UNDERSTANDING THE EVOLUTION AND SEQUENCE ARCHITECTURE OF GENE REGULATORY ELEMENTS THROUGH MACHINE LEARNING

By

Ling Chen

Dissertation

Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Biological Sciences
December 14, 2019
Nashville, Tennessee

Approved:
Antonis Rokas, Ph.D.
John Anthony Capra, Ph.D.
Maulik Patel, Ph.D.
Emily Hodges, Ph.D.
Thomas A. Lasko, M.D., Ph.D.

To my parents, who don't understand what I do most of time but let me do it anyway.

To Pumpki, who doesn't understand what I do half of time yet can recite my dissertation title.

To the cats, who totally understand what I do and always lie down between me and my work.

Acknowledgements

I would like to begin by thanking my advisor, Dr. Tony Capra. Tony introduced me to the field of genetics and computational biology and gave me both freedom and guidance to develop my research ideas. This dissertation would have not been possible without his encouragement and hard work. I also want to thank my thesis committee, Antonis Rokas, Emily Hodges, Maulik Patel, and Tom Lasko for their insightful comments, patience, and support.

I'm grateful for having my wonderful labmates who are always supportive and give constructive feedback. We had so much fun discussing science and other truly random things. I would especially like to thank Alex Fish who collaborated with me for a substantial portion of the work presented here. I cannot ask for a better lab environment.

I would like to thank my fellow BSCI graduate students, especially my cohort. You have been wonderful colleagues and friends. I also want to thank Kathy Friedman, who was our DGS, one of the professors who interviewed me, and my collaborator, for her kindness and creating an inclusive environment. And thanks to all the BSCI administrative staff for making my graduate journey, especially conference travels, go smoothly.

Finally, I would like to acknowledge my partner Pumpki Su, who is an arising young scientist and more brilliant than me in every way. Although our research differs, I am constantly inspired by her dedication and genuine interest in her field of study and benefited tremendously from our discussions of science and philosophy.

Table of contents

List of tables

Page

List of figures

Page

Chapter I

Introduction

In the earliest days of molecular biology, scientists realized that the constant value of haploid DNA content, "C-value", does not correlate with the organismal complexity (Mirsky and Hans 1951). This is referred as the "C-value paradox" (Thomas 1971), because it is contradictory to the common assumption that there is a positive correlation between organism complexity and genome size. As technology advanced and enabled the sequencing of human genome, scientists were able to determine the composition of the human genome. One of the major questions was, how many genes are there? To many scientists' surprise, the answer from the human genome project (HGP) is only about 20,000 – 25,000 genes, which is about the same size as the nematode worm (*C. elegans* sequencing consortium 1998). This further expanded the "C-value paradox": not only the total size of the genome, but also the gene count, does not correlate with organismal complexity (Hahn and Wray 2002). However, once many sequences were known from multiple species, it became possible to use genes to measure the genetic distance between species. One could ask: what genes are different and how different are they? Early efforts to compare genes between species in 1963 showed that some of the blood proteins of human and chimpanzees are virtually identical in amino acid sequences. This and other following analyses of additional proteins and DNA in the 1970s led King and Wilson to propose that non-protein-coding gene regulatory mutations account for the major biological differences between human and chimpanzees (King and Wilson 1975) rather than differences in protein coding genes. Later, sequencing of the chimpanzee genome in 2005 enabled whole genome comparison between human and chimpanzee and made this hypothesis even more plausible — the careful comparison between the human and chimpanzee typical proteins showed only two amino acid differences (Varki and Altheide 2005).

Among the 3 billion base pair human genome, less than 2% code proteins and the vast majority are non-coding sequences. The massive amount of noncoding DNA was once thought as "junk DNA" (Ohno 1972) and a large portion of them is consist of transposons, and thought to be "selfish", which means that they function for themselves but not for their host (Orgel and Crick 1980). Although the term, "junk DNA", is rarely used in modern day research of

noncoding genome, the idea that the majority of noncoding DNA is nonfunctional is maintained till today. In 2012, ENCODE project estimated 80.4% of the genome participates in at least one biochemical RNA and/or chromatin associated event in at least one cell type (Bernstein et al. 2012), strongly opposing the idea of most noncoding regions is nonfunctional. However, ENCODE's definition of "function" is based on the biomedically marks and it is arguable whether 'biochemically active' is an accurate approximation of functionality. Eddy argued that biochemically 'active' DNA does not mean that they are there primarily because they're useful for the organism, including host suppression of mobile element (Eddy 2012, 2013). Sequence conservation analyses estimated about 9% of the human genome shows detectable conservation, which also questions the claim of 80% of noncoding genome is functional. But at the mean time sequence conservation is not required for functionality because some of the genome may have human specific function and more importantly, the regulatory landscape turnover rapidly (Villar et al. 2014a, 2015b). All in all, it is still debatable how much of the noncoding genome is actually functional but at least we know that a substantial fraction of them have regulatory activity.

The regulatory regions of the genome encode the information for orchestrating the dynamic spatiotemporal patterns of gene expression required for the proper differentiation and development of multi-cellular organisms (Shlyueva et al. 2014; Kundaje et al. 2015; Villar et al. 2015b). Promoters and enhancers are two main types of regulatory elements. Promoters are in the immediate vicinity of the transcription start sites of the target gene and can initiate the transcription of the gene. However, the transcription driven by the promoters is usually at a low level. The more precise regulation of gene expression requires a class of distant regulatory elements, called enhancers. As a result of their essential role, mutations that disrupt proper enhancer activity can lead to diseases. For example, preaxial polydactyly, a frequently observed limb malformation, is due to point mutations in the *ZRS* enhancer that disrupt its proper regulation of the *Shh* gene, which is located 1 megabase (Mb) away (Lettice et al. 2003). Another example is a homozygous point mutation in an enhancer of the *TBX5* gene that leads to congenital heart disease (Smemo et al. 2012). More broadly, significant evidence for the functional importance of gene regulatory regions comes from genome-wide association studies (GWAS)—the majority of genetic variants associated with complex disease are non-protein coding, suggesting that they are influencing diseases by disrupting proper gene expression levels

(Maurano et al. 2012; Lee et al. 2018). Given the significance of enhancers, it is critical to understand the complex mechanisms of enhancers.

Enhancers were first described as a 72 base pair (bp) DNA element in the simian virus (SV40) genome that enhances the expression of a cloned rabbit hemoglobin b1 gene in HeLa cells by 200 fold (Banerji et al. 1981b) from thousands of base pairs away independent of the relative orientation of enhancer and the target gene. Since then, extensive studies have been performed to identify and investigate the function and properties of enhancers. One characteristic of enhancers is the above-mentioned distant location from their target genes. They can locate several kilobases, or even megabases, away from the target gene and can act from either upstream or downstream of the transcription start site (TSS). They are thought to regulate gene expression by looping to the proximity of target promoters in the 3D genome (Amano et al. 2009; Miguel-Escalada et al. 2015; Schoenfelder and Fraser 2019). Active enhancers are often found in accessible open chromatin regions (Boyle et al. 2008). This enables the binding of transcription factors (TFs) to the short DNA motifs in the enhancer sequences for gene regulation. The histones in the active enhancer regions show a characteristic set of modifications at their N-terminal tails, such as histone H3 lysine 4 monomethylation (H3K4me1) and H3K27 acetylation (H3K27ac) (Heintzman et al. 2009; Mendenhall et al. 2013).

Using the above characteristics as proxies for enhancer activity, many genome-wide methods for profiling of enhancers have been developed, such as DNase-seq (DNase I hypersensitive sites sequencing) (Boyle et al. 2008) and FAIRE-seq (formaldehyde-assisted isolation of regulatory elements) that identify open chromatin (Giresi et al. 2007), chromatin immunoprecipitation coupled with highthroughput sequencing (ChIP-seq) of H3K4me1 and H3K27ac histone modifications (Liu and Hauser 2007; Shlyueva et al. 2014; Villar et al. 2015c), massively parallel reporter assays (MPRA) (Melnikov et al. 2012) and bi-directionally transcribed enhancer RNAs (eRNAs) (Andersson et al. 2014). These various ways of enhancer profiling result in an abundant database of enhancers across a diverse set of cellular contexts and species, enabling the further analyses of the complex mechanisms of enhancers.

I-1 A seeming evolutionary paradox: rapid enhancer turnover and stable gene expression

There exist hundreds of thousands of sequences within non-coding regions of genomes that have identifiable evolutionary conservation across more than 400 million years of evolution. Some of these conserved non-coding elements (CNEs) have been shown to act as developmental enhancers (Polychronopoulos et al. 2017). This elevated level of evolutionary conservation has been used to as a way to computationally identify enhancers genome wide. However, many pieces of evidence suggested that enhancers may be rapidly evolving. For example, in a study of *ret* gene expression in zebrafish, researchers found that although there is no apparent sequence similarity between human and zebrafish *ret* enhancers, the majority of human *ret* enhancers drive similar gene expression profiles when introduced into the zebrafish as the zebrafish *ret* enhancers (Fisher et al. 2006), suggesting that conserved enhancers are not required for maintaining similar gene expression profiles. The comparison of transcription factor binding sites across mammals also suggested rapid evolution. In an experiment comparing four tissue specific transcription factors between human and mouse revealed 41-89% of the binding sites are species specific (Odom et al. 2007). Another more recent piece of evidence came from a direct comparison of histone modification based genome-wide liver enhancer maps across 20 mammalian species (Villar et al. 2015b). They found that among about 30,000 human liver enhancers, only 1% of them are highly conserved across mammals, suggesting rapid turnover of active enhancer regions in mammal. In contrast, 16% of the promoters are highly conserved.

On the other hand, gene expression has shown to be highly conserved. Analysis of gene expression of all known and predicted genes across twenty tissues in human, mouse, chicken, frog, and pufferfish found that more than a third of unique orthologous genes have conserved expression despite the large evolutionary distance between these species. Moreover, the conservation of expression correlates poorly with the amount of conserved non-exonic sequence (Chan et al. 2009). Brawand et al. in 2011 also showed the slow evolution of gene expression in mammals. They analyzed RNA-Seq from six organs across ten species that represent all major mammalian lineages (placentals, marsupials and monotremes). They showed that gene expression profiles cluster by tissue rather than by species, suggesting the variation of gene expression is largely between cellular contexts rather than between species (Brawand et al. 2011). Berthelot et al. also showed high correlation of gene expression in liver across 25 mammalian species (Berthelot et al. 2017). These pieces of evidence suggest the decoupling of

the conservation of gene expression and that of regulatory elements, raising the question of what mechanism is governing the evolution of enhancers so that even though the regulatory landscape is rapidly changing, there is still considerably high conservation of gene expression. These pieces of evidence raise the question that how the rapidly evolving regulatory drives largely conserved gene expression.

I-2 Enhancer sequence architecture

Transcription factors bind enhancers to exert their regulatory function. There are about 1600 TFs in the human genome, exerting control over cell differentiation, developmental patterning, and activation of specific pathways in response to environmental cues, such as immune responses (Lambert et al. 2018). These proteins have DNA-binding domains and preferentially bind to a certain set of short DNA sequences. Such TF-DNA binding specificities can be summarized in the form of sequence motifs, which are usually represented in the form of a position weight matrix indicating relative preference of the TF for each base in the binding site. TF binding motifs can be determined through various techniques both *in vivo* and *in vitro*. For example, the protein binding microarray (PBM) is an *in vitro* way for determining TF motifs where a GST-tagged TF is bound to a glass slide with arrays of short DNA sequences and the binding specificity is then determined through the fluorescence emitted from the bound DNA sequences (Mukherjee et al. 2004). Another common *in vitro* experiment method is high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX). In HT-SELEX, TF protein is mixed with a randomized pool of short DNA sequences and bound regions are selected for multiple rounds and sequenced (Jolma et al. 2013a). These *in vitro* assays deepened our understanding of TF binding specificities; however, the binding of TFs may be different *in vivo*, where chromatin accessibility, other proteins, and other factors all play a role in determining binding. In contrast, ChIP-Seq can assay TF binding sites *in vivo*. In ChIP-Seq, proteins bound to DNA are first cross-linked to the DNA. Then, the DNA is fragmented and treated with exonuclease to trim unbound sequences. Next, protein-specific antibodies are used to immunoprecipitate the DNA-protein complex. Finally, the DNA is extracted and sequenced, giving high-resolution sequences of the protein-binding sites (Johnson et al. 2007; Rhee and Pugh 2011). ChIP-Seq also has some weaknesses. For example, it does not measure equilibrium binding because of the use of cross-linkers and the data quality is highly dependent on the antibody efficiency (Lambert et al. 2018).

The above-mentioned methods have enabled characterization of binding sites for about two thirds of human TF proteins. However, consensus TF motifs are not sufficient for inferring TF binding. As shown in TF binding analyses of ChIP-seq peaks, TFs only bind to a small fraction of all motif occurrences in the genome, and some binding sites do not contain the consensus TF binding motif (Wang et al. 2012). Indeed, many additional features have been suggested to play a role in determining in vivo TF binding. For example, local variations in DNA shape have been shown to influence TF binding. Combinatorial binding is another important factor, especially in enhancer regulatory activity (Stampfel et al. 2015). Enhancer sequences have many short DNA motifs that can be bound by transcription factors. TFs often bind cooperatively as homodimers (e.g., bZIPs and bHLHs), trimers (heat shock factors) (Lambert et al. 2018), or higher-order structures, such homotypic clusters (Mathelier and Wasserman 2013), heterotypic clusters (Luna-Zurita et al. 2016) and enhanceosomes (Slattery et al. 2014). The combinatorial binding of TFs can occur through protein-protein interactions or mediated by DNA. A recent study identified 315 pairs of TFs with clear spacing and orientation preferences among 9400 tested pairs of TFs, and some of the cooperative binding is directly mediated by DNA (Jolma et al. 2015). These findings suggest that the "regulatory code" of enhancers, unlike the clear three codon genetic code, is complex not only in the sense that the binding specificities of individual TFs are degenerative and diverse, but that TF binding also has "grammar"—how multiple TF binding events in the enhancer sequence work in synergy to direct precise pattern of gene expression regulation, such as cooccurrences, ordering, spacing, and orientation of TF binding sites.

I-3 Modeling of enhancers with machine learning algorithms

Enhancers are essential to the orchestration of proper spatio-temporal gene expression in diverse cellular contexts. With the development of next generation sequencing, large amounts of genome-wide data about the location of epigenetic marks associated with enhancer activity has been produced, this has created the need for more efficient analytical methods for analyzing enhancer sequences and epigenetic properties. Machine learning algorithms can extract common patterns from large volumes of data and extrapolate the learned knowledge to predict unknowns. Therefore, machine learning algorithms are especially promising for mining meaningful biological information from the high volume of high dimensional genomic data. In recent studies, researchers have successfully applied machine learning models to enhancer sequences and enhancer-related

data. These methods can be divided to two main types: unsupervised and supervised learning methods.

I-3.1 Unsupervised learning methods

Unsupervised learning methods are the group of algorithms that draw inferences from the unlabeled input data. The goal of unsupervised learning can be clustering data points by their similarity in features, reducing their redundancy, or extracting general rules about their properties. Some common unsupervised methods are Hidden Markov models (HMMs), Gaussian mixture models, dimension reduction techniques, restricted Boltzmann machines (RBMs), self-organizing maps (SOMs), and clustering algorithms. Many unsupervised learning algorithms have been applied to solve problems in regulatory genomics (Li et al. 2015). For example, HMMs have been applied to automatically annotate the chromatin state genome-wide (ChromHMM) (Ernst and Kellis 2012) by modeling large-scale epigenetic datasets. In ChromHMM, the genome is first split into 200 bp bins, and each epigenetic mark is converted to a binary label (0 or 1) representing whether the mark is present at a particular genomic segment. Then a first-order multivariate model is trained on the whole genome to infer the hidden state of each genomic segment and the hidden states are mapped to annotations, such as enhancer, TSS, and heterochromatin. Another example is use self-organizing maps to infer the combinatorial binding rules of transcription factors (Boyle et al. 2014a). A self-organizing map is a type of artificial neural network that is trained using a neighborhood function to preserve the topological properties of the input space and visualize high-dimensional data in a low-dimensional view. Analysis of the binding of orthologous transcription factors between human-worm and human-fly identified *cis*-regulatory modules (CRMs) with at least two transcription factor binding sites. These CRMs were then fed into the SOM algorithm to find similarities of combinatorial TF binding within and between species. The co-associations of TFs were found to be mostly conserved in promoters and much less conserved in the distal regions, like enhancers.

I-3.2 Supervised learning methods

Experimental genome-wide profiling of enhancers through enhancer activity proxies has generated abundant labeled data of enhancers in various tissues and species. Supervised learning is used to model labeled data. Some common supervised algorithms include regression models, tree algorithms, support vector machines (SVMs), and the recently popular deep neural networks (DNNs). One of the first attempts to apply machine learning methods to enhancer prediction

trained a correlation-based model with histone modification profiles of established transcriptional regulatory elements and then applied the trained model to make new enhancer predictions (Heintzman et al. 2007). The majority of predicted enhancers were supported by at least one of the previous known enhancer-associated marks, such as DNaseI hypersensitivity, binding of p300, or binding of TRAP220. Tree methods have been applied to enhancer prediction as well. For instance, a vector-random-forest-based supervised model called RFECS was applied to enhancer prediction using the 24 histone modifications (Rajagopal et al. 2013). Though many methods, like those above, are based on epigenetic marks, enhancer prediction is also possible directly using DNA sequence as features. Lee et al. (Lee et al. 2011) extracted the sequence features of enhancers with a k-mer spectrum, which is the frequency spectrum of all possible short fixed-length DNA sequences (k-mer), and used that to train an SVM to predict EP300/CREBBP-bound mouse enhancers from background genomic sequences. They obtained good performing models and identified predictive k-mers that are similar to biologically relevant transcription factor motifs. Later in 2014, the same group enhanced this method using a gapped k-mer kernel (Ghandi et al. 2014). Erwin et al. integrated the DNA sequence information with evolutionary conservation, regulatory protein binding, and chromatin modifications with a multiple kernel SVM to predict developmental heart enhancers (Erwin et al. 2014).

More recently, with development of deep neural network algorithms and the availability of GPUs, deep learning models have become the dominant approach for the modeling regulatory sequences. DeepBind was developed to predict the sequence specificities of DNA- and RNA-binding proteins using a convolutional neural network (Alipanahi et al. 2015). The convolutional neural network in general has a similar architecture as an artificial neural network, that is, a layered structure of neurons. Each neuron performs a mathematical transformation of the input to output to the neurons in the next layer. The key difference is in the addition of convolution layers where the convolutional operation is used. In the case of DNA sequences, the convolutional operation in the first layer can be viewed as sliding a kernel (a weight matrix, similar to a position weight matrix) along the different positions in the DNA sequences. Intuitively, in the training of the convolutional neural network, this helps the model to learn recurring patterns, like transcription factor motifs, in the input sequences. These convolutional layers are stacked with optional pooling layers and fully connected layers, and finally connected to output neurons. The pooling layer is useful for making the neural network robust to variances in the input. The fully connected layer

can integrate all information from the previous layer and feed into the final output neurons. By using this algorithm, DeepBind achieved higher accuracy at DNA- and RNA- binding protein sequence specificity prediction than previous methods and demonstrated that the neurons in the first convolutional layer learned relevant transcription factors. Since then, there has been a burst of deep learning publications using genomic data. For example, Zhou et al. modeled genome-wide epigenetic marks identified by the ENCODE consortium with a convolutional neural network (Zhou and Troyanskaya 2015). Quang et al. improved the performance of this approach with a hybrid neural network with convolutional layers and bidirectional long short-term memory (LSTM) (Quang and Xie 2015). There are also many studies focused on enhancer prediction directly. For instance, BiRen integrates the sequence encoding power of a convolutional neural network and the benefits in learning long-term dependency of a gated recurrent unit (GRU)-based recurrent neural network (RNN) to accurately predict enhancers (Yang et al. 2017).

I-3.3 Interpretability of machine learning models

In many cases, the interpretability of machine learning models is as important, if not more important, than their accuracy. There is often a trade-off between the accuracy of the model versus the interpretability. As the model becomes more complex, it becomes more powerful at modeling complex patterns in the data and less interpretable at the same time. For example, in regression models, the importance of the features can be directly derived from their correlation coefficients. Tree models, such as random forest and gradient boosted decision trees, are better at modeling non-linear functions than regression models. Although we cannot directly obtain the feature importance from the model, there are multiple ways to calculate it, such as information gain and split counts. In SVMs, we can interpret features weights as feature importance, because the weights of features in a linear SVM can be thought as how important that feature is at determining the separation hyperplane (Iguyon and Elisseeff 2003). More complex models, like neural networks, are much harder to interpret compared to the above models. The multi-layer feed forward architecture of neural networks makes them powerful, and they have been proven to be universal function approximators (Hornik 1991; Lu et al. 2017). Neural networks are dominating performance in computer vision and natural language processing and are gaining increasing popularity in genomics and genetics. However, the millions of weights that interact in a complex way in the neural network make it hard to understand what it learns. Interpretability of models is paramount in many cases, especially in fields where researchers strive to uncover

underlying mechanisms. With the rising need for interpretability, several methods are developed to visualize the features learned by neural networks.

The main approaches for interpreting deep neural networks can be categorized as perturbation based, backpropagation gradient based, and gradient ascent based methods. Many of these methods were developed in the context of image classification and analysis. For example, occlusion is a perturbation-based method in which patches of an image are excluded in the input and then the changes in neuron activations in higher level layers and classifier output are visualized (Zeiler and Fergus 2014). The resulting maps are meaningful to humans and demonstrate that the neural network trained on large amount of image data did learn relevant features. Perturbation-based methods have also been applied to genomic sequence trained deep learning models DeepSEA uses an *in silico* mutagenesis to assign importance score to individual nucleotides based on how the perturbation of the single nucleotide influences epigenetic mark predictions (Zhou and Troyanskaya 2015). The derived scores were then used to train boosted logistic regression classifiers for predicting functional noncoding variants, and these scores achieved state-of-the-art performance. However, a recent study performed experimental saturation mutagenesis on 20 disease-associated gene promoters and enhancers for over 30000 single nucleotides showed that the variant effect prediction made by DeepSEA has low correlation with the experimental results on gene expression (Kircher et al. 2019). This suggests either the features learned by deep learning model trained on the epigenetic data is not accurate enough for explaining the single nucleotide variant effect on gene expression or the *in silico* mutagenesis method is not reliable.

The second group of neural network interpretation methods are backpropagation gradient based approaches. This group of methods is more computationally efficient than the perturbation methods because it only takes one back propagation to compute the importance score for all input positions. Simonyan et al. (Simonyan et al. 2013a) generated saliency maps by computing the gradient of the output with respect to pixels of an input image. The intuition is that if the gradient at certain pixel of the image is high, then change in the value of pixel will have a higher impact on the neuron's output. Then this approach was further developed to zero out the gradients when the input to the rectified linear unit (ReLu) during forward propagation is negative or the gradients to the ReLu during back propagation is negative (Springenberg et al. 2014). This method is referred as guided back propagation and generates sharper explanations for image trained neural networks. Layer-wise Relevance Propagation (LRP) is another approach using back propagation to calculate

feature importance (Bach et al. 2015). This computation is later found to be equivalent to multiply of gradient and input matrices (Shrikumar et al. 2017b). DeepLIFT also uses backpropagation (Shrikumar et al. 2017a). The difference is that DeepLIFT calculates the difference between the inputs of interest versus a set of reference inputs. With the help of the reference sequences, DeepLIFT can capture the meaningful difference of the inputs even when the gradients are zero and avoid the caveats that the gradients through ReLu are not continuous. DeepLIFT applied to genomic sequence was able to recover simulated transcription factor motifs in the DNA sequences better than the gradients x input and guided back propagation gradient x input method.

The last group of neural network interpretation methods are gradient ascent based methods. The logic behind this group of methods is that the feature learned by an internal neuron can be found by computing the gradients with respect to the input and iteratively tweaking the input to maximally activate the neuron. This group of methods is also referred as activation maximization or visualization by optimization. It was first proposed by Erhan and colleagues (Erhan et al. 2009). They applied this technique to interpret the neurons in Deep Belief Nets and Stacked Denoising Auto-encoders trained on the MNIST dataset, which consists of handwritten digits. They found that with random initialization, they could visualize human recognizable digits from the neurons in the hidden layers. In neural networks trained with more sophisticated image samples, the naïve version of activation maximization can give unrealistic images. Because not all possible input spaces for maximally activating a neuron have been explored by the neural network trained with the limited set of images. This problem can be addressed by constraining the optimization process. Using L2-regularization in the optimization function can improve the image quality for the final layers of a convolutional neural network (Simonyan et al. 2013a). Similarly, using natural image priors for the optimization also helped produce more realistic images (Mahendran and Vedaldi 2015a). Other regularization techniques, such as Gaussian blur (Yosinski et al. 2015), total variation (Mahendran and Vedaldi 2015b), and jitter (Mordvintsev et al. 2015), also can improve the resulting optimized image. However, the neurons in higher layers of the neural network sometimes learn multiple features. For example, a face detecting neuron in a DNN responds to both human and lion faces (Yosinski et al. 2015). Based on this observation, Yosinski et al. proposed to first perform unsupervised clustering of the input images based on their neuron activating patterns and then compute the mean image for each cluster as the prior for the activation maximization process (Nguyen et al. 2016). The gradient ascent approach has been modified for

interpretation of DNA-sequence-trained neural networks. Using activation maximization with L2-regularization, motifs for individual transcription factors were extracted from the output neurons in a neural work trained with transcription factor binding sites (Lanchantin et al. 2017).

