w(m)$ given in section 5.2.1.

Although this equation provides significant computational efficiency over the direct calculation of the original formula, it is still impractical when the number of training sequences gets larger. We further resolve this issue by adopting a suffix tree structure similar to the mismatch tree introduced by Leslie *et al.* [88] with some modifications.

### 5.2.3 Data Sets

We test this approach using both our standard EP300 bound forebrain enhancers [39] from the original kmer-SVM study (Chapter 3) and genome-wide CTCF bound regions in a lymphoblastoid cell line (GM12878) [139]. For the CTCF positive dataset, we first selected the top 5,000 ChIP-seq signal enriched regions from the reported list, and subsequently chose 2,500 sites at random to further reduce the training set size.

For the EP300 positive dataset, we further polished the original set by applying the peak refinement algorithm previously developed and described in Section 4.1.2. As a result, a new set of 1,693 common length (400 bp) peaks were obtained after removing any regions which were more than 70% repeats. Both raw data sets are available at Gene Expression Omnibus database (GSE19622 and GSE13845, respectively).

For negative sequence sets, I followed the improved negative sequence sampling method described in Section 4.1.3. Briefly, I found an equal size set of random genomic

sequences by matching length, GC and repeat fraction of the corresponding positive set. At each sampling step, we randomly selected a positive region, calculated its length, GC content and repeat fraction, and sampled an equal length genomic region that matched the GC content and the repeat fraction. We repeated sampling until we obtained the same number of sequences as the positive set.

## 5.3 Results

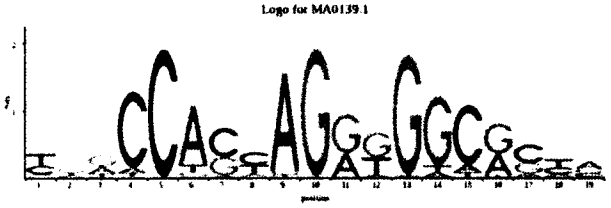


**Figure 5-1: a PWM model for CTCF binding sites**
A CTCF protein can specifically bind to a set of very long sequences via its eleven zinc finger domains, which can be effectively modeled by a PWM. The above PWM for CTCF is directly obtained from JASPAR database [84]

We first hypothesized that gkm-SVM can effectively model much longer TFBSs by using estimated *k*-mer frequencies. To test this idea, we decided to compare the classification performance of gkm-SVM to kmer-SVM by predicting genomic binding sites of CTCF proteins [139]. As shown in Figure 5-1, CTCF recognizes very long DNA sequences, and its binding specificity has been well-characterized. We further verified that the genomic CTCF binding sites [139] are almost perfectly predicted by this CTCF PWM model (Figure 5-2). In this preliminary analysis, we used as prediction scores the best matching log-odd scores to the PWM model and achieved 0.983 of auROC. Thus, we concluded that the CTCF dataset would provide an excellent opportunity to test our

hypothesis.

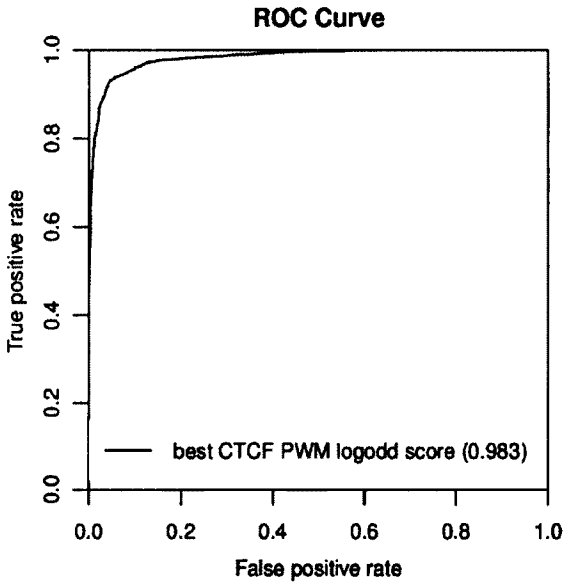



**Figure 5-2: A Classification result of CTCF binding sites using the CTCF PWM**
The CTCF bound regions and the corresponding negative regions were scored by the CTCF PWM and the best log-odd score for each sequence was then used to calculate the ROC curve. Extremely high auROC was achieved, indicating that CTCF binding is well-modeled by the PWM.

To directly measure the effect of the changes in *k*-mer length, we trained both gkm-SVM and kmer-SVM on the CTCF data set by varying *k* from 6 to 20. We then quantified the classification performances by calculating auROCs with standard five-fold cross validation techniques. Figure 5-3A shows a summary result of the comparisons. As expected, gkm-SVM performs consistently better than kmer-SVM for all *k*-mer lengths. More significantly, while the kmer-SVMs severely suffer from the overfitting problem when *k* is greater than 10, gkm-SVMs are virtually unaffected by the length *k*. In fact, gkm-SVM achieved the best result (auROC=0.970) when $k$=15, which is significantly better than the kmer-SVM (auROC=0.912 when $k$=10) as shown in Figure 5-4A. It should be noted, however, that the PWM classification result shown in Figure 5-2 was the

best among the three methods we tested in this analysis. While both kmer-SVM and gkm-SVM use entire sequences (average length is 316 bp) to calculate their prediction scores, the PWM scores are from the best PWM-matching 19 bp sub-sequences. It thus may be due to the extra ~300 bp sequences that contribute to noise in the SVM prediction scores, which in turn impair the overall classification accuracy.

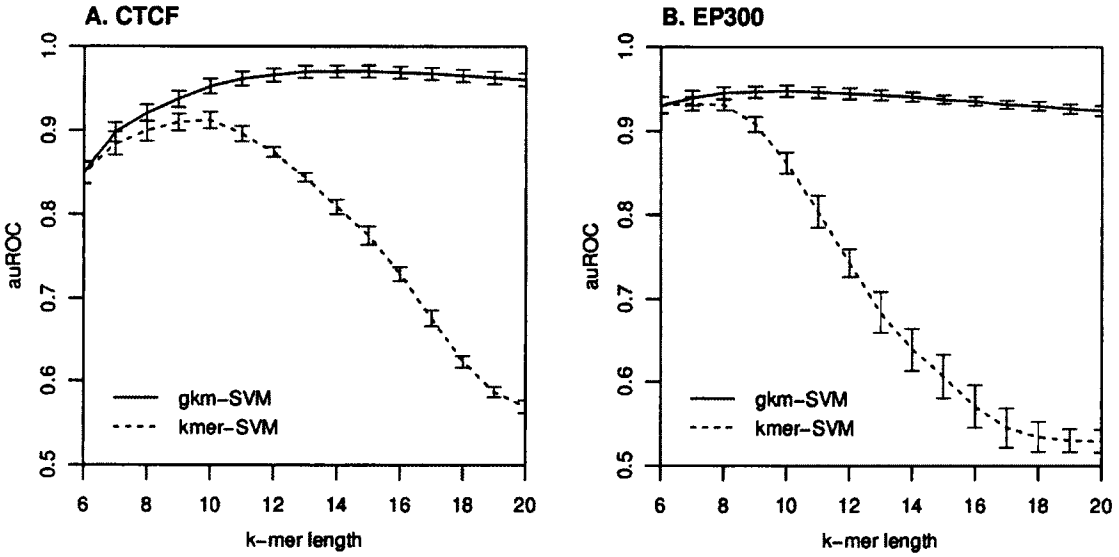



**Figure 5-3: Comparing gkm-SVM with kmer-SVM by varying *k*-mer length**
Both gkm-SVM and kmer-SVM were trained on (A) the CTCF data set and (B) the EP300 data set with different *k*-mer lengths. While auROCs of the kmer-SVMs in both cases decrease as *k* gets larger (dashed lines), gkm-SVMs are little affected by the large *k* (solid lines).

More interestingly, gkm-SVM shows consistently better performance than kmer-SVM even if $k$ is relatively small ($k < 10$) (Figure 5-3). This indicated that gkm-SVM may also be better at modeling short TFBSs than kmer-SVM. To directly test this hypothesis, we analyzed the genomic EP300 bound regions in embryonic mouse forebrain. We have previously shown that our original kmer-SVM classifiers can accurately predict the

EP300 binding and the most predicted TFBSs are relatively short (see Chapter 3). Thus, we believe that this data set would provide a direct test of the effectiveness of the more robust frequency estimation of the k-mers.

We repeated the previous analysis with the EP300 data set, and still found that gkm-SVM consistently outperforms kmer-SVMs for all *k*-mer lengths as shown in Figure 5-3B. Analogous to previous observations (Figure 5-3A), kmer-SVM accuracies also rapidly drop as *k* increases. Moreover, although the difference in performance is much smaller than the CTCF data set, the gkm-SVM achieved the best auROC (0.947) with $k$=10, while the kmer-SVM achieved 0.932 with $k$=7, suggesting that 10~15 bp *k*-mers with some flexibility may contain more complete information about TF binding.

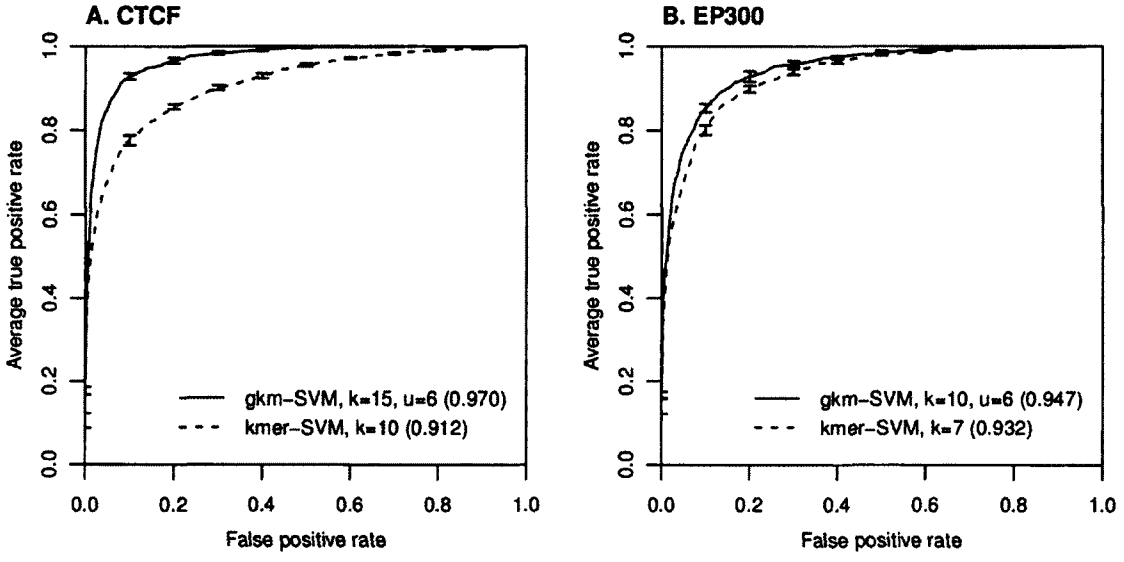



**Figure 5-4: Comparing the best classification results between gkm-SVM and kmer-SVM**
The best classification results in terms of auROC for each case are shown. Gkm-SVMs outperform kmer-SVMs on both data sets.

## 5.4 Discussion

In this chapter, we demonstrated a considerably improved approach that can effectively discern regulatory regions using estimated *k*-mer frequencies as features. We successfully incorporated the robust *k*-mer frequency estimation algorithm [8] into our SVM framework via a newly developed kernel method [9], and showed that the improved method, called gkm-SVM, significantly outperforms the original kmer-SVM method on both the EP300 bound forebrain enhancers and CTCF bound regions in a lymphoblastoid cell line. In summary, we presented that the new gkm-SVM method can use much longer k-mers as features, which in turn effectively captures long TFBSs such as CTCF as well as more robustly models short TFBSs.

# 6 Prediction of Cell-type Specific c-Myc Binding Using gkm-SVM

## 6.1 Introduction

Identifying the molecular mechanisms by which the activity of DNA regulatory elements are modulated in diverse cellular environments still remains a bottleneck in translational biomedical studies. Although great efforts have been devoted to systematically discovering genome-wide regulatory regions across multiple conditions via DNaseI-seq or ChIP-seq for several DNA interacting factors [3], [4], our understanding of the function of regulatory regions is limited. I hypothesize that the function of these regions can in principle be predicted from their primary DNA sequence. With sufficient accuracy, such sequence-based predictive models would greatly improve our understanding of regulatory elements by describing a grammar of sequence features and context rules, which could then generate specific hypotheses of the underlying molecular mechanisms.

The central element of this hypothesis is that the mechanism by which DNA sequence specifies the activity of regulatory elements is by encoding combinatorial interactions between several DNA binding factors, which together modulate specific binding site occupancy. In this chapter, I apply the comprehensive approaches developed in previous chapters to study this key mechanism directly by predicting the differential binding

patterns of the same TF in multiple cell types. In addition, I developed a systematic way to detect and summarize tissue specific co-factor regulatory vocabulary.

As a representative example, I analyzed cell-specific bindings of the c-Myc TF in five different cell types. c-Myc is an extensively studied oncogenic TF with significant clinical implications [140], [141]. Although c-Myc can specifically bind to E-box elements (CANNTG) by dimerizing with its partner protein Max, the genomic binding of c-Myc cannot usually be determined by its sequence specificity alone, and is known to be highly dependent on their biological contexts [142]. I hypothesize that diverse c-Myc binding patterns across different conditions can essentially be described by their local DNA context. I first adopted the gkm-SVM method introduced in Chapter 5 to predict genomic c-Myc bound regions, and further advanced the method to predict the fine scale structures of the c-Myc bound regions at a single base-pair resolution. I discovered that the cell-type specific c-Myc bindings can be accurately predicted by the presence of other TFBSs near the c-Myc binding sites. To summarize these predictive sequence elements, I developed an algorithm to extract the sequence features and build PWM models from a trained gkm-SVM. These *de novo* PWMs revealed several known TFBSs that may determine genomic c-Myc occupancy patterns in the cell types. The comprehensive approaches developed in this chapter are general and can be applied to a wide range of biological problems related to regulatory sequence analysis. I believe that these new methods will greatly facilitate to uncover underlying molecular mechanisms of diverse biological processes.

## 6.2 Methods

### 6.2.1 Processing of c-Myc ChIP-seq Datasets

I used genome-wide c-Myc bound regions identified by ChIP-seq in five different cell lines (HelaS3, HepG2, HUVEC, K562, and MCF7; GEO accession number is GSE32883) that had been generated as a part of the ENCODE project [3], [143]. I first defined a set of peaks for each cell type using MACS software v1.4 [128] with default settings, and selected the top 15,000 high confidence peaks (ranked by the *p*-values) from each dataset for further analysis. For the peak calling analysis, I used pooled aligned tags (i.e. all replicates were combined). I additionally obtained extended tag pileup data at every 20 bp using MACS [128] and saved them into wiggle format files to calculate ChIP-seq signal intensities, which will be referred to as "ChIP-seq signals".

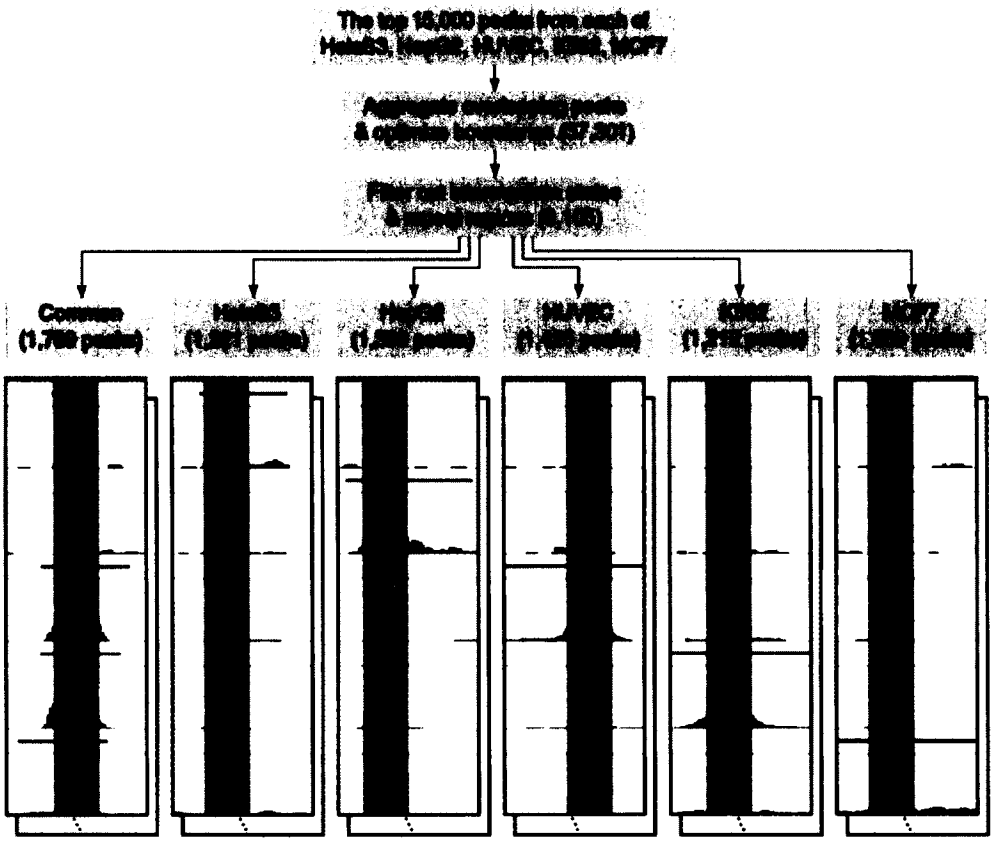


**Figure 6-1: Overview of the ChIP-seq data processing**
The flowchart of the c-Myc ChIP-seq data processing is presented. For the example peaks, a 1,000 bp window is used. Shaded areas denote 300 bp sub-regions determined by the peak refinement algorithm.

To determine sets of distinct classes of peaks (cell-type-specific or common peaks), I first aggregated the overlapping peaks from the five sets, which defined a new set of 37,301 peaks. From each of these combined peaks, I further identified a 300 bp sub-region by maximizing the ChIP-seq signals in only bound cell types as previously described in section 4.1.2. This peak refining step not only can remove less-informative flanking regions, but also can eliminate a possible bias that might be caused by systematic length differences between different datasets. I then filtered out any

intermediate peaks which exhibited a strong ChIP-seq signal in any of unbound cell types or a weak ChIP-seq signal in any of bound cell types. To determine these intermediate peaks, we independently chose the cut-off of ChIP-seq signals of the combined 37,301 peaks for each cell type by fitting the signals to sigmoid functions (Figure 6-2). It should be noted that the most frequently observed binding pattern was a common peak (i.e. a region that is bound by c-Myc in all cell types), and the second most frequently observed patterns were cell-type specific peaks (i.e. a region that is bound by c-Myc only in one cell type). After removing any region containing more than 70% of repeat elements, I finally obtained 8,105 peaks. Specifically, this final set was comprised of 1,221 (HelaS3), 1,389 (HepG2), 1,425 (HUVEC), 1,212 (K562), and 1,069 (MCF7) cell type specific peaks, and 1,789 common peaks. The overview of this procedure is displayed in the Figure 6-1.

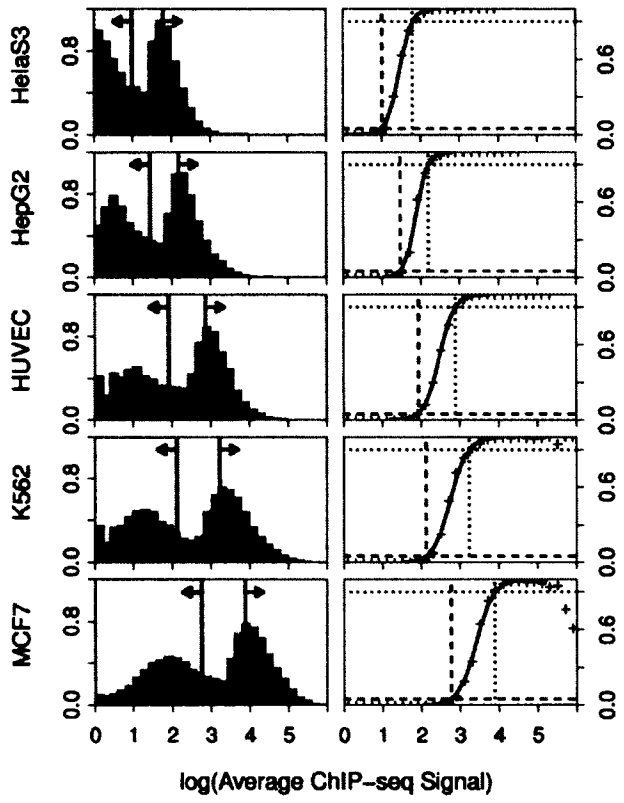



**Figure 6-2: Filtering out intermediate state peaks**
(Left column) histograms show the distributions of logarithms of average ChIP-seq signals of 37,301 refined peaks in each cell type. Red data represent bound regions and blue data represent unbound regions in the corresponding cell types. The red vertical line is the cut-off for "true" bound regions and the blue vertical line is the cut-off for "true" unbound regions. (Right column) The signals are fitted to sigmoid functions to estimate posterior probabilities. 90% of the posterior probabilities is selected as the cut-off of bound regions (dotted lines), equivalent to the red line in the right panel, 5% is selected as the cut-off of unbound regions (dashed lines), equivalent to the blue line.

### 6.2.2 Differential Expression Analysis of c-Myc Target Genes

To investigate the connection between c-Myc binding and gene expression levels, I examined the expression of the nearest gene for each c-Myc bound region. This is a rough approach, as we do not usually know the actual target genes of the distal regulatory regions. The RefSeq database [144] accessed from the UCSC genome browser [93] were

used as gene identifiers for this analysis. The right column in Figure 6-3A shows the distribution of distances to the nearest genes for each set. While virtually all common c-Myc bound regions are located at promoters, most cell-type specific c-Myc bound regions are located at a far distance (~50k bp) from the nearest genes. To test whether these cell-type specific c-Myc bound regions can act as distal transcription enhancers in the specific cell type, I further analyzed gene expression profiles of the nearest genes in the five cell types. I used publically available gene expression datasets (measured by Affymetrix Human Exon 1.0), "Duke Affy Exon (hg19)," downloaded from the ENCODE data coordination center at UCSC (http://genome.ucsc.edu/ENCODE/). It should be noted that cells that were analyzed by the c-Myc ChIP-seq experiments were also used to extract RNA samples for the microarray experiments. Thus, the gene expression datasets are directly relevant to our study. More importantly, these datasets have been uniformly processed, and probes linked to genes have been aggregated to the gene expression levels. Thus, additional data processing was minimally required.

Figure 6-3B shows the distribution of the expression levels of the nearest genes for each set. The nearest genes were further divided into two categories; genes within 2,000 bp of the corresponding c-Myc bound regions (left column), and genes at least 2,000 bp away from the corresponding c-Myc bound regions (right column). Significant differences in expression levels between bound (red) and unbound cell types (blue) are observed in both cases, although proximal genes shown in left column are usually more significant. The Kolmogorov-Smirnov test was applied to calculate the indicated significant *p*-values.