In summary, many machine learning algorithms have been applied to model enhancers. The current state-of-the-art methods are based on deep neural networks. Although much effort has been made to understand the features learned by neural networks in image classification, our understanding of neural networks trained on DNA sequences, especially complex regulatory elements like enhancers, is limited. This raises the question of what these neural networks with superior performance learned about enhancer sequence architecture and what are the benefits and limitations of neural network at modeling enhancer sequences. More specifically, whether the neural network can learn the complex combinatorial binding pattern of transcription factors when trained to perform common enhancer prediction tasks, such as predicting enhancers from difference cellular contexts.

In this dissertation, I present the work I have done investigating evolution of enhancers and dissecting their sequence architectures. In this Chapter, I outline the background and the motivation for studying these questions. In Chapter II, I show that conserved enhancers in mammalian species are more pleiotropic than species-specific enhancers, suggesting evolutionary constraint underpinning the loss and gain of enhancers. In Chapter III, I demonstrate the conservation of enhancer sequence properties despite the rapid turnover of the location of active enhancers in mammalian species through a machine learning based, cross-species prediction framework. In Chapter IV, I investigate the power of a state-of-art enhancer prediction algorithm, deep neural networks, at modeling enhancer architecture. Finally, in Chapter V, I summarize the conclusions of the proceeding chapters and discuss future work that could be done to answer questions raised by the findings in this dissertation.

Chapter II

Gene regulatory enhancers with evolutionarily conserved activity are more pleiotropic than those with species-specific activity

This chapter is a collaboration with Alex Fish. I was co-first author on the manuscript published in *Genome Biology and Evolution* in 2017 (Fish et al. 2017).

II-1 Introduction

Mammalian genomes harbor hundreds of thousands of regulatory enhancer sequences that are essential for directing spatiotemporal patterns of gene expression during development and differentiation (Shlyueva et al. 2014; Consortium et al. 2015). Enhancers contain binding sites for transcription factors (TFs), the binding patterns of which regulate gene expression. Genetic variants that disrupt the functionality of enhancer sequences, and thereby alter gene expression levels, are major contributors to both speciation events (Romero et al. 2012) and risk for complex disease (Maurano et al. 2012; Corradin and Scacheri 2014). Given their functional importance, there is considerable interest in better understanding the evolutionary processes underlying both enhancer sequence conservation and, more importantly, enhancer activity conservation.

Much of the transcriptional machinery responsible for regulating gene expression levels is conserved across species. For example, TFs and the sequence motifs they recognize are often conserved between human and fly (Amoutzias et al. 2007; Wei et al. 2010b; Cheng et al. 2014; Nitta et al. 2015a). Consequently, a sequence's TF binding profile across different species is typically similar (Wilson et al. 2008); however, the enhancer activity of orthologous sequences is less consistent. Ritter et al. 2010 examined the activity profiles of 41 pairs of conserved regulatory elements between human and zebrafish. Roughly a third of these pairs demonstrated consistent activity patterns between species, but the majority did not (Ritter et al. 2010). Villar et al. 2015 demonstrated that regulatory activity turnover is pervasive between even more closely related species; only 1% of human liver enhancers had conserved activity across 20 mammals (Villar et al. 2015a). Thus, despite similarity in TFs and their binding motifs, orthologous sequences can have highly variable enhancer activity across species.

Pleiotropy—broadly defined as a single genetic locus influencing multiple traits (Paaby and Rockman 2013) —has been proposed to contribute to the evolutionary conservation of both genes and regulatory activity (Galis et al. 2002; He and Zhang 2006; Cheng et al. 2014; Papakostas et al. 2014; Chesmore et al. 2016; Huang et al. 2017). Mutations in pleiotropic regions face a trade-off: Variants potentially advantageous to one function may be deleterious for others (Guillaume and Otto 2012). Consequently, pleiotropic regions may be more likely to be constrained by selection than nonpleiotropic regions. The relationship between pleiotropy and conservation has been demonstrated on several scales. Highly pleiotropic genes are more likely to have conserved orthologs in other species (He and Zhang 2006) and are more likely to have constrained expression levels (Papakostas et al. 2014). In the context of regulatory functions, binding sites for transcription factors that are observed in multiple cellular contexts, and are therefore presumed to be more pleiotropic, are more likely to be conserved between human and mouse (Cheng et al. 2014). Thus, we predicted a positive relationship between pleiotropy and enhancer activity conservation across species.

In this study, we investigated whether enhancers with conserved regulatory activity between species were more likely to be pleiotropic than enhancers with similarly alignable sequences, but species-specific regulatory activity. We quantified pleiotropy at several stages of human gene regulation: Density and diversity of TF binding motifs, extent of regulatory activity across cellular contexts, and number of target genes. We investigated these measures of pleiotropy in liver enhancers recently identified from genome-wide histone modification profiles across ten diverse mammalian species. We compared two groups of sequences present and alignable across all ten species: 1) those with liver activity in all ten mammals considered (conserved-activity) and 2) those with liver enhancer activity in only one species (species-specific-activity). We found that the conserved-activity enhancers consistently had stronger evidence of more, and more diverse, regulatory functions than the species-specific-activity enhancers. We also demonstrated that machine learning classifiers can accurately distinguish these two classes of enhancers using these measures of functional potential and diversity. Overall, our results argue that conserved-activity enhancers are more pleiotropic than species-specific-activity enhancers with similar levels of sequence alignability. This suggests that more diverse functional activity contributes to conserved activity across species, and that conserved activity may facilitate acquisition of additional functions.

II-2 Materials and Methods

II-2.1 Identifying Enhancers and Cross-Species Alignments

Enhancers were previously identified by Villar et al. (Villar et al. 2015a) in primary liver tissues collected from 20 mammalian species. Using ChIP-seq, Villar et al. (2015) identified H3K27ac and H3K4me3 peaks across the entire genome; putative enhancers were defined as genomic regions exclusively containing H3K27ac peaks (i.e., H3K27ac peaks that did not overlap H3K4me3 peaks) found in at least two representatives of the species. We restricted our analysis to the following ten species with high quality genome builds: Human (Homo sapiens), macaque (Macaca mulatta), marmoset (Callithrix jacchus), mouse (Mus musculus), rat (Rattus norvegicus), rabbit (Oryctolagus cuniculus), cow (Bos taurus), pig (Sus scrofa), dog (Canis familiaris), and cat (Felis catus). Cross-species comparisons to identify whether or not a sequence was present and active in other species were performed in reference to the eutherian mammal EPO alignment. To determine enhancer activity in cellular contexts other than the liver, we used enhancers identified by CAGE by the FANTOM Consortium (http://enhancer.binf.ku.dk/presets/; last accessed September 21, 2017) (Andersson et al. 2014).

II-2.2 Standardizing Enhancer Length

To avoid confounding by length, we restricted enhancer sequences used in the majority of analyses to 5 kb centered on the middle of the enhancer. If a putative enhancer was shorter than 5 kb, we extended the enhancer boundaries symmetrically in both directions until it was 5 kb. The length of 5 kb was selected as it the intermediate point between the average length of the conserved-activity enhancers (7,895 bp) and species-specific-activity enhancers (2,545 bp). For sequences shorter than 5 kb, standardizing length could potentially dilute the density of TF binding motifs; however, it would increase the likelihood of overlapping enhancers in multiple cellular contexts or mapping to additional gene targets. In other words, it would have inconsistent effects on measures of pleiotropy; only in the TF binding motif analysis would the potential for pleiotropy for shorter sequences possibly be reduced. We demonstrated that the decreased density of TF binding motifs in human-specific-activity enhancers was not a product of length standardization (Figure I-1). Consequently, any influence of the length standardization in the subsequent analyses of breadth of activity and gene targets would only increase the likelihood of pleiotropic effects in species-

specific-activity enhancers. This would reduce our power to detect increased evidence for pleiotropy in conserved-activity enhancers, but would not result in false positives.

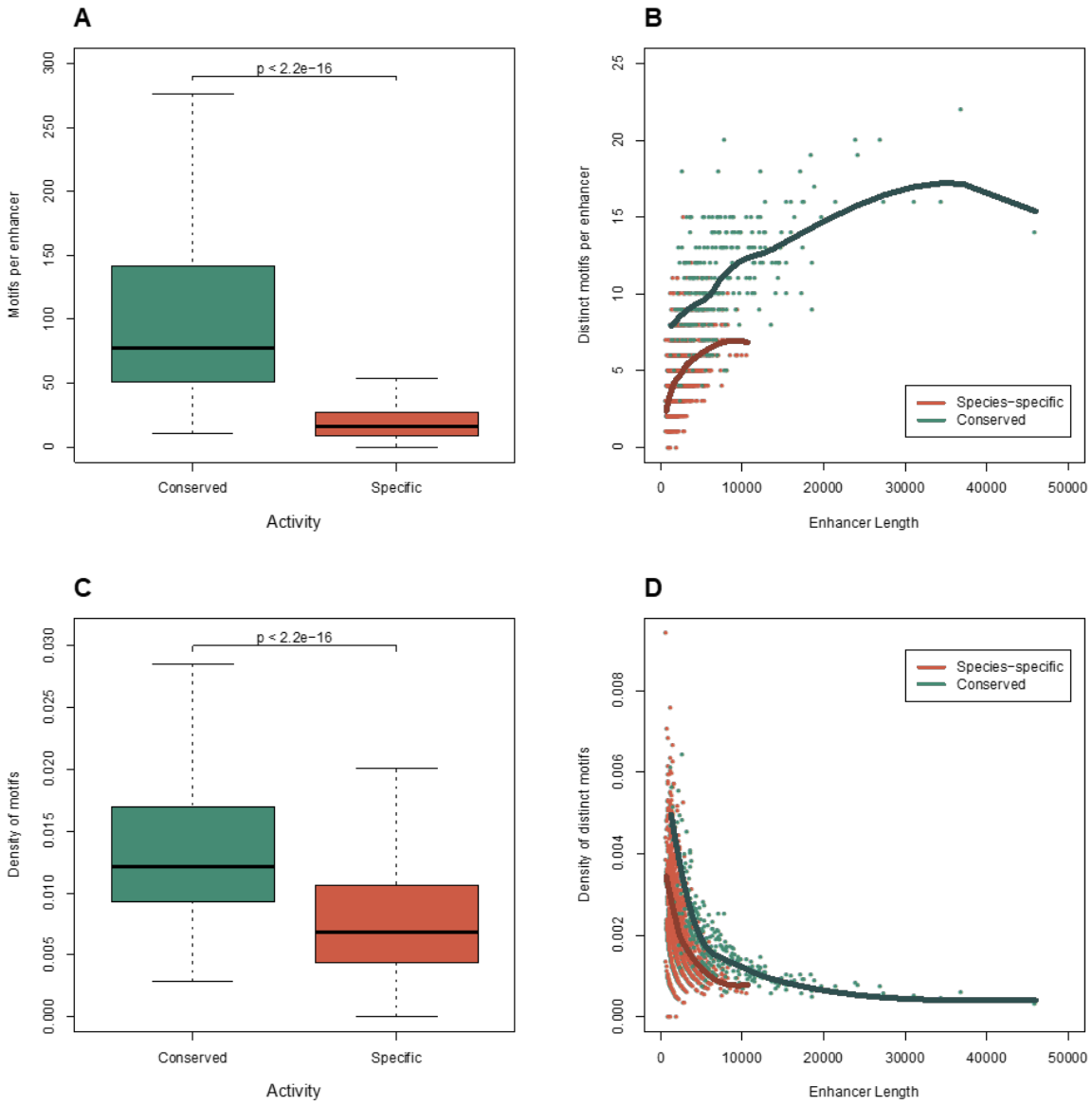

Figure II-1. Conserved-activity enhancers have greater regulatory potential and diversity relative to human-specific-activity enhancers when lengths are not standardized. We investigated whether standardizing enhancer sequences to 5 kb—the intermediate point between the average length of human-specific-activity enhancers and conserved-activity enhancers—diluted the frequency of TF binding motifs in human-specific-activity enhancers because they were, on average, shorter. To do so, we examined the frequency of TF binding motifs (JASPAR, Core Vertebrates) in sequences prior to length standardization. We first determined the total count of TF binding motifs (A), and observed that conserved-activity enhancers had a significantly greater number of TF binding motifs ($p < 2.2\text{x}10^{-16}$, Mann-Whitney *U* test). We then investigated the diversity of TF binding motifs by counting the number of distinct TF binding motifs per enhancer sequence (B). In comparing distinct motifs, we do not expect the number of

distinct motifs to scale linearly with enhancer length. Essentially, the longer a sequence is, the more likely it is to contain a greater number of motifs, meaning that each subsequent motif is less likely to be unique. This results in a non-linear relationship with distinct TF motifs and enhancer length. Consequently, we examined the number of distinct TF motifs as a function of enhancer length. The LOESS curve fit for the conserved-activity enhancers consistently remained above the LOESS curve for human-specific-activity enhancers, indicating that conserved-activity enhancers have a greater diversity of TF binding motifs across all observed enhancer lengths. We additionally examined the density of TF binding motifs per base pair of the enhancer sequence. (C) Conserved-activity enhancers have a significantly greater density of total TF binding motifs per base pair ($p < 2.2\text{x}10^{-16}$, Mann-Whitney *U* test) than their human-specific-activity counterparts. We similarly compared the density of distinct TF binding motifs using the same approach previously described (B). (D) The LOESS curve for conserved-activity enhancers remained above that of human-specific-activity enhancers, indicating that they had a greater density of distinct TF binding motifs across all observed enhancer lengths.

II-2.3 Identification of TF Binding Motifs

We identified TF binding motifs using four databases derived across diverse sets of species and using different experimental approaches: Motifs derived from ChIP-Seq peaks in human by the ENCODE Project ($n = 2{,}065$) (Kapur et al. 2011); motifs derived from ChIP-Seq peaks and HT-SELEX in vertebrates by JASPAR (Core Vertebrates) ($n = 519$) (Mathelier et al. 2016); motifs derived from ChIP-Seq peaks in human and HT-SELEX by HOCOMOCOv9 ($n = 426$) (Kulakovskiy et al. 2016); and motifs derived from ChIP-Seq peaks and HT-SELEX from human and mouse ($n = 843$) (Jolma et al. 2013a). For each of these data sets, we scanned the putative enhancer sequences for motif occurrences using FIMO (Grant et al. 2011), using the default settings and requiring a *q*-value of <0.1 to be considered a match.

II-2.4 SVM Classifiers

We trained SVM classifiers to distinguish between conserved-activity enhancers and species-specific-activity enhancers using three different kinds of features: TF binding motif frequencies, *k*-mer spectra, and functional genomics annotations. For the TF motif-based classifiers, each enhancer was associated with a feature vector that included the frequency of all possible TF motifs in its sequence. We then trained a linear SVM to distinguish the two classes of enhancers. The kernel was normalized using the square root diagonal kernel normalizer. All training and testing was done in the EnhancerFinder framework (Erwin et al. 2014).

K-mer spectra quantify sequence content with the frequency of each unique nucleotide combination of length *k* in the enhancer sequence. We determined the *k*-mer spectra of each

enhancer sequence using EnhancerFinder (Erwin et al. 2014);the kernel was normalized using the square root diagonal kernel normalizer. The reverse complement of the sequence was considered (i.e., counts for ATG and CAT were combined). We examined various *k* (4, 5, 6, 7, 8) and found consistent results across settings (Figure I-2).

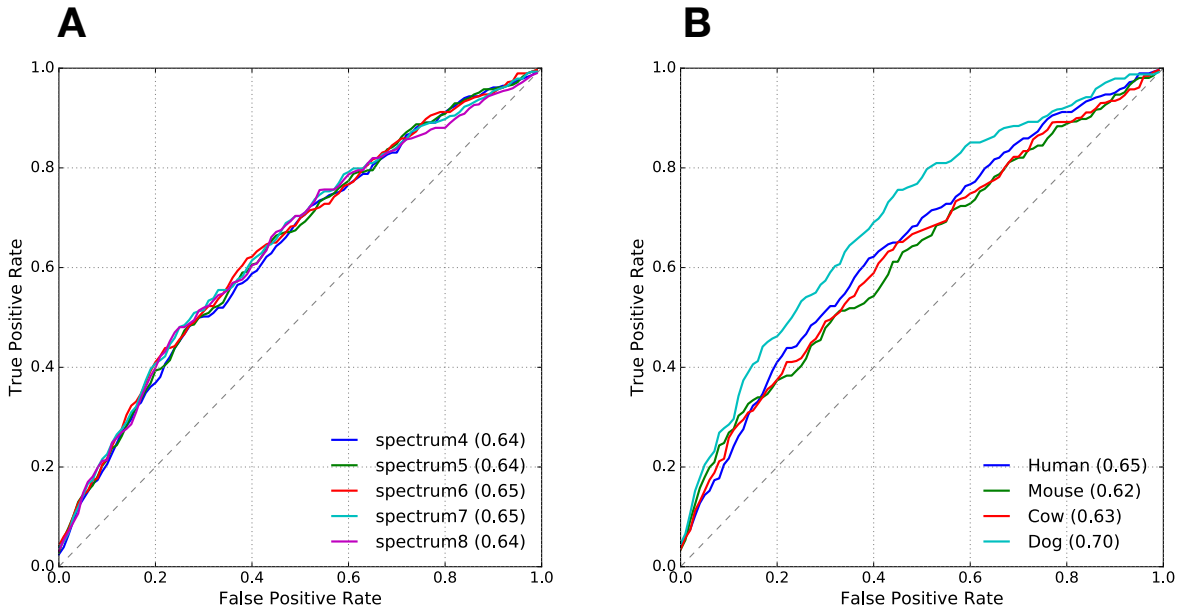

Figure II-2. ***k*-mer spectrum classifiers distinguish conserved-activity from species-specific-activity enhancers with moderate accuracy.** Performance of SVM classifiers trained to distinguish conserved-activity enhancers from human-specific-activity enhancers based on their *k*-mer spectra, the frequency of each unique nucleotide combination of length *k*, is similar across a range of possible *k* (4, 5, 6, 7, 8). There was little variation in the classifiers' ability to distinguish between the two enhancer categories across this range. (B) Performance of SVM classifiers trained to distinguish conserved-activity enhancers from species-specific-activity enhancers based on their 6-mer spectra is similar across human, mouse, dog, and cow. The area under each ROC curve is provided in the parentheses in the legend.

To investigate whether functional genomics annotations were predictive of enhancer activity conservation, we used data collected by the ENCODE Project. Specifically, we used DNase-Seq, histone modifications, and TFBS Peaks (SPP) curated by the ENCODE Analysis Hub at the European Bioinformatics Institute (https://genome.ucsc.edu/ENCODE/downloads.html; last accessed September 21, 2017). We considered each genome-wide annotation as a binary feature, and each enhancer was assigned 0 if it did not overlap an element of the annotation set or 1 if it

did overlap. Training and testing of the functional genomics classifier was also carried out in the EnhancerFinder framework (Erwin et al. 2014) .

II-2.5 Target Gene Mapping and Analysis of Gene Expression across Contexts

We used two methods to map the enhancers to their target genes: 1) GTEx eQTL association based target gene mapping. We first identified SNPs in the enhancer regions of interest and SNPs in high linkage disequilibrium with them ($r2 > 0.9$, based the 1000 Genomes EUR super population). Then, using expression data from GTEx, we considered genes for which these SNPs were eQTL to be putative target genes (The GTEx Consortium et al. 2015). 2) FANTOM enhancer–TSS associations. The FANTOM consortium released a set of target predictions for each of their predicted transcribed enhancers based on the coexpression of the enhancer and genes across tissues (Andersson etal. 2014). We overlapped each liver enhancer of interest with the FANTOM enhancers. We then considered any genes associated with an overlapping FANTOM enhancer as putative target genes. To analyze the breadth of activity of target genes, we used the median Reads Per Kilobase of transcript per Million mapped reads (RPKMs) for genes from the GTEx v6 RNA-Seq data, which includes 53 types of tissue (The GTEx Consortium 2015).

II-3 Results and Discussion

In this study, we explored attributes that distinguish genomic regions with both alignable sequence and regulatory activity across diverse mammals from those with similarly alignable sequences, but regulatory activity isolated to a single species. We analyzed genome-wide maps of histone modifications in primary liver tissue from ten mammals to quantify the regulatory activity conservation spectrum for liver regulatory sequences. Following Villar et al. (2015), we defined regulatory activity as peaks of H3K27ac histone modifications without the H3K4me3 modification. As histone modifications are correlated with enhancer activity in reporter assays (Creyghton et al. 2010; Nord et al. 2013; Villar et al. 2015b), we refer to these sequences as enhancers for brevity. As illustrated in Figure II-3, we considered two enhancer sets of interest: Sequences that can be aligned across the genomes of ten mammalian species with evidence of enhancer activity in each species (conserved-activity enhancers; $n = 283$) and sequences that can be aligned across the ten species with evidence of enhancer activity exclusively in a single species (species-specific-activity enhancers). We examined species-specific-activity liver enhancer sets

across four different mammalian species: Human (n = 1,913), mouse (n = 1,526), dog (n = 1,894), and cow (n = 3,093).

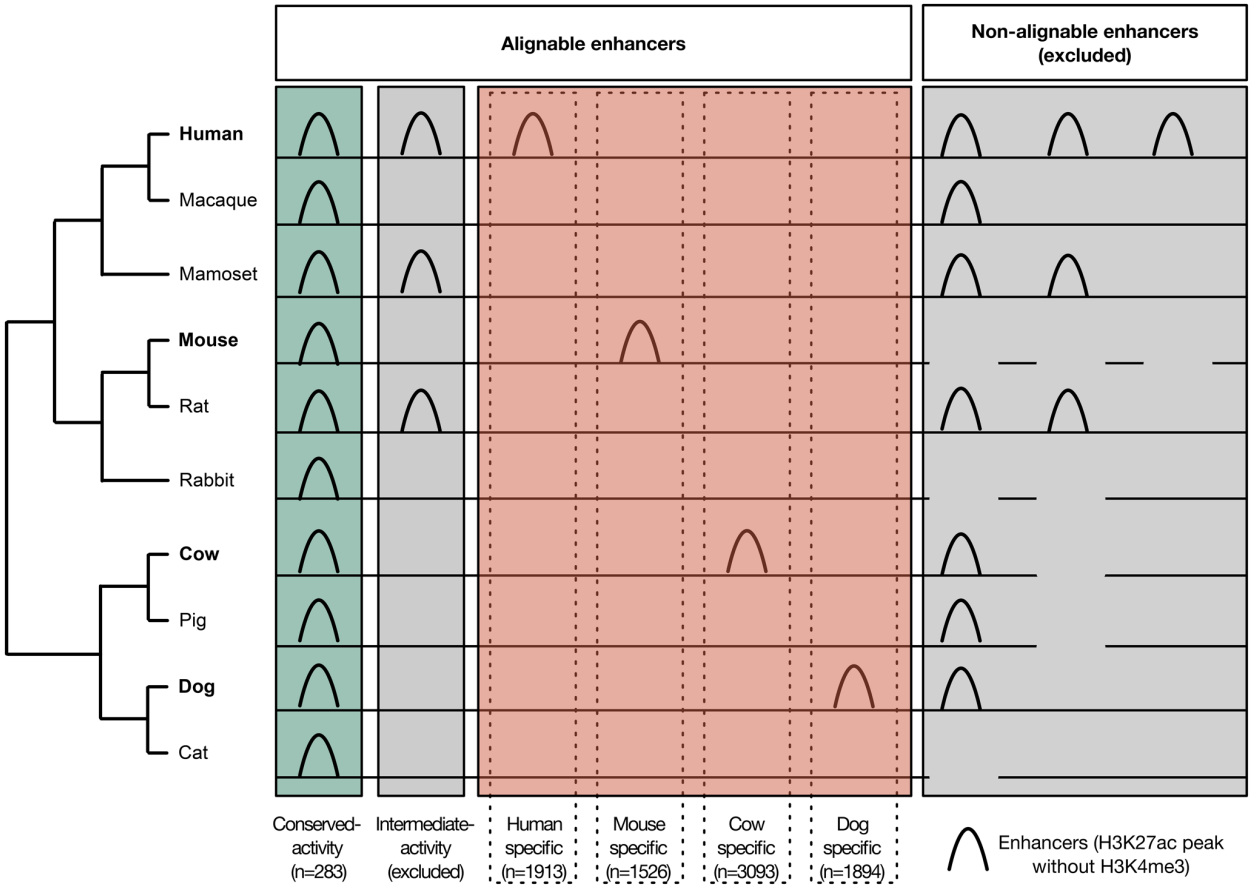

Figure II-3. Defining the enhancers analyzed in this study. We analyzed enhancers previously identified in primary liver tissue from ten mammals (Villar et al. 2015b). Enhancers were defined on the basis of histone modifications (presence of H3K27ac peaks, absence of H3K4me3 peak). We analyzed sequences alignable across all ten species to disentangle sequence conservation from regulatory activity conservation across species. To identify trends that would hold across all mammals, we restricted our analysis to extremes of the activity conservation spectrum: conserved-activity enhancers (green), and species-specific-activity enhancers (red). We considered species-specific-activity enhancers from human, mouse, cow, and dog to represent a diverse array of clades.