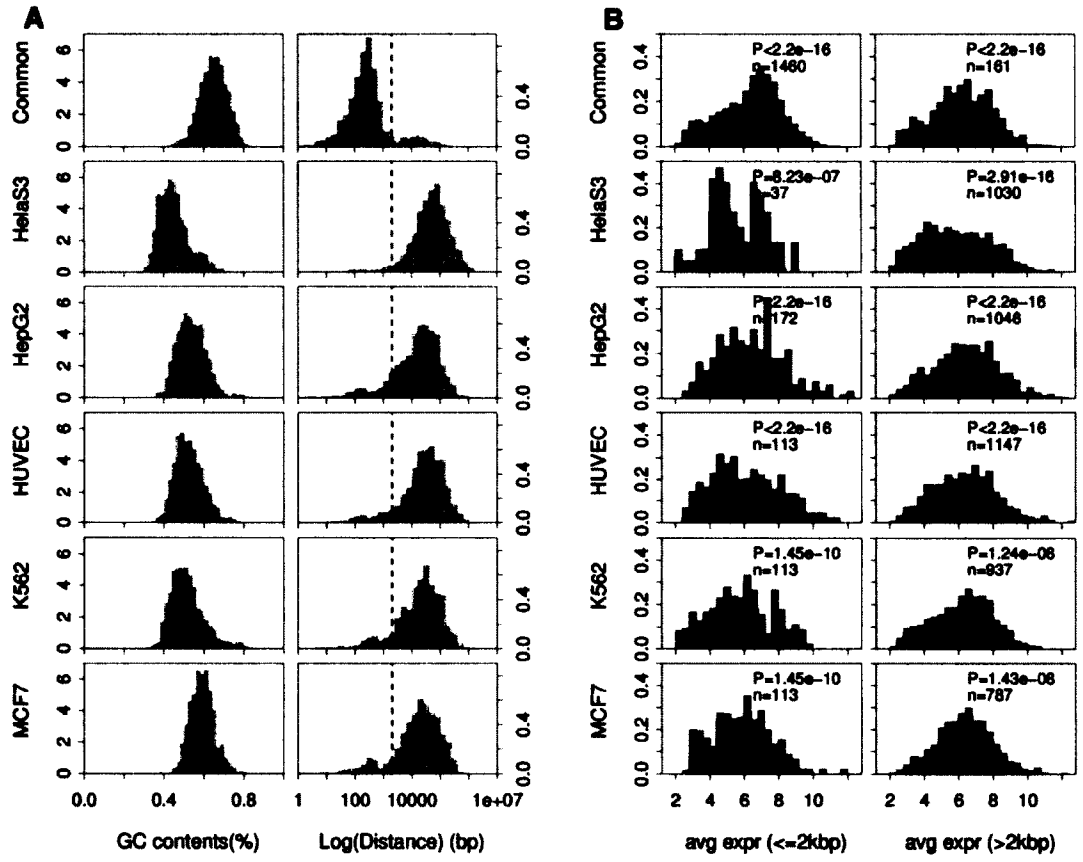



**Figure 6-3: Differential expression of c-Myc target genes**

(A) Common c-Myc bound regions are enriched for promoters and GC enriched, while cell specific binding is distal and less GC enriched. Most common regions are within 2k bp of TSS (dashed line) while cell specific elements are ~50k bp from TSS. (B) Distributions of expression levels of the nearest genes in the corresponding cell types (red) vs. the control (blue) of all other cell types combined are shown, except for the common peaks which use randomly selected genes as a control. Genes were divided into two subsets; proximal genes (left) and distal genes (right) to the c-Myc bound regions.

### 6.2.3 Gene Ontology Enrichment Analysis of c-Myc Target Genes

To further support the biological relevance of the cell-type specific c-Myc bound regions, I performed the gene ontology (GO) enrichment analysis of the nearest genes using DAVID [145], [146]. Table 6-1 summarizes top three clusters of similar GO terms for each set of genes reported by DAVID. In general, biologically relevant genes are enriched near the different classes of c-Myc bound regions. For example, genes near the common c-Myc bound regions are significantly enriched in essential biological processes, such as ribosome biogenesis, consistent with previous findings [147]. HUVEC specific c-Myc bound regions are also significantly associated with angiogenesis, consistent with the origin of the cell type (isolated from normal human umbilical vein). However, the overall significance is relatively low, partly because of the very approximate way genes are associated with the c-Myc bound regions, especially for cell-specific distal binding sites.

| GO Category | GO Term | *P*-value |
|---|---|---|
| | **(A) Common** | |
| **Cluster 1 (Enrichment Score: 47.45)** | | |
| Cellular Component | intracellular organelle lumen | 2.8e-64 |
| | membrane-enclosed lumen | 3.6e-63 |
| | organelle lumen | 1.2e-61 |
| **Cluster 2 (Enrichment Score: 34.76)** | | |
| Biological Process | ribonucleoprotein complex biogenesis | 1.8e-53 |
| | ribosome biogenesis | 2.1e-39 |
| | ncRNA metabolic process | 1.0e-34 |
| **Cluster 3 (Enrichment Score: 34.04)** | | |
| Cellular Component | ribonucleoprotein complex | 1.0e-75 |
| | ribosome | 1.0e-44 |
| Biological Process | translation | 3.6e-61 |
| | **(B) HelaS3** | |
| **Cluster 1 (Enrichment Score: 5.77)** | | |
| Biological Process | positive regulation of developmental process | 2.6e-8 |
| | positive regulation of cell differentiation | 4.5e-8 |
| | regulation of cell development | 3.9e-3 |
| **Cluster 2 (Enrichment Score: 4.46)** | | |
| Cellular Component | extracellular matrix | 9.7e-6 |
| | extracellular region part | 4.0e-5 |
| | proteinaceous extracellular matrix | 9.7e-5 |
| **Cluster 3 (Enrichment Score: 4.30)** | | |
| Biological Process | embryonic limb morphogenesis | 1.8e-6 |
| | embryonic appendage morphogenesis | 1.8e-6 |
| | limb morphogenesis | 2.0e-6 |
| | **(C) HepG2** | |
| **Cluster 1 (Enrichment Score: 4.02)** | | |
| Cellular Component | extracellular space | 7.1e-6 |
| | extracellular region part | 2.4e-5 |
| | extracellular region | 4.6e-3 |
| **Cluster 2 (Enrichment Score: 3.58)** | | |
| Biological Process | response to hormone stimulus | 3.1e-5 |
| | response to corticosteroid stimulus | 1.7e-4 |
| | response to endogenous stimulus | 2.3e-4 |
| **Cluster 3 (Enrichment Score: 2.82)** | | |
| Biological Process | response to nutrient | 3.5e-4 |
| | response to nutrient levels | 4.6e-4 |
| | response to extracellular stimulus | 8.1e-4 |

| GO Category | GO Term | *P*-value |
|---|---|---|
| | **(D) HUVEC** | |
| **Cluster 1 (Enrichment Score: 11.07)** | | |
| Biological Process | angiogenesis | 5.8e-13 |
| | blood vessel development | 1.1e-11 |
| | vasculature development | 2.3e-11 |
| **Cluster 2 (Enrichment Score: 5.63)** | | |
| Biological Process | cell motion | 2.9e-8 |
| | cell migration | 4.2e-6 |
| | localization of cell | 1.5e-5 |
| **Cluster 3 (Enrichment Score: 5.07)** | | |
| Cellular Component | plasma membrane part | 7.3e-7 |
| | intrinsic to plasma membrane | 1.4e-5 |
| | integral to plasma membrane | 5.2e-5 |
| | **(E) K562** | |
| **Cluster 1 (Enrichment Score: 3.91)** | | |
| Biological Process | regulation of transcription from RNA polymerase II promoter | 4.5e-6 |
| | negative regulation of transcription, DNA-dependent | 5.1e-5 |
| | negative regulation of biosynthetic process | 6.9e-5 |
| **Cluster 2 (Enrichment Score: 2.91)** | | |
| Biological Process | regulation of transcription from RNA polymerase II promoter | 4.5e-6 |
| | positive regulation of transcription, DNA-dependent | 8.2e-5 |
| | positive regulation of RNA metabolic process | 9.8e-5 |
| **Cluster 3 (Enrichment Score: 2.77)** | | |
| Biological Process | glycoprotein metabolic process | 9.2e-5 |
| | glycoprotein biosynthetic process | 2.2e-4 |
| | biopolymer glycosylation | 4.4e-3 |
| | **(F) MCF7** | |
| **Cluster 1 (Enrichment Score: 2.95)** | | |
| Cellular Component | negative regulation of transcription, DNA-dependent | 2.8e-5 |
| | negative regulation of RNA metabolic process | 3.9e-4 |
| | negative regulation of transcription from RNA polymerase II promoter | 1.3e-4 |
| **Cluster 2 (Enrichment Score: 2.36)** | | |
| Biological Process | plasma membrane part | 9.3e-5 |
| | intrinsic to plasma membrane | 2.1e-2 |
| | integral to plasma membrane | 3.9e-2 |
| **Cluster 3 (Enrichment Score: 1.95)** | | |
| Biological Process | cell morphogenesis involved in differentiation | 1.5e-3 |
| | cell morphogenesis involved in neuron differentiation | 3.6e-3 |
| | cellular component morphogenesis | 3.9e-3 |

**Table 6-1: Gene ontology enrichment analysis of the nearest genes in each set**

### 6.2.4 Histone Modification Patterns in The c-Myc Bound Loci

To directly support the idea that the majority of these c-Myc bound regions act as either promoters (in case of commonly bound regions) or distal enhancers (in case of cell-type specific bound regions), I analyzed histone modification patterns of the c-Myc bound regions using publicly available ENCODE ChIP-seq datasets (Broad Histone) of three representative histone modifications; H3K4me3, H3K4me1, and H3K27ac. It should be noted that ChIP-seq datasets for MCF7 was not available at the time of the data acquisition and was therefore excluded. Figure 6-4 shows heatmaps of the ChIP-seq signals in the c-Myc bound regions. Consistent with previous findings, H3K4me3 (a marker for promoters) are specifically enriched in the common c-Myc bound regions, H3K4me1 (a marker for enhancers) are also specifically enriched in the cell-type specific c-Myc bound regions only for the relevant cell types, and H3K27ac (a marker for both "active" promoters and enhancers) are enriched in both common and cell-type specific regions in the relevant cell types.

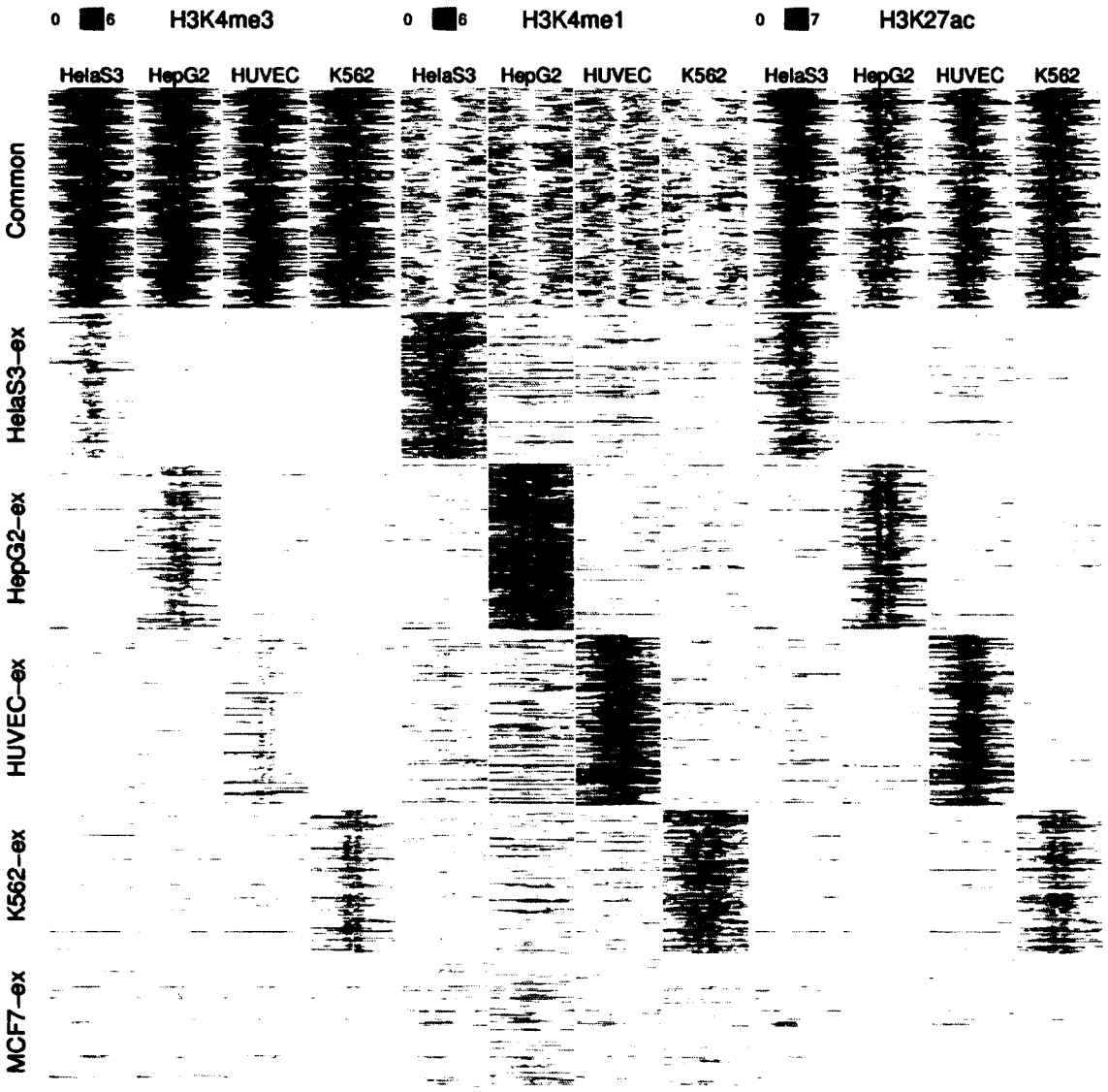



**Figure 6-4: Heatmaps of histone modification ChIP-seq signals in c-Myc bound loci** Histone modification patterns identified by ChIP-seq further support the biological function of the c-Myc bound regions. A 2,000 bp window centered at the midpoint of the c-Myc bound regions is presented. The ChIP-seq signals are calculated from extended tag pileup generated by MACS.

### 6.2.5 Gkm-SVM And a Multiclass Classifier

For the classification analysis, I trained gkm-SVM classifiers on each set of the c-Myc

bound regions (as a positive set) with two different types of negative sets. First, 10X larger random genomic regions were used as negative sequence sets for gkm-SVM training. Since each positive set has distinct distributions of GC content as shown in Figure 6-3A, a negative set was independently generated for each positive set by following the method described in the section 4.1.3. In a separate SVM classification, the union of all other c-Myc bound regions was used as a negative set. These alternative negative sets can reveal sequence features specifically important for discrimination between differential c-Myc binding patterns. More importantly, the gkm-SVM classifiers trained on these negative sets can be directly used to build a multiclass classifier, commonly known as one-vs-all methods [148]. To regularize raw SVM scores for the multiclass classification, I further converted the SVM scores to posterior probabilities by fitting them to sigmoid functions [132], [133]. As suggested by Platt [132], a five-fold cross-validation method was employed to eliminate possible overfitting issues.

One important aspect of gkm-SVM training is the choice of parameters. The length of gapped *k*-mer (*k*) and the number of informative positions of gapped *k*-mers (*u*) are the two most important parameters as they define the feature space. For this study, I chose to use k=10 and u=6 based on previous observations that this specific combination performed well across diverse datasets.

### 6.2.6 Fine Scale Structure Prediction

Since gkm-SVM methods normally use much longer *k*-mers as features than typical kmer-SVM methods, extracting predictive sequence features and interpreting

classification results have become major obstacles in the interpretation of the gkm-SVM classifier results. In this study, estimated 10-mer frequencies are used as features, and the number of all possible 10-mers is $4^{10}$ = 1,048,576 features! Since we generally run gkm-SVM with significantly longer *k*, the top *k*-mers ranked by SVM weights from a trained gkm-SVM are much more difficult to interpret compared to those from the kmer-SVM. However, the more robust gkm-SVM classifier has now made it possible to identify individual TFBSs *within* regulatory regions directly. By scoring all oligomers (with a length similar to the length of typical TFBSs, e.g. 12~20 bp) appearing in the regions, putative TFBSs can now be identified as high scoring oligomers.

These peaks of the gkm-SVM scoring function within the positive and negative set regions define the set of most predictive sequence features for the classifier. We refer to these oligomers as 'gkm-peaks'. To systematically identify gkm-peaks in the c-Myc bound regions, I scanned both positive and negative sets using a 14 bp window and compared the distributions of the oligomer scores. For efficient computations, I first generated a SVM score table for all possible 10-mers, and then calculated 14 bp oligomer scores, denoted as S($s$), using the following equation:

$$S(s) = \sum_{i=1}^{4} f_{SVM}\,(s(i, 10))$$

where $s$ is a 14 bp oligomer, $s(i, 10)$ is the 10 bp sub-sequence of the oligomer $s$ starting at the $i^{th}$ position, and $f_{SVM}(s)$ is the SVM score of the oligomer $s$. Since I simply sum up the 10-mer SVM scores without normalization, there is a systematic scale difference between S($s$) and the exact $f_{SVM}(s)$, but the overall distributions of the two scoring

systems are the same. The left column in Figure 6-5 directly compares the distributions of oligomer scores between the positive set (red) and the corresponding negative set (blue) for each dataset. As expected, significant score shift of positive set toward the right side relative to negative set are observed. A SVM score "cutoff" for determining gkm-peaks is then defined as the minimum SVM score such that:

$$\frac{\mathrm{P}(x > cutoff|positives)}{\mathrm{P}(x > cutoff|negatives)} > ODDRATIO\_CUTOFF$$

where *ODDRATIO_CUTOFF* is a predetermined parameter, and $x$ is a SVM score. The SVM score probabilities, P( · |*positives*) and P( · |*negatives*), are empirically calculated from actual score distributions. We chose *ODDRATIO_CUTOFF* = 5 for common c-Myc bound regions and *ODDRATIO_CUTOFF* = 4 for all other cell-type specific c-Myc bound regions to obtain high confidence sets of gkm-peaks. The right column in Figure 6-5 demonstrates the relationship between the SVM cutoffs vs. $\frac{\mathrm{P}(x>cutoff\ |positives\ )}{\mathrm{P}(x>cutoff\ |negatives\ )}$ for each dataset. Ultimately, non-overlapping oligomers above the determined SVM cutoff were collected as gkm-peaks, resulting in 8,684 gkm-peaks from Common (4.84 gkm-peaks/region), 3,062 from HelaS3 (2.51 gkm-peaks/region), 3,616 from HepG2 (2.60 gkm-peaks/region), 3,759 from HUVEC (2.64 gkm-peaks/region), 3,255 from K562 (2.69 gkm-peaks/region), and 1,360 from MCF7 (1.27 gkm-peaks/region). These sets of gkm-peaks will be used to find *de novo* PWMs, and the detailed algorithm will be described in the next section.

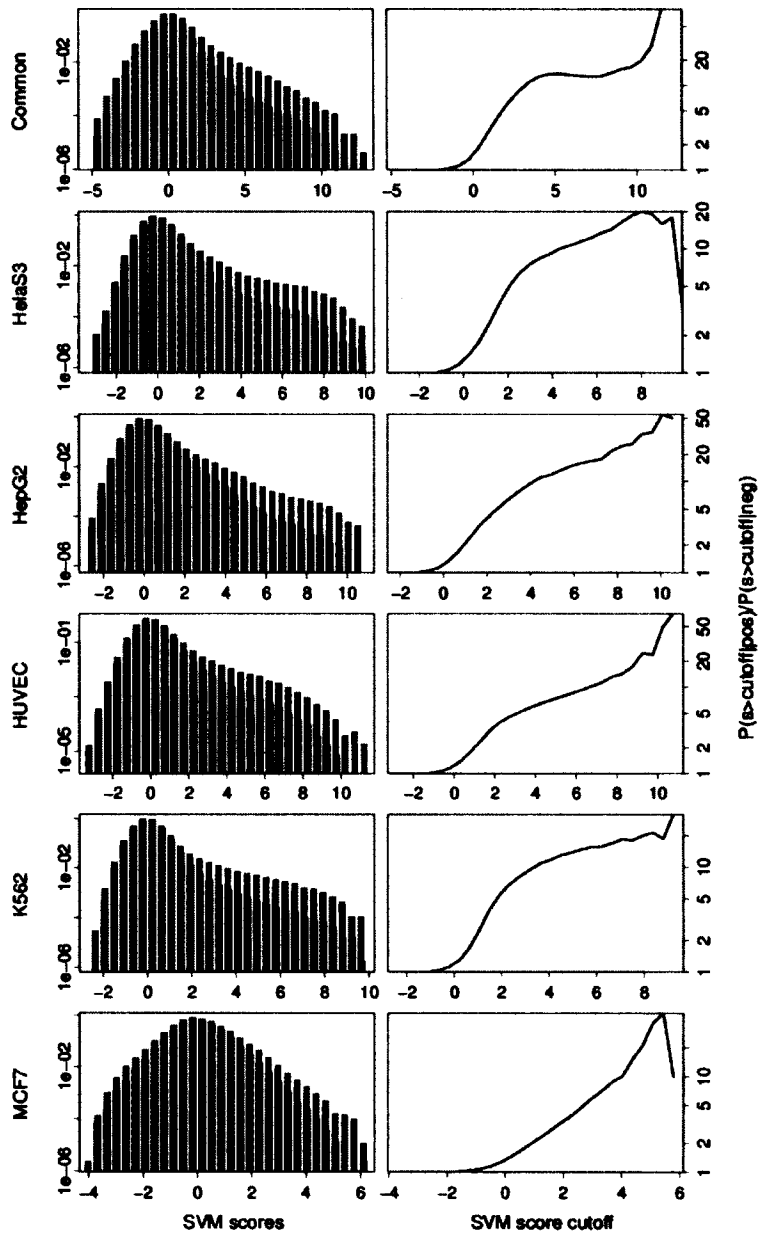



**Figure 6-5: Distributions of SVM scores of 14 bp oligomers**
(*Left column*) Oligomer score distributions are compared between the positive set (red) and the negative set (blue). To emphasize the tail distribution, the Y-axis is displayed in a logarithmic scale. (*Right column*) The odd-ratio vs. SVM score cutoff is shown. The dashed line is *ODDRATIO_CUTOFF*, where the SVM score cutoff is determined.

### 6.2.7 *De novo* PWM Finding Algorithm

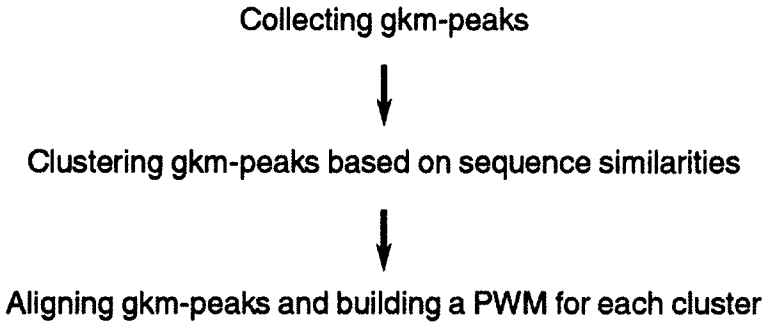



**Figure 6-6: Overview of the *de novo* PWM finding algorithm**
A basic strategy for motif finding based on a set of oligomers (gkm-peaks) is introduced. Essentially, gkm-peaks are clustered and aligned to build PWM models.

To further summarize the predictive sequence features and to facilitate interpretation of the classification results, I developed a new *de novo* PWM finding algorithm using gkm-peaks, which were introduced in the previous section 6.2.6. Figure 6-6 shows a basic strategy for motif finding, and details are described as follows.

First, I adopted the affinity propagation clustering algorithm [149] to cluster the gkm-peaks. This algorithm takes as input a set of similarity scores between all pairs of data points, and identifies clusters and the corresponding exemplars. To calculate the similarity scores, I used a gkm-kernel function previously introduced in the section 5.2.2, and the parameters used for the kernel function are the same as the ones used in the gkm-SVM classification analysis (i.e. $k$=10 and $u$=6). I then systematically searched for an optimal set of clusters by varying the "preference" parameter of the affinity propagation algorithm. This parameter is directly related to the likelihood of each data point being

selected as an exemplar. Thus, the number of clusters can be modulated by changing the "preference", and the smaller preference the smaller number of clusters. I tested six different preferences ranging between -50 and -300, and selected -300 as the optimal parameter for subsequent analysis. In fact, the number of clusters was almost unchanged when preference < -200.

After obtaining a set of clusters, I further developed a method to align the gkm-peaks and build a PWM model for each of the clusters since the earlier clustering step does not provide any alignment information between the gkm-peaks. A simple Expectation-Maximization (EM) algorithm is developed to align the gkm-peaks as follows:

1. Initial step: build an initial PWM model using the exemplar of a cluster
2. E step: given a current PWM, find the best alignment of each gkm-peak
    a. Log-odd ratio is used to evaluate alignments
    b. Only the best matching 10 bp sub-sequence is considered
3. M step: update a current PWM using the best alignments found in M step
    a. Any gkm-peaks with low alignment score are filtered out
4. Repeat the previous two E-M steps (2 and 3) until no changes are made

The above EM algorithm is guaranteed to find a local maximum. After the multiple alignments between all gkm-peaks are identified, the frequency of each nucleotide at each of the aligned columns was directly used to build a PWM model. In addition, PWMs were iteratively compared and merged if they are similar to each other (measured by Pearson correlation coefficient > 0.9). This additional step usually removes at most one or two PWMs for each dataset. Ultimately, 17 PWMs were obtained from Common (Figure 6-7), 12 PWMs from HelaS3 (Figure 6-8), 12 PWMs from HepG2 (Figure 6-9), 8 PWMs

from HUVEC (Figure 6-10), 11 PWMs from K562 (Figure 6-11), and 9 PWMs from MCF7 (Figure 6-12). For comparison, I also found the best matching known PWM for each of the *de novo* PWMs by scoring it for known motifs available in public databases (JASPAR [84], TRANSFAC [83], and UniPROBE [100]) combined with recently reported PWMs as a part of ENCODE project [150] using the TOMTOM package [101]. I used a stringent threshold (q-value < 0.05) to consider only PWMs matched with high confidence. Thus, in the following figures, the part of the best matching known PWM is left blank if there is no such PWM that meets the q-value threshold.

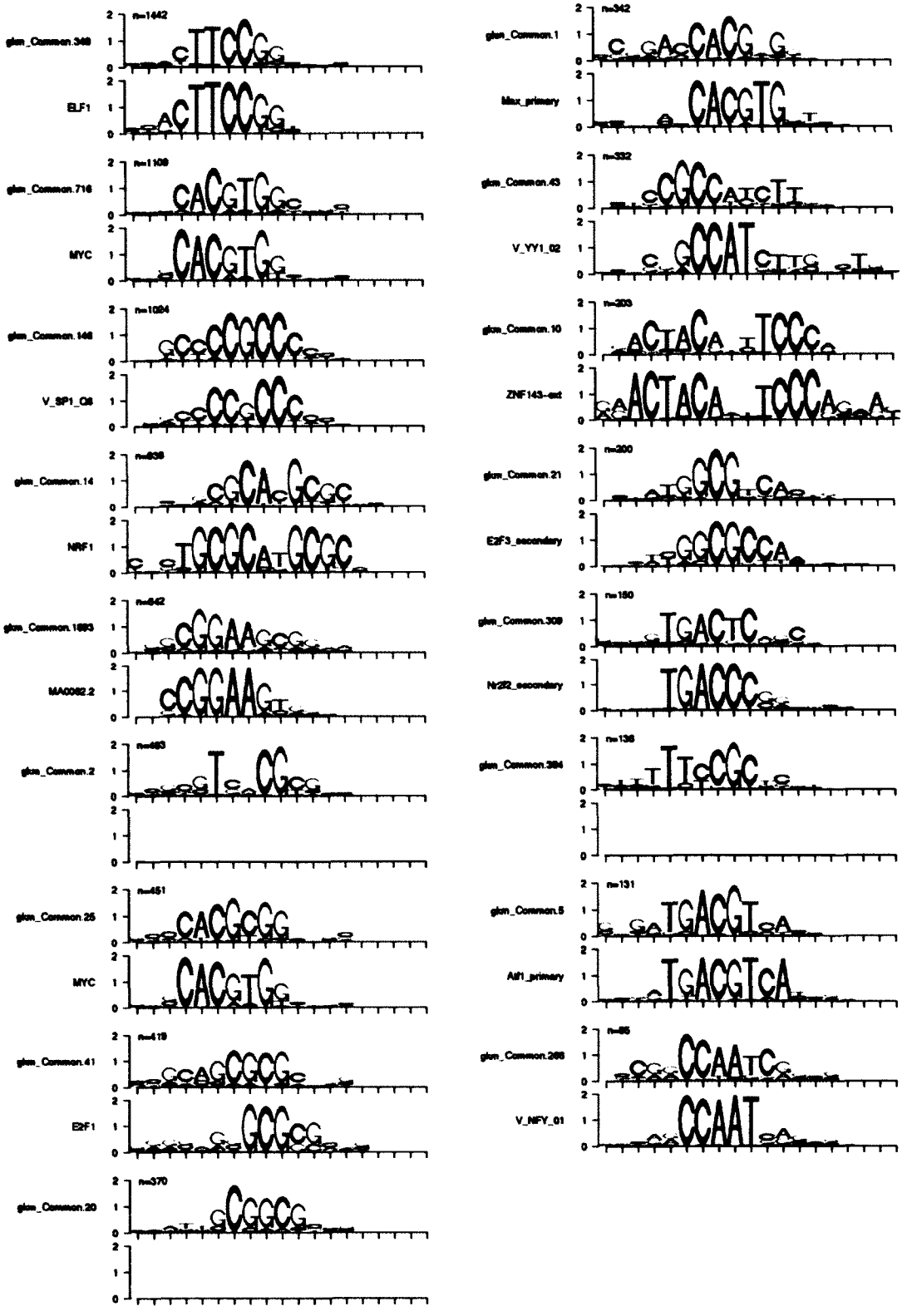


**Figure 6-7: *de novo* PWMs and the best matched known PWMs from common c-Myc bound regions**

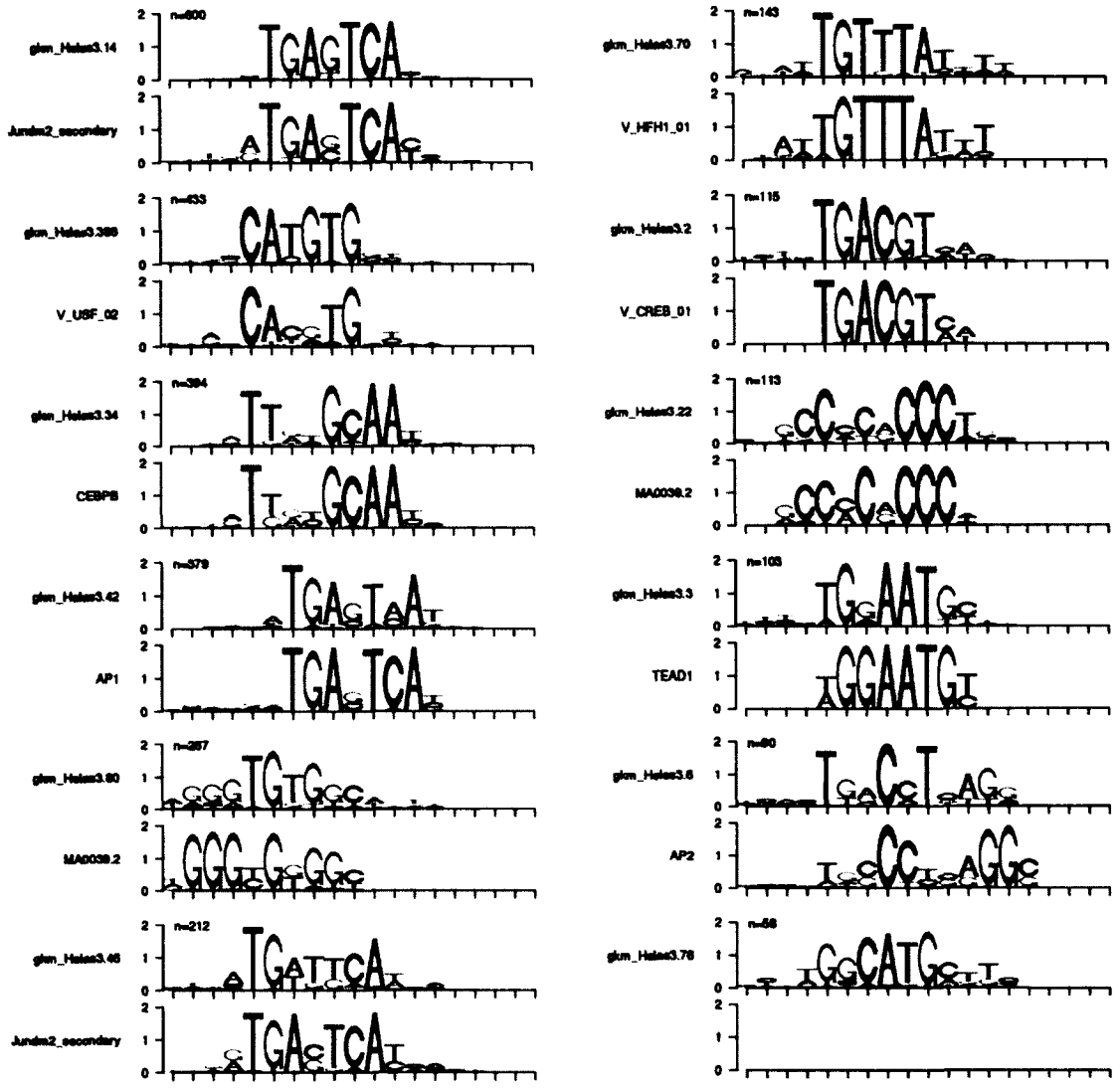


**Figure 6-8:** ***de novo*** **PWMs and the best matched known PWMs from HelaS3 specific c-Myc bound regions**

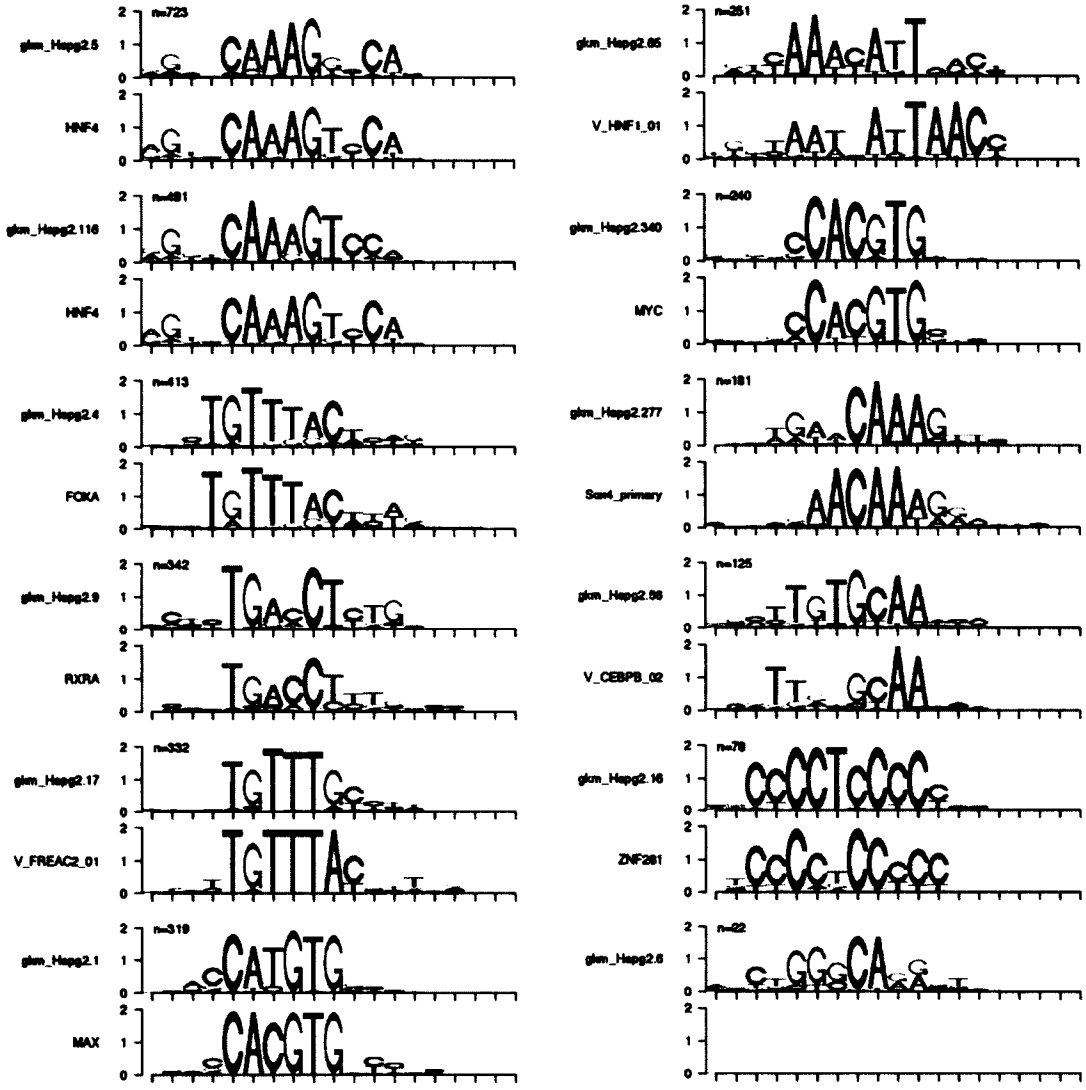


**Figure 6-9:** ***de novo*** **PWMs and the best matched known PWMs from HepG2 specific c-Myc bound regions**

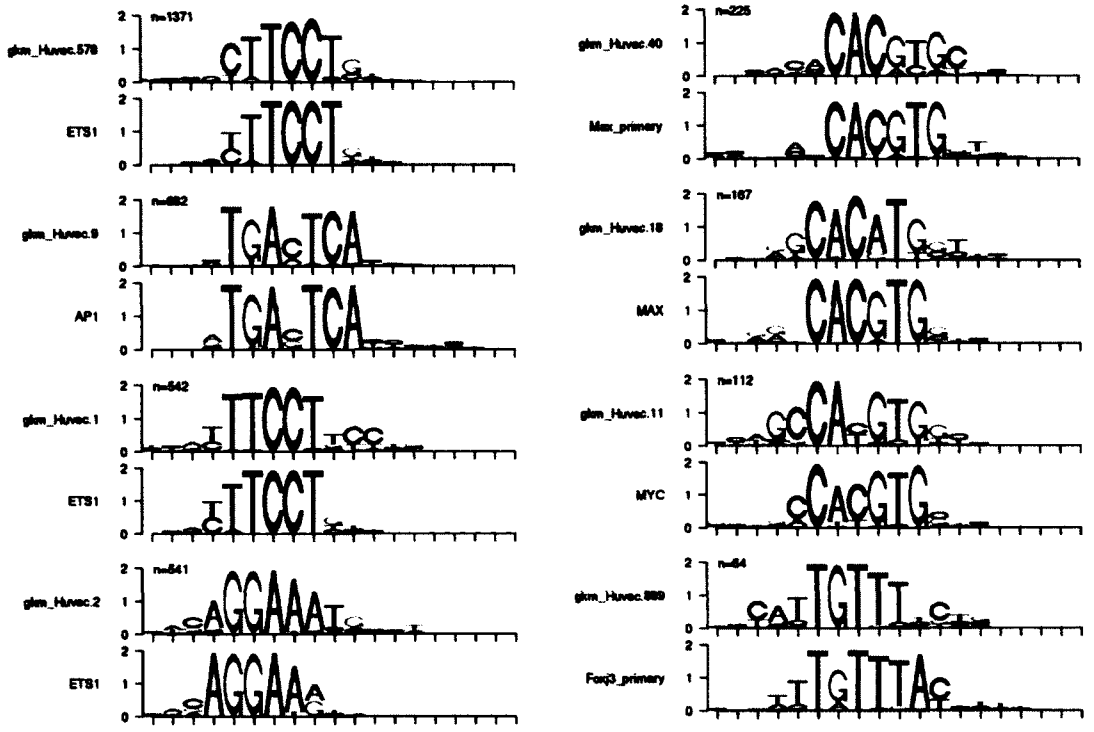



**Figure 6-10:** ***de novo*** **PWMs and the best matched known PWMs from HUVEC specific c-Myc bound regions**

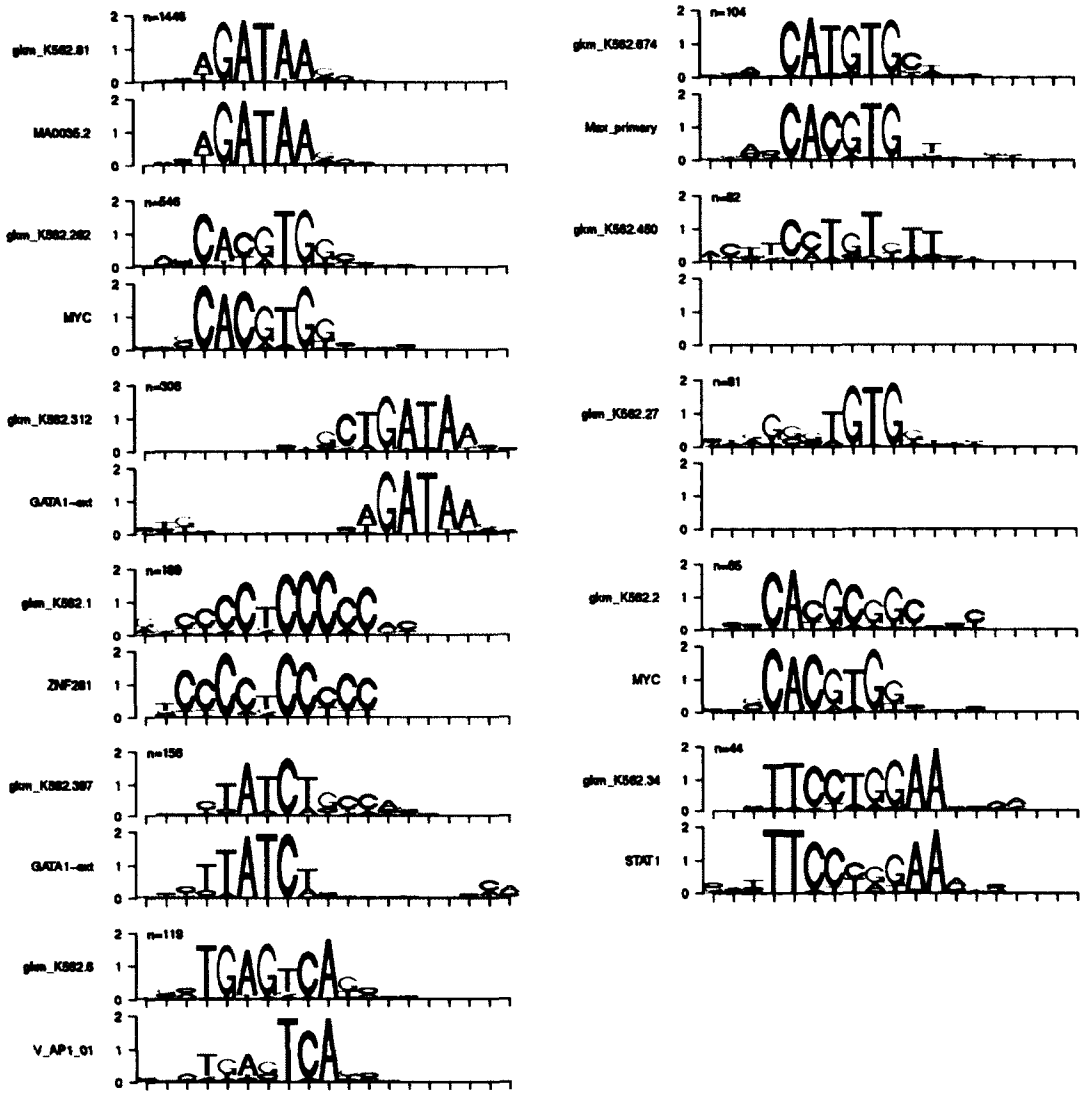


**Figure 6-11:** ***de novo*** **PWMs and the best matched known PWMs from K562 specific c-Myc bound regions**

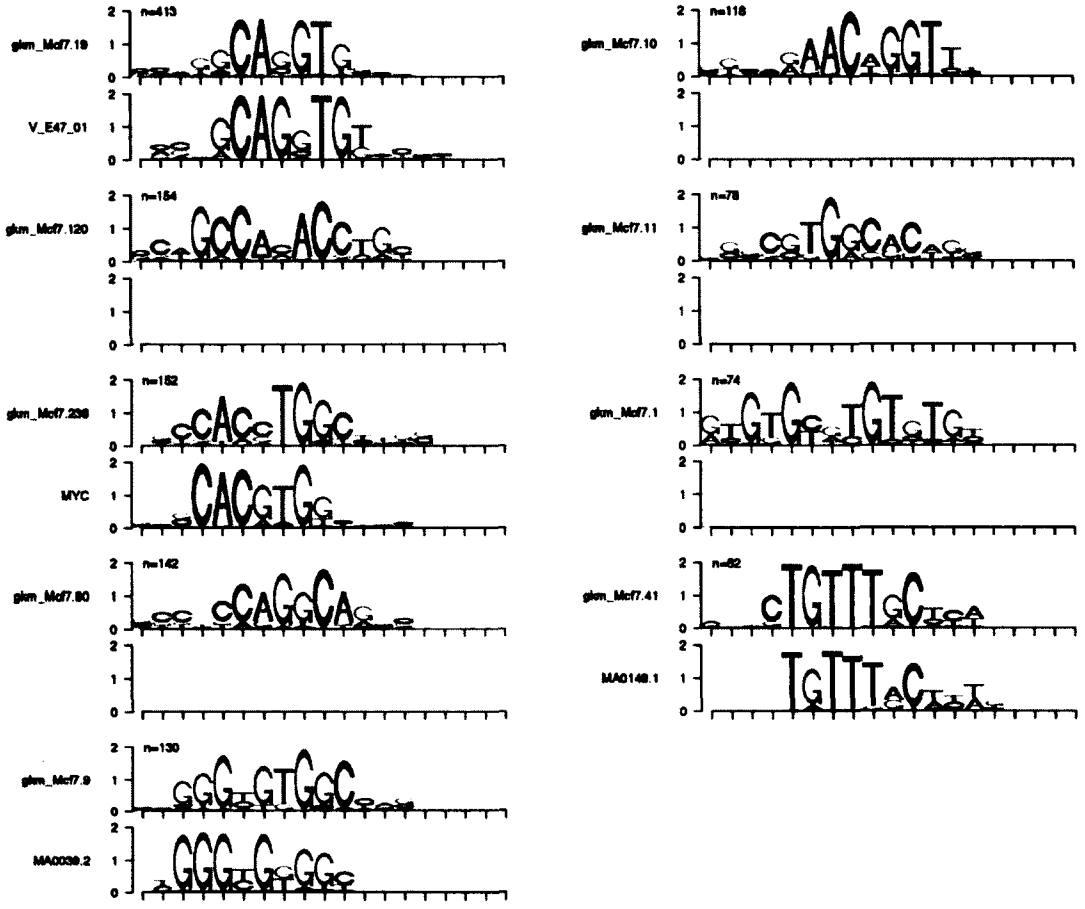



**Figure 6-12: *de novo* PWMs and the best matched known PWMs from MCF7 specific c-Myc bound regions**

### 6.2.8 ChIP-seq Signal Profiles of the c-Myc Bound Loci for other TFs

To test whether the gkm-peaks that match to known PWMs are bound by the cognate TFs *in vivo* in the corresponding cell types, I systematically analyzed ChIP-seq signal profiles of other TFs using publically available datasets from the ENCODE project [3]. Among the known PWMs found to be enriched in the common c-Myc bound regions (Figure 6-7), ChIP-seq datasets for Nrf1, Sp1, Elf1, Gabp, Yy1, and Znf143 are available from the database for the multiple cell types. Similarly, c-Jun (AP1), c-Fos (AP1), and Cebp ChIP-seq datasets are available for HelaS3 (Figure 6-8); Hnf4a, Hnf4g, and Rxra ChIP-seq for HepG2 (Figure 6-9); c-Jun (AP1) and c-Fos (AP1) ChIP-seq for HUVEC (Figure 6-10); Gata1 and Gata2 ChIP-seq for K562 (Figure 6-11). For each ChIP-seq dataset, the c-Myc bound regions containing the corresponding binding site were further segregated from the ones without the binding site. I calculated the average signal intensity of each set of the c-Myc bound regions for all aforementioned ChIP-seq datasets.