II-3.1 Conserved-Activity Enhancers Have Greater Density and Diversity of TF Binding Motifs than Species-Specific-Activity Enhancers

The differential enhancer activity of alignable sequences may be attributable to differences in sequence properties that determine their regulatory potential, as quantified by both the density of

TF binding motifs and the diversity of distinct TFs with binding motifs. Within a species, enhancers with a greater density of TF binding motifs are both stronger (Erceg et al. 2014) and more robust to disruptive genetic variation (Ludwig et al. 2011). We hypothesized that these principles generalize to enhancer conservation between species. To investigate this, we scanned all enhancer sequences for matches to a curated set of nonredundant TF binding motifs from the JASPAR database (Mathelier et al. 2016). Unless otherwise noted, enhancers were length standardized (Materials and Methods) to avoid confounding.

Conserved-activity liver enhancers have a greater density of TF binding motifs than human-specific-activity enhancers (Figure II-4A; median: 61 versus 44 per enhancer; Mann–Whitney *U* (MWU) test, $P < 2.2 \times 10^{-16}$). Moreover, conserved-activity enhancers contain binding sites for almost double the number of distinct TFs (Figure II-4B; median: 10 versus 6 per enhancer; MWU test, $P < 2.2 \times 10^{-16}$). This finding is robust across other databases of TF binding motifs, including motifs from the ENCODE Project (Kapur et al. 2011), HOCOMOCO (Kulakovskiy et al. 2016), and SELEX (Jolma et al. 2013a) studies. Furthermore, the other species-specific-activity (mouse, dog, and cow) enhancers also had both lower density and diversity of TF binding motifs relative to conserved-activity enhancers. This trend also was consistent when enhancers were not length standardized. Thus, conserved-activity enhancers have both a greater density and diversity of TF binding motifs than species-specific-activity enhancers, across multiple species and TF motif databases.

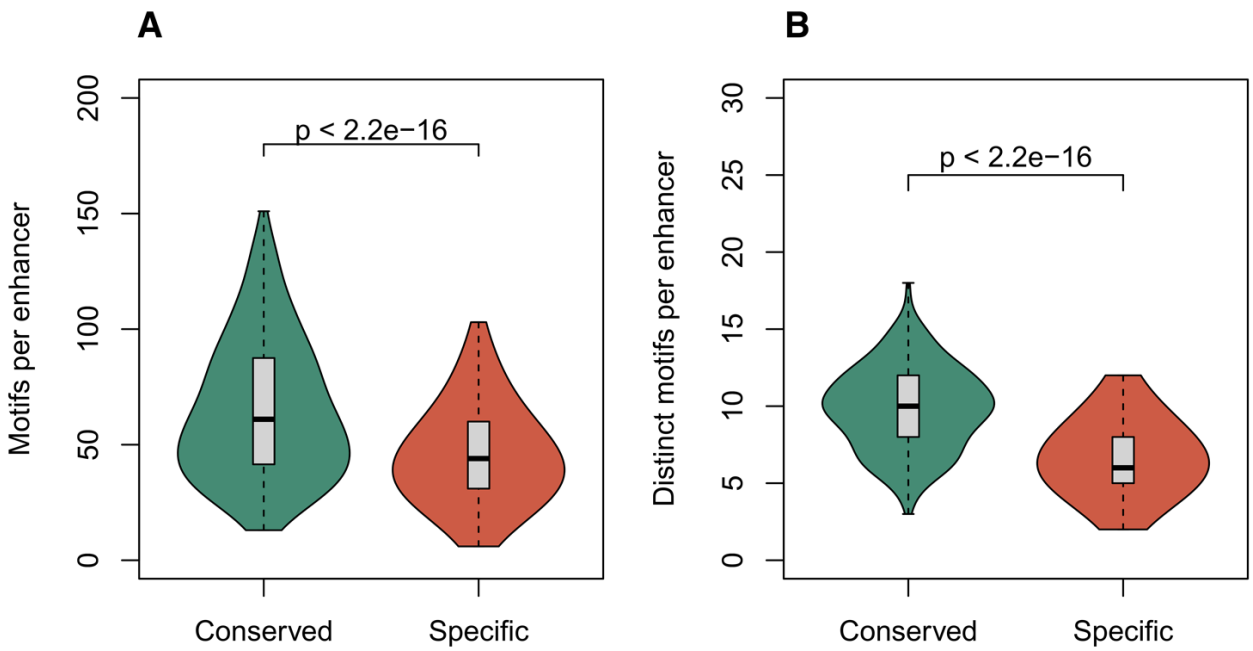

Figure II-4. Conserved-activity enhancers have greater TF binding motif density and diversity than species-specific-activity enhancers. We quantified the total number of TF binding motifs (JASPAR, Core Vertebrates (Mathelier et al. 2016)) and of distinct TF motifs within the enhancer sequences. The 283 conserved-activity enhancers have (A) significantly more TF binding motifs per enhancer (median of 61 vs. 44) and (B) binding motifs for significantly more distinct TFs (median of 10 vs. 6) than the 1,913 human-specific activity enhancers. Each box covers the first through third quartiles, and the whiskers extend to the most extreme data point within 1.5 times the interquartile range of the full distribution. The distributions are summarized using the R vioplot package with default parameters applied to each distribution with boxplot outliers removed. The enhancers were standardized by length and compared with the Mann–Whitney U test.

We next examined whether differences in TF binding motif profiles were sufficient to distinguish conserved-activity enhancers from species-specific-activity enhancers in a machine learning framework. First, we trained linear support vector machine (SVM) classifiers with conserved-activity enhancers as positives and species-specific-activity enhancers as negatives using the frequency of each distinct TF binding motif in the enhancer sequence as features. We performed 10-fold cross validation, computed receiver operator characteristic (ROC) curves, and evaluated classifier performance by the area under the ROC curve (auROC).

The classifiers accurately discriminated the conserved-activity enhancers from the species-specific-activity enhancers in each species (auROC: 0.88–0.97, Figure II-5A). We hypothesize that the particularly strong performance of the mouse classifier may be due to rodent-specific

differences in the genomic GC content distribution compared with other mammals (Romiguier et al. 2010). To benchmark the performance of the classifiers, we ranked enhancers by the density of TF binding motifs in the sequence and evaluated the predictive ability of this single feature (Figure II-5B). Performance notably decreased when only considering the density of TF binding motifs (auROC: 0.53–0.74) for all species, especially mouse. Other approaches for quantifying enhancer sequence properties, such as *k*-mer spectra (Materials and Methods), were not as effective at predicting the conservation of regulatory activity. These results indicate that not only do conserved-activity enhancers have a greater density and diversity of TF binding motifs than species-specific-activity enhancers, but the occurrence patterns of specific TF binding motifs are informative about the conservation of enhancer activity across species.

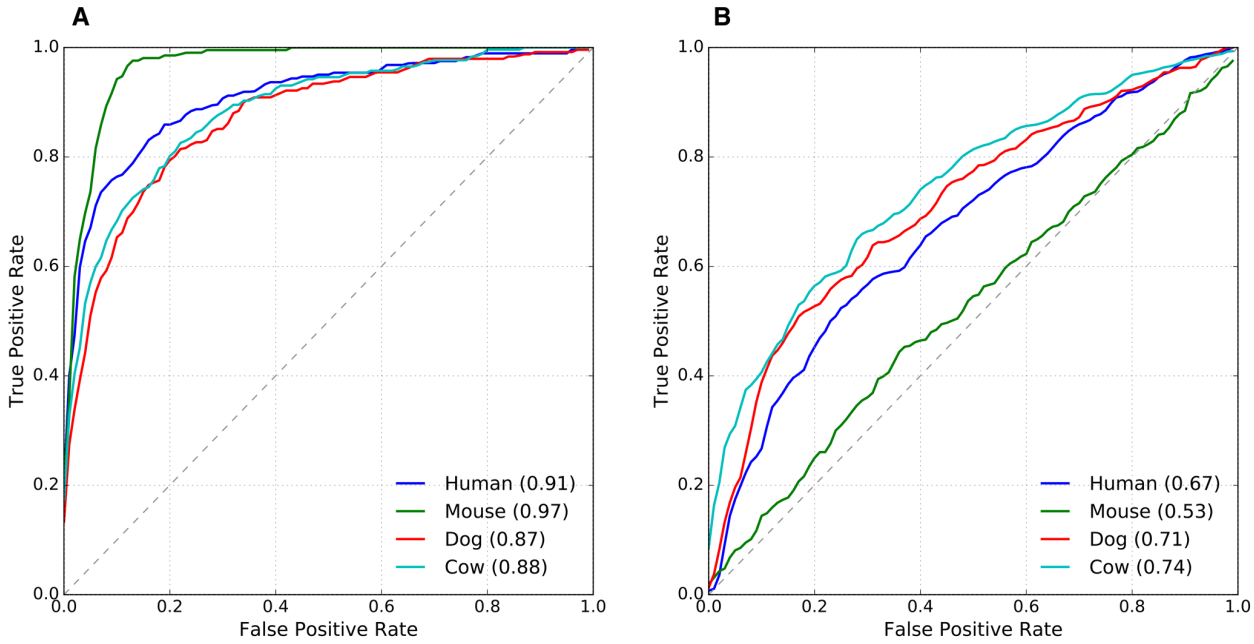

Figure II-5. TF binding motif patterns can distinguish between conserved-activity and species-specific-activity enhancers. (A) In each species (human, mouse, dog, and cow), we trained SVM classifiers to distinguish conserved-activity from species-specific-activity enhancers using the frequency of each TF binding motif individually as features. (B) For comparison, we attempted to distinguish the two classes of enhancers based only on the total number of TF binding motifs. The ROC curves display classifier performance, and the area under the curves is provided in the legend.

II-3.2 Conserved-Activity Enhancers Are Active in More Cellular Contexts than Human-Specific-Activity Enhancers

We next investigated whether the greater density and diversity of TF binding motifs of conserved-activity enhancers translated to increased regulatory activity across biological contexts within a species. We focused on human, as enhancers have been identified in a more diverse set of biological contexts for human than other species. We used enhancers in 108 cellular contexts identified by the FANTOM consortium, which used cap analysis of gene expression (CAGE) assays to identify bi-directionally transcribed "eRNA" transcripts (Andersson et al. 2014). On average, conserved-activity enhancers overlapped an active FANTOM enhancer in more than seven cellular contexts, which was double the number of cellular contexts expected from all human liver enhancers (mean: 7.2 vs. 3.6 per enhancer; $P = 1.09 \times 10^{-6}$, MWU test) and almost quadruple the number of active contexts for human-specific-activity enhancers (mean: 7.2 vs. 1.9 per enhancer; $P = 2.22 \times 10^{-12}$, MWU test) (Figure II-6A). We next tested whether specific cellular contexts in FANTOM drove this enrichment by evaluating the overlap with enhancers from each FANTOM context separately. Conserved-activity enhancers were significantly more likely to overlap FANTOM enhancers relative to all human liver enhancers in 36 of 108 cellular contexts, and relative to human-specific-activity enhancers in 72 of 108 cellular contexts ($P < 0.05$ after Bonferroni correction, Fisher's exact test). The converse—significant depletion of conserved-activity enhancers relative to all enhancers, or human-specific-activity enhancers—was never observed. These results demonstrate that conservation of activity across species

within one cellular context (the liver) is positively correlated with the breadth of activity across other cellular contexts in humans.

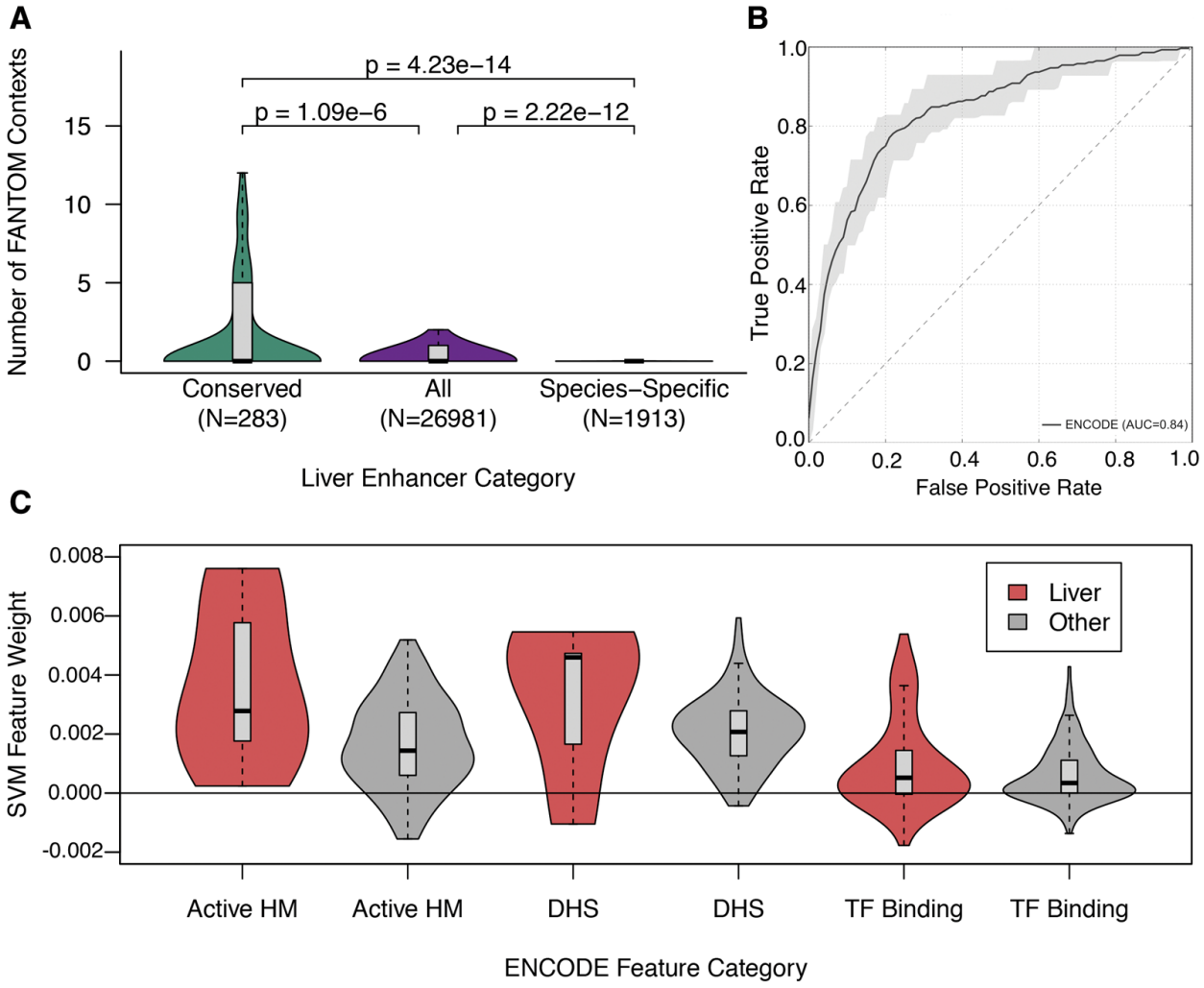

Figure II-6. Breadth of activity across cellular contexts within species is positively associated with conservation of regulatory activity across species. We identified the overlap of conserved-activity liver enhancers, all human liver enhancers (regardless of their conservation status), and human-specific-activity enhancers with enhancers identified across 108 human cellular contexts via CAGE by the FANTOM consortium. In addition to being less active across contexts than conserved-activity enhancers ($P = 4.23 \times 10{-}14$, MWU test), the human-specific-activity enhancers were active in significantly fewer cellular contexts than expected based on all liver enhancers ($P = 2.22 \times 10{-}12$). (B) We then trained an SVM on DNase I hypersensitivity sites (DHS), histone modifications, and TF binding profiles identified genome-wide in 125 cellular contexts by ENCODE in human to distinguish conserved-activity enhancers from human-specific-activity-enhancers (auROC = 0.84). We considered each genome-wide annotation as a binary feature, and each enhancer was assigned 0 if it did not overlap an element of the annotation set or 1 if it did overlap. Shaded areas are bounded by the max and min ROC obtained across 10-fold cross validation. (C) We examined the

weights assigned to each feature by the classifier; the absolute value of a feature's weight indicates its overall importance, and the sign indicates whether it is more associated with conserved-activity enhancers (positive) or human-specific-activity enhancers (negative). The weights for features associated with active genomic regions (i.e., active histone modifications (HM), DHS, and TF binding) are positively skewed regardless of the cellular context, indicating they are generally more associated conserved-activity enhancers. The weights for features from liver contexts (red) are consistently more positive than similar features from other contexts (gray). Distributions are summarized as described for Figure II-4.

We next examined whether patterns in functional genomics data indicative of regulatory activity (and inactivity) across diverse cellular contexts within a species could predict the activity conservation of liver enhancers across species. We used human data collected by the ENCODE Project for this component of the analysis (Bernstein et al. 2012; Sloan et al. 2016). In contrast to FANTOM, ENCODE performed a broad array of functional genomic assays, including DNase I hypersensitivity sites (DHS), histone modifications, and TF binding profiles genome-wide in 125 cellular contexts. In 10-fold cross-validation, the ENCODE classifier was able to distinguish many conserved-activity enhancers from human-specific-activity enhancers (auROC = 0.84; Figure II-6B); however, it was not as accurate as the TF binding motif based classifier (Figure II-5A; auROC = 0.91). As anticipated, the majority of the most predictive features for both conserved-activity and human-specific-activity enhancers were from liver contexts (Table II-1 and II-2). Additionally, there was a general association between active functional annotations and conserved-activity enhancers (Figure II-6C), regardless of the cellular context in which the annotation was identified. This association between conserved-activity enhancers and active annotations across contexts argues that they are active in a broader range of cellular contexts.

Table II-1. The features most associated with conserved-activity liver enhancers are primarily associated with active regions and are from liver cells. The top five features most associated with conserved-activity enhancers, along with the weight assigned to them by the SVM classifier, are shown.

Feature	Weight	Tissue	Activity Status
HEPG2 H3K4me2	0.0076	Liver	Active
HEPG2 H3K27ac	0.0075	Liver	Active
CMK DNase	0.0059	Blood	Active
HEPG2 H3K79me2	0.0058	Liver	Active
Huh-7 DNase	0.0055	Liver	Active

Table II-2. The features most associated with human-specific enhancers are primarily indicative of inactive genomic regions. The top five features most associated with human-specific-activity enhancers, along with the weight assigned to them by the SVM classifier, are shown.

Feature	Weight	Tissue	Activity Status
NT2-D1 H3K9me3	-0.0022	Testis	Inactive
HEPG2 H3K27me3	-0.0020	Liver	Inactive
HELA-S3 H3K27me3	-0.0018	Cervix	Inactive
HEPG2 FOSL2-Peak	-0.0018	Liver	Active
HUVEC H3K27me3	-0.0016	Blood Vessel	Inactive

Motivated by these differences in the breadth of activity of the conserved-activity and human-specific liver enhancers, we analyzed the weights assigned to each TF motif by the trained human enhancer SVM classifier from the previous section. Three transcription factors' motifs (SP1, SP2, and EWSR1-FLI) were assigned notably higher weights than others (Figure II-7). SP1 and SP2 are broadly expressed zinc finger TFs from the Sp/XKLF family that recognize common GC box motifs and carry out diverse functions across many tissues (Philipsen and Suske 1999). EWSR1-FLI1 is a fusion of EWSR1 and FLI1, an ETS family TF, that is involved in oncogenesis in Ewing tumors (Guillon et al. 2009). The ETS family is a diverse family of TFs with broad functions, but all members have a conserved DNA-binding domain that recognizes diverse motifs with a core GGA(A/T) sequence. While many specific ETS family TFs are present in the motif database, the

EWSR1-FLI1 motif consists of GGAA repeats, so we believe that this motif is likely highly weighted as a proxy for this family of broadly active factors. Thus, motifs useful in distinguishing the conserved-activity from human-specific-activity enhancers can be bound by diverse, broadly expressed TFs, further supporting their potential for activity in many cellular contexts.

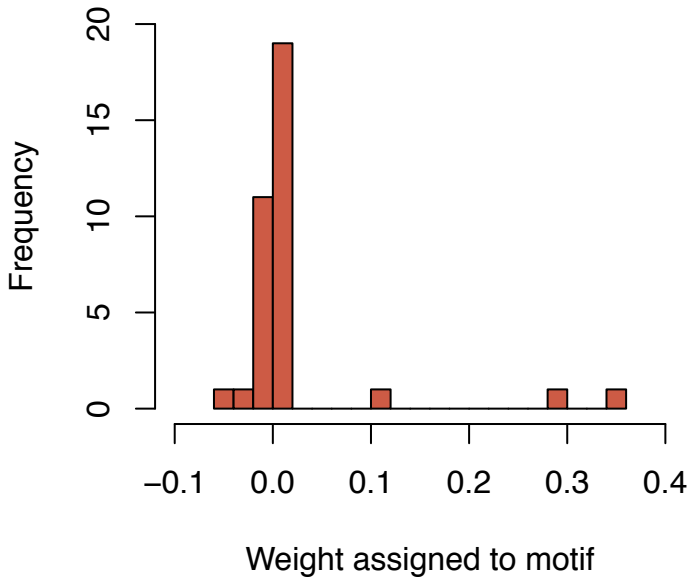

Figure II-7. Distribution of weights assigned to TF motifs in the trained SVM classifier. We trained a SVM classifier to distinguish conserved-activity enhancers from human-specific-activity enhancers using the frequencies of all TF binding motifs in the JASPAR database. The classifier assigned weights to each motif. The absolute values indicate the magnitude of the contribution of the motif to the prediction, and the sign indicates whether the motif was associated with conserved-activity enhancer (positive) or human-specific-activity enhancers (negative). Three motifs were assigned notably larger weights than the others: SP2, SP1, and EWSR1-FLI, in decreasing order.

II-3.3 Conserved-Activity Enhancers Have More Target Genes and Their Target Genes Are More Broadly Active than Human-Specific-Activity Enhancers

Conserved-activity enhancers have a higher density and diversity of TF binding sites, and they exhibit regulatory activity in more genomic contexts than human-specific-activity enhancers. Given their greater regulatory potential and function, we hypothesized that conserved-activity enhancers may regulate more genes, genes with more diverse functions, or both, than do human-specific-activity enhancers. To investigate this, we mapped enhancers to target genes using two

complementary approaches. First, we considered the enhancer–gene pairs predicted by the FANTOM project, which are derived from the coexpression patterns of eRNA and mRNA across many cellular contexts in human (Andersson et al. 2014). FANTOM target data are available for 89 of 283 (31.4%) conserved-activity enhancers and 317 of 1,913 (16.6%) human-specific-activity enhancers. Second, we mapped enhancers to genes using genotype and expression data from the GTEx project (The GTEx Consortium et al. 2015). We identified genetic variants within the enhancer sequences that were significantly associated with gene expression levels, and we then used these expression quantitative trait loci (eQTLs) to match enhancers to potential target genes (Materials and Methods). Using this approach, 174 out of 283 (61.4%) conserved-activity enhancers and 1,250 of 1,913 (65.0%) human-specific-activity enhancers mapped to at least one target gene.