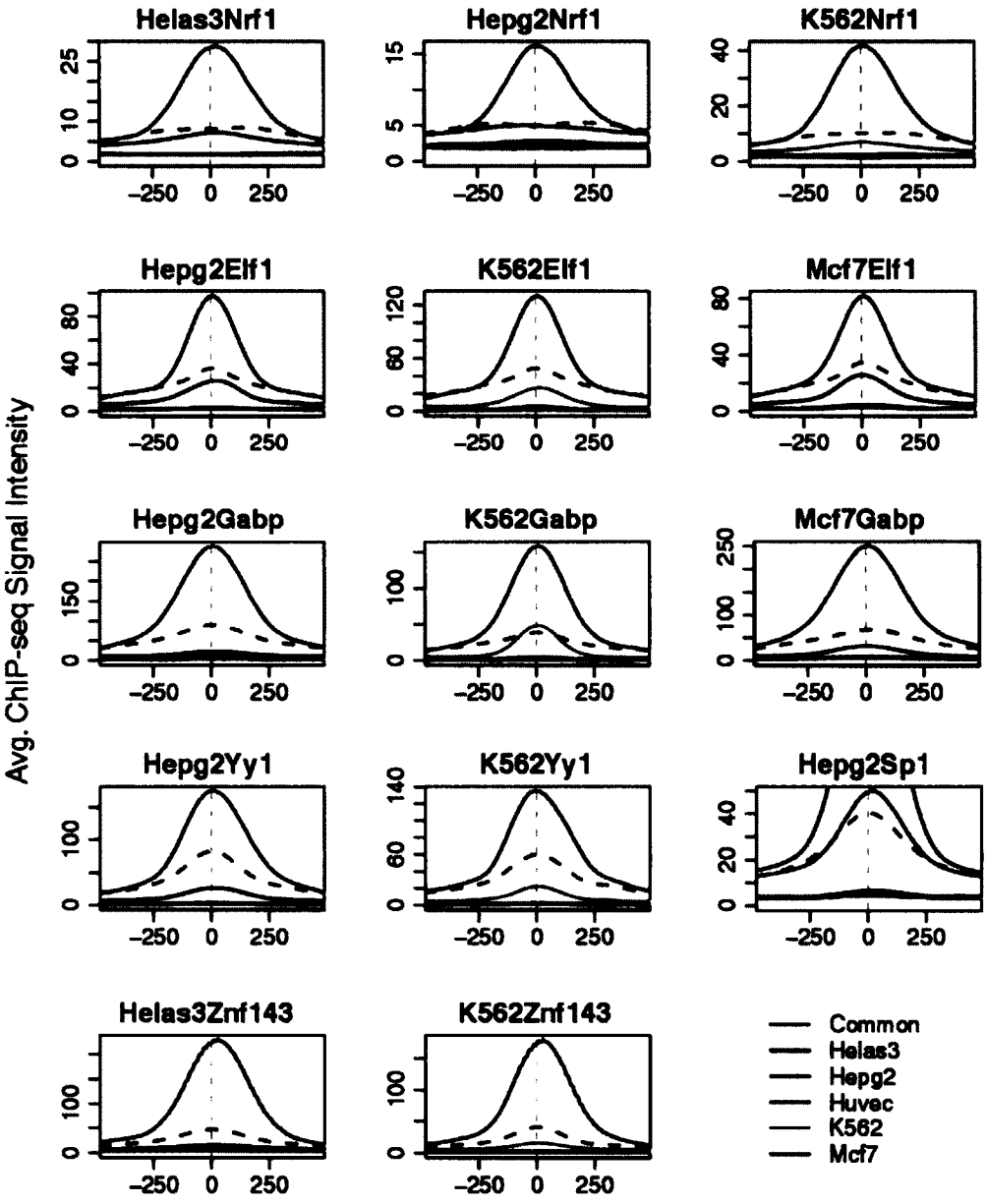



**Figure 6-13: Average ChIP-seq signal profiles of the c-Myc bound regions for TFs enriched in the common bound loci**

Binding sites of the TFs presented here are specifically enriched in the common c-Myc bound regions. 1,000 bp window centered at the midpoint of the c-Myc bound regions are shown. Red "Dashed" lines are the common c-Myc bound regions without the cognate binding sites. In all cases, ChIP-seq signals are specifically enriched only in the common c-Myc bound regions containing the cognate binding sites.

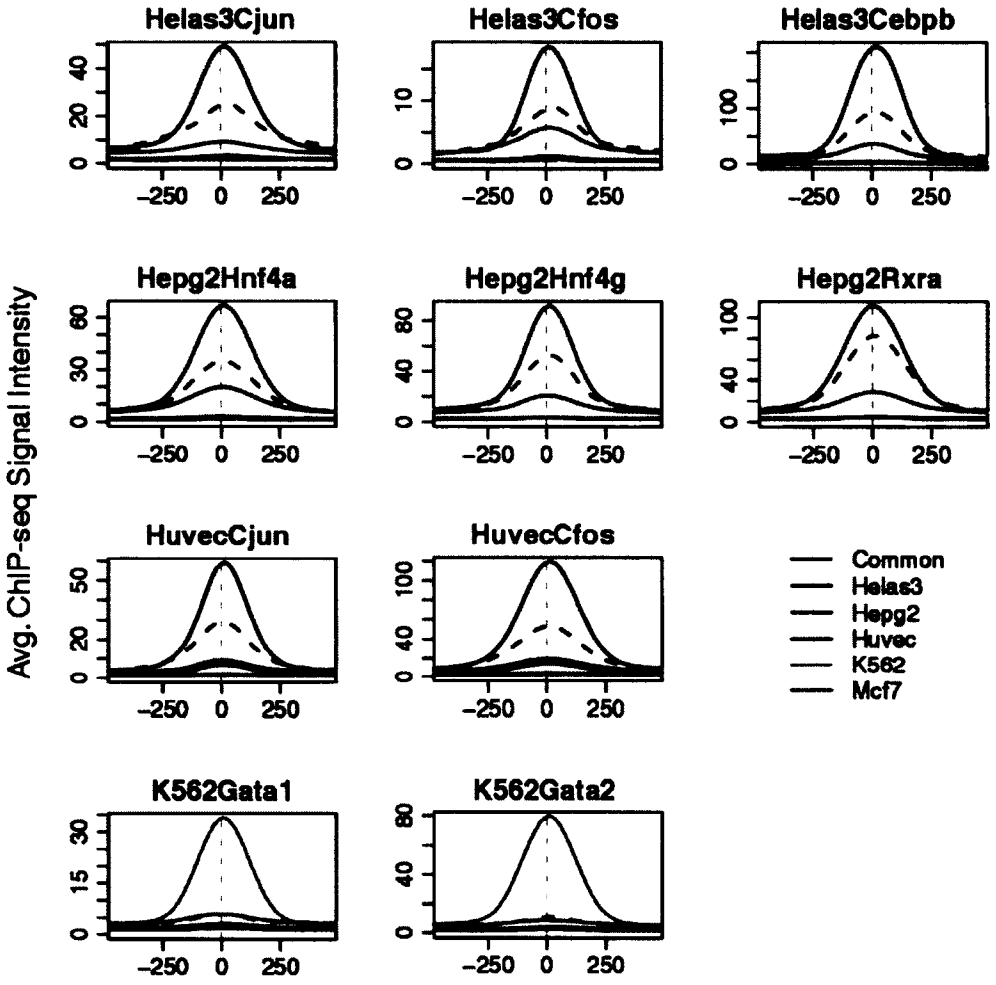
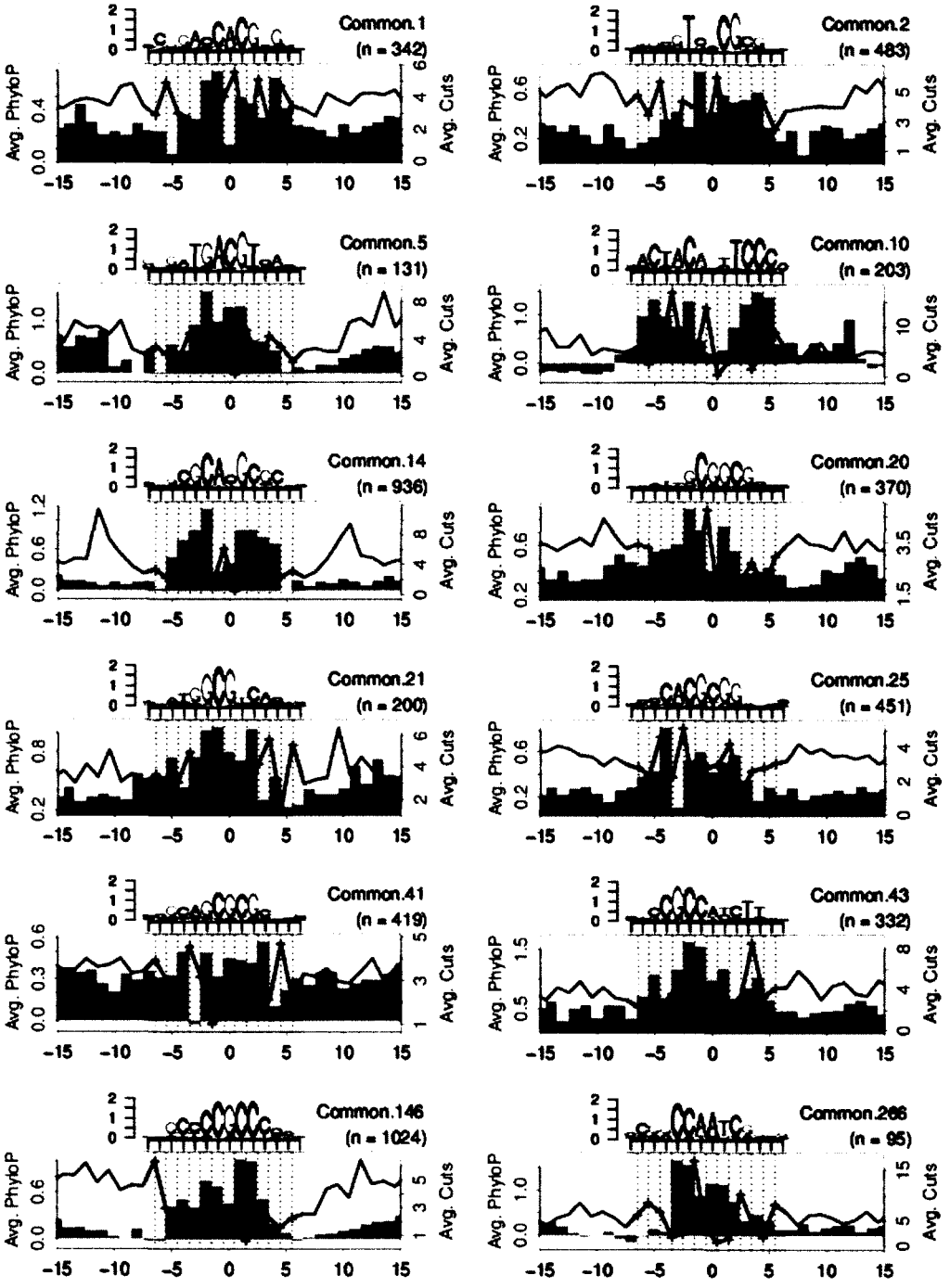



**Figure 6-14: Average ChIP-seq signal profiles of the c-Myc bound regions for TFs enriched in the cell-type specific bound loci**

Binding sites of the TFs are specifically enriched in the corresponding cell-type specific c-Myc bound regions. Dashed lines are the corresponding cell-type specific c-Myc bound regions without the cognate binding sites. Similar to the previous observation, ChIP-seq signals are specifically enriched only in the corresponding c-Myc bound regions containing the cognate binding sites.

### 6.2.9 PhyloP Scores and DNaseI Footprints of *de novo* PWMs

It is generally known that functional TFBSs are more evolutionarily conserved than non-TFBSs. Also, characteristic DNaseI cut (or DNaseI footprint) patterns are commonly observed in genome-wide *in vivo* TFBSs [151]. I analyzed these two properties of the *de novo* PWMs to further support their biological relevance. I used PhyloP scores [152] (calculated from the multiple alignments of 46 vertebrate species, accessed from UCSC genome browser [93]) for conservation, and Digital Genomic Footprints datasets [151] for DNaseI cut patterns. For each column of each *de novo* PWM, the average PhyloP score from the aligned gkm-peaks was measured. Similarly, the average number of DNaseI cuts from the aligned gkm-peaks was measured in the corresponding cell type. It should be noted that K562 DNaseI cut profile was used for Common PWMs. Also, DNaseI cut profiles of HelaS3 and MCF7 was not available at the time of the analysis. Therefore, those two profiles were not included.




(Continued on the next page)

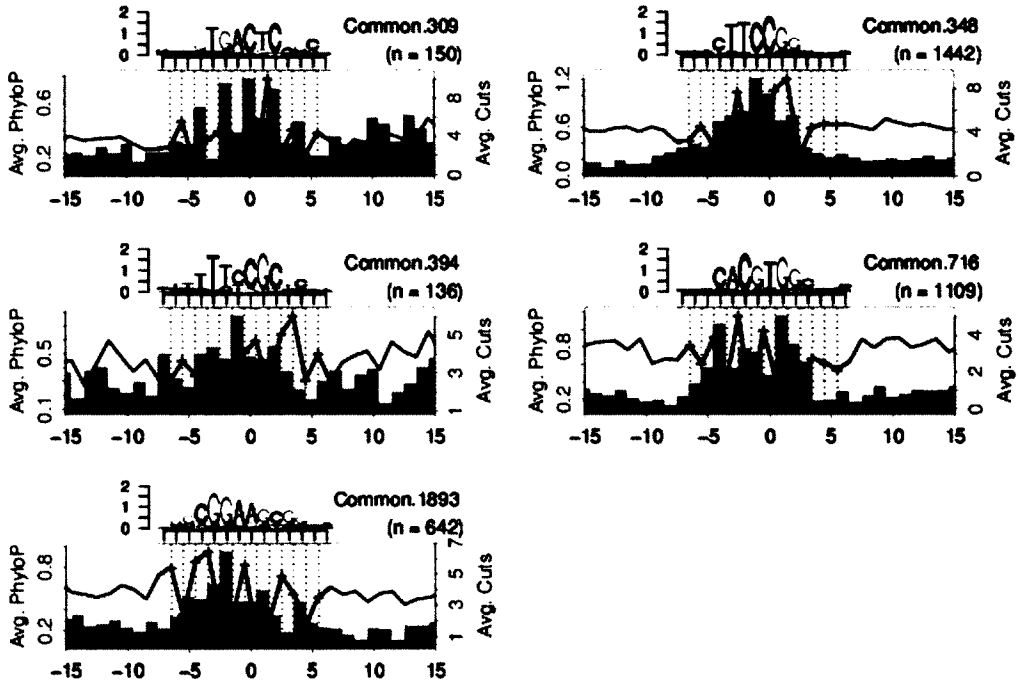



**Figure 6-15: Conservation and DNaseI cut profiles of Common PWMs**

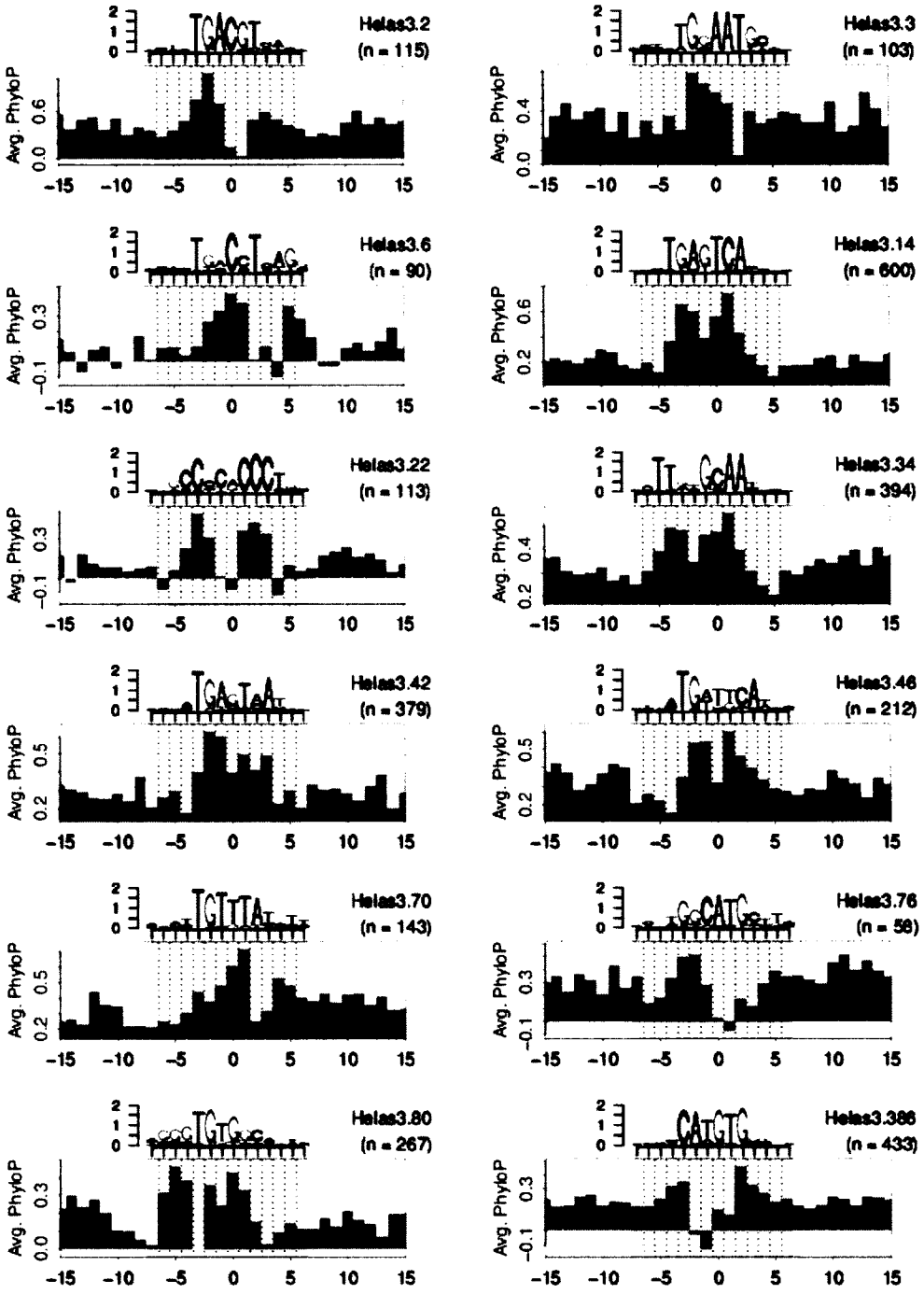



**Figure 6-16: Conservation and DNaseI cut profiles of HelaS3 PWMs**

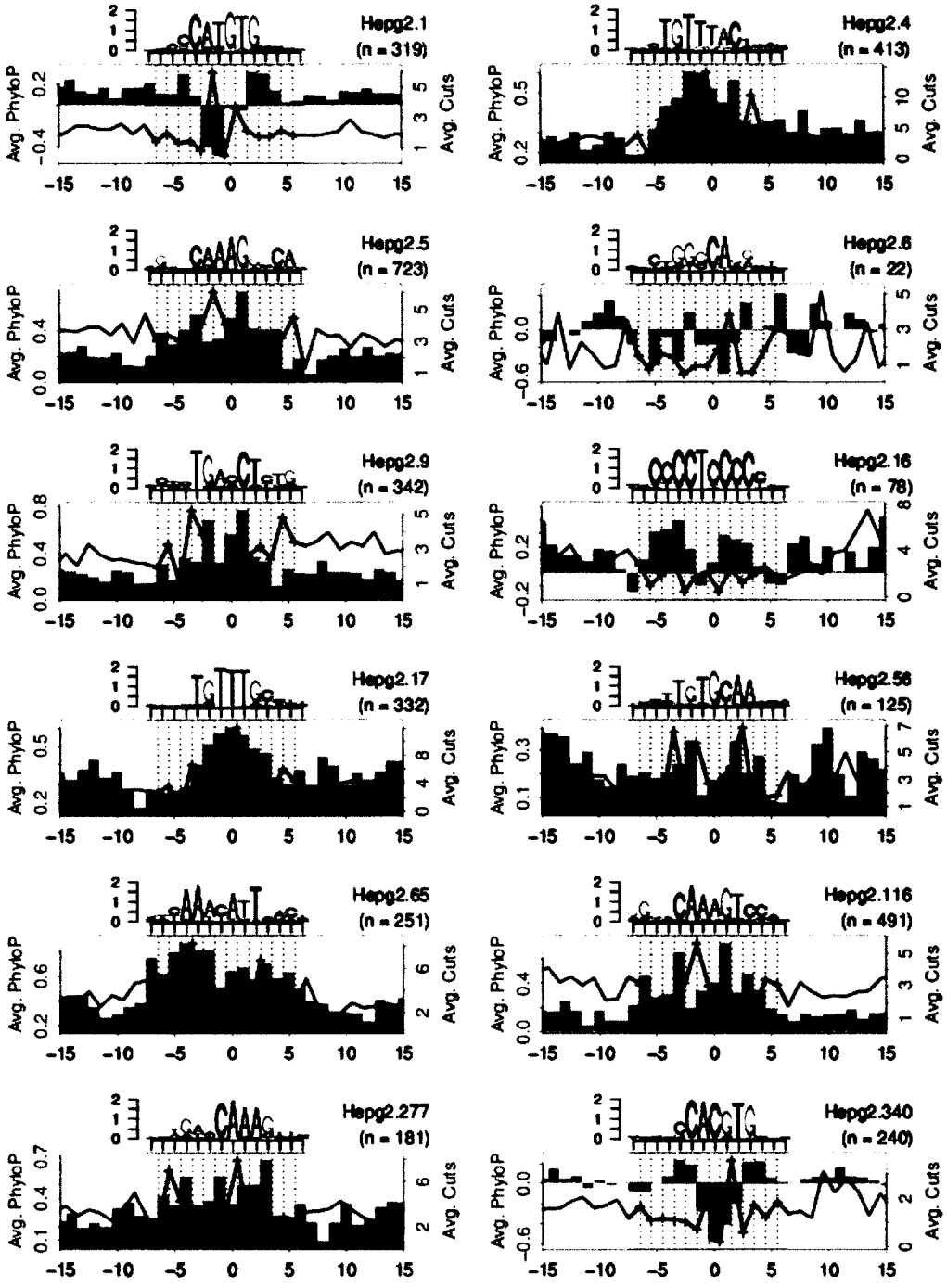



**Figure 6-17: Conservation and DNaseI cut profiles of HepG2 PWMs**

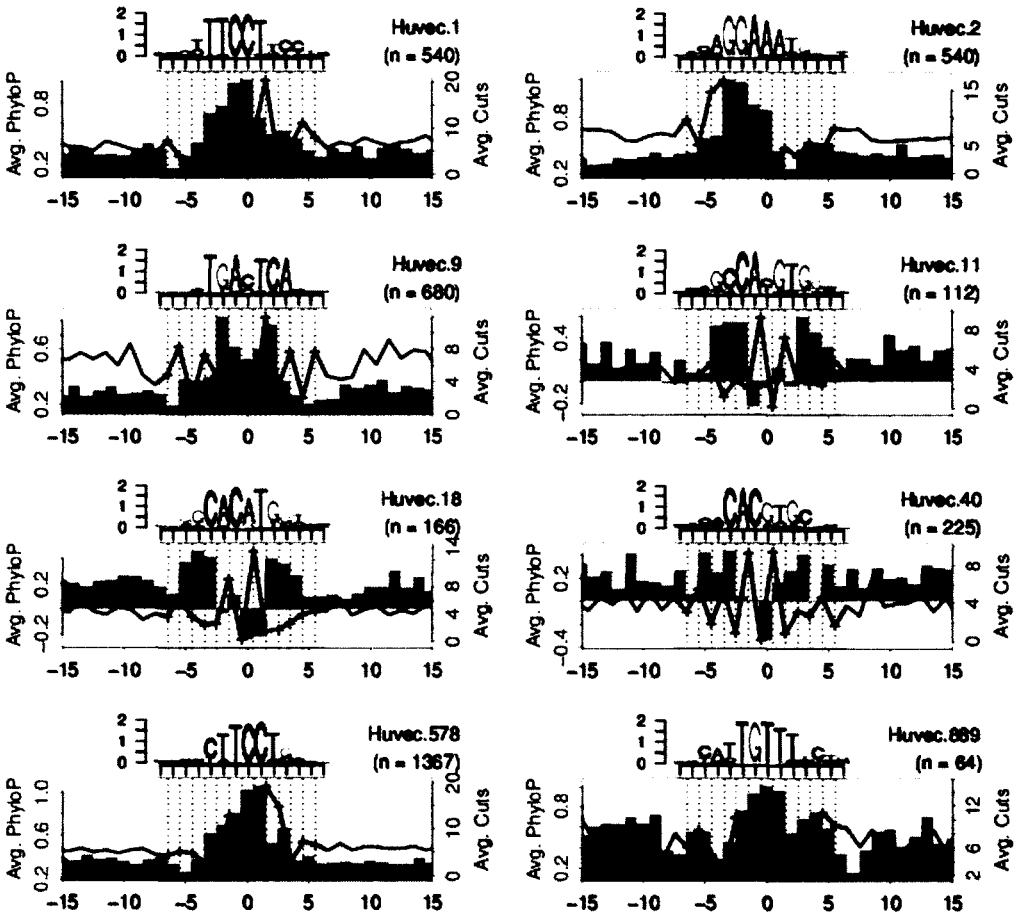



**Figure 6-18: Conservation and DNaseI cut profiles of HUVEC PWMs**

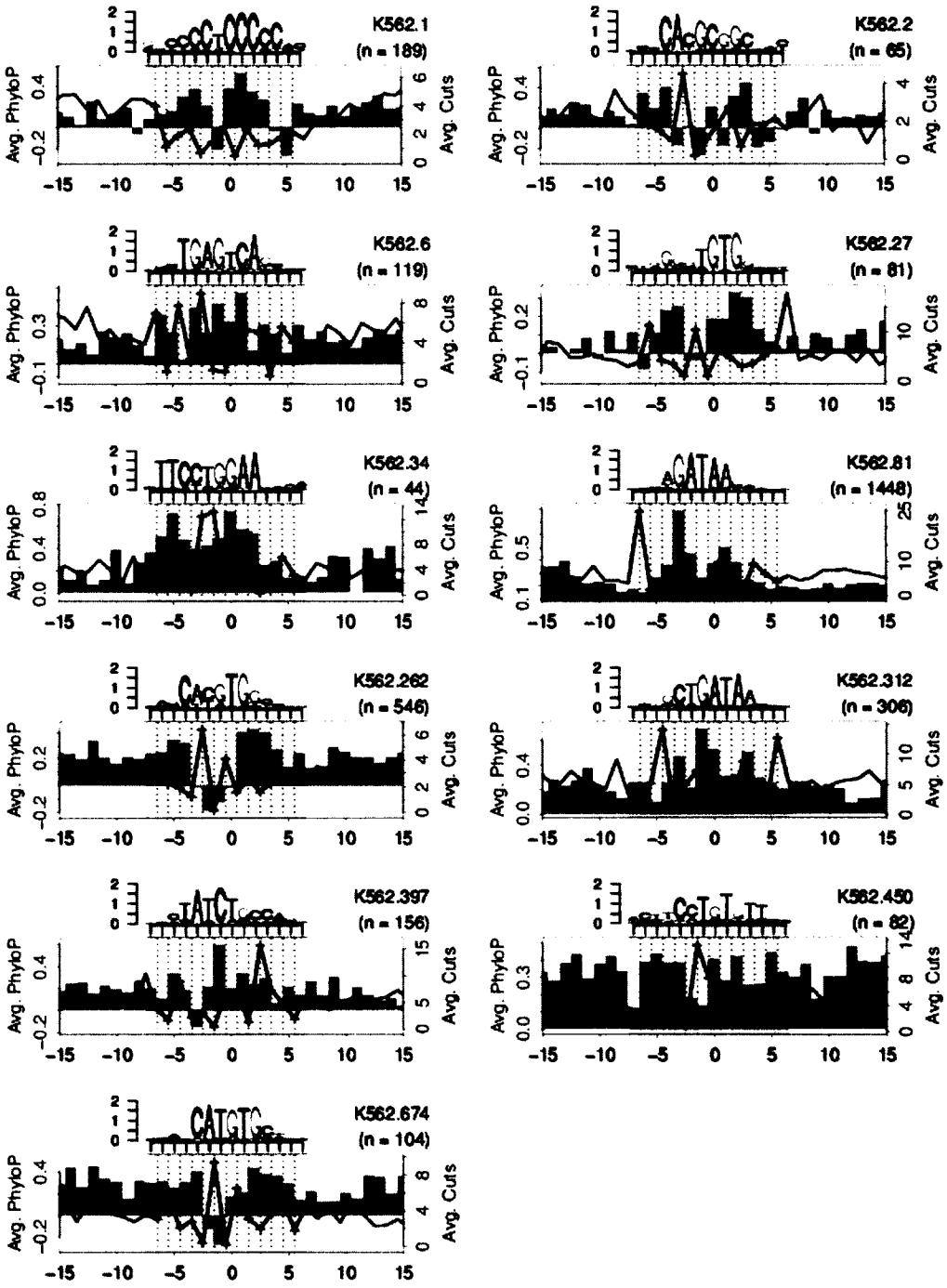



**Figure 6-19: Conservation and DNaseI cut profiles of K562 PWMs**

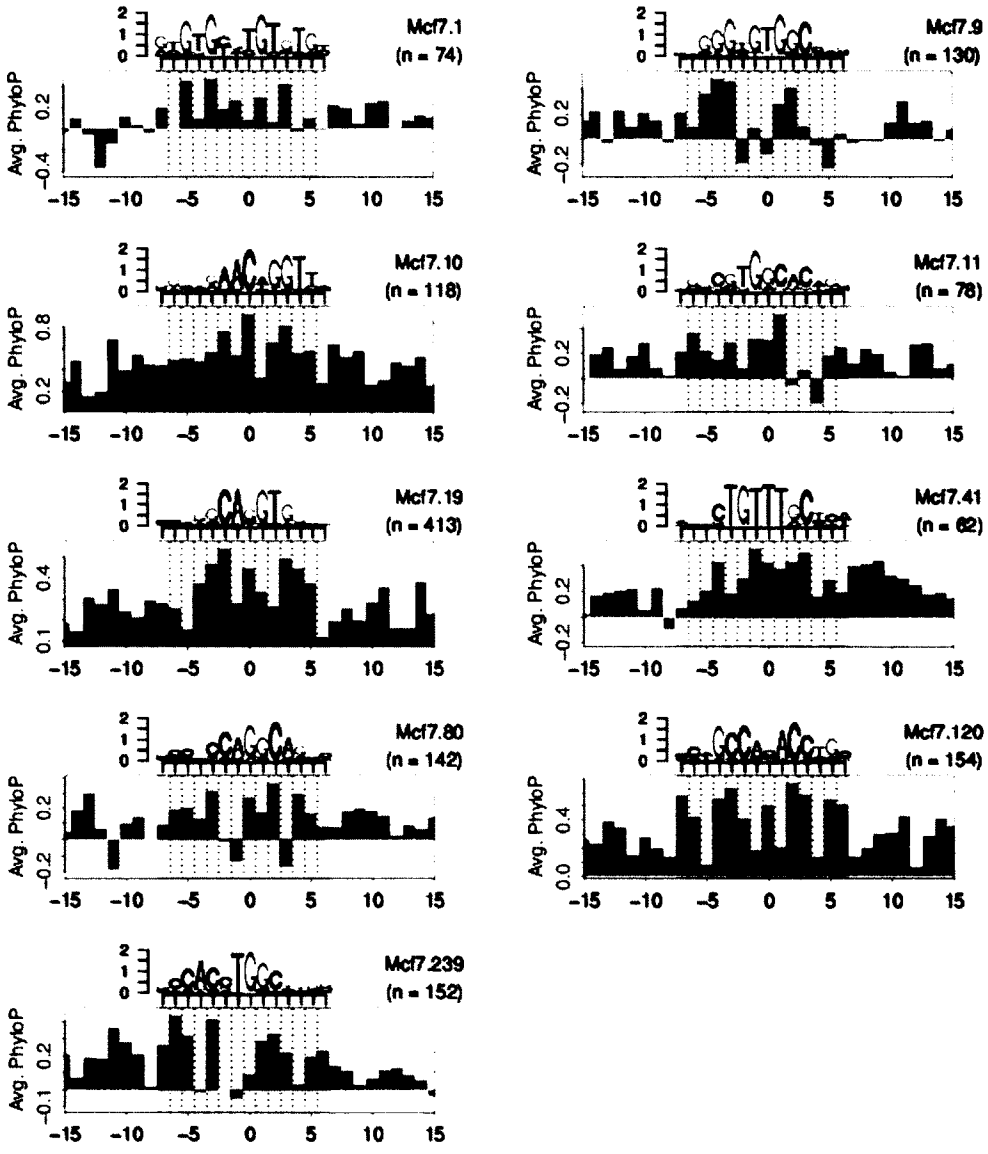



**Figure 6-20: Conservation and DNaseI cut profiles of MCF7 PWMs**

## 6.3 Results

### 6.3.1 Gkm-SVM Accurately Predicts Genomic c-Myc Bound Regions

A crucial underlying hypothesis of this study is that local primary DNA context of the c-Myc bound regions contains a sufficient amount of information to distinguish between differential genomic binding patterns of the c-Myc TF across diverse conditions. I directly test this hypothesis using a sequence-based discriminative method that can accurately predict regulatory sequences from the genome (see Chapter 5). I first define the 8,105 distinct regions that are either exclusively bound by c-Myc in just one cell type or commonly bound in all cell types by integrating c-Myc ChIP-seq datasets from five distinct cell lines; HelaS3, HepG2, HUVEC, K562, and MCF7, as described in detail in the section 6.2.1 (Figure 6-1).

Interestingly, while most of the common c-Myc bound regions are located at promoters consistent with the well-established role of c-Myc as a promoter associated TF, most of the cell-type specific c-Myc bound regions are distant from any known transcription start sites (Figure 6-3A). I hypothesized that these distal c-Myc bound regions might actually be cell-type specific enhancers, and indeed found that the expressions of the nearest genes of the cell-type specific c-Myc bound regions in the bound cell-type is significantly higher than the expressions of the same genes in the unbound cell-types (Figure 6-3B). I further performed gene ontology (GO) enrichment analysis of those genes, and discovered several biologically relevant GO terms enriched

in these sets of genes (Table 6-1). Although the associations between the c-Myc bound regions and their nearest genes are by no means perfect, these positive correlations suggests that many of these distal regions can act as cell-type specific distal enhancers. This idea can also be directly assessed by analyzing relevant ChIP-seq datasets of histone modification markers. Figure 6-4 shows heatmaps of three representative histone modification markers for either enhancers or promoters or both, and further support the previous analysis.

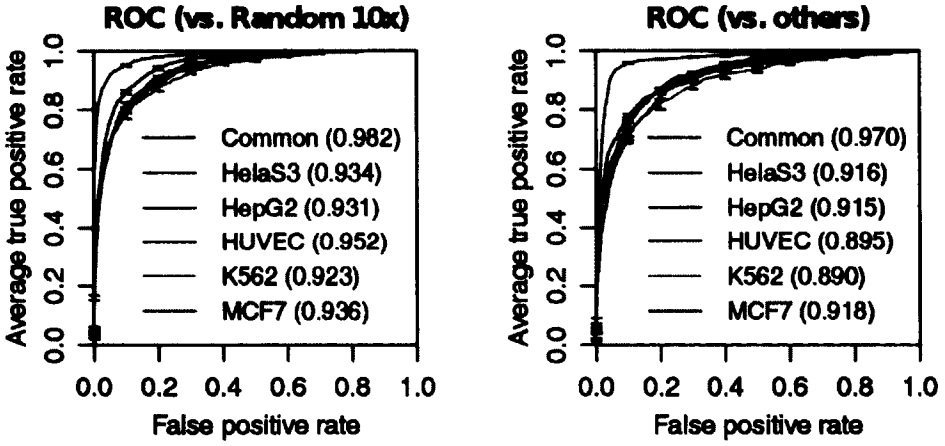



**Figure 6-21: Classification results of cell-type specific c-Myc bindings**
(*Left*) Classification of c-Myc bound regions of each class vs. 10X larger random genomic regions. (*Right*) Classification of c-Myc bound regions of each type vs. c-Myc bound regions of combined remaining types. Each dataset is classified with high accuracy (auROC > 0.9).

| | | Predicted | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Comm. | HelaS3 | HepG2 | HUVEC | K562 | MCF7 | acc |
| Actual | Comm. | **1629** | 18 | 34 | 42 | 27 | 39 | **.911** |
| | HelaS3 | 11 | **845** | 81 | 125 | 83 | 76 | **.692** |
| | HepG2 | 41 | 87 | **966** | 144 | 86 | 95 | **.696** |
| | HUVEC | 81 | 118 | 95 | **978** | 68 | 85 | **.686** |
| | K562 | 109 | 101 | 106 | 75 | **765** | 56 | **.631** |
| | MCF7 | 87 | 26 | 112 | 96 | 47 | **701** | **.656** |

**Table 6-2: Classification accuracy of the six-class gkm-SVM classifier**

I then asked whether these differential occupancies of c-Myc in different cell types could be predicted by their local DNA sequences. We first trained a gkm-SVM on each set of (cell-type-specific or common) c-Myc bound regions against 10X larger random genomic sequences and measured the performance using ROC curves with a standard five-fold cross validation technique. We also trained a gkm-SVM on each set against the others to determine whether those regions are directly separable from each other. Figure 6-21 shows that gkm-SVM can discriminate each set from the others with reasonable accuracy (auROC > 0.89) as well as from random sequences with high accuracy (auROC > 0.92). By combining the six binary SVM classifiers trained on one against others, I further constructed a multi-class classifier and achieved 72.7% overall classification accuracy (Table 6-2). It should be noted that the accuracy achieved here is significantly higher than a random classification accuracy (1/6 = 16.7%). Figure 6-22 further highlights how well gkm-SVM predictions (blue and red) can recapitulate other genomic signals from original experiments (c-Myc ChIP-seq, orange) and other relevant data sets (DNaseI-seq, green). Remarkable resemblance between gkm-SVM predictions and

experimental results strongly support our finding that local DNA sequence determines cell-specific c-Myc binding, via co-binding of distinct TFs flanking the c-Myc binding sites.

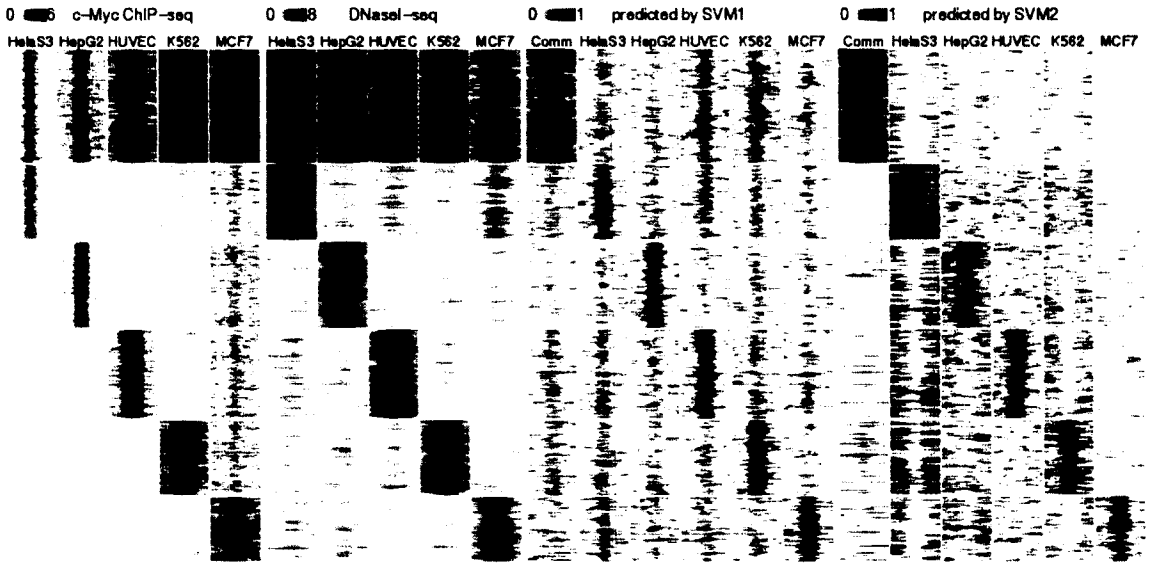



**Figure 6-22: Comparison of ChIP-seq, DNaseI, and gkm-SVM predictions in c-Myc bound loci**

Heatmaps of various types of genomic signals are compared. 1,000 bp regions are centered at the midpoint of the c-Myc bound regions. The first (c-Myc ChIP-seq) and second (DNaseI-seq) panels show logarithm of ChIP-seq and DNaseI signals of every 20 bp region. The third (SVM1) and fourth (SVM2) panels show the posterior probabilities estimated from SVM scores of 300 bp sliding window at every 10 bp interval. SVM1 was trained using 10X random genomic sequences as a negative set, whereas SVM2 used union of c-Myc bound regions in other cell-types as a negative set. SVM scores calculated from primary DNA sequence can remarkably recapitulate the original experiment signals.

### 6.3.2 Gkm-SVM Predicts Fine Scale Structures of c-Myc Bound Loci

The robustness of the gkm-SVM also enables us to identify fine structure of regulatory regions with single base pair resolution. Figure 6-23 shows a specific example of the fine

scale prediction at the well-studied CAD gene promoter known to be regulated by c-Myc in several cell types [153]. As demonstrated by the ChIP-seq experiments, this promoter is strongly occupied by c-Myc in all the five cell types and correctly predicted by the gkm-SVMs. Furthermore, we show that the prediction of the fine scale structure of this promoter precisely identifies TFBSs that mostly contribute to the SVM prediction. This analysis reveals several distinct high scoring peaks, some of which overlap with previously characterized Sp-1 binding sites [154] as well as c-Myc binding sites. More significantly, our prediction in this specific locus generally agrees with DNaseI genomic footprint profiles [151] measured in the three cell types.

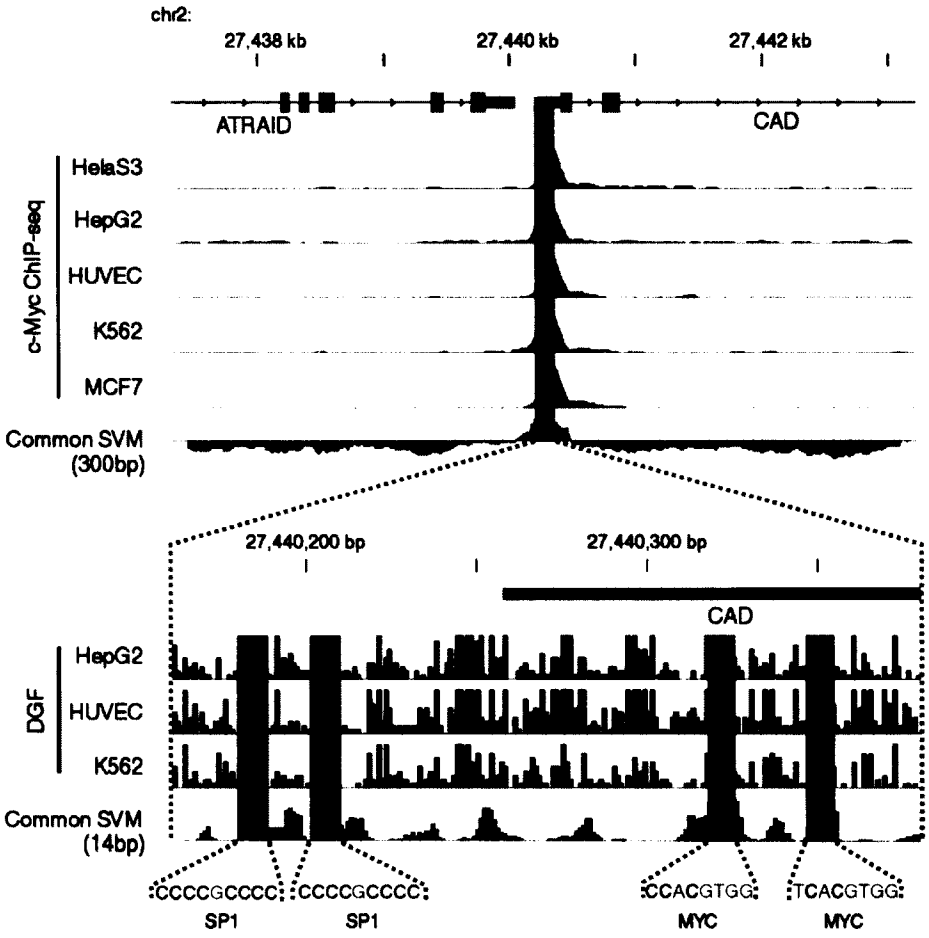



**Figure 6-23: An example of fine scale structure prediction in the CAD promoter**
Individual TFBSs are precisely predicted by gkm-SVM. (*Top*) The "c-Myc ChIP-seq" tracks show that the CAD promoter is strongly bound by c-Myc in all cell types. The "Common SVM (300 bp)" track shows the SVM score of 300 bp sub-sequence at each position. (*Bottom, Zoom-in*) "DGF" tracks shows invariant DNaseI cut profiles across the three cell types, and the last track "Common SVM" shows predicted TFBSs in this locus by scoring 14 bp oligomer at every position.