In both the FANTOM and GTEx target sets, conserved-activity enhancers map to significantly more gene targets than human-specific-activity enhancers (Figure II-8A ; mean FANTOM: 2.4 vs. 1.9, $P = 0.01$; mean GTEx: 3.9 vs. 3.0, $P = 0.05$; MWU test). Conserved-activity enhancers target a similar number of genes as human liver enhancers in general (mean FANTOM: 2.4 vs. 2.5, $P = 0.6$; mean GTEx: 3.9 vs. 3.7, $P = 0.8$), and thus the lower number of targets predicted for human-specific-activity enhancers suggests that human-specific-activity enhancers are depleted of targets. However, the gene targets of conserved-activity enhancers are expressed in a more diverse array of cellular contexts than both all liver enhancers (mean FANTOM: 23.6 vs. 16.4, $P = 1.84 \times 10^{-7}$; mean GTEx: 17.8 vs. 15.9, $P = 0.02$; MWU test) and human-specific-activity enhancers (mean FANTOM: 23.6 vs. 11.6, $P = 1.14 \times 10^{-13}$; mean GTEx: 17.8 vs. 14.7, $P = 0.002$; MWU test), for both FANTOM and GTEx mappings (Figure II-8B). Ultimately, conserved-activity enhancers appear to regulate the expression of more genes than human-specific-activity enhancers, and their gene targets are more broadly expressed than those of human liver enhancers collectively.

We next hypothesized that the gene targets of conserved-activity enhancers would be more likely to be evolutionarily constrained. To investigate this, we analyzed probability of loss-of-function intolerance (pLI) scores computed by the Exome Aggregation Consortium (ExAC) to quantify constraint on genes. Using the FANTOM enhancer-gene mappings, conserved-activity enhancers mapped to genes with significantly higher pLI scores than the target genes of all human liver enhancers (mean pLI: 0.53 vs. 0.41, $P = 1.6 \times 10{-}3$; MWU test) (Figure II-8C). This finding generalized to the GTEx enhancer-gene mappings ($P = 2.0 \times 10{-}3$). As expected, the human-

specific-activity enhancer targets had lower median pLI than the conserved-activity targets from FANTOM, but surprisingly, the pLI of human-specific-activity enhancers was greater than for liver enhancers overall (Figure II-8C). However, this was not true among the GTEx targets; the pLI scores for the human-specific-activity enhancer targets from GTEx were not significantly different from the targets of all human liver enhancers. This will require further study as enhancer–target mappings improve.

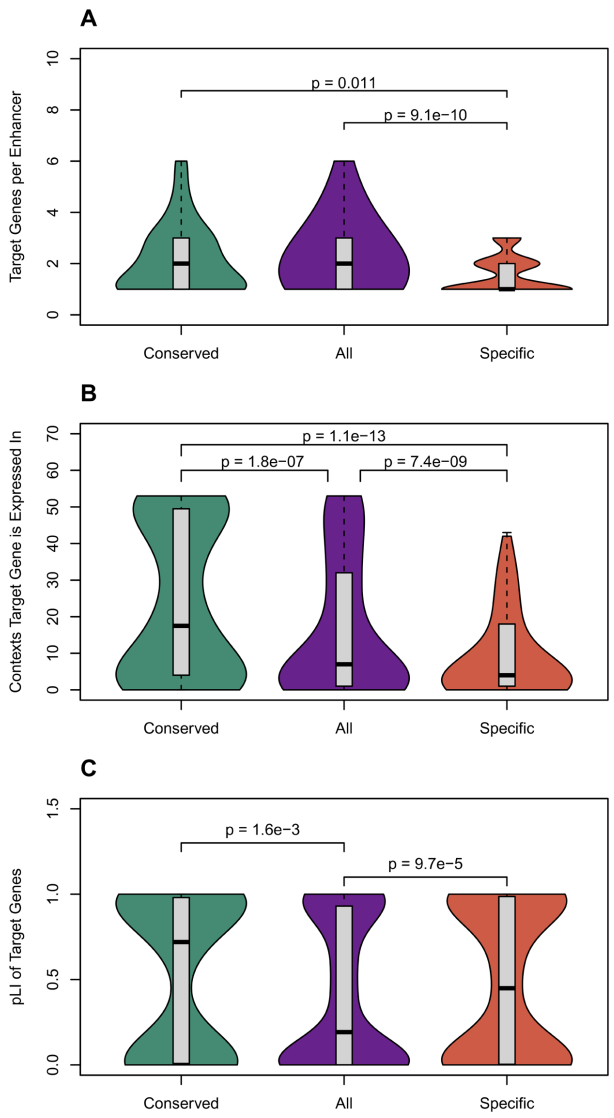

Figure II-8. Conserved-activity enhancers have more target genes than human-specific-activity enhancers; their targets are expressed in more cellular contexts; and their targets are under stronger evolutionary constraint. Conserved-activity enhancers have significantly more target genes than human-specific-activity enhancers, but a similar number as all human liver enhancers. (B) The gene targets of conserved-activity enhancers were active in significantly more contexts than either human-specific-activity or all liver enhancers. Target gene expression was determined from GTEx (The GTEx Consortium et al. 2015). (C) The target genes for conserved-activity enhancers have significantly higher probabilities of being loss-of-function intolerant (pLI) according to ExAC than target genes for all human liver enhancers. Thus, the target genes of conserved-activity enhancers are under more evolutionary constraint. The targets of human-specific-activity enhancers were also less tolerant of loss of function than liver enhancers overall. Enhancers were mapped to target genes based on coexpression patterns by the FANTOM Consortium. Results were similar when identifying targets based on eQTL (Figure II-9). The Mann–Whitney U test was used for all comparisons. Distributions are summarized as described for Figure II-2.

Overall, these results indicate that conserved-activity enhancers regulate the expression of more genes than human-specific-activity enhancers, and that these genes are both more broadly expressed and experience stronger constraint than the gene targets of all human liver enhancers.

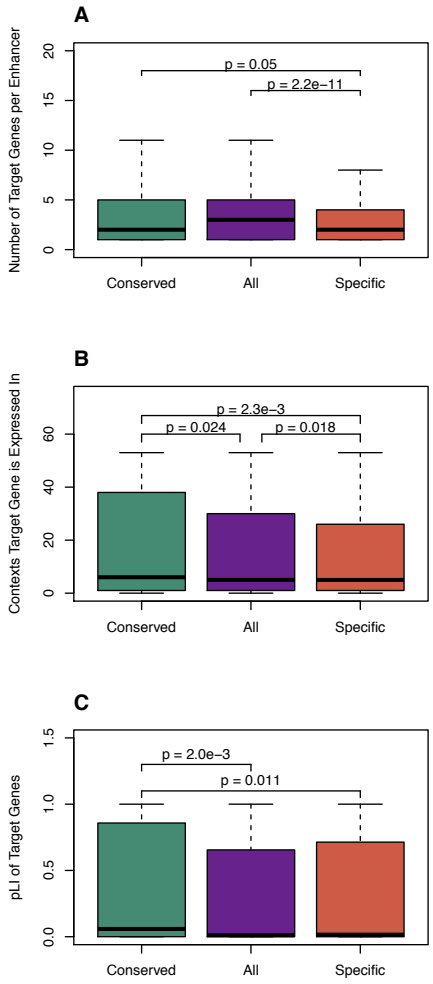

Figure II-9. Conserved-activity enhancers have more target genes, which are expressed in more cellular contexts, than human-specific-activity enhancers. Using the GTEx enhancer–gene mappings (Methods), we observed that (A) conserved-activity enhancers had significantly more target genes than human-specific-activity enhancers, but a similar number as all human liver enhancers. Thus, conserved-activity enhancers appear to regulate more genes than human-specific-activity enhancers, but they do not regulate more genes than would be anticipated by virtue of being a human liver enhancer. (B) Genes targeted by conserved-activity enhancers are expressed in more cellular contexts than human-specific-activity or human liver enhancers overall. (C) The pLI score of target genes identified by GTEx for conserved-activity enhancers was significantly greater than that of target genes for all human liver enhancers, illustrating that their target genes are under more constraint. The Mann-Whitney *U* test was used for all comparisons.

II-4 Conclusion and Discussion

In this study, we demonstrated that liver enhancers with conserved activity across mammals have greater evidence for pleiotropy than similarly alignable sequences with only species-specific activity across three levels of regulatory function: TF binding potential, enhancer activity across tissues, and downstream gene targets. We first found that conserved-activity enhancers have both significantly more TF binding motifs and binding motifs for more distinct TFs, illustrating a greater potential for diverse regulatory activity. We then demonstrated that this increased potential is realized: Conserved-activity liver enhancers are active enhancers in significantly more cellular contexts than species-specific-activity liver enhancers. Furthermore, these differences in activity are also apparent in the attributes of their gene targets; conserved-activity enhancers have more gene targets, and their targets are both more broadly expressed and under greater levels of constraint than species-specific-activity enhancers. These overall differences are sufficiently large that we could accurately classify conserved-activity and human-specific-activity enhancers in a machine learning framework.

Several previous studies have suggested that pleiotropy may play a role in the conservation of regulatory activity across species, but the relationship between pleiotropy and regulatory conservation has not been comprehensively evaluated. For example, the conservation of TF binding at orthologous sequences was positively correlated with the number of cellular contexts in which the sequence had an open chromatin conformation (Cheng et al. 2014). Similarly, an enhancer's breadth of activity across cellular contexts is positively correlated with the predicted deleteriousness of variants within the enhancer sequence, suggesting that breadth of enhancer activity across contexts is associated with stronger purifying selection (Huang et al. 2017). Our results significantly expand these previous findings beyond the breadth of enhancer activity to other dimensions of regulatory activity, including TF binding density and diversity and gene targets. Additionally, we demonstrate that these trends generalize across mammalian species. Thus, our results provide consistent evidence that enhancers with conserved activity are more pleiotropic than other enhancers.

Given the fast turnover of liver enhancers relative to species divergence (Villar et al. 2015b), we anticipate that the majority of species-specific enhancers are young, rather than being remnants of ancestral enhancer elements lost in other lineages. Newly created enhancers likely vary in their regulatory potential. For example, an enhancer that first gains activity in a genomic

region that is accessible in many cellular contexts or by gaining a binding site for a broadly expressed TF is likely to have greater pleiotropic potential than an enhancer that arises in a more context-specific region. Over time, the first enhancer would have an easier path to expanding its regulatory role, and thus its constraint. However, constrained activity in one context could also promote pleiotropy by providing a stable functional substrate for developing regulatory activity in additional contexts. Furthermore, enhancers are a diverse and heterogeneous assortment of DNA elements, and other factors likely contribute to their evolutionary dynamics. For instance, our results on the density and diversity of TF binding sites suggest that the robustness of enhancer sequences to disruptive genetic variation may influence activity conservation. More work on the interactions of pleiotropy, activity, and constraint is needed to shed light on the development and evolution of regulatory sequences. Comprehensive mapping of enhancer activity across multiple tissues and species will help resolve these questions.

Several technical limitations may impact the interpretation of our results. Genome-wide profiles of histone modifications have a limited resolution to identify the boundaries of enhancer elements (Shlyueva et al. 2014). As a consequence, it is possible that separate enhancers in close proximity to one another might not be distinguished as separate elements; if multiple enhancers were merged together, this could result in apparent signatures of pleiotropy. However, we demonstrate pleiotropy for these genomic regions at the finest resolution achievable using current, high-throughput techniques. Second, we focused on deeply alignable sequences with extreme differences in activity conservation—those active in all species versus those active only in one. We focused on these extremes to increase our likelihood of detecting differences and to identify patterns that hold across mammals. However, it is possible this may have obfuscated lineage-specific (e.g., primate-specific) patterns underlying conservation of regulatory activity in some clades. Third, mapping enhancers to target genes is a challenging problem, and our current knowledge of gene targets is incomplete. Many enhancers do not have any predicted target genes identified and those that do are likely to include false positives. To account for this uncertainty, we considered two independent mapping strategies and found consistently more targets for the conserved-activity enhancers compared with those with species-specific-activity. Despite these caveats, our findings are consistent across multiple methods of defining TF binding motifs, breadth of enhancer activity, and downstream gene targets.

Finally, the identification of enhancers is an imperfect process, and no genome-wide identification strategy is completely accurate. Histone modification profiles, in particular H3K27ac without H3K4me3, are strongly correlated with enhancer activity in reporter assays, and their use has enabled fundamental studies of enhancer activity genome-wide (Creyghton et al. 2010; Nord et al. 2013; Villar et al. 2015b). Nonetheless, there is the possibility of both false positives and false negatives using this approach. False negatives could potentially result in some sequences with true enhancer activity in multiple species being considered species-specific-activity enhancers. Their inclusion would be unlikely to create spurious results as they would likely diminish differences between the species-specific and conserved-activity-enhancer categories. In contrast, false positives are more concerning as they could include nonenhancers in the species-specific-activity enhancer category. However, we demonstrate that all human enhancers, regardless of conservation, demonstrate reduced pleiotropy relative to conserved-activity enhancers, which suggests this finding is not a product of false positives within the species-specific-activity enhancer category. Furthermore, observation of the histone modification signature was required in two biological replicates to define an enhancer, decreasing the risk for false positives. Thus, while the use of histone modifications to identify putative enhancers has caveats, the difference in pleiotropy between enhancer categories is unlikely to be a product of false positives or negatives.

Overall, our work argues that pleiotropy influences the conservation of enhancer activity of noncoding sequences across mammalian evolution. The functional diversity of regulatory sequences must be integrated into models of their evolution. In addition to improving our theoretical understanding of evolutionary constraint on regulatory regions, better understanding the evolutionary forces acting upon the genomic regulatory landscape will also have practical benefits. For example, we demonstrate that machine-learning classifiers can be trained to distinguish conserved-activity from species-specific-activity enhancers using features that reflect their pleiotropy. In the future, these classifiers could be adapted to predict which enhancers will generalize between species, prioritize new tissues for genome-wide assays, and estimate the effects of mutations on enhancer activity.

Chapter III

Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties

This chapter is a collaboration with Alex Fish. I was co-first author on the manuscript published in *PLoS Computational Biology* (Chen et al. 2018).

III-1 Introduction

Enhancers are genomic regions distal to promoters that bind transcription factors (TFs) to regulate the dynamic spatiotemporal patterns of gene expression required for proper differentiation and development of multi-cellular organisms (Shlyueva et al. 2014; Consortium et al. 2015). It is critical to understand the mechanisms underlying enhancer evolution and function, as alterations in their activity influence both speciation and disease (Maurano et al. 2012; Corradin and Scacheri 2014; Brazel and Vernimmen 2016). Recent genome-wide profiling of TF occupancy and histone modifications associated with enhancer activity revealed that the regulatory landscape changes dramatically between species—both enhancer activity and TF occupancy at orthologous regions distal to promoters are extremely variable across closely related mammals (Taher et al. 2011; Woo and Li 2012; Cotney et al. 2013; Hsu and Ovcharenko 2013; Villar et al. 2014b, 2015a; Reilly et al. 2015). However, the gene regulatory circuits (Stergachis et al. 2014) and expression of orthologous genes in similar tissues are largely conserved across mammals (Chan et al. 2009; Brawand et al. 2011; Merkin et al. 2012). Much of the gene regulatory machinery is also conserved; TFs and the short DNA motifs they bind are highly similar between human, mouse, and fly (Amoutzias et al. 2007; Wei et al. 2010a; Cheng et al. 2014; Nitta et al. 2015b). In short, there is considerable change in the enhancer activity of orthologous regions across mammals, despite the relative conservation of gene expression and TF binding preferences.

The rapid turnover in enhancer activity between orthologous regions in different species has largely been attributed to differences in the DNA sequences of the elements involved, rather than differences in the broader nuclear context (Wilson et al. 2008; Ritter et al. 2010; Schmidt et al. 2010; Li and Ovcharenko 2015; Prescott et al. 2015a). Genome-wide profiles of TF binding have shown that 60–85% of binding differences in human, mouse, and dog for the TFs CEBPB and

HNF4A can be explained by genetic variation that disrupts their binding motifs (Schmidt et al. 2010). Genetic differences are also often responsible for differential enhancer activity between more closely related species; for example, variation in TF motifs at orthologous enhancers was predictive of activity differences between human and chimp neural crest enhancers (Prescott et al. 2015a). This suggests that, while there is turnover at orthologous sequences, sequence properties predictive of enhancer activity may still be conserved.

Until recently, investigation of the conservation of enhancer sequence properties across mammalian evolution has been hampered by a lack of known enhancers across diverse species within the same cellular context. The canonical definition of enhancer activity is the ability to drive expression in transgenic reporter assays (Banerji et al. 1981a; Shlyueva et al. 2014), which cannot currently be scaled to assess regulatory potential genome-wide. However, high-throughput assays such as ChIP-seq can assess histone modifications associated with enhancer activity (Creyghton et al. 2010; Nord et al. 2013) to identify putative enhancers genome-wide in many tissues and species (Cotney et al. 2012; Villar et al. 2015a). Using known enhancers, machine learning approaches have learned their sequence properties and successfully distinguished enhancers active in specific cellular contexts from both the genomic background and enhancers active in other tissues (Lee et al. 2011, 2015; Burzynski et al. 2012; Taher et al. 2012; Erwin et al. 2014; Ghandi et al. 2014; Quang and Xie 2015; Zhou and Troyanskaya 2015; Min et al. 2017; Yang et al. 2017). Moreover, some of these studies suggested the potential for cross-species enhancer prediction. For instance, the similarity of co-occurrence of sequence patterns can be used to identify orthologous enhancers in distantly related Drosophila species (Arunachalam et al. 2010), and annotated cis-regulatory modules (CRMs) in Drosophila can predict CRMs in highly diverged insect species based on binding site composition similarity (Kazemian et al. 2014). However, TF binding sites have been suggested to evolve and turnover much more rapidly between closely related mammals than Drosophila species (Stefflova et al. 2013; Villar et al. 2014b). Nonetheless, a comprehensive analysis across clades suggests that transcriptional networks and gene regulatory sequences evolve at similar rates across animals (Carvunis et al. 2015). Indeed, in mammals, a machine learning model trained on mouse enhancers accurately predicted orthologous regions of the human genome (Lee et al. 2011). However, due to the rapid turnover of enhancer activity between human and mouse, the majority of orthologous regions are not active human enhancers (Villar et al. 2015a). Overall, these previous studies suggest the

potential for evolutionary conservation of sequence properties of mammalian enhancers, but comprehensive genome-wide quantification of the degree and dynamics of this conservation is needed.

In this study, we investigate the degree of regulatory sequence property conservation by applying machine learning classifiers to genome-wide enhancer datasets across diverse mammals. We first confirm that SVM classifiers trained using short DNA sequence patterns can accurately identify many enhancers genome-wide in the adult liver, developing limb and developing brain. Then, by using classifiers trained in one species to predict enhancers in the others, we demonstrate that many enhancer sequence properties are conserved across species, even though the enhancer activity of specific loci is not. We establish the robustness of this conservation to different enhancer identification techniques by showing that classifiers trained using high-confidence human and mouse enhancer sequences validated in transgenic assays also generalize across species and are similar to classifiers trained on histone-modification-defined enhancers. Furthermore, the short DNA patterns most predictive of enhancer activity in each species matched a common set of binding motifs for TFs enriched for expression in relevant tissues. This suggests the patterns learned by classifiers capture biologically relevant sequences that influence TF binding. In addition to SVM classifiers, we also trained CNNs on liver enhancers in each species. The multilayer structures of CNNs are promising for modeling more complex sequence patterns beyond short DNA motifs (Alipanahi et al. 2015; Quang and Xie 2015, 2019; Zhou and Troyanskaya 2015; Kelley et al. 2016; Min et al. 2017; Yang et al. 2017). The CNNs predicted enhancers with higher accuracy than SVM models, but the CNNs generalized less well across species, suggesting less conservation of some patterns they learned. Together, our results argue that, though there is rapid change of active gene regulatory sequences between mammalian species, many of the short sequence patterns encoding enhancer regulatory activity have been conserved over 180 million years of mammalian evolution. Our findings also suggest avenues for identifying enhancers in species without genome-wide enhancer-associated histone modification data and establish a framework for future exploration of the conservation and divergence of regulatory sequence properties between species.

III-2 Materials and Methods

III-2.1 Genomic data

All work presented in this paper is based on hg19, rheMac2, mm10 (mouse liver dataset), mm9 (mouse limb and brain dataset), bosTau6, canFam3 and monDom5 DNA sequence data from the UCSC Genome Browser. For consistency with the original studies, liver gene annotations are from Ensembl v73, limb and brain gene annotations are from Ensembl v67 (Flicek et al. 2014). The sequence divergence between each pair of species was computed from the tree model built from fourfold degenerate sites in the 100-way multiple species alignment from UCSC Genome Browser (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons100way/hg19.100way.phastCons.mod).

III-2.2 Enhancer and genomic background datasets

We evaluated the ability of machine learning models to distinguish different sets of enhancers (positives) from sets of matched regions from the genomic background (negatives). In this section, we describe the collection and processing of the enhancer and genomic background sets. In the next section, we describe the training and evaluation of the SVM classifiers.

We analyzed three multi-species histone-modification-defined enhancer datasets in this study. The first consisted of liver enhancers identified by genome-wide ChIP-seq profiling of histone modifications (H3K27ac without H3K4me3) in 20 species from five mammalian orders (Villar et al. 2015a). These regions are almost entirely distal to coding regions (i.e., more than 1kb away from the nearest TSS) (Villar et al. 2015c). We use the definition of "high-quality genomes" from Villar et al. 2015 (Villar et al. 2015b). We selected a member of each order with a high-quality genome build for analysis when possible; however, the most diverged order—marsupials—did not have a species with a high-quality genome build. We consequently selected opossum, as it was the most diverged from humans. For all analyses, we did not consider enhancers or random regions that fell in genome assembly gaps (UCSC gap track) when generating negatives. For human and mouse, we also excluded the ENCODE blacklist regions (Bernstein et al. 2012) (https://sites.google.com/site/anshulkundaje/projects/blacklists). This resulted in the following number of observed enhancers in each species: human (N=29152), macaque (N=22911), mouse (N=18517), cow (N=30892), dog (N=18966), and opossum (N=23160) (Villar et al. 2015a). A small fraction of liver enhancers overlapped with one another (3.0% in human, 2.0% in macaque, 3.0% in mouse, 6.2% in cow and 1.6% in dog), and the

overlaps were mostly under 10% of the enhancers' lengths. Thus, these overlaps are unlikely to cause overfitting during within-species cross-validation runs. We also performed cross-species analyses both with and without orthologous sequences.

We generated four different sets of matched genomic background regions for use as negatives in the training and evaluation of the liver classifiers for each of the six species. The first are random genomic regions matched on length and chromosome to the observed enhancers. Second, for the GC-controlled analyses, we generated genomic background regions matched to the enhancers on length, chromosome, and GC-content. Third, for the repeat controlled analysis, we obtained repetitive elements identified by RepeatMasker for each species (Smit et al. 2013) and generated random regions from the genomic background matched on length, chromosome, GC-content, and proportion overlap with repetitive elements. Finally, we generated negatives using flanking regions of enhancers. We define the flanking region of an enhancer is 10 times of its length on either side. We then randomly select 10 negative regions of same length as the enhancer that do not overlap other enhancers from the candidate flanking regions. To reflect the fact that enhancers make up a small portion of the genome, we chose an imbalanced data design with 10 times as many of the genomic background (negative) regions as there were enhancers.

The second enhancer dataset contained human (N=25304), macaque (N=88560), and mouse (N=87406) enhancers identified from profiling the H3K27ac modification in developing limb tissue (Cotney et al. 2013). The third enhancer dataset contained human (N=48853), macaque (N=57446), and mouse (N=51888) enhancers identified from profiling the H3K27ac modification in developing brain tissue (Reilly et al. 2015). For limb and brain enhancers, we excluded regions within 1 kb of a transcription start site. For each species, we combined the enhancer regions from different development stages. The genomic background regions for each species were defined following the same procedure as for the liver enhancers.

To determine how well classifiers generalized across additional tissue types, we used human enhancers identified by the Roadmap Epigenomics Project (Consortium et al. 2015) in nine tissues from diverse body systems: liver (GI, E066), hippocampus middle (brain, E071), pancreas (exocrine-endocrine, E098), gastric (GI, E094), left ventricle (heart, E095), lung (E096), ovary (reproductive, E097), bone marrow derived mesenchymal stem cell cultured cells (stromal-connective, E026) and CD14 primary cells (white blood, E029). We defined enhancers in these tissues as H3K27ac without H3K4me3 regions. For each tissue, we generated not-GC-

controlled and GC-controlled negative training examples as described for the liver enhancers above.

In addition to the histone-modification-defined enhancers, we also analyzed enhancers validated in transgenic reporter assays in embryonic day 11.5 mouse embryos from VISTA (Visel et al. 2007). We investigated all six tissues with at least 50 positive enhancer elements in both species: forebrain, midbrain, hindbrain, limb, heart and branchial arch. These enhancers comprised the positive training examples. For each positive, we generated 10 length and chromosome matched random genomic regions as negative training examples. There are not enough failed reporter assays across all selected tissues to generate ten sets of negatives, and there are biases in how the human and mouse regions were selected for testing in VISTA. Thus, we did not use classifiers trained on regions with failed reporter assays as negatives for cross-species analyses.

To demonstrate the histone-defined enhancer classifier can predict VISTA enhancers, we removed the regions of VISTA limb enhancers that overlap Cotney et al. 2013 limb H3K27ac regions from the VISTA set and the regions of limb H3K27ac regions that overlap VISTA from the H3K27ac set to ensure no overlapping regions between training and testing. There are 96 human VISTA limb enhancers left and 32 mouse VISTA limb enhancers. Because of the small number of mouse enhancers, we only applied the human limb H3K27ac classifier to predict the human VISTA limb enhancers.

III-2.3 Spectrum kernel SVM classification

An SVM is a discriminative classifier that learns a hyperplane to separate the positive and negative training data in feature space. We used the *k*-mer spectrum kernel to quantify sequence features for the SVM (Leslie et al. 2002). Training, classification, evaluation, and the computation of features weights were performed with the kebabs R package (v1.4.1) (Palme et al. 2015). We used the default kernel normalization to the unit sphere, considered reverse complements separately, used the cosine similarity. We initially performed a grid search with k=5 and C in the range of 1, 15, 50, 100, 1000 using the human liver enhancer dataset. We found that the performance of the SVMs in cross-validation is robust in C=1,15 with cross validation errors of 0.2610, 0.2608 and ROC AUCs of 0.8213, 0.8209. We chose C=15 for training SVMs in human and other species. The good performance in cross-validation runs suggest the SVMs are well regularized with C=15 and any slight over-estimation of performance would result in an underestimation of cross-species

generalization. Due to the imbalanced training dataset, we set class weights of 10 for the positives and 1 for the negatives to increase the penalty on misclassification of positives. We report all analyses with $k = 5$, but classifier performance and generalization were similar for $k = 4$–7 (0.81, 0.82, 0.82, 0.82, respectively for liver).

To evaluate classifier performance within-species, we performed ten-fold cross validation. In other words, for each set of positives and negatives, the entire data set was randomly partitioned into ten independent sets that maintained the ratio of positives and negatives. Positives and negatives from nine of the ten sets were then used to train the classifier, the trained classifier was then applied to the remaining partition, and these predictions were used to evaluate the classifier. This process was performed ten times, testing each partition once. To summarize performance, we averaged the auROC and auPR over the ten runs. For cross-species classification, we trained on the whole dataset in the training species and evaluated the performance on the test species.

We also evaluated more flexible models, such as the mismatch (Leslie et al. 2002; Palme et al. 2015) and gappy pair kernels (Mahrenholz et al. 2011; Palme et al. 2015), These *k*-mer-based prediction models are similar to the spectrum kernel, but the mismatch kernel allows a maximum mismatch of *m* nucleotides in the *k*-mer and the gappy pair kernel considers pairs of *k*-mers with maximum gap of length *m* between them. For comparison, we trained the gappy pair kernel with $k = 2$, $m = 1$ and mismatch kernel with $k = 5$, $m = 1$ to compare with the 5-mer spectrum kernel. The mismatch and gappy pair kernels did not significantly increase the performance (auROCs of 0.82 and 0.82, respectively for liver) and are less interpretable than the *k*-mer spectra. It is possible that other parameter settings could yield slightly improved performance, but the resulting models would be more difficult to interpret, and optimizing performance was not the goal of our study.

III-2.4 Transcription factor motif analysis

5-mers were matched to known TF binding motifs in the JASPAR 2014 Core vertebrate database (Mathelier et al. 2014) using the TOMTOM package with default parameters (Gupta et al. 2007). The sharing of 5-mers and TFs across species was visualized using UpSetR (Conway et al. 2017).

III-2.5 Transcription factor expression data

For the human TF expression analysis, we obtained RNA-seq data for TFs across 12 tissues from the Gene Expression Atlas (https://expressionatlas.org/hg19/adult/). Genes with non-zero FPKM (Fragments Per Kilobase of transcript per Million mapped reads) in a tissue were considered as expressed. For all the other species, we obtained the expression of TFs from Berthelot et al. 2017

(Berthelot et al. 2017). The mouse TF expression in Berthelot et al. 2017 was first reported in Rudolph et al. 2016 (Rudolph et al. 2016), so we obtained the mouse gene expression from in Rudolph et al. 2016.

III-2.6 Convolutional neural network (CNN) classifier training and interpretation

Because of the fixed-length input of CNNs and the challenges of training CNNs using unbalanced datasets, we used the center 3000 bp (approximately the median length) of liver enhancers in six selected species as the positive training sequences and the same number of length matched random genomic regions in the corresponding species as negative training sequences During data preparation, we partitioned the data into training (80%), validation (10%), and hold-out testing sets (10%).

A typical convolutional neural network consists of convolutional layers, max-pooling layers, fully connected layers, and an output layer. To determine the CNN structure, we defined a hyperparameter space, including a range of learning rates (0.0001, 0.0005, 0.001), number of convolutional layers (3 to 5), number of neurons in each layer (32, 64, 128, 256, 512) of the window size of the filters (4, 8, 16), the window size of pooling (0, 4), and the regularization strength (dropout fraction 0–1). We trained 100 CNN models on human liver enhancers with the training dataset and selected the structure of CNN (Figure III-1) based on the smallest loss on the validation set using keras 2.0.8 (Chollet and others 2015) with hyperparameters suggested by the Tree-structured Parzen Estimator (TPE) approach implemented in the hyperopt (Bergstra et al. 2013) library. Then, we trained the enhancer CNN model with the best human CNN structure in the other five species, but different regularization strengths, 30 times in order to find the best performing CNN model for each species based on the loss of validation set. The performance of within-species prediction is reported based on the auROC of predicting the hold-out testing set of the training species and the performance of cross-species prediction is reported based on the auROC of predicting all data in the testing species. To prevent the model overfitting the training data, we used an early stopping strategy during the training, together with dropout layers, and data partitioning. More specifically, we monitored the loss on the validation set and stopped the training process if the validation loss ceased decreasing.

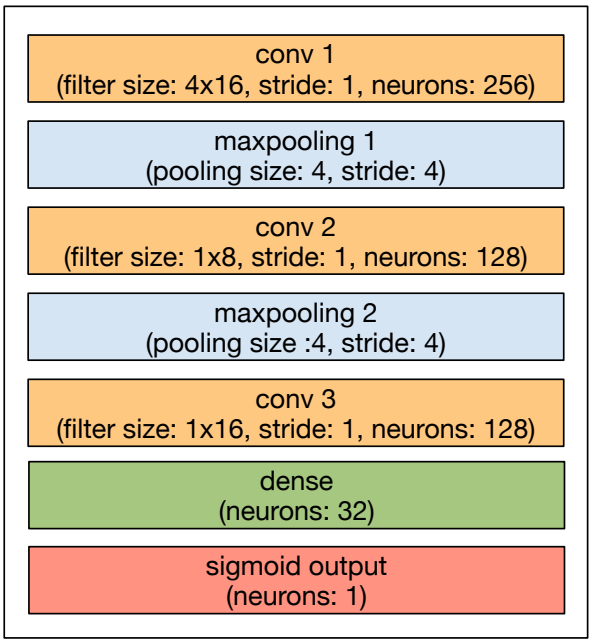

Figure III-1. The convolutional neural network (CNN) structure for training CNN classifiers of liver enhancers.

To interpret the first layer of the human liver CNN, we forward propagated sequences in the human liver validation dataset through the CNN and selected the sequence patches that maximally activate each neuron (> 0.5 maximum activation value of the neuron) in the first layer. Then, we converted the resulting sets of sequence patches to position weight matrices (PWMs) and mapped the PWMs to human TF motifs from the HOCOMOCO v11 (Kulakovskiy et al. 2016) database using TOMTOM with default parameters (Gupta et al. 2007).

III-2.7 Comparison of CNNs to k-mer SVM, polynomial kernel SVM, and gkm-SVM models

For comparison to the performance of CNNs, we trained gkm-SVM (Gupta et al. 2007), polynomial kernel SVM and a 5-mer spectrum kernel SVM on the same balanced dataset as the CNNs. For gkm-SVM, we split the training data into 90% training set and 10% testing set. Then we trained gkm-SVMs with default parameters (wgkm kernel, l=11, k=7, d=3) for 2 different Cs (0.1, 1). With C of 0.1, the training of gkm-SVM took 15.5 hours on a machine with a 2.4 GHz Intel Xeon CPU E5-2630 v3, 8 cores, and 2 CPUs; with C of 1, the training took 2 days and 13.5 hours. We report the performance of gkm-SVM on prediction of the testing set. For the polynomial kernel SVM, we split the training data into 90% training set and 10% testing set for each species. Then, we trained 5-mer 2nd degree polynomial kernel SVMs on the *k*-mer spectrum of the training sequences in human. We selected C of 0.001 and performed the training of the polynomial kernel

SVMs for every species. We report the performance of the polynomial kernel SVMs on the prediction of testing set. For the 5-mer spectrum SVMs, the performance of within-species prediction is reported based on the average auROC of ten-fold cross validation and the performance of cross-species prediction is reported based on the auROC of predicting all data in the testing species. The better performance of CNNs compared to the SVMs is not driven by differences in the testing set. When using the exact same training and testing set of the human liver enhancer dataset, the 5-mer SVM achieved ROC AUC of 0.782 and PR AUC of 0.756, which are very similar to the average performance over cross-validation folds: ROC AUC of 0.783, PR AUC of 0.761. Similarly, the gkm-SVM achieved ROC AUC of 0.767 and PR AUC of 0.749, which are similar to the reported performance of gkm-SVM in cross validation: ROC AUC of 0.763 and PR AUC of 0.745.

III-3 Results

III-3.1 Enhancers can be predicted from short DNA sequence patterns in mammals

Genome-wide enhancer activity across many mammalian species has been assayed via ChIP-seq profiling of enhancer-associated histone modifications in the adult liver (Villar et al. 2015a), developing limb (Cotney et al. 2013) and developing brain (Reilly et al. 2015). Certain chemical modifications to histones, such as acetylation of lysine 27 of histone H3 (H3K27ac) and lack of trimethylation of lysine 4 of H3 (H3K4me3), are associated with active enhancers and provide a genome-wide proxy for the active enhancer landscape (Creyghton et al. 2010; Nord et al. 2013). For brevity, we refer to genomic regions with enhancer-associated histone modification combinations identified in these previous studies as “enhancers.”

For this study, we selected six representative diverse mammals with cross-species enhancer data and high-quality genome builds: human, macaque, mouse, cow, dog, and opossum (Methods). Liver enhancers were available for all species; developing limb and brain enhancers were available for human, macaque, and mouse. For each species and tissue, we evaluated how well short DNA sequence patterns identified enhancers. We trained two machine learning algorithms, k-mer SVMs and CNNs, on raw DNA sequence patterns. This approach has the advantage that it is not dependent on previous knowledge of TF motifs. For the k-mer SVMs, we quantified DNA sequence patterns present in each genomic region by computing its k-mer spectrum—the observed frequencies of all possible nucleotide substrings of length k. We then

trained SVM classifiers on the k-mer spectra to distinguish enhancers from random genomic regions matched to the enhancers on various attributes, such as length, GC-content, and repeat-content, as appropriate. To reflect the fact that most of the genome does not have enhancer activity, we trained and evaluated the SVM classifiers on positive and negative sets containing ten times as many negative non-enhancer regions as enhancers and weighted misclassification costs. We used ten-fold cross validation to evaluate the classifiers, and we quantified performance by computing the average area under receiver operating characteristic (auROC) and precision-recall (auPR) curves over the ten cross-validation folds (Figure III-2; Methods). We also trained CNN models for this problem. Due to the challenges of training CNNs, these analyses were performed on balanced training, validation, and testing sets. For all comparisons with SVMs, we compared CNN performance to both the average SVM performance over cross-validation folds and the performance of the SVM on the single CNN test set. (See CNN results and Methods for details.) To document the training setup and performance, we assigned each prediction task an experiment number (Appendix A). We report the experiment number for results throughout the paper for clarity.

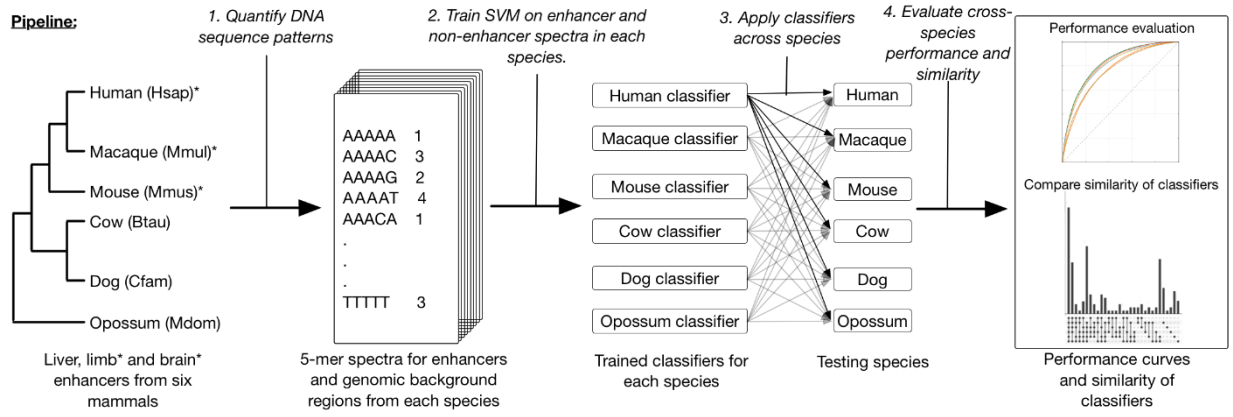

Figure III-2. Overview of the framework for evaluating DNA patterns predictive of enhancer activity across diverse mammals. Starting with liver, limb and brain enhancers and genomic background regions from six mammals, the first step of the pipeline quantified each of these genomic regions by their 5-mer spectrum—the frequency of occurrence of all possible length five DNA sequence patterns. Using the spectra as features, we trained a spectrum kernel support vector machine (SVM) to distinguish enhancers from non-enhancers in each species and evaluated their performance with ten-fold cross validation. Then, we applied classifiers trained on one species to predict enhancer activity in all other species. Finally, we evaluated the performance of cross-species prediction compared to within species prediction and quantified the similarity of different species' classifiers by the sharing of TF motifs among the most predictive 5-mers. Limb and brain enhancer data were only available for human, macaque, and mouse.

We first evaluated the ability of SVM classifiers trained on 5-mer spectra to identify liver enhancers in the six selected mammals: human, macaque, mouse, cow, dog and opossum (experiments 1, 8, 15, 22, 29, 36). As expected from previous work (Lee et al. 2011; Burzynski et al. 2012; Gorkin et al. 2012), all classifiers could distinguish active liver enhancers from length-matched random background regions; auROCs ranged from 0.78 in dog to 0.84 in mouse (Figure III-3a; auPRs ranged from 0.27 to 0.35, Figure III-4a). Next, we trained 5-mer spectrum SVM classifiers to predict enhancers active in limb (experiments 147, 151, 155) and brain (experiment 165, 169, 173) for human, macaque, and mouse. Again, classifiers accurately distinguished enhancers from the background with even stronger performance than the liver classifiers. The limb classifiers achieved auROCs of ~0.89 in each species (Figure III-3b; auPRs from 0.43 to 0.46, Figure III-4b), and the brain classifiers had auROCs from 0.90–0.93 (Figure III-3c; auPRs from 0.54 to 0.56, Figure III-4c). However, we note that the auPRs are lower than auROCs due to the unbalanced training set.

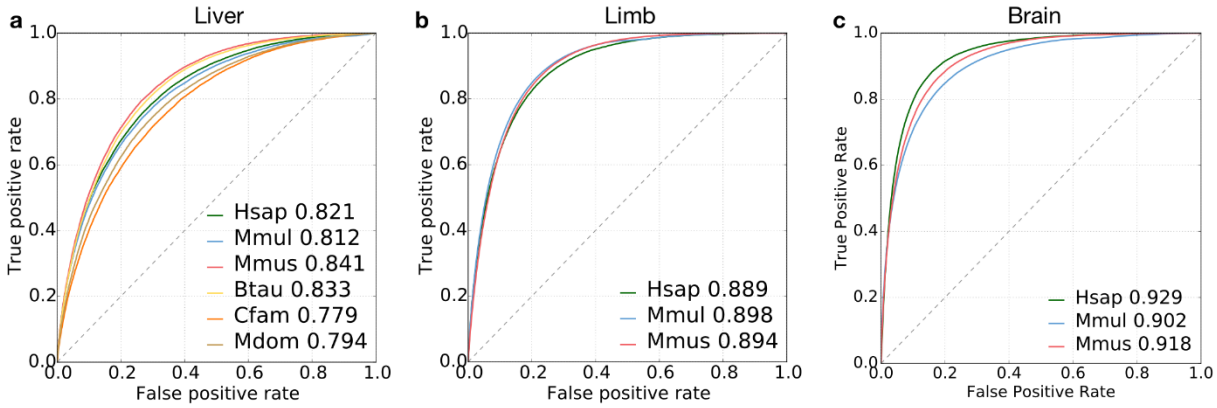

Figure III-3. Performance of DNA sequence-based enhancer identification in diverse mammals.ROC curves for classification of liver enhancers vs. the genomic background in six diverse mammals: human (Hsap), macaque (Mmul), mouse (Mmus), cow (Btau), dog (Cfam), and opossum (Mdom). (b) ROC curves for classification of developing limb enhancers in human, macaque, and mouse. (c) ROC curves for classification of developing brain enhancers in human, macaque, and mouse. Area under the curve (AUC) values are given after the species name. Ten-fold cross validation was used to generate all ROC and PR curves (Figure III-4a-c).

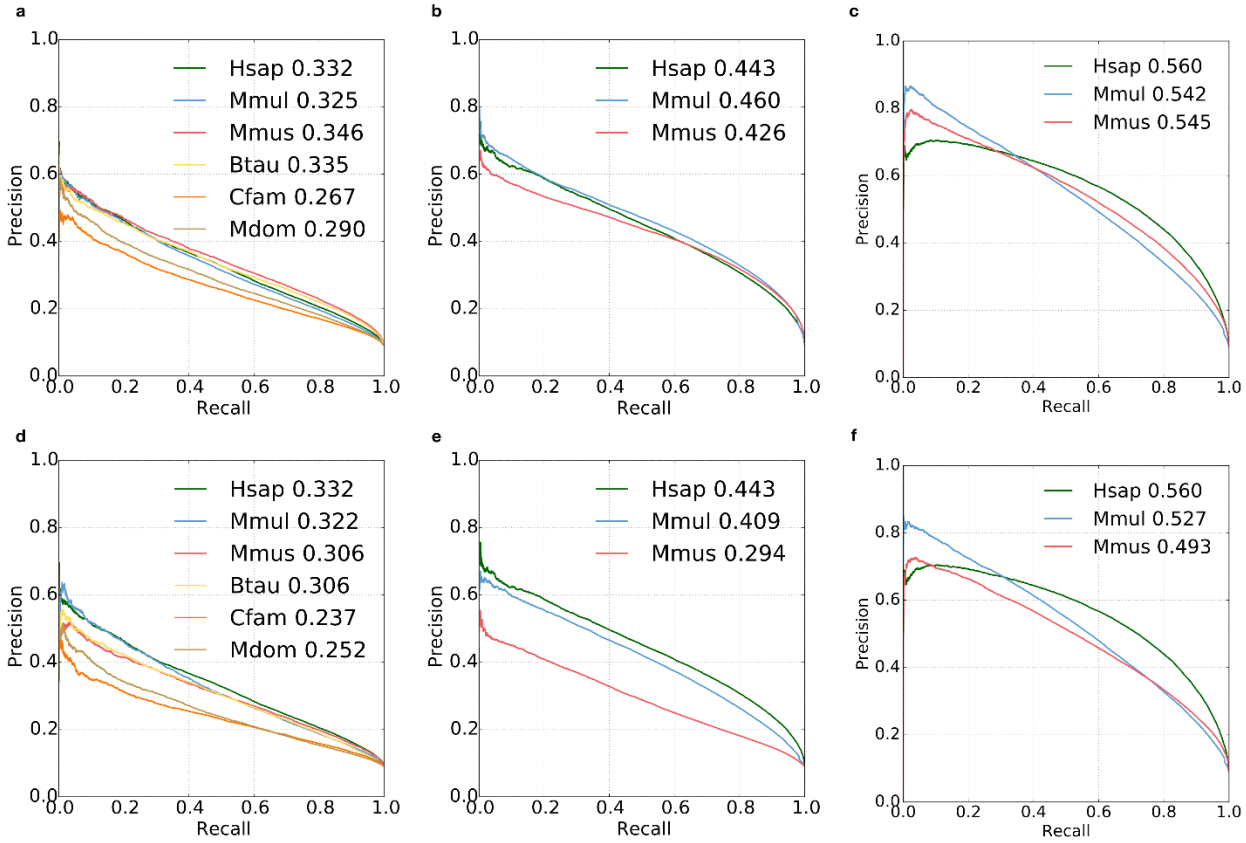

Figure III-4. Precision-recall (PR) curves for the classification of enhancers vs. the genomic background (non-GC-controlled). (a) Classification of liver enhancers in six diverse mammals: human (Hsap, experiment 1), macaque (Mmul, experiment 8), mouse (Mmus, experiment 15), cow (Btau, experiment 22), dog (Cfam, experiment 29), and opossum (Mdom, experiment 36). (b) Classification of developing limb enhancers in human (experiment 147), macaque (experiment 151), and mouse (experiment 155). (c) Classification of developing brain enhancers in human (experiment 165), macaque (experiment 169), and mouse (experiment 173). (d) Generalization of the human-trained liver enhancer classifier to the other five mammals (experiment 1s-6). The cross-validation PR curve for a classifier trained and tested on human is included for reference. (e) Generalization of the human-trained limb enhancer classifier to macaque and mouse (experiment 147-149). (f) Generalization of the human-trained brain enhancer classifier to macaque and mouse (experiment 165-167). AUC values are given after the species name. The cross-validation PR curve for a classifier trained and tested on human is included for reference.

The choice of k did not substantially influence performance; the auROCs for human liver classifiers are 0.81, 0.82, 0.82, 0.82, respectively across k of 4, 5, 6, and 7. We also explored the application of classifiers based on more flexible k-mer features, i.e., the gappy and mismatch k-mer kernels (experiments 145, 146) (Palme et al. 2015), but they did not improve performance (auROCs of 0.82 and 0.82). The gkm-SVM approach also performed similarly to the k-mer SVM

on the liver enhancers (auROC 0.76); because of the long computation time of gkm-SVM (experiment 347, Methods), we could only compare it on the balanced liver enhancer set (5-mer SVM auROC of 0.78). These results illustrate that SVMs trained only on DNA sequence patterns can distinguish many enhancers from background sequences across a variety of mammals for three tissues and two developmental time-points.

III-3.2 Short sequence properties predictive of enhancers are conserved across species

We then investigated whether learned DNA sequence patterns predictive of enhancer activity were conserved across mammals by testing whether classifiers trained in one species could distinguish enhancers from the genomic background in another species. First, we applied the human liver classifier to the five other species (experiments 2–6). We quantified cross-species performance using the relative AUCs—the auROC or auPR of the enhancer classifier trained on species A and applied to species B, divided by the average auROC or auPR over cross-validation folds obtained by the classifier trained and tested on species B. In other words, the relative auROC is the proportion of within-species performance achieved by a classifier trained in a different species. The classifier trained on human liver enhancers predicted liver enhancers in other mammals nearly as accurately as classifiers trained in each species (Figure III-5a, PR curves in Figure III-4d), and its relative performance decreased only slightly across species (Figure III-5b, relative auROCs > 95.5%, relative auPRs > 87%). Furthermore, the scores from the human classifier applied to human enhancers were significantly positively correlated with the scores from non-human classifiers (Figure III-6; Spearman's ρ between 0.90 for macaque and 0.66 for opossum). When expanded to all pair-wise combinations of species (experiments 1–36), classifiers accurately predicted enhancers in every mammalian species tested, regardless of the specific species they were trained in; the average relative auROC was 96.0% (Figure III-5b; average relative auPR was 85%, raw AUCs in Appendix A). The human classifier was generally the best at cross-species prediction; this is likely due to the higher genome assembly quality and other biases towards human sequences.