We hypothesized that these high scoring peaks, referred to as gkm-peaks, may be clustered into a few distinct TFBSs that determine the cell-type specificity of the c-Myc binding patterns. To test this idea, I scanned all c-Myc bound regions with the corresponding gkm-SVM to identify all potential TFBSs (as gkm-peaks) (see the section 6.2.6 for details), and then identified a set of clusters using affinity propagation clustering

method [149] for each dataset. I further aligned the gkm-peaks and generated a position weight matrix (PWM) model for each of the clusters using an expectation maximization algorithm introduced in the section 6.2.7. For each set of c-Myc bound regions, I found *de novo* 10~15 PWMs, most of which almost perfectly match to known PWMs (Figure 6-7 ~ Figure 6-12). I asked whether these identified known TFBSs are also occupied by the cognate TFs *in vivo* in the corresponding cell types. By analyzing ChIP-seq datasets for those identified TFs, I found a striking correlation between the predicted TFBSs and their actual binding patterns (Figure 6-13 and Figure 6-14). Moreover, as revealed by PhyloP [152] and DNaseI cut profile analysis [151], these *de novo* identified TFBSs are more evolutionarily conserved and displayed specific DNaseI cut patterns consistent with previous observations [151]. Taken together, this evidence from multiple independent genomic datasets establishes the biological relevance of the *de novo* PWMs for accessory transcription factors.

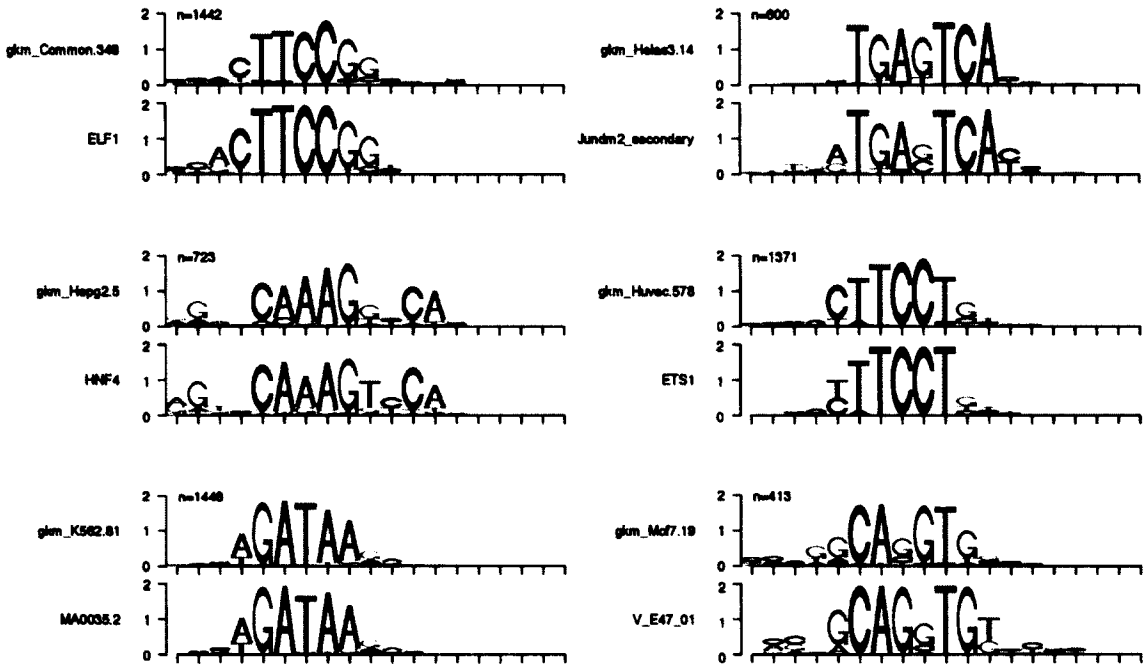



**Figure 6-24: The most enriched *de novo* PWM for each set**
The most frequently observed TFBS for each dataset is presented. For each pair (top and bottom) of PWMs, the top one is a *de novo* PWM and the bottom one is the known PWM that matches to the top PWM.

Although some PWMs are shared between multiple types of c-Myc bound regions, the most enriched PWM for each dataset is distinct, as shown in Figure 6-24. More significantly, many of these TFBSs are recognized by TFs that are known to play critical roles in the corresponding cell types. For example, the GATA1 TF identified from K562 specific c-Myc bound regions is known to play essential roles in erythroid differentiation [155]. The HNF4 TF which is specific to HepG2 is also one of the best known regulators for liver specific cell lineages [156]. The ETS1 TF from HUVEC is another extensively studied TF and members of the ETS family are known to be important for angiogenesis [157].

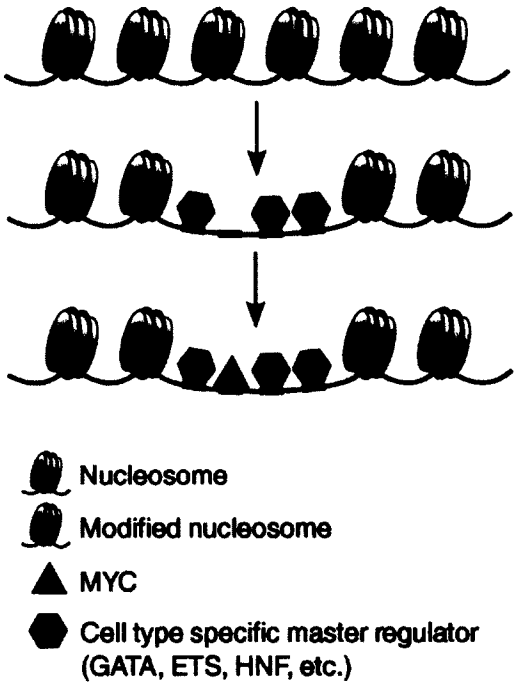



**Figure 6-25: Cell-type specific c-Myc binding model**

This observation led us to hypothesize that cell-specific regulatory regions might largely be predetermined by a few major TFs, and that c-Myc is then bound to those activated or poised regions which also contain E-box elements (Figure 6-25). To directly test this idea, I repeated the same analysis using DNaseI-seq datasets from the five cell types, and asked whether the most enriched *de novo* PWMs are still the same. Since DNaseI hypersensitive sites (DHSs) capture a broad range of regulatory elements and are not limited to a specific TF, predictive sequence features for the cell-type specific DHSs would further provide us new insights into the mechanisms of cell-specific c-Myc bound regions. Figure 6-26 shows the most enriched PWM for each set of the DHSs. Strikingly, these *de novo* PWMs are almost identical to the originally found PWMs from the cell-type specific c-Myc bound regions, but the Myc binding site is itself notably absent. For

the common DHSs, the SP1 binding site is now identified as the most frequently observed TFBS, which is also significantly enriched in the common c-Myc bound regions (the 3rd most enriched PWM). For the MCF7 specific DHSs, a novel PWM was identified as the most enriched PWM, which is also observed in the MCF7 specific c-Myc bound regions as Mcf7.10 in Figure 6-12 and Figure 6-20. This novel PWM finding highlights the potential use of the method developed in this study.

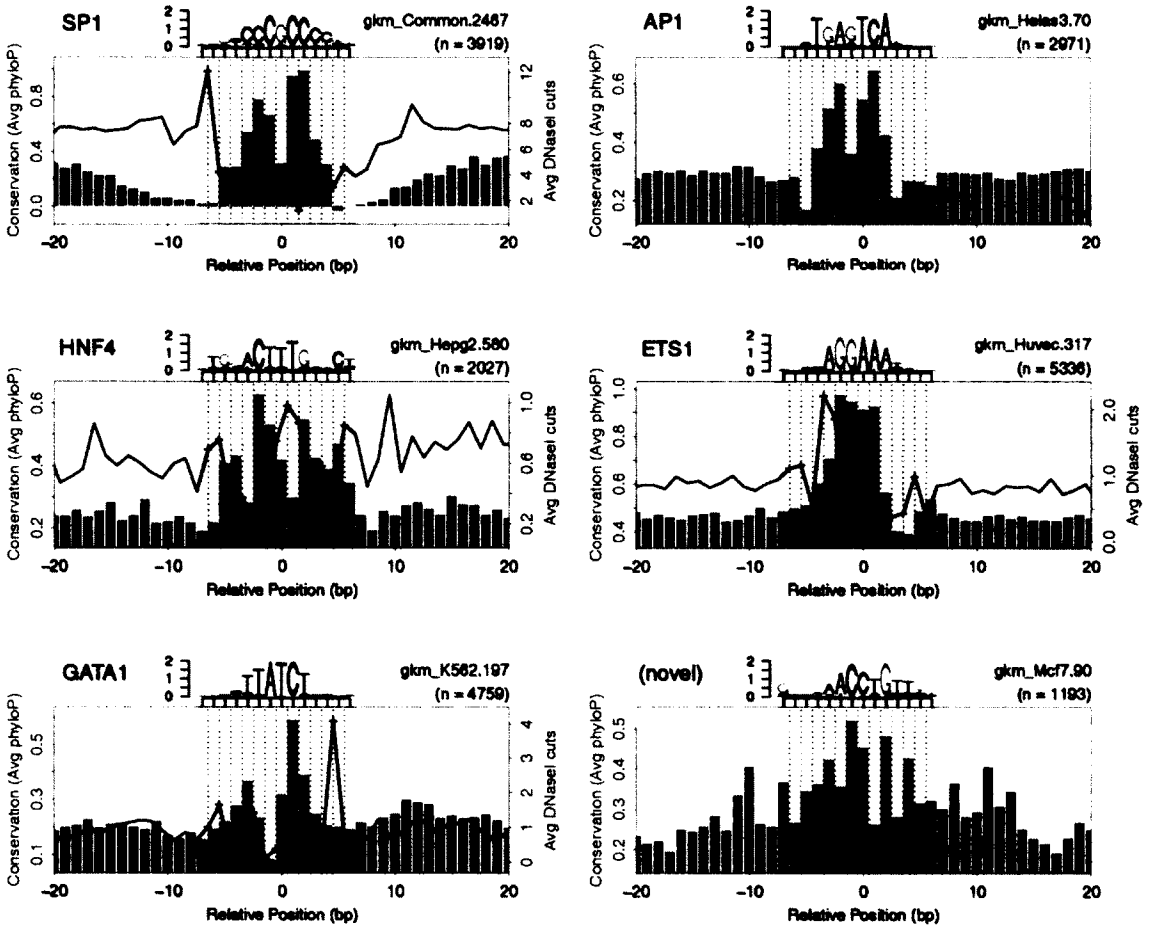



**Figure 6-26: Top *de novo* PWM for each set of cell-specific DHSs**

## 6.4 Discussion

In this chapter, I demonstrated a comprehensive computational approach to identify underlying molecular mechanisms by which the activity of cell-specific regulatory elements is modulated. As a paradigm of this, I identified several thousands of cell-type specific c-Myc bound regions and successfully predicted them from the primary DNA sequences by applying sequence-based discriminative models developed in previous chapters. I further advanced the method to precisely identify putative TFBSs within the regulatory regions, and discovered that the presence of distinct cofactor TFBSs nearby the c-Myc binding sites determines the binding of the c-Myc TF in the specific cell type. I also developed a new method that systematically analyze and summarize these predictive TFBSs, and found several key known regulators that determine the cell-type specific c-Myc binding. I believe that the demonstrated approach in this chapter will greatly facilitate to broaden our understanding of regulatory biology.

# 7 General Discussion

The field of regulatory genomics is undergoing rapid and exciting growth, spurred by the combination of many factors. It seems almost unnecessary to mention that the human reference genome has provided the framework for the development of several technologies which are coming together to revolutionize our understanding of the function and role of gene regulation. Key directions among these are: (1) The development of new experimental technologies for producing genomic maps of chromatin accessibility, histone modification state, and DNA binding by regulator proteins in an ever growing number of cell types, environmental conditions, and disease states; (2) The development of DNA sequence based machine learning approaches to detect regulatory elements; (3) The assessment of common human sequence variation and associations with disease. However, this enterprise is still in a relatively early stage, and is sure to yield many surprises as we develop a more complete understanding of regulatory mechanisms and evolution.

As discussed throughout my dissertation, the accuracy of enhancer prediction algorithms has improved dramatically, but there is still significant room for improvement, and the development of higher precision classifiers would have a substantial impact on the rate at which biologically and medically relevant enhancers can be identified. Progress will likely come in the area of development of better kernel functions or distance measures used by the classifiers. A key ingredient that has not been fully

exploited is detailed information about the spatial and configurational constraints between and within clusters of binding sites. While several kernel approaches which incorporate positional information have been introduced, most have been developed in the context of positional constraints relative to a single preferred genomic location or anchor point. In applications to other problems, positional information relative to a transcription start site [65], to a splice site [89], [90] or to a translational start site [158] 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.*[159]. 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.

While current approaches are improving in their ability to detect combinations of cofactor binding sites, most scoring functions are invariant to arbitrary reshuffling of these binding sites. It seems unlikely that this type of variation would completely preserve enhancer function. A challenge here will be that high order grammatical structures and more complex statistical learning models require more data for training, and there is limited variation in the existing data (evolution tends to discard its mistakes). Therefore, generation of large synthetic enhancer datasets might be necessary to fully test the space of regulatory variation.

Ultimately, enhancer prediction tools should be able to predict the impact of DNA variants on cell-type specific enhancer function. This would be a significant advance, but

it is worth pointing out that even being able to precisely predict the strength of a mutated enhancer in isolation may not be sufficient to predict the phenotypic consequence of the mutation. Each enhancer has multiple inputs and operates within a highly connected regulatory network. A mutation, which strengthens an enhancer in one individual, may have a stronger or weaker effect in another individual because of nonlinear interactions with other variants. Nevertheless, our biological networks are extremely robust, so there may be simple design principles which help quantify these interactions. The critical mutations are those which most dramatically affect the overall output of the regulatory element in the context of its biological circuit.

# 8 Bibliography


[1] L. A. Hindorff, P. Sethupathy, H. A. Junkins, E. M. Ramos, J. P. Mehta, F. S. Collins, and T. A. Manolio, "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits," *PNAS*, vol. 106, no. 23, pp. 9362–9367, Jun. 2009.

[2] M. A. Beer and S. Tavazoie, "Predicting Gene Expression from Sequence," *Cell*, vol. 117, pp. 185–198, Apr. 2004.

[3] The ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012.

[4] Mouse ENCODE Project Consortium, J. A. Stamatoyannopoulos, M. Snyder, R. Hardison, B. Ren, T. Gingeras, D. M. Gilbert, M. Groudine, M. Bender, R. Kaul, T. Canfield, E. Giste, A. Johnson, M. Zhang, G. Balasundaram, R. Byron, V. Roach, P. J. Sabo, R. Sandstrom, A. S. Stehling, R. E. Thurman, S. M. Weissman, P. Cayting, M. Hariharan, J. Lian, Y. Cheng, S. G. Landt, Z. Ma, B. J. Wold, and J. Dekker, "An encyclopedia of mouse DNA elements (Mouse ENCODE)," *Genome Biology*, vol. 13, no. 8, p. 418, Aug. 2012.

[5] D. Lee, R. Karchin, and M. A. Beer, "Discriminative prediction of mammalian enhancers from DNA sequence," *Genome Research*, vol. 21, no. 12, pp. 2167 – 2180, Dec. 2011.

[6] D. U. Gorkin, D. Lee, X. Reed, C. Fletez-Brant, S. L. Bessling, S. K. Loftus, M. A. Beer, W. J. Pavan, and A. S. McCallion, "Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes," *Genome Res.*, vol. 22, no. 11, pp. 2290–2301, Nov. 2012.

[7] C. Fletez-Brant, D. Lee, A. S. McCallion, and M. A. Beer, "kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic datasets," *Submitted.*

[8] M. Ghandi, M. Mohammad-Noori, and M. A. Beer, "Robust k-mer Frequency Estimation Using Gapped k-mers," *Submitted.*

[9] M. Ghandi, D. Lee, M. Mohammad-Noori, and M. A. Beer, "Enhanced Sequence Classification using Robust k-mer Frequency Estimation," *In Preparation.*

[10] D. Lee and M. A. Beer, "Mammalian Enhancer Prediction," *in press.*

[11] F. Jacob and J. Monod, "Genetic regulatory mechanisms in the synthesis of proteins," *Discovering Molecular Genetics*, p. 192, 1961.

[12] M. Ptashne, *A genetic switch: phage lambda revisited.* Cold Spring Harbor Laboratory Pr, 2004.

[13] J. Banerji, "Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences," *Cell*, vol. 27, no. 2, pp. 299–308, Dec. 1981.

[14] E. M. Blackwood and J. T. Kadonaga, "Going the Distance: A Current View of Enhancer Action," *Science*, vol. 281, no. 5373, pp. 60–63, Jul. 1998.

[15] M. Bulger and M. Groudine, "Looping versus linking: toward a model for long-distance gene activation," *Genes Dev.*, vol. 13, no. 19, pp. 2465–2477, Oct. 1999.

[16] L. Patthy, "Genome evolution and the evolution of exon-shuffling — a review," *Gene*, vol. 238, no. 1, pp. 103–114, Sep. 1999.

[17] N. D. Heintzman, G. C. Hon, R. D. Hawkins, P. Kheradpour, A. Stark, L. F. Harp, Z. Ye, L. K. Lee, R. K. Stuart, C. W. Ching, K. A. Ching, J. E. Antosiewicz-Bourget, H. Liu, X. Zhang, R. D. Green, V. V. Lobanenkov, R. Stewart, J. A. Thomson, G. E. Crawford, M. Kellis, and B. Ren, "Histone modifications at human enhancers reflect global cell-type-specific gene expression," *Nature*, vol. 459, no. 7243, pp. 108–112, May 2009.

[18] Y. Shen, F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, and B. Ren, "A map of the cis-regulatory sequences in the mouse genome," *Nature*, vol. 488, no. 7409, pp. 116–120, Aug. 2012.

[19] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, Oct. 2009.

[20] T. A. Manolio, "Genomewide Association Studies and Assessment of the Risk of Disease," *New England Journal of Medicine*, vol. 363, no. 2, pp. 166–176, 2010.

[21] F. Grosveld, G. B. van Assendelft, D. R. Greaves, and G. Kollias, "Position-independent, high-level expression of the human β-globin gene in transgenic mice," *Cell*, vol. 51, no. 6, pp. 975–985, Dec. 1987.

[22] G. B. van Assendelft, O. Hanscombe, F. Grosveld, and D. R. Greaves, "The β-globin dominant control region activates homologous and heterologous promoters in a tissue-specific manner," *Cell*, vol. 56, no. 6, pp. 969–977, Mar. 1989.

[23] J. H. Chung, M. Whiteley, and G. Felsenfeld, "A 5′ element of the chicken β-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila," *Cell*, vol. 74, no. 3, pp. 505–514, Aug. 1993.

[24] L. A. Lettice, T. Horikoshi, S. J. H. Heaney, M. J. van Baren, H. C. van der Linde, G. J. Breedveld, M. Joosse, N. Akarsu, B. A. Oostra, N. Endo, M. Shibata, M. Suzuki, E. Takahashi, T. Shinka, Y. Nakahori, D. Ayusawa, K. Nakabayashi, S. W. Scherer, P. Heutink, R. E. Hill, and S. Noji, "Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly," *PNAS*, vol. 99, no. 11, pp. 7548–7553, May 2002.

[25] L. A. Lettice, S. J. H. Heaney, L. A. Purdie, L. Li, P. de Beer, B. A. Oostra, D. Goode, G. Elgar, R. E. Hill, and E. de Graaff, "A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly," *Hum. Mol. Genet.*, vol. 12, no. 14, pp. 1725–1735, Jul. 2003.

[26] J. D. Lauderdale, J. S. Wilensky, E. R. Oliver, D. S. Walton, and T. Glaser, "3′ deletions cause aniridia by preventing PAX6 gene expression," *PNAS*, vol. 97, no. 25, pp.

13755–13759, Dec. 2000.

[27] D. A. Kleinjan and V. van Heyningen, "Long-Range Control of Gene Expression: Emerging Mechanisms and Disruption in Disease," *The American Journal of Human Genetics*, vol. 76, no. 1, pp. 8–32, Jan. 2005.

[28] E. S. Emison, A. S. McCallion, C. S. Kashuk, R. T. Bush, E. Grice, S. Lin, M. E. Portnoy, D. J. Cutler, E. D. Green, and A. Chakravarti, "A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk," *Nature*, vol. 434, no. 7035, pp. 857–863, Apr. 2005.

[29] E. A. Grice, E. S. Rochelle, E. D. Green, A. Chakravarti, and A. S. McCallion, "Evaluation of the RET regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer," *Hum. Mol. Genet.*, vol. 14, no. 24, pp. 3837–3845, Dec. 2005.

[30] X. Zhang, R. Cowper-Sal·lari, S. D. Bailey, J. H. Moore, and M. Lupien, "Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus," *Genome Res.*, vol. 22, no. 8, pp. 1437–1446, Aug. 2012.

[31] R. J. Britten and E. H. Davidson, "Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty," *Q Rev Biol*, vol. 46, no. 2, pp. 111–138, Jun. 1971.

[32] M. C. King and A. C. Wilson, "Evolution at two levels in humans and chimpanzees," *Science*, vol. 188, no. 4184, pp. 107–116, Apr. 1975.

[33] A. H. Brand and N. Perrimon, "Targeted gene expression as a means of altering cell fates and generating dominant phenotypes," *Development*, vol. 118, no. 2, pp. 401–415, Jun. 1993.