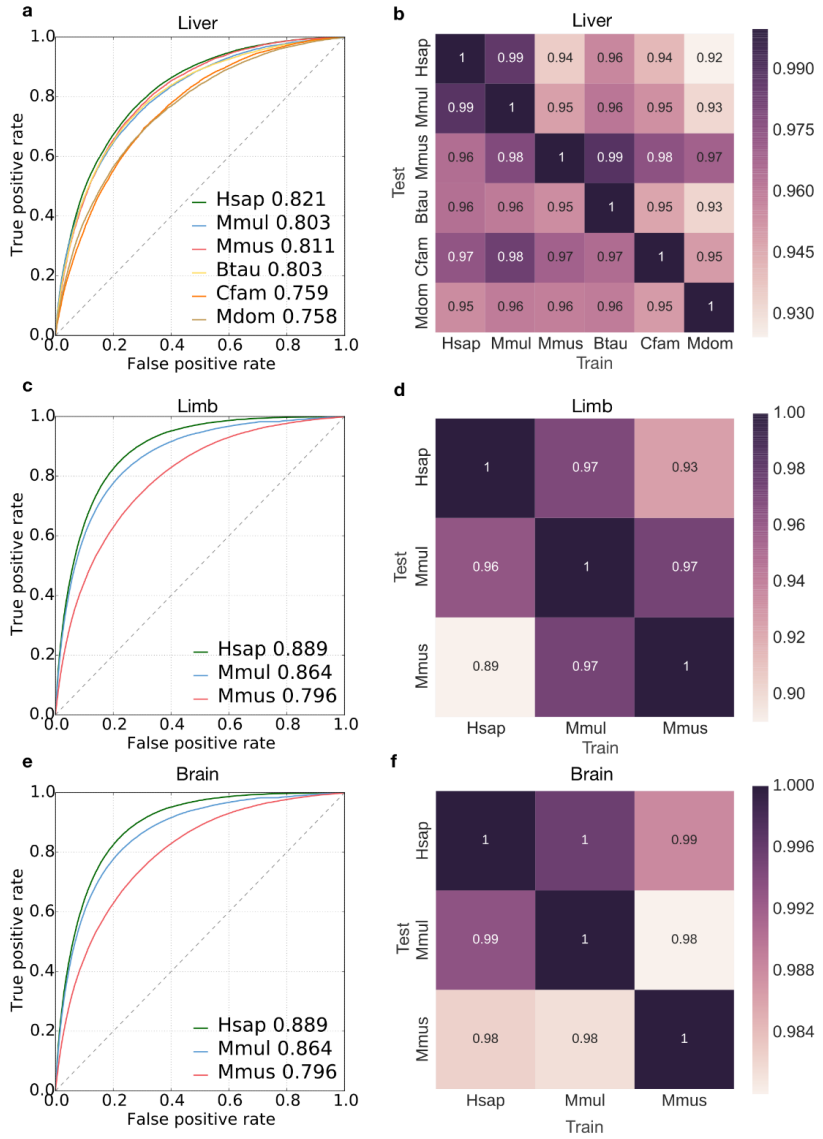

Figure III-5. Human-trained enhancer classifiers accurately predicted liver, limb and brain enhancers in diverse mammals. ROC curves of the performance of the human liver enhancer classifier applied to the human (Hsap), macaque (Mmul), mouse (Mmus), cow (Btau), dog (Cfam) and opossum (Mdom) datasets. Area under the curve (auROC) values are given after the species name. (b) Heat map showing the relative auROC of liver enhancer classifiers applied across species compared to the performance of classifiers trained and evaluated on the same species (Figure III-3a). The classifiers were trained on the species listed on the x-axis and tested on species on the y-axis. (c) ROC curves showing the performance of the human limb enhancer classifier on human, macaque and mouse. (d) Heat map showing the relative auROC of limb enhancer classifiers applied across species compared to the performance of classifiers trained and evaluated on the same species (Fig III-3b). (e) ROC curves showing the performance of the human brain enhancer classifier on human, macaque and mouse. (f) Heat map showing the relative auROC of brain enhancer classifiers applied across species compared to the performance of classifiers trained and evaluated on the same species (Fig III-3c). The raw auROC and auPR values for all comparisons are given in Appendix A.

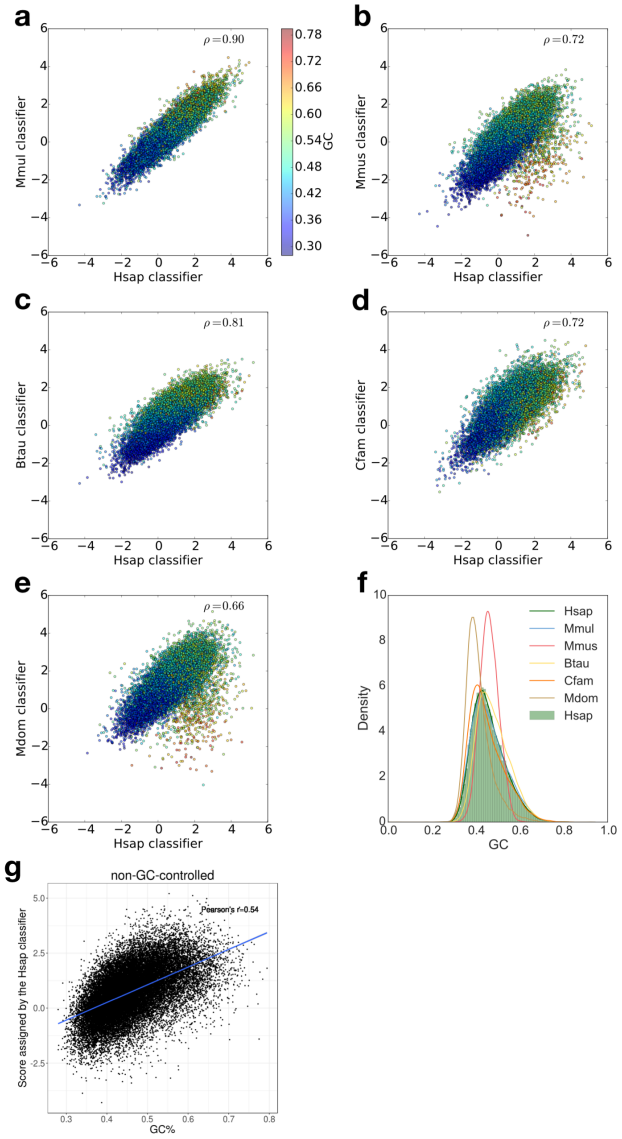

Figure III-6. The predictions of enhancer classifiers (not-GC-controlled) trained in different species were strongly correlated. Scatter plots showing the correlation between scores assigned to human enhancers by the human-trained classifier and the classifiers trained on other species: (a) Human (Hsap, experiment 1) vs. Macaque (Mmul, experiment 7). (b) Human vs. Mouse (Mmus, experiment 13) (c) Human vs. Cow (Btau, experiment 19) (d) Human vs. Dog (Cfam, experiment 25) (e) Human vs. Opossum (Mdom, experiment 31). Each dot represents a human liver enhancer sequence. The enhancer score assigned by the human-trained classifier is plotted on the x-axis, and the score assigned by the classifier trained on the other specified species is plotted on the y-axis. The color indicates the GC content. Correlation is quantified by Spearman's rank correlation coefficient (ρ). (f) The GC content distribution of liver enhancers in human, macaque, mouse, cow, dog, and opossum. Human, macaque, cow and dog enhancers have a similar GC distribution. Mouse and opossum have less variation in GC content and are depleted of high GC enhancers compared to the other species. (g) The GC content of human enhancers is positively correlated with the scores assigned by the human-trained classifier (Pearson's r=0.54, P<2.2e-16).

Classifiers generalized better to more closely related species; generalization was inversely correlated with the species' evolutionary divergence, as quantified by substitutions per neutrally evolving site (Figure III-7, Spearman's rho = –0.4, P = 0.14; Methods). This trend became even stronger when controlling for differences in GC content between species (Spearman's rho = –0.72, P < 2.2e–16; S15 Fig).

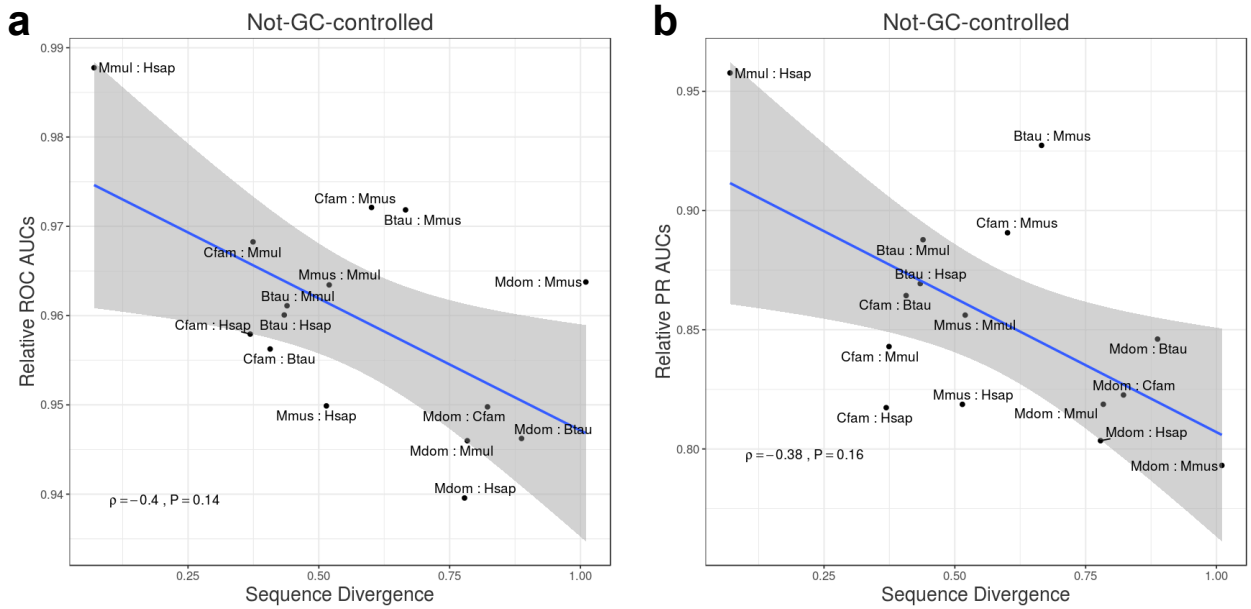

Figure III-7. Neutral sequence divergence is inversely correlated with the cross-species prediction accuracy. Correlation of relative auROCs from the non-GC-controlled classifiers (experiments 1–36) with sequence divergence. Spearman's rho is –0.4 (P = 0.14). (b) Correlation of relative auPRs from the non-GC-controlled classifiers (experiments 1–36) with sequence divergence. Spearman's rho is –0.38 (P = 0.16). Both correlations increased significantly when accounting for GC-content in the classifiers (Fig. S15). Sequence divergence is quantified as the expected number of substitutions per neutrally evolving site as derived from four-fold degenerate sites in codons in the UCSC Genome Browser's100-way multiple species alignments (Methods). To determine the relative auROC/PR for each pair of species, the mean was taken across the two classifiers when applied cross-species (i.e., the relative auROC/PR from the human classifier applied to mouse and the relative auROC/PR mouse classifier applied to human were averaged).

Classifiers trained to identify enhancers in developing limb and brain also accurately generalized across species. The average relative auROC for the developing limb classifiers was 95.0% across all species pairs (Fig III-5c-d; raw AUCs in Appendix A), and the average relative

auROC for the developing brain classifiers was 98.6% (Fig III-3e-f; raw AUCs in Appendix A). The ability of classifiers to generalize to other species illustrates the conservation of sequence properties predictive of enhancers across mammals.

To ensure that the small fraction of liver enhancers shared between pairs of species were not driving performance, we identified human liver enhancers that overlapped enhancers from three other mammalian species with genome-wide multiple sequence alignments (mouse: 13.6%; cow: 20.0%; dog: 16.7%) and vice versa. For each pair of species, the overlapping enhancers were removed from both the human training set and the other species' testing set, and then new human classifiers were trained and evaluated (experiments 183–188). The classifiers achieved relative auROCs of 0.962 (mouse), 0.957 (cow) and 0.968 (dog), very similar to the analyses that did not remove shared enhancers (mouse: 0.964, cow: 0.964, and dog: 0.974), suggesting that the shared enhancers do not drive the cross-species generalization.

III-3.3 Enhancers validated in transgenic assays show similar cross-species patterns

Genome-wide mapping of enhancer-associated histone modifications is a cost-effective means to identify putative enhancers; however, the presence of these modifications does not guarantee enhancer activity. Many experimental and computational approaches have been used to identify enhancers (Shlyueva et al. 2014; Kleftogiannis et al. 2016), and there is considerable disagreement among different strategies (Benton et al. 2017). To investigate the generality of our conclusions drawn from histone-modification-derived enhancers, we also analyzed enhancers validated *in vivo* via transgenic assays from the VISTA enhancer database. We included six tissues (limb, forebrain, midbrain, hindbrain, heart and branchial arch) with a sufficient number of validated enhancers (>=50) in human and mouse. Consistent with the results from classifiers trained on histone-modification defined enhancers, the classifiers trained and evaluated on VISTA human enhancers accurately predicted VISTA mouse enhancers in the corresponding tissue from genomic background, and vice versa (experiments 189–212; Figure III-8; average relative auROC = 96.3%, average auPR = 81.6%). This suggests that sequence patterns in enhancers confirmed via reporter assays are conserved between human and mouse. Moreover, the limb classifier trained on H3K27ac regions (excluding VISTA overlaps) accurately predicted VISTA enhancers (auROC = 0.82, auPR=0.35 in human; experiment 237; Methods) and was competitive with the VISTA-trained limb classifier itself (auROC = 0.80, auPR=0.39 in human; experiment 238). This suggests that sequence properties predictive of histone-modification defined enhancers are also predictive

of enhancers validated in transgenic assays. Thus, in spite of the limited number and biases present in the sequences tested for enhancer activity by VISTA, our models capture conserved sequence attributes of these functionally validated enhancers.

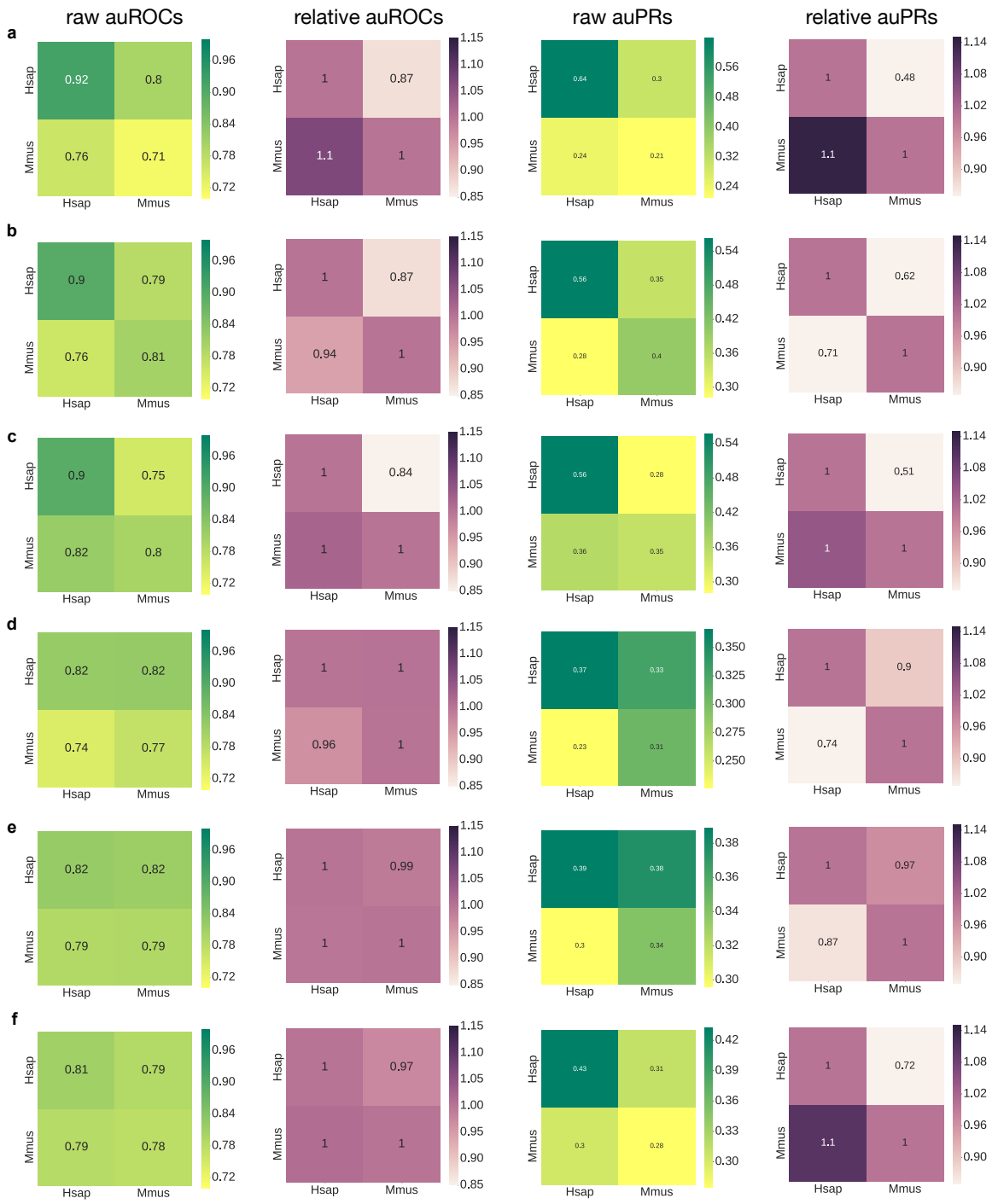

Figure III-8. Evaluation of between human (Hsap) and mouse (Mmus) VISTA enhancer classification tasks (non-GC-controlled, experiment 189-232). The number of enhancers in each tissue is indicated in brackets. (a) Forebrain enhancers (Human, 312; Mouse, 85) (b) Midbrain (Human, 259; Mouse 69) (c) Hindbrain (Human, 239; Mouse 58) (d) Heart (Human, 97; Mouse, 120) (e) Branchial arch (Human, 73; Mouse, 73). (f) Limb (Human 168; Mouse, 84). The human classifier usually generalized better than mouse classifiers. This may be due to the larger sample size of human enhancers in most of the tissues.

Overall, these results show that the DNA sequence profiles of enhancer sequences captured by species-specific 5-mer spectrum SVM classifiers are predictive of enhancers in other mammalian species in corresponding tissues. The strong generalization of performance and correlation of the predictions for specific sequences by classifiers trained in different species indicates that many sequence properties predictive of enhancers are conserved across mammals.

Short DNA sequence patterns remain predictive of enhancer activity after controlling for GC content and repetitive elements

III-3.4 Short DNA sequence patterns remain predictive of enhancer activity after controlling for GC content and repetitive elements

Enhancer activity is positively correlated with GC content (Figure III-6), and enhancers are often born from repetitive sequences derived from transposable elements (Rebollo et al. 2012; Chuong et al. 2013; Su et al. 2014; Simonti et al. 2017). Thus, we sought to evaluate the extent to which these properties influenced the generalization of our enhancer prediction models across species. First, we trained GC-controlled classifiers using negative sets of random genomic regions matched on GC content (experiments 37¬–72, 156–164, and 174–182). The predictive power of the GC-controlled classifiers was substantial (average auROC of 0.75 for liver, 0.79 for limb and 0.81 for brain; average auPR of 0.24 for liver, 0.28 for limb and 0.34 for brain; Appendix A), but as expected, less than the corresponding classifiers without GC-control (average auROC of 0.81 for liver, 0.89 for limb and 0.92 for brain; Fig III-3). Nevertheless, GC-controlled classifiers maintained strong cross-species generalization: liver classifiers had an average relative auROC of 94.8% (average relative auPR of 86.3%) when applied to the other five species (Figure III-9); limb classifiers had an average relative auROC of 95.0% (relative auPR of 82.4%) when applied across species; brain classifier had an average relative auROC of 94.8% (relative auPR of 84.2%). We observed similar cross-species generalization with classifiers trained on VISTA enhancers and GC-controlled negatives as well (experiment 214-237). The enhancer predictions for individual sequences by the GC-controlled classifiers were significantly correlated, and as expected, high GC-content sequences no longer received consistently high scores (Figure III-10). Ultimately, the strong cross-species generalization of the GC-controlled classifiers suggests that enhancers differ from the genomic background in sequence patterns beyond GC-content, and that those patterns are conserved. In addition, we trained classifiers to distinguish enhancers from their flanking regions (within 10x enhancer length) (experiments 372–407). These classifiers also performed similarly

and generalized across species (Figure III-11); this suggests that the conserved sequence properties are specific to enhancers, not just regulatory genomic neighborhoods.

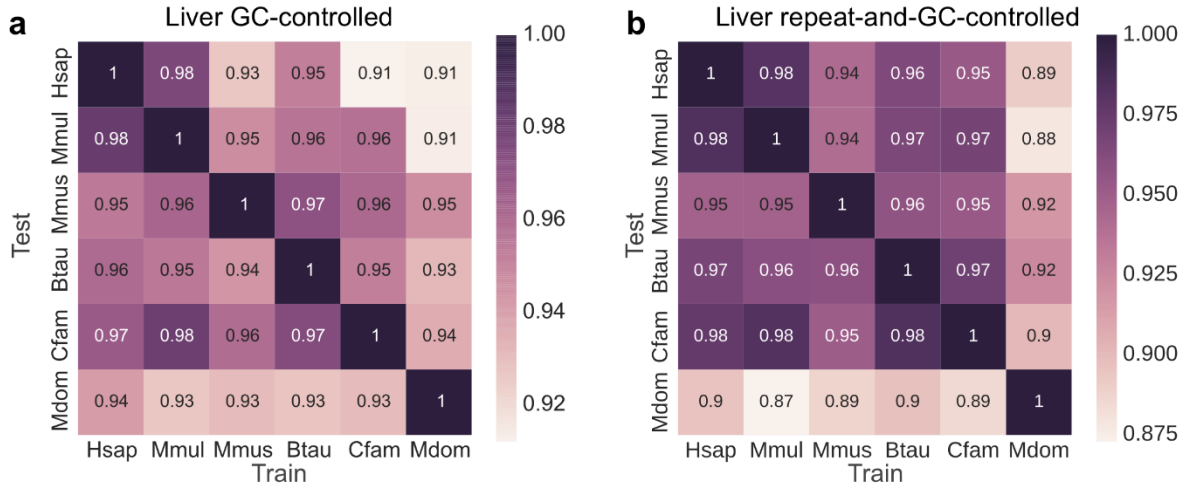

Figure III-9. Enhancer sequence properties remain conserved across diverse mammals after controlling for both GC-content and repetitive elements. The heat maps give the cross-species relative auROCs for SVM classifiers trained on 5-mer spectra to identify enhancers in the species along the x-axis, and then used to predict enhancers in the species on the y-axis. The "negative" training regions from the genomic background were matched to the enhancers': (a) GC-content, and (b) GC-content and proportion overlap with repetitive elements.