[34] R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. E. Antonarakis, J. Attwood, R. Baertsch, J. Bailey, K. Barlow, S. Beck, E. Berry, B. Birren, T. Bloom, P. Bork, M. Botcherby, N. Bray, M. R. Brent, D. G. Brown, S. D. Brown, C. Bult, J. Burton, J. Butler, R. D. Campbell, P. Carninci, S. Cawley, F. Chiaromonte, A. T. Chinwalla, D. M. Church, M. Clamp, C. Clee, F. S. Collins, L. L. Cook, R. R. Copley, A. Coulson, O. Couronne, J. Cuff, V. Curwen, T. Cutts, M. Daly, R. David, J. Davies, K. D. Delehaunty, J. Deri, E. T. Dermitzakis, C. Dewey, N. J. Dickens, M. Diekhans, S. Dodge, I. Dubchak, D. M. Dunn, S. R. Eddy, L. Elnitski, R. D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, G. A. Fewell, P. Flicek, K. Foley, W. N. Frankel, L. A. Fulton, R. S. Fulton, T. S. Furey, D. Gage, R. A. Gibbs, G. Glusman, S. Gnerre, N. Goldman, L. Goodstadt, D. Grafham, T. A. Graves, E. D. Green, S. Gregory, R. Guigó, M. Guyer, R. C. Hardison, D. Haussler, Y. Hayashizaki, L. W. Hillier, A. Hinrichs, W. Hlavina, T. Holzer, F. Hsu, A. Hua, T. Hubbard, A. Hunt, I. Jackson, D. B. Jaffe, L. S. Johnson, M. Jones, T. A. Jones, A. Joy, M. Kamal, E. K. Karlsson, D. Karolchik, A. Kasprzyk, J. Kawai, E. Keibler, C. Kells, W. J. Kent, A. Kirby, D. L. Kolbe, I. Korf, R. S. Kucherlapati, E. J. Kulbokas, D. Kulp, T. Landers, J. P. Leger, S. Leonard, I. Letunic, R. Levine, J. Li, M. Li, C. Lloyd, S. Lucas, B. Ma, D. R. Maglott, E. R. Mardis, L. Matthews, E. Mauceli, J. H. Mayer, M. McCarthy, W. R. McCombie, S. McLaren, K. McLay, J. D. McPherson, J. Meldrim, B. Meredith, J. P. Mesirov, W. Miller, T. L. Miner, E. Mongin, K. T. Montgomery, M. Morgan, R. Mott, J. C. Mullikin, D. M. Muzny, W. E. Nash, J. O. Nelson, M. N. Nhan, R. Nicol, Z. Ning, C. Nusbaum, M. J. O'Connor, Y. Okazaki, K. Oliver, E.

Overton-Larty, L. Pachter, G. Parra, K. H. Pepin, J. Peterson, P. Pevzner, R. Plumb, C. S. Pohl, A. Poliakov, T. C. Ponce, C. P. Ponting, S. Potter, M. Quail, A. Reymond, B. A. Roe, K. M. Roskin, E. M. Rubin, A. G. Rust, R. Santos, V. Sapojnikov, B. Schultz, J. Schultz, M. S. Schwartz, S. Schwartz, C. Scott, S. Seaman, S. Searle, T. Sharpe, A. Sheridan, R. Shownkeen, S. Sims, J. B. Singer, G. Slater, A. Smit, D. R. Smith, B. Spencer, A. Stabenau, N. Stange-Thomann, C. Sugnet, M. Suyama, G. Tesler, J. Thompson, D. Torrents, E. Trevaskis, J. Tromp, C. Ucla, A. Ureta-Vidal, J. P. Vinson, A. C. Von Niederhausern, C. M. Wade, M. Wall, R. J. Weber, R. B. Weiss, M. C. Wendl, A. P. West, K. Wetterstrand, R. Wheeler, S. Whelan, J. Wierzbowski, D. Willey, S. Williams, R. K. Wilson, E. Winter, K. C. Worley, D. Wyman, S. Yang, S.-P. Yang, E. M. Zdobnov, M. C. Zody, and E. S. Lander, "Initial sequencing and comparative analysis of the mouse genome," *Nature*, vol. 420, no. 6915, pp. 520–562, Dec. 2002.

[35] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young, "Transcriptional regulatory code of a eukaryotic genome," *Nature*, vol. 431, no. 7004, pp. 99–104, 2004.

[36] K. I. Zeller, X. Zhao, C. W. H. Lee, K. P. Chiu, F. Yao, J. T. Yustein, H. S. Ooi, Y. L. Orlov, A. Shahab, H. C. Yong, Y. Fu, Z. Weng, V. A. Kuznetsov, W.-K. Sung, Y. Ruan, C. V. Dang, and C.-L. Wei, "Global mapping of c-Myc binding sites and target gene networks in human B cells," *PNAS*, vol. 103, no. 47, pp. 17834–17839, Nov. 2006.

[37] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigó, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, J. A. Stamatoyannopoulos, R. E. Thurman, M. S. Kuehn, C. M. Taylor, S. Neph, C. M. Koch, S. Asthana, A. Malhotra, I. Adzhubei, J. A. Greenbaum, R. M. Andrews, P. Flicek, P. J. Boyle, H. Cao, N. P. Carter, G. K. Clelland, S. Davis, N. Day, P. Dhami, S. C. Dillon, M. O. Dorschner, H. Fiegler, P. G. Giresi, J. Goldy, M. Hawrylycz, A. Haydock, R. Humbert, K. D. James, B. E. Johnson, E. M. Johnson, T. T. Frum, E. R. Rosenzweig, N. Karnani, K. Lee, G. C. Lefebvre, P. A. Navas, F. Neri, S. C. J. Parker, P. J. Sabo, R. Sandstrom, A. Shafer, D. Vetrie, M. Weaver, S. Wilcox, M. Yu1, F. S. Collins, J. Dekker, J. D. Lieb, T. D. Tullius, G. E. Crawford, S. Sunyaev, W. S. Noble, I. Dunham, A. Dutta, R. Guigó, F. Denoeud, A. Reymond, P. Kapranov, J. Rozowsky, D. Zheng, R. Castelo, A. Frankish, J. Harrow, S. Ghosh, A. Sandelin, I. L. Hofacker, R. Baertsch, D. Keefe, P. Flicek, S. Dike, J. Cheng, H. A. Hirsch, E. A. Sekinger, J. Lagarde, J. F. Abril, A. Shahab, C. Flamm, C. Fried, J. Hackermüller, J. Hertel, M. Lindemeyer, K. Missal, A. Tanzer, S. Washietl, J. Korbel, O. Emanuelsson, J. S. Pedersen, N. Holroyd, R. Taylor, D. Swarbreck, N. Matthews, M. C. Dickson, D. J. Thomas, M. T. Weirauch, J. Gilbert, J. Drenkow, I. Bell, X. Zhao, K. G. Srinivasan, W.-K. Sung, H. S. Ooi, K. P. Chiu, S. Foissac, T. Alioto, M. Brent, L. Pachter, M. L. Tress, A. Valencia, S. W. Choo, C. Y. Choo, C. Ucla, C. Manzano, C. Wyss, E. Cheung, T. G. Clark, J. B. Brown, M. Ganesh, S. Patel, H. Tammana, J. Chrast, C. N. Henrichsen, C. Kai, J. Kawai, U. Nagalakshmi, J. Wu, Z. Lian, J. Lian, P. Newburger, X. Zhang, P. Bickel, J. S. Mattick, P. Carninci, Y. Hayashizaki, S. Weissman, E. T. Dermitzakis, E. H. Margulies, T. Hubbard, R. M. Myers, J. Rogers, P. F. Stadler, T. M. Lowe, C.-L. Wei, Y. Ruan, M. Snyder, E. Birney, K. Struhl, M. Gerstein, S. E. Antonarakis, T. R. Gingeras, J. B. Brown, P. Flicek, Y. Fu, D. Keefe, E. Birney, F. Denoeud, M. Gerstein, E. D. Green, P. Kapranov, U. Karaöz, R. M. Myers, W. S. Noble, A. Reymond, J. Rozowsky, K. Struhl, A. Siepel, J. A. Stamatoyannopoulos, C. M. Taylor, J. Taylor, R. E. Thurman, T. D. Tullius, S. Washietl, D. Zheng, L. A. Liefer, K. A. Wetterstrand, P. J. Good,

E. A. Feingold, M. S. Guyer, F. S. Collins, E. H. Margulies, G. M. Cooper, G. Asimenos, D. J. Thomas, C. N. Dewey, A. Siepel, E. Birney, D. Keefe, M. Hou, J. Taylor, S. Nikolaev, J. I. Montoya-Burgos, A. Löytynoja, S. Whelan, F. Pardi, T. Massingham, J. B. Brown, H. Huang, N. R. Zhang, P. Bickel, I. Holmes, J. C. Mullikin, A. Ureta-Vidal, B. Paten, M. Seringhaus, D. Church, K. Rosenbloom, W. J. Kent, E. A. Stone, M. Gerstein, S. E. Antonarakis, S. Batzoglou, N. Goldman, R. C. Hardison, D. Haussler, W. Miller, L. Pachter, E. D. Green, A. Sidow, Z. Weng, N. D. Trinklein, Y. Fu, Z. D. Zhang, U. Karaöz, L. Barrera, R. Stuart, D. Zheng, S. Ghosh, P. Flicek, D. C. King, J. Taylor, A. Ameur, S. Enroth, M. C. Bieda, C. M. Koch, H. A. Hirsch, C.-L. Wei, J. Cheng, J. Kim, A. A. Bhinge, P. G. Giresi, N. Jiang, J. Liu, F. Yao, W.-K. Sung, K. P. Chiu, V. B. Vega, C. W. H. Lee, P. Ng, A. Shahab, E. A. Sekinger, A. Yang, Z. Moqtaderi, Z. Zhu, X. Xu, S. Squazzo, M. J. Oberley, D. Inman, M. A. Singer, T. A. Richmond, K. J. Munn, A. Rada-Iglesias, O. Wallerman, J. Komorowski, G. K. Clelland, S. Wilcox, S. C. Dillon, R. M. Andrews, J. C. Fowler, P. Couttet, K. D. James, G. C. Lefebvre, A. W. Bruce, O. M. Dovey, P. D. Ellis, P. Dhami, C. F. Langford, N. P. Carter, D. Vetrie, P. Kapranov, D. A. Nix, I. Bell, S. Patel, J. Rozowsky, G. Euskirchen, S. Hartman, J. Lian, J. Wu, A. E. Urban, P. Kraus, S. V. Calcar, N. Heintzman, T. H. Kim, K. Wang, C. Qu, G. Hon, R. Luna, C. K. Glass, M. G. Rosenfeld, S. F. Aldred, S. J. Cooper, A. Halees, J. M. Lin, H. P. Shulha, X. Zhang, M. Xu, J. N. S. Haidar, Y. Yu, E. Birney*, S. Weissman, Y. Ruan, J. D. Lieb, V. R. Iyer, R. D. Green, T. R. Gingeras, C. Wadelius, I. Dunham, K. Struhl, R. C. Hardison, M. Gerstein, P. J. Farnham, R. M. Myers, B. Ren, M. Snyder, D. J. Thomas, K. Rosenbloom, R. A. Harte, A. S. Hinrichs, H. Trumbower, H. Clawson, J. Hillman-Jackson, A. S. Zweig, K. Smith, A. Thakkapallayil, G. Barber, R. M. Kuhn, D. Karolchik, D. Haussler, W. J. Kent, E. T. Dermitzakis, L. Armengol, C. P. Bird, T. G. Clark, G. M. Cooper, P. I. W. de Bakker, A. D. Kern, N. Lopez-Bigas, J. D. Martin, B. E. Stranger, D. J. Thomas, A. Woodroffe, S. Batzoglou, E. Davydov, A. Dimas, E. Eyras, I. B. Hallgrímsdóttir, R. C. Hardison, J. Huppert, A. Sidow, J. Taylor, H. Trumbower, M. C. Zody, R. Guigó, J. C. Mullikin, G. R. Abecasis, X. Estivill, E. Birney, G. G. Bouffard, X. Guan, N. F. Hansen, J. R. Idol, V. V. B. Maduro, B. Maskeri, J. C. McDowell, M. Park, P. J. Thomas, A. C. Young, R. W. Blakesley, D. M. Muzny, E. Sodergren, D. A. Wheeler, K. C. Worley, H. Jiang, G. M. Weinstock, R. A. Gibbs, T. Graves, R. Fulton, E. R. Mardis, R. K. Wilson, M. Clamp, J. Cuff, S. Gnerre, D. B. Jaffe, J. L. Chang, K. Lindblad-Toh, E. S. Lander, M. Koriabine, M. Nefedov, K. Osoegawa, Y. Yoshinaga, B. Zhu, and P. J. de Jong, "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project," *Nature*, vol. 447, no. 7146, pp. 799–816, Jun. 2007.

[38] N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. A. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford, and B. Ren, "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome," *Nat Genet*, vol. 39, no. 3, pp. 311–318, Mar. 2007.

[39] A. Visel, M. J. Blow, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, B. Ren, E. M. Rubin, and L. A. Pennacchio, "ChIP-seq accurately predicts tissue-specific activity of enhancers," *Nature*, vol. 457, no. 7231, pp. 854–858, Feb. 2009.

[40] G. E. Crawford, I. E. Holt, J. Whittle, B. D. Webb, D. Tai, S. Davis, E. H. Margulies, Y. Chen, J. A. Bernat, D. Ginsburg, D. Zhou, S. Luo, T. J. Vasicek, M. J. Daly, T. G. Wolfsberg, and F. S. Collins, "Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS)," *Genome Res.*, vol. 16, no. 1, pp. 123–131,

Jan. 2006.

[41] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford, "High-Resolution Mapping and Characterization of Open Chromatin across the Genome," *Cell*, vol. 132, no. 2, pp. 311–322, Jan. 2008.

[42] R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B.-K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford, and J. A. Stamatoyannopoulos, "The accessible chromatin landscape of the human genome," *Nature*, vol. 489, no. 7414, pp. 75–82, Sep. 2012.

[43] M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K.-K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj, A. P. Boyle, P. Cayting, A. Charos, D. Z. Chen, Y. Cheng, D. Clarke, C. Eastman, G. Euskirchen, S. Frietze, Y. Fu, J. Gertz, F. Grubert, A. Harmanci, P. Jain, M. Kasowski, P. Lacroute, J. Leng, J. Lian, H. Monahan, H. O'Geen, Z. Ouyang, E. C. Partridge, D. Patacsil, F. Pauli, D. Raha, L. Ramirez, T. E. Reddy, B. Reed, M. Shi, T. Slifer, J. Wang, L. Wu, X. Yang, K. Y. Yip, G. Zilberman-Schapira, S. Batzoglou, A. Sidow, P. J. Farnham, R. M. Myers, S. M. Weissman, and M. Snyder, "Architecture of the human regulatory network derived from ENCODE data," *Nature*, vol. 489, no. 7414, pp. 91–100, Sep. 2012.

[44] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein, "Mapping and analysis of chromatin state dynamics in nine human cell types," *Nature*, vol. 473, no. 7345, pp. 43–49, May 2011.

[45] M. R. Brent, "Steady progress and recent breakthroughs in the accuracy of automated genome annotation," *Nature Reviews Genetics*, vol. 9, no. 1, pp. 62–73, Jan. 2008.

[46] J. Su, S. A. Teichmann, and T. A. Down, "Assessing Computational Methods of Cis-Regulatory Module Prediction," *PLoS Comput Biol*, vol. 6, no. 12, p. e1001020, Dec. 2010.

[47] B. P. Berman, Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin, and M. B. Eisen, "Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome," *Proc Natl Acad Sci*, vol. 99, no. 2, pp. 757–762, Jan. 2002.

[48] T. L. Bailey and W. S. Noble, "Searching for statistically significant regulatory modules," *Bioinformatics*, vol. 19, no. Suppl 2, pp. ii16–ii25, Oct. 2003.

[49] O. Johansson, W. Alkema, W. W. Wasserman, and J. Lagergren, "Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm," *Bioinformatics*, vol. 19, no. Suppl 1, pp. i169–i176, Jul. 2003.

[50] S. Sinha, E. van Nimwegen, and E. D. Siggia, "A probabilistic method to detect regulatory modules," *Bioinformatics*, vol. 19, no. Suppl 1, pp. i292–i301, Jul. 2003.

[51] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements," *Nat Rev Genet*, vol. 5, no. 4, pp. 276–287, Apr. 2004.

[52] S. Fisher, E. A. Grice, R. M. Vinton, S. L. Bessling, and A. S. McCallion, "Conservation of RET Regulatory Function from Human to Zebrafish Without Sequence Similarity," *Science*, vol. 312, no. 5771, pp. 276 –279, Apr. 2006.

[53] D. M. McGaughey, R. M. Vinton, J. Huynh, A. Al-Saif, M. A. Beer, and A. S. McCallion, "Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b," *Genome Res*, vol. 18, no. 2, pp. 252 –260, Feb. 2008.

[54] M. J. Blow, D. J. McCulley, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, J. Bristow, B. Ren, B. L. Black, E. M. Rubin, A. Visel, and L. A. Pennacchio, "ChIP-Seq identification of weakly conserved heart enhancers," *Nat Genet*, vol. 42, no. 9, pp. 806–810, 2010.

[55] S. Sinha, M. D. Schroeder, U. Unnerstall, U. Gaul, and E. D. Siggia, "Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila," *BMC Bioinformatics*, vol. 5, p. 129, 2004.

[56] S. Sinha and X. He, "MORPH: Probabilistic Alignment Combined with Hidden Markov Models of cis-Regulatory Modules," *PLoS Comput Biol*, vol. 3, no. 11, p. e216, Nov. 2007.

[57] X. He, X. Ling, and S. Sinha, "Alignment and Prediction of cis-Regulatory Modules Based on a Probabilistic Model of Evolution," *PLoS Comput Biol*, vol. 5, no. 3, p. e1000299, Mar. 2009.

[58] X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis, "Systematic discovery of regulatory motifs in human promoters and 3[prime] UTRs by comparison of several mammals," *Nature*, vol. 434, no. 7031, pp. 338–345, Mar. 2005.

[59] O. Hallikas, K. Palin, N. Sinjushina, R. Rautiainen, J. Partanen, E. Ukkonen, and J. Taipale, "Genome-wide Prediction of Mammalian Enhancers Based on Analysis of Transcription-Factor Binding Affinity," *Cell*, vol. 124, no. 1, pp. 47–59, Jan. 2006.

[60] L. A. Pennacchio, G. G. Loots, M. A. Nobrega, and I. Ovcharenko, "Predicting tissue-specific enhancers in the human genome," *Genome Res*, vol. 17, no. 2, pp. 201 –211, Feb. 2007.

[61] L. A. Pennacchio, N. Ahituv, A. M. Moses, S. Prabhakar, M. A. Nobrega, M. Shoukry, S. Minovitsky, I. Dubchak, A. Holt, K. D. Lewis, I. Plajzer-Frick, J. Akiyama, S. De Val, V. Afzal, B. L. Black, O. Couronne, M. B. Eisen, A. Visel, and E. M. Rubin, "In vivo enhancer analysis of human conserved non-coding sequences," *Nature*, vol. 444, no. 7118, pp. 499–502, Nov. 2006.

[62] L. Elnitski, R. C. Hardison, J. Li, S. Yang, D. Kolbe, P. Eswara, M. J. O'Connor, S. Schwartz, W. Miller, and F. Chiaromonte, "Distinguishing Regulatory DNA From Neutral Sites," *Genome Res*, vol. 13, no. 1, pp. 64 –72, Jan. 2003.

[63] D. Kolbe, J. Taylor, L. Elnitski, P. Eswara, J. Li, W. Miller, R. Hardison, and

F. Chiaromonte, "Regulatory Potential Scores From Genome-Wide Three-Way Alignments of Human, Mouse, and Rat," *Genome Res.*, vol. 14, no. 4, pp. 700–707, Apr. 2004.

[64] D. C. King, J. Taylor, L. Elnitski, F. Chiaromonte, W. Miller, and R. C. Hardison, "Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences," *Genome Res*, vol. 15, no. 8, pp. 1051 –1060, 2005.

[65] S. Sonnenburg, A. Zien, and G. Ratsch, "ARTS: accurate recognition of transcription starts in human," *Bioinformatics*, vol. 22, no. 14, pp. e472–480, Jul. 2006.

[66] M. Megraw, F. Pereira, S. T. Jensen, U. Ohler, and A. G. Hatzigeorgiou, "A transcription factor affinity-based code for mammalian transcription initiation," *Genome Research*, vol. 19, no. 4, pp. 644 –656, Apr. 2009.

[67] L. Narlikar, N. J. Sakabe, A. A. Blanski, F. E. Arimura, J. M. Westlund, M. A. Nobrega, and I. Ovcharenko, "Genome-wide discovery of human heart enhancers," *Genome Res*, vol. 20, no. 3, pp. 381 –392, Mar. 2010.

[68] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, New York, NY, 1992, pp. 144–152.

[69] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY: Springer, 1995.

[70] B. Schölkopf, K. Tsuda, and J. P. Vert, *Kernel methods in computational biology*. Cambridge, MA: The MIT press, 2004.

[71] A. Ben-Hur, C. S. Ong, S. Sonnenburg, B. Schölkopf, and G. Rätsch, "Support Vector Machines and Kernels for Computational Biology," *PLoS Comput Biol*, vol. 4, no. 10, p. e1000173, Oct. 2008.

[72] C. Leslie, E. Eskin, and W. S. Noble, "The spectrum kernel: a string kernel for SVM protein classification," *Pac Symp Biocomput*, pp. 564–75, 2002.

[73] H. E. Peckham, R. E. Thurman, Y. Fu, J. A. Stamatoyannopoulos, W. S. Noble, K. Struhl, and Z. Weng, "Nucleosome positioning signals in genomic DNA," *Genome Research*, vol. 17, no. 8, pp. 1170 –1177, 2007.

[74] J. T. Kadonaga, "Regulation of RNA Polymerase II Transcription by Sequence-Specific DNA Binding Factors," *Cell*, vol. 116, no. 2, pp. 247–257, Jan. 2004.

[75] D. Carter, L. Chakalova, C. S. Osborne, Y. Dai, and P. Fraser, "Long-range chromatin regulatory interactions in vivo," *Nat Genet*, vol. 32, no. 4, pp. 623–626, Dec. 2002.

[76] J. P. Noonan and A. S. McCallion, "Genomics of Long-Range Regulatory Elements," *Annu Rev Genom Human Genet*, vol. 11, no. 1, pp. 1–23, Sep. 2010.

[77] A. Woolfe, M. Goodson, D. K. Goode, P. Snell, G. K. McEwen, T. Vavouri, S. F. Smith, P. North, H. Callaway, K. Kelly, K. Walter, I. Abnizova, W. Gilks, Y. J. K. Edwards, J. E. Cooke, and G. Elgar, "Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development," *PLoS Biol*, vol. 3, no. 1, p. e7, Nov. 2004.

[78] A. Visel, S. Prabhakar, J. A. Akiyama, M. Shoukry, K. D. Lewis, A. Holt, I. Plajzer-Frick, V. Afzal, E. M. Rubin, and L. A. Pennacchio, "Ultraconservation identifies a

small subset of extremely constrained developmental enhancers," *Nat Genet*, vol. 40, no. 2, pp. 158–160, Feb. 2008.

[79] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones, "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing," *Nat Meth*, vol. 4, no. 8, pp. 651–657, 2007.

[80] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold, "Genome-Wide Mapping of in Vivo Protein-DNA Interactions," *Science*, vol. 316, no. 5830, pp. 1497–1502, Jun. 2007.

[81] T.-K. Kim, M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, E. Markenscoff-Papadimitriou, D. Kuhl, H. Bito, P. F. Worley, G. Kreiman, and M. E. Greenberg, "Widespread transcription at neuronal activity-regulated enhancers," *Nature*, vol. 465, no. 7295, pp. 182–187, May 2010.

[82] H. M. Chan and N. B. La Thangue, "p300/CBP proteins: HATs for transcriptional bridges and scaffolds," *J Cell Sci*, vol. 114, no. 13, pp. 2363–2373, Jul. 2001.

[83] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D.-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, "TRANSFAC(R): transcriptional regulation, from patterns to profiles," *Nucl Acids Res*, vol. 31, no. 1, pp. 374–378, Jan. 2003.

[84] J. C. Bryne, E. Valen, M.-H. E. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, and A. Sandelin, "JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update," *Nucl Acids Res*, vol. 36, no. suppl_1, pp. D102–106, Jan. 2008.

[85] G. Leung and M. B. Eisen, "Identifying Cis-Regulatory Sequences by Word Profile Similarity," *PLoS ONE*, vol. 4, no. 9, p. e6901, 2009.

[86] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," *Bioinformatics*, vol. 16, no. 10, pp. 906–914, Oct. 2000.

[87] R. Karchin, K. Karplus, and D. Haussler, "Classifying G-protein coupled receptors with support vector machines," *Bioinformatics*, vol. 18, no. 1, pp. 147–159, Jan. 2002.

[88] C. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble, "Mismatch string kernels for discriminative protein classification," *Bioinformatics*, vol. 20, no. 4, pp. 467–76, Mar. 2004.

[89] G. Rätsch, S. Sonnenburg, and B. Schölkopf, "RASE: recognition of alternatively spliced exons in C.elegans," *Bioinformatics*, vol. 21 Suppl 1, pp. i369–77, Jun. 2005.

[90] S. Sonnenburg, G. Schweikert, P. Philips, J. Behr, and G. Ratsch, "Accurate splice site prediction using support vector machines," *BMC Bioinformatics*, vol. 8, no. Suppl 10, p. S7, 2007.

[91] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y.-H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W.-K. Sung, N. D. Clarke, C.-L. Wei, and H.-H. Ng, "Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells," *Cell*, vol. 133, no. 6, pp. 1106–1117, Jun. 2008.

[92] C. Vandewalle, F. Roy, and G. Berx, "The role of the ZEB family of transcription factors in development and disease," *Cell Mol Life Sci*, vol. 66, no. 5, pp. 773–787, Nov. 2008.

[93] D. Karolchik, R. M. Kuhn, R. Baertsch, G. P. Barber, H. Clawson, M. Diekhans, B. Giardine, R. A. Harte, A. S. Hinrichs, F. Hsu, K. M. Kober, W. Miller, J. S. Pedersen, A. Pohl, B. J. Raney, B. Rhead, K. R. Rosenbloom, K. E. Smith, M. Stanke, A. Thakkapallayil, H. Trumbower, T. Wang, A. S. Zweig, D. Haussler, and W. J. Kent, "The UCSC Genome Browser Database: 2008 update," *Nucl Acids Res*, vol. 36, no. suppl_1, pp. D773–779, Jan. 2008.

[94] S. J. Schultheiss, W. Busch, J. U. Lohmann, O. Kohlbacher, and G. Rätsch, "KIRMES: kernel-based identification of regulatory modules in euchromatic sequences," *Bioinformatics*, vol. 25, no. 16, pp. 2126 –2133, 2009.

[95] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large Scale Multiple Kernel Learning," *J Mach Learn Res*, vol. 7, pp. 1531–1565, 2006.

[96] S. Sonnenburg, G. Rätsch, S. Henschel, C. Widmer, J. Behr, A. Zien, F. de Bona, A. Binder, C. Gehl, and V. Franc, "The SHOGUN Machine Learning Toolbox," *Journal of Machine Learning Research*, vol. 11, p. 1799–1802, Jun. 2010.

[97] T. Joachims, "Making large-scale support vector machine learning practical," Cambridge, MA: MIT Press, 1999, pp. 169–184.

[98] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler, "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome Res*, vol. 15, no. 8, pp. 1034 –1050, 2005.

[99] V. Gotea, A. Visel, J. M. Westlund, M. A. Nobrega, L. A. Pennacchio, and I. Ovcharenko, "Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers," *Genome Res*, vol. 20, no. 5, pp. 565 –577, May 2010.

[100] D. E. Newburger and M. L. Bulyk, "UniPROBE: an online database of protein binding microarray data on protein-DNA interactions," *Nucl Acids Res*, vol. 37, no. suppl_1, pp. D77–82, Jan. 2009.

[101] S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble, "Quantifying similarity between motifs," *Genome Biology*, vol. 8, no. 2, p. R24, Feb. 2007.

[102] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proc Natl Acad Sci*, vol. 100, no. 16, pp. 9440 –9445, 2003.

[103] M. F. Berger, G. Badis, A. R. Gehrke, S. Talukder, A. A. Philippakis, L. Peña-Castillo, T. M. Alleyne, S. Mnaimneh, O. B. Botvinnik, E. T. Chan, F. Khalid, W. Zhang, D. Newburger, S. A. Jaeger, Q. D. Morris, M. L. Bulyk, and T. R. Hughes, "Variation in

Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences," *Cell*, vol. 133, no. 7, pp. 1266–1276, Jun. 2008.

[104] A. Bulfone, L. Puelles, M. Porteus, M. Frohman, G. Martin, and J. Rubenstein, "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*, vol. 13, no. 7, pp. 3155–3172, Jul. 1993.

[105] I. Matsuo, S. Kuratani, C. Kimura, N. Takeda, and S. Aizawa, "Mouse Otx2 functions in the formation and patterning of rostral head.," *Genes Dev*, vol. 9, no. 21, pp. 2646–2658, Nov. 1995.

[106] T. Zerucha, T. Stühmer, G. Hatch, B. K. Park, Q. Long, G. Yu, A. Gambarotta, J. R. Schultz, J. L. R. Rubenstein, and M. Ekker, "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*, vol. 20, no. 2, pp. 709–721, Jan. 2000.

[107] J. E. Lee, "Basic helix-loop-helix genes in neural development," *Curr Opin Neurobiol*, vol. 7, no. 1, pp. 13–20, Feb. 1997.

[108] N. Bertrand, D. S. Castro, and F. Guillemot, "Proneural genes and the specification of neural cell types," *Nat Rev Neurosci*, vol. 3, no. 7, pp. 517–530, Jul. 2002.

[109] S. E. Ross, M. E. Greenberg, and C. D. Stiles, "Basic Helix-Loop-Helix Factors in Cortical Development," *Neuron*, vol. 39, no. 1, pp. 13–25, Jul. 2003.

[110] A. Erives and M. Levine, "Coordinate enhancers share common organizational features in the Drosophila genome," *Proc Natl Acad Sci*, vol. 101, no. 11, pp. 3851–3856, Mar. 2004.

[111] N. Ghanem, O. Jarinova, A. Amores, Q. Long, G. Hatch, B. K. Park, J. L. R. Rubenstein, and M. Ekker, "Regulatory Roles of Conserved Intergenic Domains in Vertebrate Dlx Bigene Clusters," *Genome Res*, vol. 13, no. 4, pp. 533 –543, Apr. 2003.

[112] J. T. Wigle and D. D. Eisenstat, "Homeobox genes in vertebrate forebrain development and disease," *Clin Genet*, vol. 73, no. 3, pp. 212–226, 2008.

[113] D. Kurokawa, H. Kiyonari, R. Nakayama, C. Kimura-Yoshida, I. Matsuo, and S. Aizawa, "Regulation of Otx2 expression and its functions in mouse forebrain and midbrain," *Development*, vol. 131, no. 14, pp. 3319–3331, Jul. 2004.

[114] M. D. Wilson, N. L. Barbosa-Morais, D. Schmidt, C. M. Conboy, L. Vanes, V. L. J. Tybulewicz, E. M. C. Fisher, S. Tavare, and D. T. Odom, "Species-Specific Transcription in Mice Carrying Human Chromosome 21," *Science*, vol. 322, no. 5900, pp. 434–438, Oct. 2008.

[115] S. W. Flavell and M. E. Greenberg, "Signaling Mechanisms Linking Neuronal Activity to Gene Expression and Plasticity of the Nervous System," *Annu Rev Neurosci*, vol. 31, no. 1, pp. 563–590, Jul. 2008.

[116] S. Mason, M. Piper, R. M. Gronostajski, and L. J. Richards, "Nuclear Factor One Transcription Factors in CNS Development," *Mol Neurobiol*, vol. 39, no. 1, pp. 10–23, Dec. 2008.

[117] M. Wilson and P. Koopman, "Matching SOX: partner proteins and co-factors

of the SOX family of transcriptional regulators," *Curr Opin Genetics Dev*, vol. 12, no. 4, pp. 441–446, Aug. 2002.

[118] M. Demarque and N. C. Spitzer, "Activity-Dependent Expression of Lmx1b Regulates Specification of Serotonergic Neurons Modulating Swimming Behavior," *Neuron*, vol. 67, no. 2, pp. 321–334, Jul. 2010.

[119] G. Pavesi, G. Mauri, and G. Pesole, "An algorithm for finding signals of unknown length in DNA sequences," *Bioinformatics*, vol. 17, no. Suppl 1, pp. S207–S214, Jun. 2001.

[120] J. D. Hughes, P. W. Estep, S. Tavazoie, and G. M. Church, "Computational identification of Cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae," *J Mol Biol*, vol. 296, no. 5, pp. 1205–1214, Mar. 2000.

[121] T. Bailey and C. Elkan, "Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers," *Proc Int Conf Intell Syst Mol Biol*, vol. 2, pp. 28–36, 1994.

[122] S. Frietze, X. Lan, V. X. Jin, and P. J. Farnham, "Genomic Targets of the KRAB and SCAN Domain-containing Zinc Finger Protein 263," *J Biol Chem*, vol. 285, no. 2, pp. 1393 –1403, Jan. 2010.

[123] S. Sonnenburg, A. Zien, P. Philips, and G. Ratsch, "POIMs: positional oligomer importance matrices--understanding support vector machine-based signal detectors," *Bioinformatics*, vol. 24, no. 13, pp. i6–14, Jul. 2008.

[124] S. J. Schultheiss, "Kernel-Based Identification of Regulatory Modules," in *Computational Biology of Transcription Factor Binding*, vol. 674, I. Ladunga, Ed. Totowa, NJ: Humana Press, 2010, pp. 213–223.

[125] K. Koh, S.-J. Kim, and S. Boyd, "An Interior-Point Method for Large-Scale l1-Regularized Logistic Regression," *J Mach Learn Res*, vol. 8, pp. 1519–1555, Dec. 2007.

[126] D. Thanos and T. Maniatis, "Virus induction of human IFNbeta gene expression requires the assembly of an enhanceosome," *Cell*, vol. 83, no. 7, pp. 1091–1100, Dec. 1995.

[127] T. Maniatis, J. V. Falvo, T. H. Kim, T. K. Kim, C. H. Lin, B. S. Parekh, and M. G. Wathelet, "Structure and Function of the Interferon-β Enhanceosome," *Cold Spring Harb Symp Quant Biol*, vol. 63, pp. 609 –620, Jan. 1998.

[128] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nussbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu, "Model-based Analysis of ChIP-Seq (MACS)," *Genome Biol*, vol. 9, no. 9, p. R137, 2008.

[129] L. Song, Z. Zhang, L. L. Grasfeder, A. P. Boyle, P. G. Giresi, B.-K. Lee, N. C. Sheffield, S. Gräf, M. Huss, D. Keefe, Z. Liu, D. London, R. M. McDaniell, Y. Shibata, K. A. Showers, J. M. Simon, T. Vales, T. Wang, D. Winter, Z. Zhang, N. D. Clarke, E. Birney, V. R. Iyer, G. E. Crawford, J. D. Lieb, and T. S. Furey, "Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity," *Genome Research*, vol. 21, no. 10, pp. 1757 –1767, Oct. 2011.

[130] J. Goecks, A. Nekrutenko, J. Taylor, and $author firstName $author.lastName, "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent

computational research in the life sciences," *Genome Biology*, vol. 11, no. 8, p. R86, Aug. 2010.

[131] V. Kecman, *Learning and soft computing: support vector machines, neural networks, and fuzzy logic models*. MIT press, 2001.

[132] J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," *Advances in Large Margin Classifiers*, pp. 61–74, 1999.

[133] H. Lin, C.-J. Lin, and R. C. Weng, "A Note on Platt's Probabilistic Outputs for Support Vector Machines," 2003.

[134] N. Ivanova, R. Dobrin, R. Lu, I. Kotenko, J. Levorse, C. DeCoste, X. Schafer, Y. Lun, and I. R. Lemischka, "Dissecting self-renewal in stem cells with RNA interference," *Nature*, vol. 442, no. 7102, pp. 533–538, Jun. 2006.

[135] G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole, "Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes," *Nucl Acids Res*, vol. 32, no. suppl_2, pp. W199–203, Jul. 2004.

[136] S. John, P. J. Sabo, R. E. Thurman, M.-H. Sung, S. C. Biddie, T. A. Johnson, G. L. Hager, and J. A. Stamatoyannopoulos, "Chromatin accessibility pre-determines glucocorticoid receptor binding patterns," *Nat Genet*, vol. 43, no. 3, pp. 264–268, Mar. 2011.

[137] M. Patel, J. M. Simon, M. D. Iglesia, S. B. Wu, A. W. McFadden, J. D. Lieb, and I. J. Davis, "Tumor-Specific Retargeting of an Oncogenic Transcription Factor Chimera Results in Dysregulation of Chromatin and Transcription," *Genome Res.*, vol. 22, no. 2, pp. 259–270, Feb. 2012.

[138] N. U. Rashid, P. G. Giresi, J. G. Ibrahim, W. Sun, and J. D. Lieb, "ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions," *Genome Biology*, vol. 12, no. 7, p. R67, 2011.

[139] R. McDaniell, B.-K. Lee, L. Song, Z. Liu, A. P. Boyle, M. R. Erdos, L. J. Scott, M. A. Morken, K. S. Kucera, A. Battenhouse, D. Keefe, F. S. Collins, H. F. Willard, J. D. Lieb, T. S. Furey, G. E. Crawford, V. R. Iyer, and E. Birney, "Heritable Individual-Specific and Allele-Specific Chromatin Signatures in Humans," *Science*, vol. 328, no. 5975, pp. 235 –239, Apr. 2010.

[140] N. Meyer and L. Z. Penn, "Reflecting on 25 years with MYC," *Nature Reviews Cancer*, vol. 8, no. 12, pp. 976–990, Dec. 2008.

[141] C. V. Dang, "MYC on the Path to Cancer," *Cell*, vol. 149, no. 1, pp. 22–35, Mar. 2012.

[142] E. Guccione, F. Martinato, G. Finocchiaro, L. Luzi, L. Tizzoni, V. D. Olio, G. Zardo, C. Nervi, L. Bernard, and B. Amati, "Myc-binding-site recognition in the human genome is determined by chromatin context," *Nature Cell Biology*, vol. 8, no. 7, pp. 764–770, 2006.

[143] B.-K. Lee, A. A. Bhinge, A. Battenhouse, R. M. McDaniell, Z. Liu, L. Song, Y. Ni, E. Birney, J. D. Lieb, T. S. Furey, G. E. Crawford, and V. R. Iyer, "Cell-Type Specific and Combinatorial Usage of Diverse Transcription Factors Revealed by Genome-Wide

Binding Studies in Multiple Human Cells," *Genome Res.*, vol. 22, no. 1, pp. 9–24, Jan. 2012.

[144] K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott, "NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy," *Nucleic Acids Research*, vol. 40, no. D1, pp. D130–D135, Nov. 2011.

[145] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nat. Protocols*, vol. 4, no. 1, pp. 44–57, Dec. 2008.

[146] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucl. Acids Res.*, vol. 37, no. 1, pp. 1–13, Jan. 2009.

[147] H. Ji, G. Wu, X. Zhan, A. Nolan, C. Koh, A. De Marzo, H. M. Doan, J. Fan, C. Cheadle, M. Fallahi, J. L. Cleveland, C. V. Dang, and K. I. Zeller, "Cell-Type Independent MYC Target Genes Reveal a Primordial Signature Involved in Biomass Accumulation," *PLoS ONE*, vol. 6, no. 10, p. e26057, Oct. 2011.

[148] R. Rifkin and A. Klautau, "In Defense of One-Vs-All Classification," *J. Mach. Learn. Res.*, vol. 5, pp. 101–141, Dec. 2004.

[149] B. J. Frey and D. Dueck, "Clustering by Passing Messages Between Data Points," *Science*, vol. 315, no. 5814, pp. 972–976, Feb. 2007.

[150] J. Wang, J. Zhuang, S. Iyer, X. Lin, T. W. Whitfield, M. C. Greven, B. G. Pierce, X. Dong, A. Kundaje, Y. Cheng, O. J. Rando, E. Birney, R. M. Myers, W. S. Noble, M. Snyder, and Z. Weng, "Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors," *Genome Res.*, vol. 22, no. 9, pp. 1798–1812, Sep. 2012.

[151] S. Neph, J. Vierstra, A. B. Stergachis, A. P. Reynolds, E. Haugen, B. Vernot, R. E. Thurman, S. John, R. Sandstrom, A. K. Johnson, M. T. Maurano, R. Humbert, E. Rynes, H. Wang, S. Vong, K. Lee, D. Bates, M. Diegel, V. Roach, D. Dunn, J. Neri, A. Schafer, R. S. Hansen, T. Kutyavin, E. Giste, M. Weaver, T. Canfield, P. Sabo, M. Zhang, G. Balasundaram, R. Byron, M. J. MacCoss, J. M. Akey, M. A. Bender, M. Groudine, R. Kaul, and J. A. Stamatoyannopoulos, "An expansive human regulatory lexicon encoded in transcription factor footprints," *Nature*, vol. 489, no. 7414, pp. 83–90, Sep. 2012.

[152] K. S. Pollard, M. J. Hubisz, K. R. Rosenbloom, and A. Siepel, "Detection of nonneutral substitution rates on mammalian phylogenies," *Genome Res.*, vol. 20, no. 1, pp. 110–121, Jan. 2010.

[153] R. J. Miltenberger, K. A. Sukow, and P. J. Farnham, "An E-box-mediated increase in cad transcription at the G1/S-phase boundary is suppressed by inhibitory c-Myc mutants.," *Mol. Cell. Biol.*, vol. 15, no. 5, pp. 2527–2535, May 1995.

[154] R. Kollmar, K. A. Sukow, S. K. Sponagle, and P. J. Farnham, "Start site selection at the TATA-less carbamoyl-phosphate synthase (glutamine-hydrolyzing)/aspartate carbamoyltransferase/dihydroorotase promoter.," *J. Biol. Chem.*, vol. 269, no. 3, pp. 2252–2257, Jan. 1994.

[155] L. Pevny, M. C. Simon, E. Robertson, W. H. Klein, S.-F. Tsai, V. D'Agati, S. H. Orkin, and F. Costantini, "Erythroid differentiation in chimaeric mice blocked by a

targeted mutation in the gene for transcription factor GATA-1," *Nature*, vol. 349, no. 6306, pp. 257–260, Jan. 1991.

[156] F. M. Sladek, W. M. Zhong, E. Lai, and J. E. Darnell, "Liver-enriched transcription factor HNF-4 is a novel member of the steroid hormone receptor superfamily.," *Genes Dev.*, vol. 4, no. 12b, pp. 2353–2365, Dec. 1990.

[157] E. Lelièvre, F. Lionneton, F. Soncin, and B. Vandenbunder, "The Ets family contains transcriptional activators and repressors involved in angiogenesis," *The International Journal of Biochemistry & Cell Biology*, vol. 33, no. 4, pp. 391–407, Apr. 2001.

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

[159] J. Keilwagen, J. Grau, I. A. Paponov, S. Posch, M. Strickert, and I. Grosse, "De-Novo Discovery of Differentially Abundant Transcription Factor Binding Sites Including Their Positional Preference," *PLoS Comput Biol*, vol. 7, no. 2, p. e1001070, Feb. 2011.

CURRICULUM VITAE

The Johns Hopkins University School of Medicine

**DONGWON LEE** March 20, 2013

## EDUCATION

| | | | |
|---|---|---|---|
| **Ph.D.** expected | 2013 | Biomedical Engineering<br>Advisor: Michael A. Beer, Ph.D. | Johns Hopkins School of Medicine |
| **B. S.** | 2007 | Computer Science<br>(Double Major: Biological Science) | KAIST |

## PROFESSIONAL EXPERIENCE

| | | |
|---|---|---|
| Software Developer | Jan 2002-Jan 2005 | Softwise, Seoul, Korea |
| Software Developer | Jan 2001-Dec 2001 | Synapsoft, Seoul, Korea |

## SCHOLARSHIPS

1999-2005 Undergraduate Studies Fellowship, KAIST

## HONORS AND AWARDS

| | | |
|---|---|---|
| 2013 | Nupur Dinesh Thekdi Award | Johns Hopkins School of Medicine |
| 2007 | magna cum laude | KAIST |

## PUBLICATIONS

### Original research:

**Lee D**, Beer MA. (2013) DNA sequence predicts cell-type-specific transcription factor binding regulation. *In preparation.*

Ghandi M, **Lee D**, Mohammad-Noori M, Beer MA. (2013) Enhanced sequence classification using robust k-mer frequency estimation. *In preparation.*

Fletez-Brant C*, **Lee D***†, McCallion AS, Beer MA†. (2012) kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic datasets. *Submitted.*

Gorkin DU, **Lee D**, Reed X, Fletez-Brant C, Blessling SL, Loftus SK, Beer MA, Pavan WJ, McCallion AS. (2012) Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes. Genome Res 22:2290-2301.

**Lee D**, Karchin R, Beer MA. (2011) Discriminative prediction of mammalian enhancers from DNA sequence. Genome Res 21:2167-2180.

- *This paper was selected as one of the most influential papers of 2011 in the fields of systems biology and regulatory genomics at 5th Annual RECOMB Conference on Regulatory and Systems Genomics*

*co-first authors
†co-corresponding authors

**Book chapters:**

**Lee D** and Beer MA. (2013) Mammalian Enhancer Prediction. Genome Analysis: Current procedures and Applications. *In press.*

**Posters & Abstracts:**

**Lee D**, Ghandi M, and Beer MA. (2012) Predicting the find structure of cis-regulatory elements. 9th Annual RECOMB Satellite on Regulatory Genomics, Redwood city, CA, November 12-15, 2012.

**Lee D**, Ghandi M, and Beer MA. (2012) Predicting cell specific transcription factor occupancy from DNA sequence. 9th Annual RECOMB Satellite on Regulatory Genomics, Redwood city, CA, November 12-15, 2012.

**Lee D**, Karchin R, Beer MA. (2009) Successful enhancer prediction from DNA sequence. 6th Annual RECOMB Satellite on Regulatory Genomics, Cambridge, MA, December 2-6, 2009

## TEACHING EXPERIENCE

Sep 2011-Dec 2011, Teaching Assistant, Statistical Mechanics and Thermodynamics, Department of Biomedical Engineering, Johns Hopkins University

- Taught 1 hour session per week

Feb 2010-May 2010, Teaching Assistant, Foundation of Computational Biology and Bioinformatics II (FCBBII), Department of Biomedical Engineering, Johns Hopkins University

- Taught 1 hour lecture per week