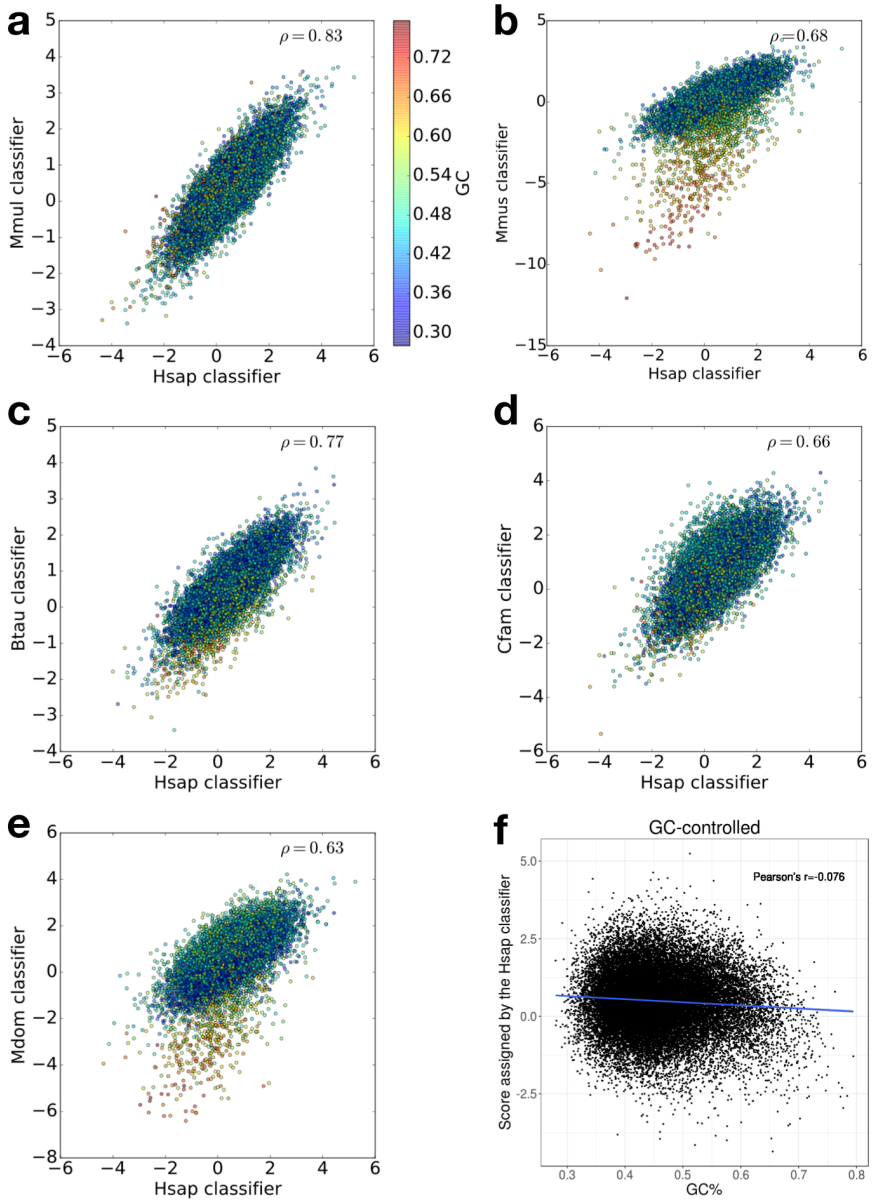

Figure III-10. The predictions of enhancer classifiers trained in different species are strongly correlated (GC-controlled analysis). Scatter plots showing the correlation between scores assigned to human enhancers by the human-trained classifier and the classifiers trained on other species in GC-controlled analysis: (a) Human (experiment 37) vs. Macaque (experiment 43). (b) Human vs. Mouse (experiment 49) (c) Human vs. Cow (experiment 55) (d) Human vs. Dog (experiment 61) (e) Human vs. Opossum (experiment 67). Each dot represents a human liver enhancer sequence. The enhancer score assigned by the human-trained classifier is plotted on the x-axis, and the score assigned by the classifier trained on the other specified species is plotted on the y-axis. The color indicates the GC content. The correlation between enhancer scores produced by different species classifiers is quantified by Spearman's rank correlation coefficient (ρ). (f) The GC content of human enhancers has low correlation with the scores assigned by the human-trained classifier (Pearson's r=–0.076, P<2.2e-16).

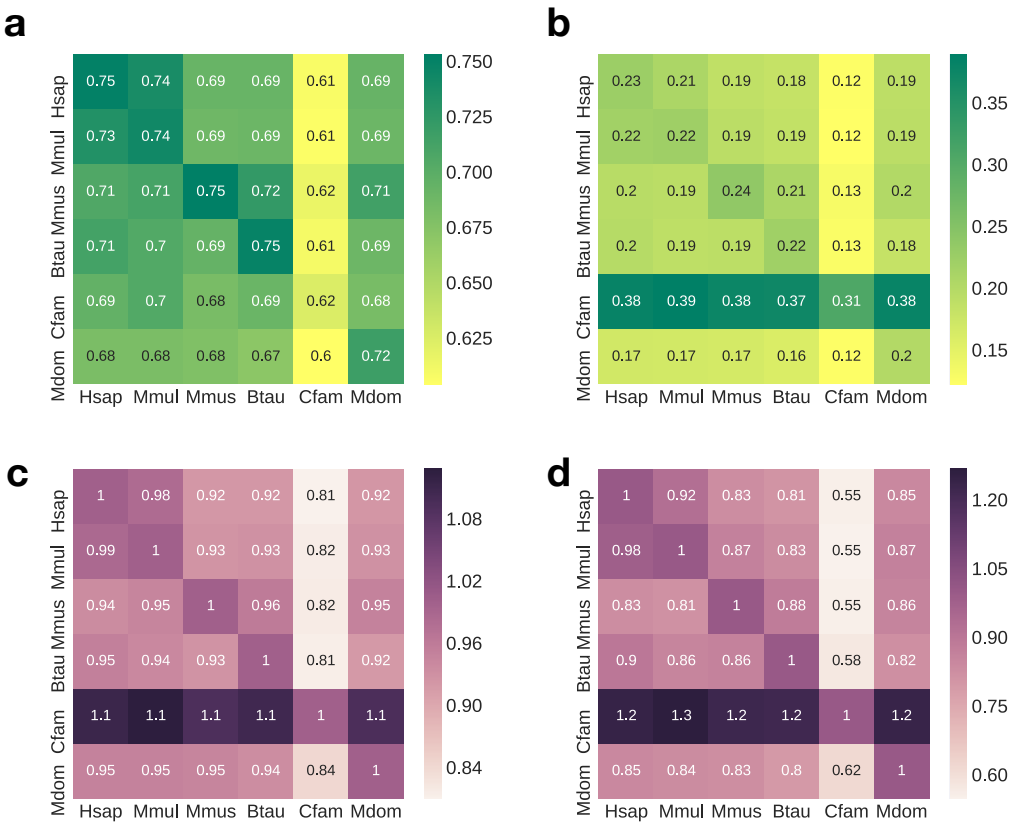

Figure III-11. Evaluation of between species liver enhancer classification using flanking regions as negatives (experiments 372–407). We evaluated the ability of the 5-mer spectrum classifier to distinguish enhancers from flanking regions and the ability of these classifiers to generalize across species: (a) auROC, (b) auPR, (c) relative auROC, (d) relative auPR. We defined the flanking region of an enhancer as 10 times its length on either side. We then randomly selected 10 negative regions of same length as the enhancer that did not overlap other enhancers from the candidate flanking regions. Classifiers were then applied across species. The classifiers performed similarly to the GC-controlled classifiers and generalized very well across species. The dog classifier had much lower performance and generalization than the other classifiers. This could indicate differences in the sequence similarity of regulatory neighborhoods in dogs or be due to the quality of the dog genome assembly.

The generalization of each liver GC-controlled classifier across species had the same pattern as the classifiers without GC-control: the human classifier had the best generalization (average relative auROC = 96.1%), while the opossum had the worst (average relative auROC = 92.8%), which is likely due to the quality of genome assembly. In these GC-controlled analyses, we observed a stronger inverse correlation between the relative performance across species and sequence divergence (Figure III-12, Spearman's rho = –0.72, P < 2.2e–16) than in the non-GC-

controlled analysis (Figure III-7). This indicates that both genomic differences in GC content distribution and overall evolutionary divergence influence the conservation of the sequence patterns predictive of putative enhancers.

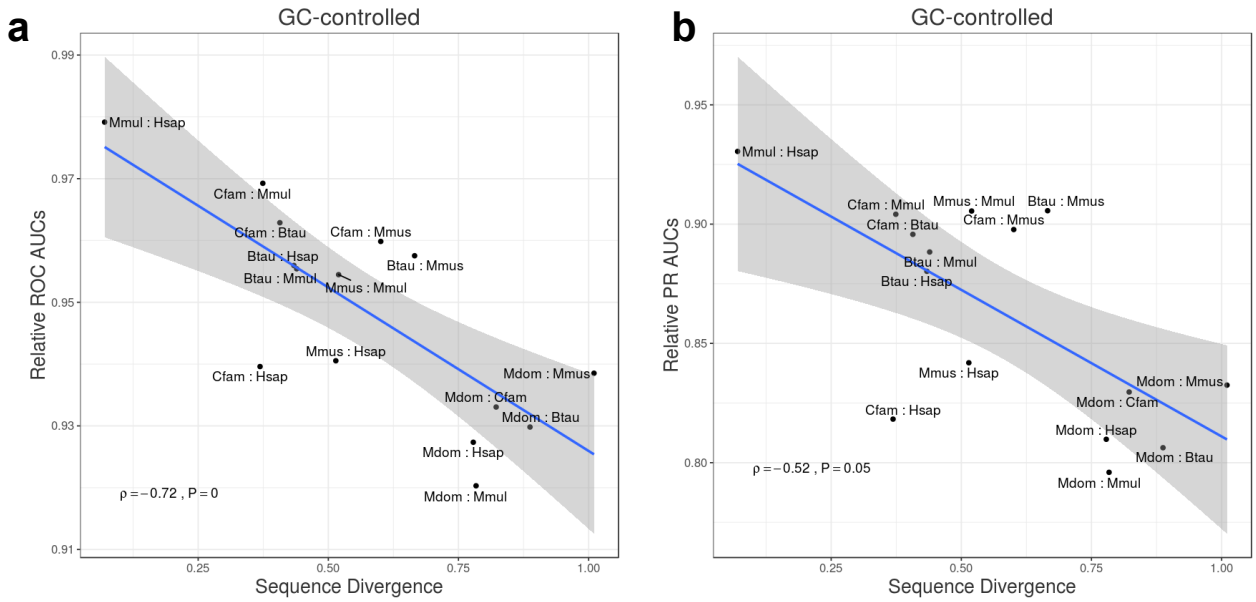

Figure III-12. Neutral sequence divergence is significantly inversely correlated with the GC-controlled cross-species prediction accuracy. Correlation of relative auROCs from the GC-controlled classifiers (experiments 37–72) with sequence divergence. Spearman's rho is –0.72 ($P = 0$). (b) Correlation of relative auPRs from the GC-controlled classifiers (experiments 37–72) with sequence divergence. Spearman's rho is –0.52 ($P = 0.05$). Sequence divergence is quantified as the number of substitutions per neutrally evolving site as derived from four-fold degenerate sites in codons in the UCSC Genome Browser's100-way multiple species alignments (Methods). To determine the relative auROC/PR for each pair of species, the mean was taken across the two classifiers when applied cross-species (i.e., the relative auROC/PR from the human classifier applied to mouse and the relative auROC/PR mouse classifier applied to human were averaged).

(b)

To evaluate the influence of repetitive elements on the ability to distinguish enhancers from the background and the observed conservation of sequence properties across species, we trained classifiers to distinguish enhancers that did not overlap a repetitive element (only 3.3% of all enhancers in human) from matched non-repetitive regions from the genomic background (experiments 73–108). Neither the ability to distinguish enhancers from the background in a species, nor the ability of predictive sequence properties to generalize across species, was

substantially reduced (Figure III-13). This demonstrates that, while repetitive elements contribute to enhancer activity, the conservation of sequence properties predictive of enhancers is not contingent on their presence.

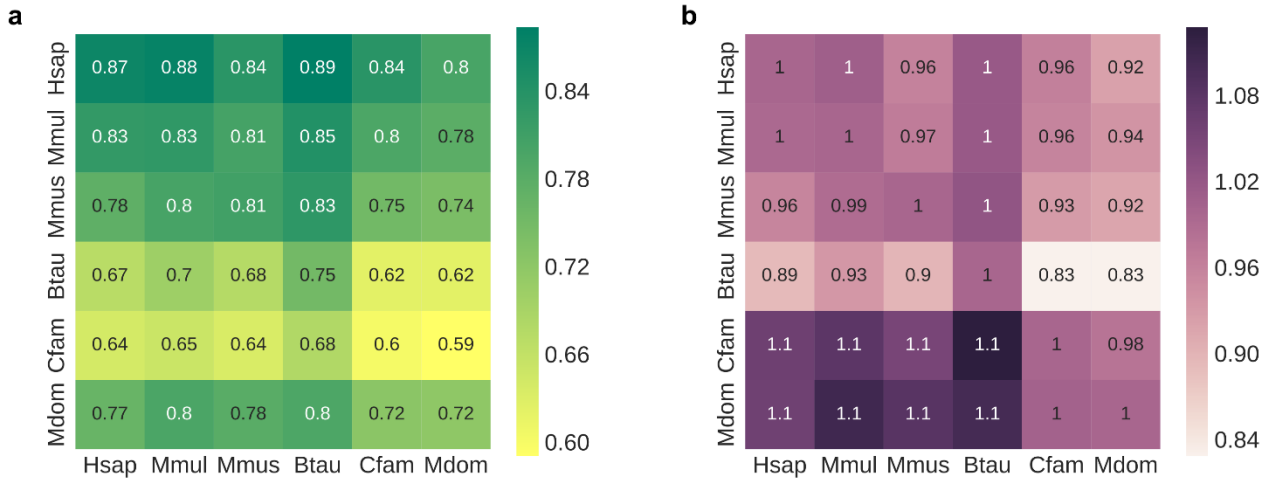

Figure III-13. Classifiers trained on enhancers lacking repetitive elements generalize across species (experiments 73–108). In liver enhancers from each species, we identified those that did not overlap a repetitive element (Methods). The vast majority of enhancers overlapped at least one repetitive element, leaving at total of 966 (human), 1321 (macaque), 914 (mouse), 2772 (cow), 451 (dog), 556 (opossum) enhancers. Classifiers trained on these 'repeat-free' enhancers generalized well across species as measured by (a) raw auROC and (b) relative auROC. Surprisingly, classifiers trained in other species better predicted dog and opossum enhancers than the dog and opossum trained classifiers. This is likely a consequence of the small training sets remaining for dog and opossum; these two species had the fewest liver enhancers without repeat overlap (451 and 556, respectively, while the other species each had at least 900 remaining).

To further examine the influence of GC-content and repetitive elements across all observed enhancer sequences, we also trained classifiers to distinguish all enhancer regions from genomic background regions matched for both GC-content and the proportion of overlap with a repeat element (experiments 109–144). The performance of these classifiers slightly decreased (average auROC of 0.73, auPR of 0.21, Appendix A) relative to when not controlling for repeat overlap (average auROC of 0.75, auPR of 0.24) or neither repeats or GC-content (average auROC of 0.81, auPR of 0.32; Figure III-3). This indicates that, as expected, both features are informative about enhancer function. However, the repeat and GC-controlled classifiers still generalized across

species (average relative auROC = 94.0%, Figure III-9b; average relative auPR = 85.4%); this demonstrates that enhancer sequence properties beyond both GC and repeat content are conserved across species.

III-3.5 Enhancer sequence properties are more similar across the same tissue in different species than across different tissues in the same species

Gene expression patterns are significantly more similar in corresponding tissues across species than between different tissues in the same species (Chan et al. 2009; Brawand et al. 2011; Merkin et al. 2012), and we demonstrated that enhancer sequence properties are strongly conserved in the same tissue across species (Figure III-3). Thus, we hypothesized that, as for gene expression, enhancer sequence properties would be more similar in the same tissue across species (cross-species) than between different tissues in the same species (cross-tissue). To test this, we performed cross-tissue analysis using human enhancers identified in nine diverse cellular contexts, including liver, by the Roadmap Epigenomics Project (Consortium et al. 2015) (experiments 239–274; Methods). We applied the classifier trained on human liver enhancers (from Villar et al. 2015) to Roadmap Epigenomics enhancers from: liver, brain hippocampus middle, pancreas, gastric, left ventricle, lung, ovary, CD14 cells, and bone marrow. We compared the relative auROC between the cross-tissue and cross-species prediction tasks (Figure III-14). In the non-GC-controlled analysis, the human liver enhancer classifier predicts enhancers in macaque, mouse, cow, dog and opossum better than all non-liver Roadmap tissues. In the GC-controlled analysis, we observed the same trend. The cross-species predictions are more accurate than cross-tissue predictions, with the exception of the Roadmap gastric tissue (dark green), which is also a digestive tissue. When compared to the relative auROCs of all pairwise cross-species analyses in liver, limb and brain, those of human liver to non-liver Roadmap tissues are significantly lower (all $P < 0.008$, Mann-Whitney U test; Figure III-14b). In addition to the human cross-tissue analysis, we also examined the cross-tissue performance of the liver, limb and brain classifiers over all three species with enhancer data: human, macaque and mouse. For each species, we applied the classifiers trained in liver, limb, and brain to that species' enhancers in the other two tissues. Again, cross-species performance (all pairwise relative auROCs) was significantly higher than cross-tissue performance in both GC-controlled and non-GC-controlled analyses (Figure III-14b). We observed the same trend in relative auPRs (Figure III-15). The ability of enhancers to regulate gene expression is often contingent on both cell-type specific attributes, such as expression patterns of TFs

(Vaquerizas et al. 2009), and properties that are shared across active enhancers in general. The stronger performance of the trained classifiers in the cross-species compared to cross-tissue prediction tasks suggests that they capture cell-type-specific sequence attributes and that these features are conserved across species.

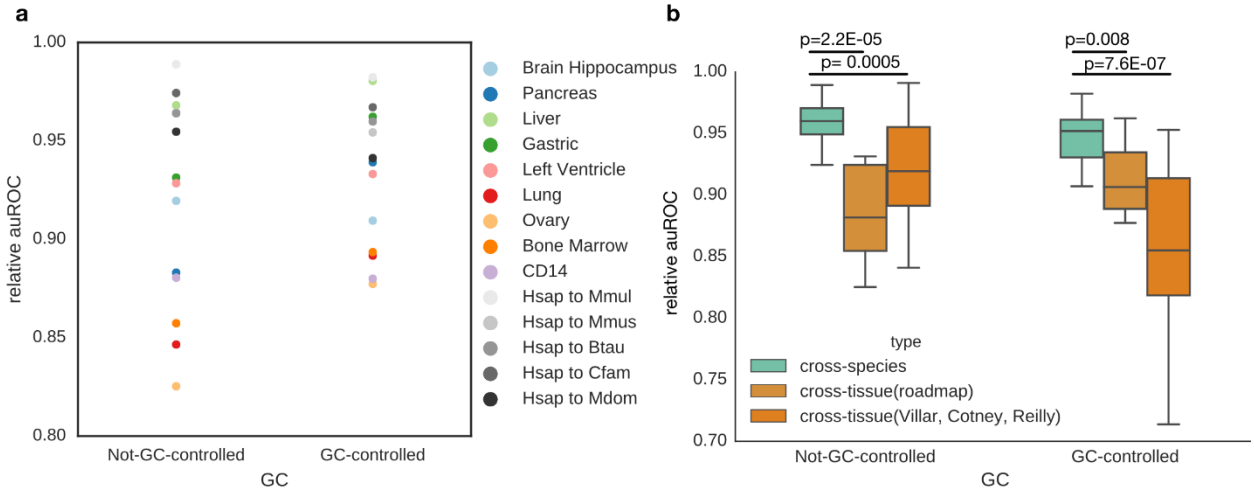

Figure III-14. Enhancer classifiers generalize more accurately across the same tissue in different species than across different tissues in the same species. The human-trained liver classifier obtains better performance when applied to liver enhancers from other species (gray dots) than when applied to enhancers from other human tissues. This also holds for GC-controlled analyses, with the exception of predicting enhancers active in the gastric mucosa. (b) In the not-GC-controlled analysis, the cross-species performance (average relative auROC = 96.2%) is significantly better than the cross-tissue (roadmap) performance (88.4%, Mann Whitney U test, $P = 0.00005$) and the cross-tissue (Villar, Cotney, Reilly) performance (92.0%, Mann Whitney U test, $P = 2.2E-05$). This also holds true for the GC-controlled analysis. The cross-species performance (average relative auROC = 94.6%) is significantly better than the cross-tissue (roadmap) performance (91.2%, Mann Whitney U test, $P = 0.008$) and the cross-tissue (Villar, Cotney, Reilly) performance (85.8%, Mann Whitney U test, $P = 7.6E-07$).

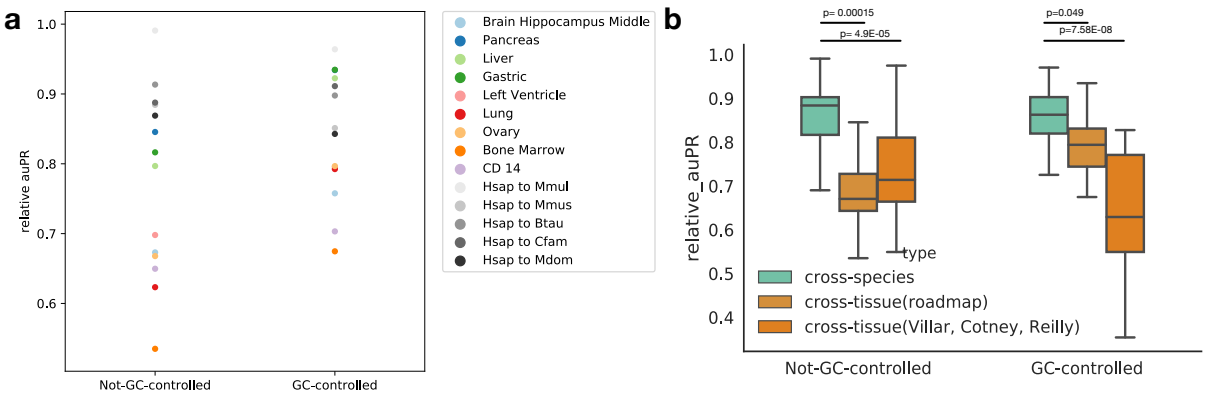

Figure III-15. Enhancer classifiers generalize more accurately across the same tissue in different species than across different tissues in the same species (relative auPRs). The human-trained liver classifier obtains better performance when applied to liver enhancers from other species (gray dots) than when applied to enhancers from other human tissues. This also holds for GC-controlled analyses, with the exception of predicting enhancers active in the gastric mucosa. (b) In the not-GC-controlled analysis, the cross-species performance is significantly better than the cross-tissue (roadmap) performance (P = 0.00015, Mann Whitney U test) and the cross-tissue (Villar, Cotney, Reilly) performance (P = 4.9E-05). This also holds true for the GC-controlled analysis. The cross-species performance is significantly better than the cross-tissue (roadmap) performance (P = 0.049) and the cross-tissue (Villar, Cotney, Reilly) performance (P = 7.58E-08).

III-3.6 The most predictive sequence patterns in different species match binding motifs for many of the same transcription factors

To interpret the biological relevance of the sequence patterns learned by the trained SVM enhancer prediction models in each species, we analyzed the similarity of the sequence properties in a functional context: TF binding motifs. First, we matched the 5% (n = 52) most enhancer-associated 5-mers learned by the human GC-controlled liver classifier to a database of 205 known TF motifs (Mathelier et al. 2014) using TOMTOM (Figure III-16a). The enhancer-associated 5-mers were significantly more likely to match at least one TF motif than expected at random (46.1% vs. 27.7%; one-tailed P = 0.0035, binomial test). The 5% (n=52) most background-associated 5-mers were not significantly different from random (21.6% matched at least one TF, two-tailed P = 0.43, binomial test). This illustrates that the classifiers learned sequence patterns with regulatory potential.

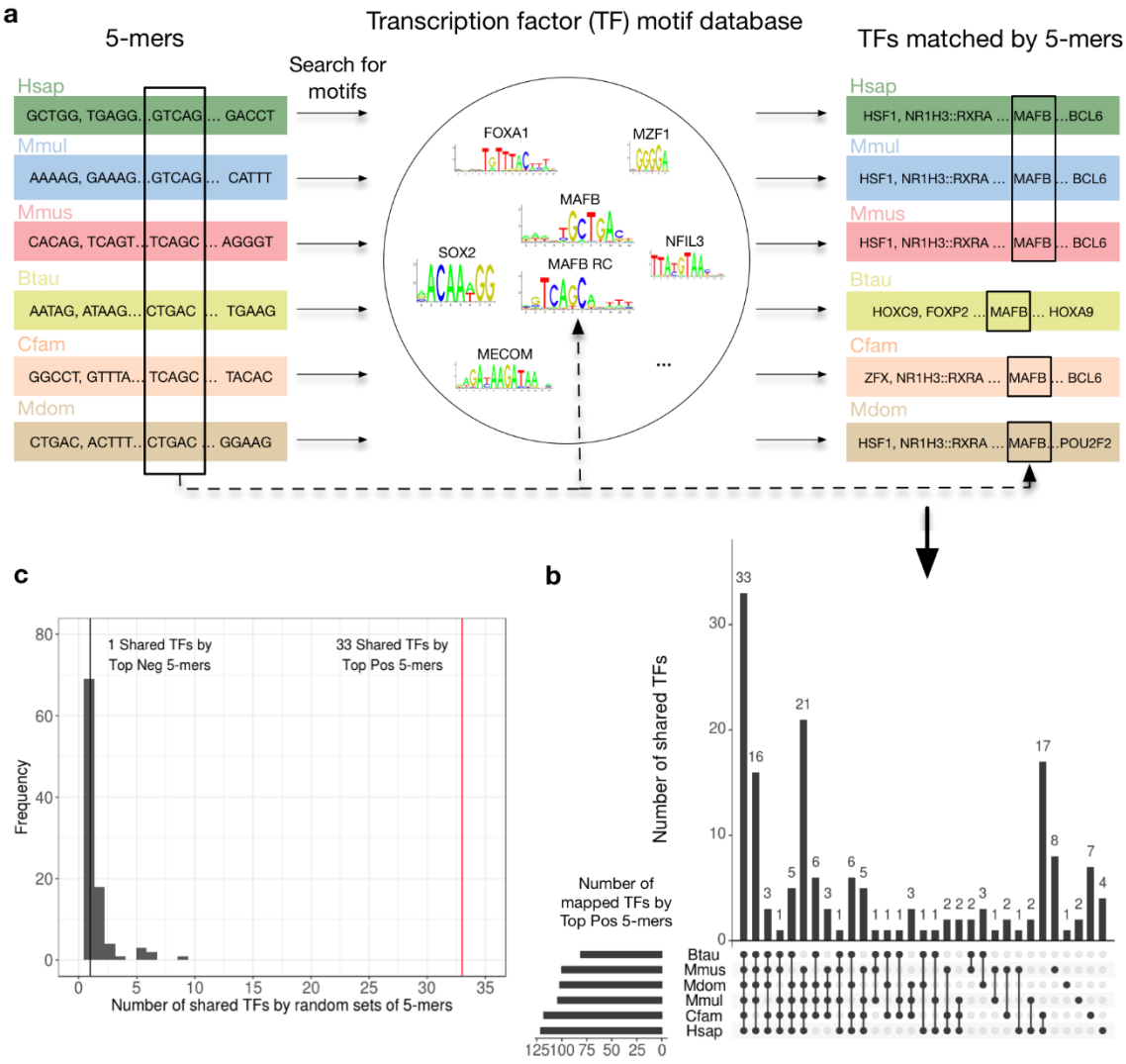

Figure III-16. The DNA sequence patterns most predictive of liver activity across species matched a common set of transcription factors. Transcription factor analysis workflow. For each species enhancer classifier, we found TF motifs matched by the top 5% positively weighted 5-mers. Note that different 5-mers (marked with black box on the left) can match the same motif, e.g., MAFB and its reverse complement (RC). The overlap of matched TFs were then compared across each species' classifier. (b) 33 of the TF motifs matched by the top 5% positive 5-mers from each GC-controlled liver classifier are shared in all species. The total number of TFs matched by top 5-mers in each species was: 121 (human), 104 (macaque), 100 (mouse), 81 (cow), 118 (dog), 102 (opossum). Similar results were observed for the non-GC-controlled classifier (Figure III-17). (c) The number of TFs matched by all species based on 5-mers in top positive, top negative, and 100 random sets of 5% of all possible 5-mers. The 33 TF motifs shared among the high-weight set for each species is thus significantly more than expected.

Next, we investigated whether the TF binding motifs matched by enhancer-associated 5-mers were shared between species. The highly weighted 5-mers in the human-trained classifier matched 121 TF motifs. Of these, the binding motifs for 33 TF were also matched by enhancer-associated 5-mers in all other species (Fig III-16b). This is significant enrichment for shared TF motifs among the enhancer-associated 5-mers compared to the number of TF motifs shared across all species on average over 100 random sets of 5% of 5-mers from each species (Figure III-16c, $P = 0$). Similarly, only one TF motif (MZF1) was shared among all the species' most background-associated 5-mers; this is not significantly different from the number expected at random ($P = 0.97$). Moreover, the sharing of TFs matched by the top positive 5-mers between the human liver classifier and other species' liver classifiers is significantly higher than that between the human liver classifier and classifiers for other human tissues ($P = 0.019$, Mann Whitney *U* test; Figure III-17). This also suggests more conservation of enhancer sequence properties across species within the same tissue than within the same species across different tissues. We obtained similar results when comparing the TFs matched by 5-mers from non-GC-controlled liver SVM models (27 shared TFs by enhancer-associated 5-mers in all species, $P = 0$, Figure III-18). The limb and brain classifiers also shared more TFs among the top 5% of enhancer-associated 5-mers than expected from random sets: 12 TFs ($P = 0.33$, GC-controlled) and 20 TFs ($P = 0.1$) were shared among the limb classifiers; 22 TFs ($P = 0.05$, GC-controlled) and 16 TFs ($P = 0.14$, non-GC-controlled) were shared among the brain classifiers. However, it is likely that the smaller number of available species for developing limb and brain enhancers, our limited knowledge of binding motifs for TFs active in developing limb and brain, and the heterogeneity of developing limb and brain tissue reduced power to detect sharing compared to liver.

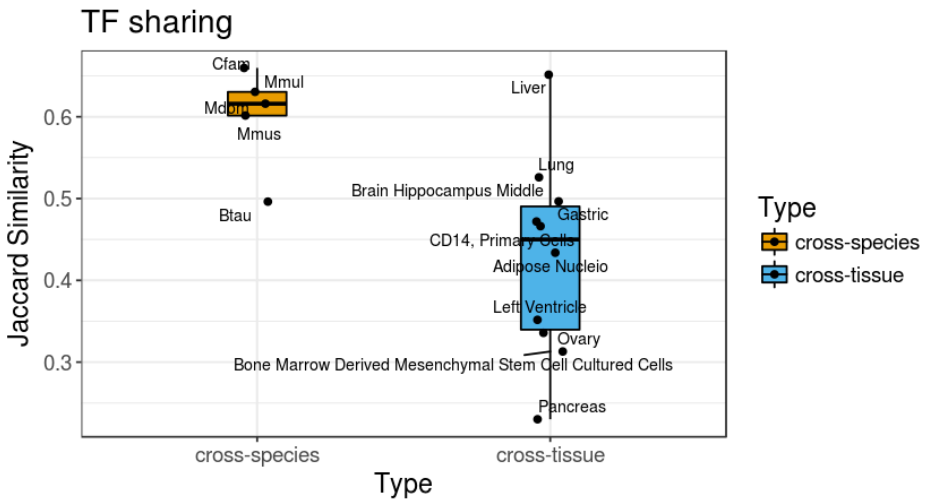

Figure III-17. TFs matched by the top positive k-mers between the human classifier and other species are more similar than those between the human liver tissue and other Roadmap tissues (GC-controlled negatives). For each pair of SVM classifiers, the Jaccard similarity of the top positive k-mer-mapped TFs is plotted.

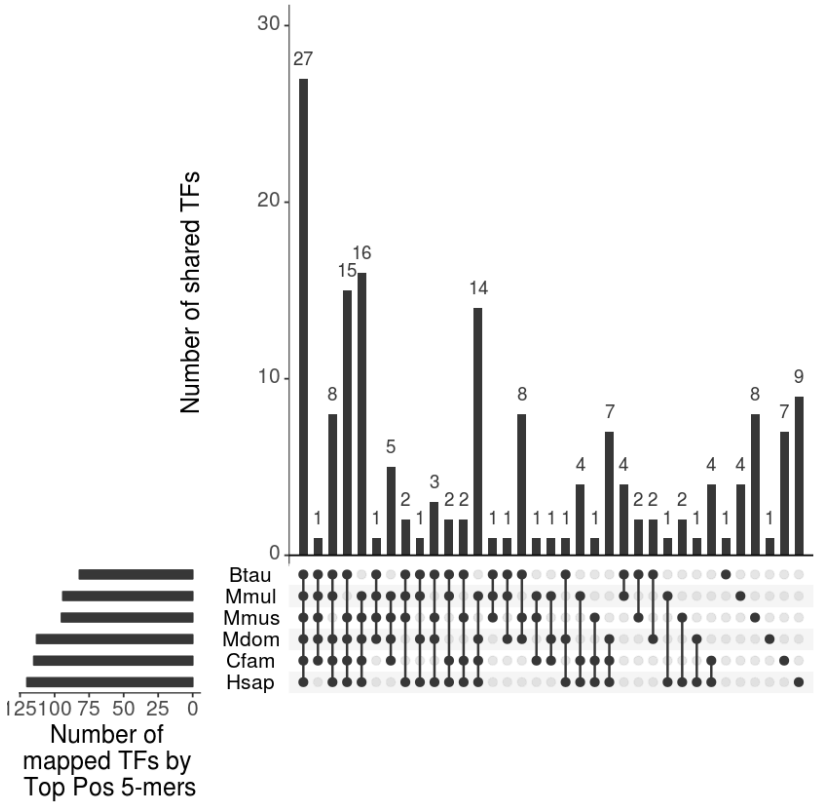

Figure III-18.The DNA sequence patterns most predictive of liver enhancer activity across species matched a common set of transcription factors (non-GC-controlled). Of the TFs matched by the top k-mers from each non-GC-controlled liver classifier (experiments 1, 8, 15, 22, 29, 36), 27 are shared by all six species.

To evaluate the relevance of the shared TF motifs to liver function, we analyzed expression patterns of the TFs across 12 tissues (Bernstein et al. 2012). Shared TFs among liver enhancer-associated 5-mers were significantly enriched for liver expression (Table III-1, P = 0.011, one-tailed Fisher's exact test), and 6 out of the 7 shared TFs not expressed in liver have a liver-expressed TF in the same subfamily (Appendix B). Many of the shared TFs play essential roles in liver function. For instance, they are enriched for activity in the TGF-β signaling pathway compared to non-shared TFs; the enrichment is mainly due to members of the AP-1 (JUN, FOS, and MAF subfamilies) and SMAD families (Methods) (The Gene Ontology Consortium 2000, 2015). TGF-4 signaling is a central regulatory mechanism that is disrupted in all stages of chronic liver disease (Dooley and ten Dijke 2012). Further, mice deficient in c-JUN or MAF have an embryonic lethal liver phenotype (Eferl et al. 1999; Yamazaki et al. 2012). The only TF shared among the background-associated 5-mers, MZF1, is lowly expressed in the liver and not detected at the protein level (Uhlen et al. 2015). We also searched for matches to the binding motifs of known liver master regulators among the highly weighted motifs. While none of them were shared among all models, several including, HNF1A, HNF4A, and FOXA1 matched highly weighted motifs in three or more species. This demonstrates that the sequence patterns learned in each species capture similar motifs that are recognized by TFs important to the relevant tissue context.

Table III-1. TFs motifs shared among the top 5-mers across all species' liver enhancer SVM classifiers are significantly enriched for liver expression (P = 0.011, one-tailed Fisher's exact test).

	Shared TFs	*Not shared TFs*
Liver expressed	26	89
Not liver expressed	7	70
Percent Liver expressed	78.8%	56.0%

III-3.7 Convolutional neural networks predict enhancers more accurately than SVMs, but generalize less well across species

Convolutional neural networks have recently achieved the state-of-art performance at predicting regulatory sequences (Quang and Xie 2015; Zhou and Troyanskaya 2015) and may be better at capturing more complex sequence patterns than *k*-mer SVMs. To investigate the performance and

generalization of CNNs at identifying enhancers across species, we trained CNNs to distinguish liver enhancers from the random genomic background in each species. Here, we used the center 3000 bp of enhancers and a balanced negative set due to the fixed-length input of CNNs and the challenges of training CNNs on unbalanced sets (Methods). To compare the performance of CNNs with the SVM models, we trained a CNN model (experiment 275), a 5-mer spectrum SVM classifier (experiment 325), a 5-mer polynomial kernel SVM (Methods, experiment 351) and an 11-mer gkm-SVM (Methods, experiment 347) on the same human dataset using training, validation, and testing partitions to avoid overfitting (Methods). We found that the k-mer SVM, polynomial kernel SVM, 11-mer gkm-SVM achieved similar performance on this human dataset, with auROCs of 0.78, 0.78, 0.76 and auPRs of 0.75, 0.76, 0.76, respectively (Figure III-19). The CNN performed considerably better, achieving an auROC of 0.86 and auPR of 0.84. Although we did not explore the full hyperparameter space for SVMs, CNN out-performed SVMs by a substantial margin. This was true for the three SVM algorithms we tested across a range of Cs (Figure III-19). Because the prohibitive runtime of gkm-SVM (Methods) and small difference in performance, we continued the CNN comparison with the 5-mer spectrum SVMs. We trained CNNs (experiment 275-310) and 5-mer spectrum SVMs (experiment 311-346) on the 3,000bp long, balanced enhancer datasets for each species and performed cross-species enhancer predictions. The CNN model is substantially better than the SVMs at distinguishing enhancers from genomic background in each species (Figure III-20a), suggesting that the ability to model more complex sequence patterns improves predictions. Moreover, the first layer of the human liver CNN learned many binding motifs for TFs relevant to liver biology, including CEBPB, HNF4A, and HNF1A (Figure III-20b).

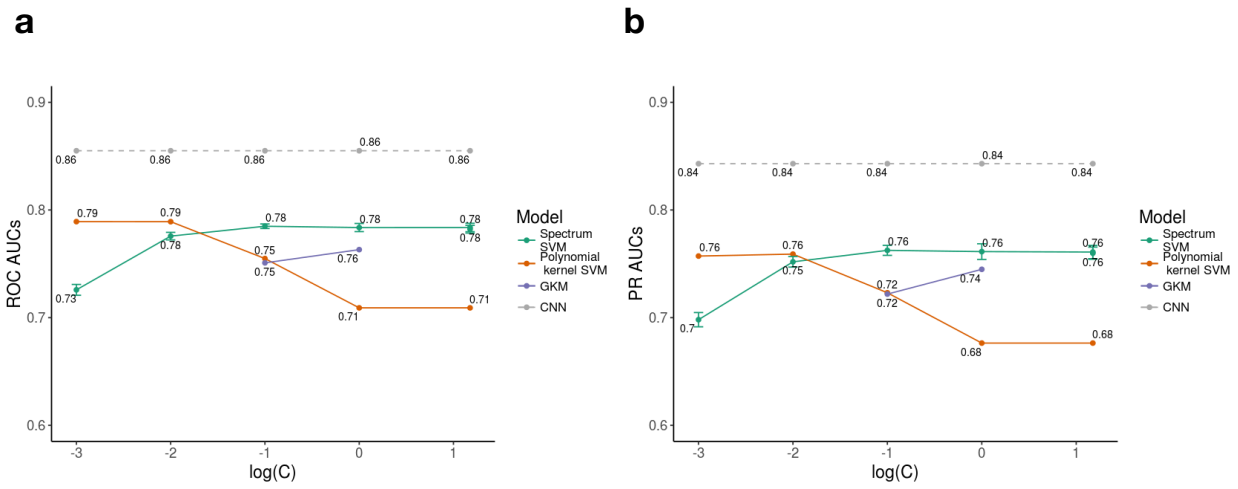

Figure III-19. CNNs perform substantially better than 5-mer spectrum SVMs, 5-mer polynomial kernel SVMs, and 11-mer gkm-SVMs across C values. We evaluate the performance of SVMs across a range of C values (0.001 to 15, x-axis, experiments 348–354, 366–371) and compare it with the CNN model (experiment 275, hyper-parameter selection is described in the Methods). (a) Comparison of auROCs between different classifiers. (b) Comparison of auPRs between different classifiers.

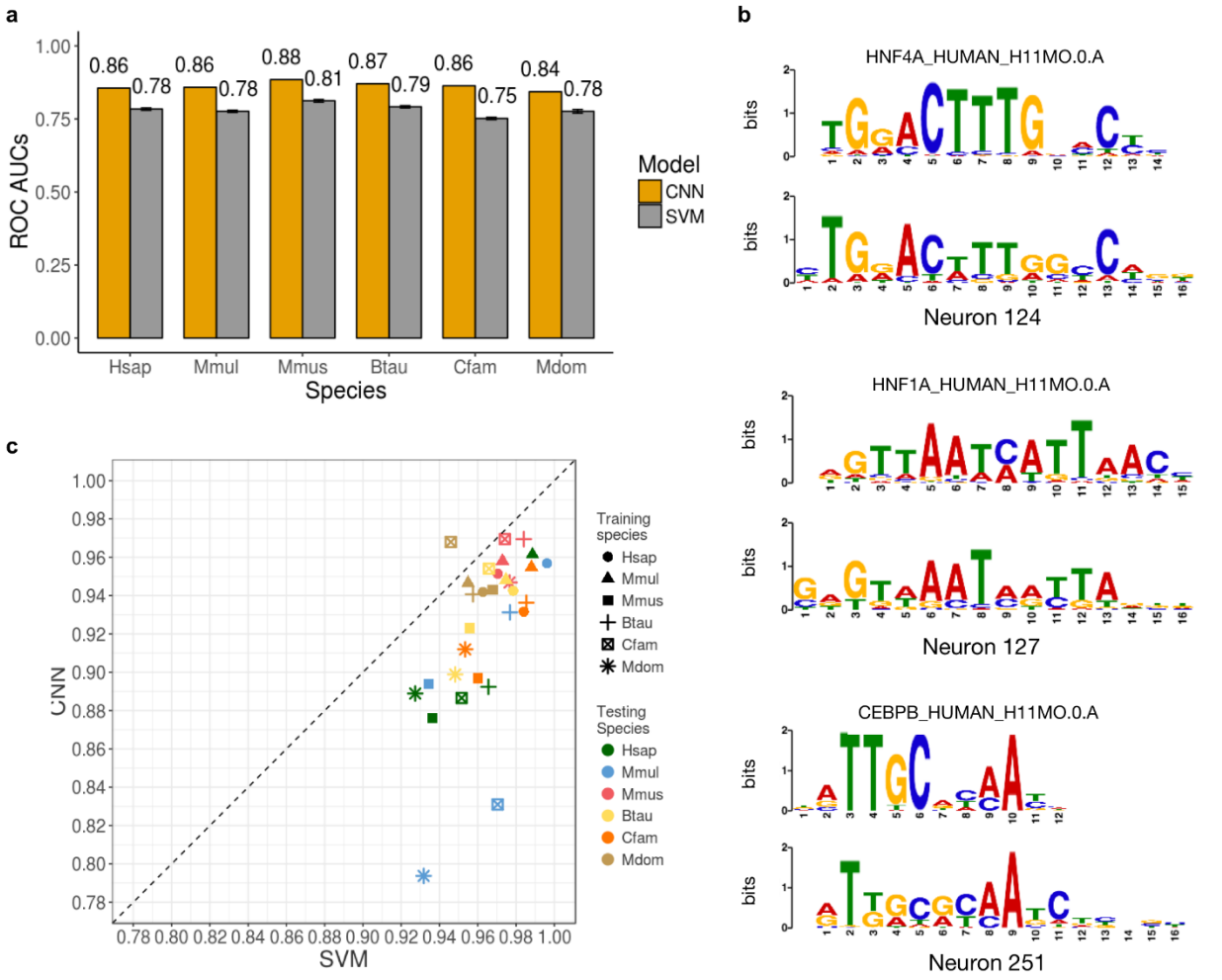

Figure III-20. CNNs identify enhancers more accurately than 5-mer spectrum SVM models, but generalize less well across species. The auROCs of CNN models perform substantially better than the 5-mer SVM model in each species. The error bars give the standard error of ten-fold cross-validation for the SVM models. (b) Neurons in the first layer of the CNN learned the motifs of important liver TFs, including HNF4A, HNF1A, and CEBPB. (c) The relative auROCs of the CNN models applied across species are consistently lower than for the 5-mer SVMs applied across the same species. This suggests that the CNN models do not generalize as well across species as the SVM models.

Next, we performed the cross-species enhancer prediction with the CNNs. The CNN models generalize well across species (relative auROC from 0.79 to 0.97), but their generalization is consistently worse than the SVM models (Figure III-20c; Raw auROCs and auPRs is Appendix A). We observed similar results with repeat and GC-control, and the removal of shared orthologous

enhancers (Figure III-21). In addition, we applied the 5-mer polynomial kernel SVM across species to test if the worse generalization of the CNNs could be explained by its ability to capture k-mer interactions (experiments 408–443). The polynomial kernel SVMs perform similarly to the k-mer SVMs within species and do not generalize substantially worse than k-mer SVMs across species, suggesting little influence of global k-mer interaction patterns on enhancer identification (Figure III-22). This is consistent with the finding that the co-binding patterns of TFs are mostly conserved between human and mouse (Boyle et al. 2014a). These results suggest that the sequence properties learned by the CNNs are less conserved across species than those learned by the k-mer spectrum SVMs. These could include better representations of TF motifs or more sophisticated interactions between TFs, such as their orientation, spacing and ordering. However, developing clear biological interpretations for the patterns learned by the CNNs is challenging.

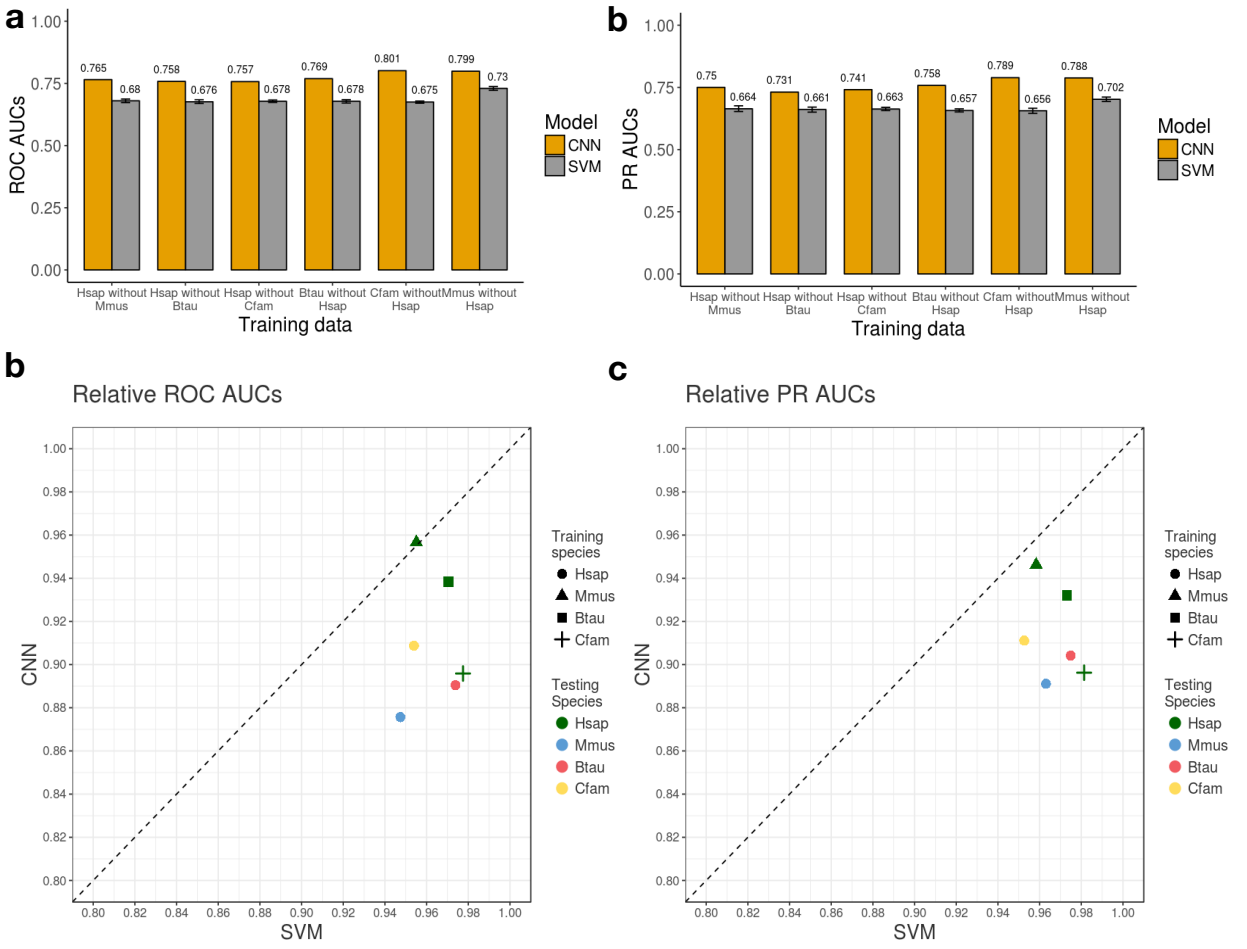

Figure III-21. The CNNs trained on GC-controlled, repeat-controlled enhancer datasets with orthologous enhancers removed performed better than the 5-mer spectrum SVMs trained on the same data and generalized worse across species (experiments 354–365). (a) The auROCs of CNN models were substantially better than the 5-mer SVM models in each species. The error bars give the standard error of ten-fold cross-validation for the SVM models. We removed the enhancer orthologs between each pair of human and another species. For instance, "Hsap without Mmus" means human enhancers with mouse enhancer orthologs removed from consideration. (b) The auPRs of CNN models were substantially better than the 5-mer SVM models in each species. The error bars give the standard error of ten-fold cross-validation for the SVM models. (c) The relative auROCs of the CNN models applied across species are consistently lower than for the 5-mer spectrum SVMs applied across the same species. (d) The relative auPRs of the CNN models applied across species are consistently lower than for the 5-mer spectrum SVMs applied across the same species. This suggests that the CNN models did not generalize as well across species as the SVM models.

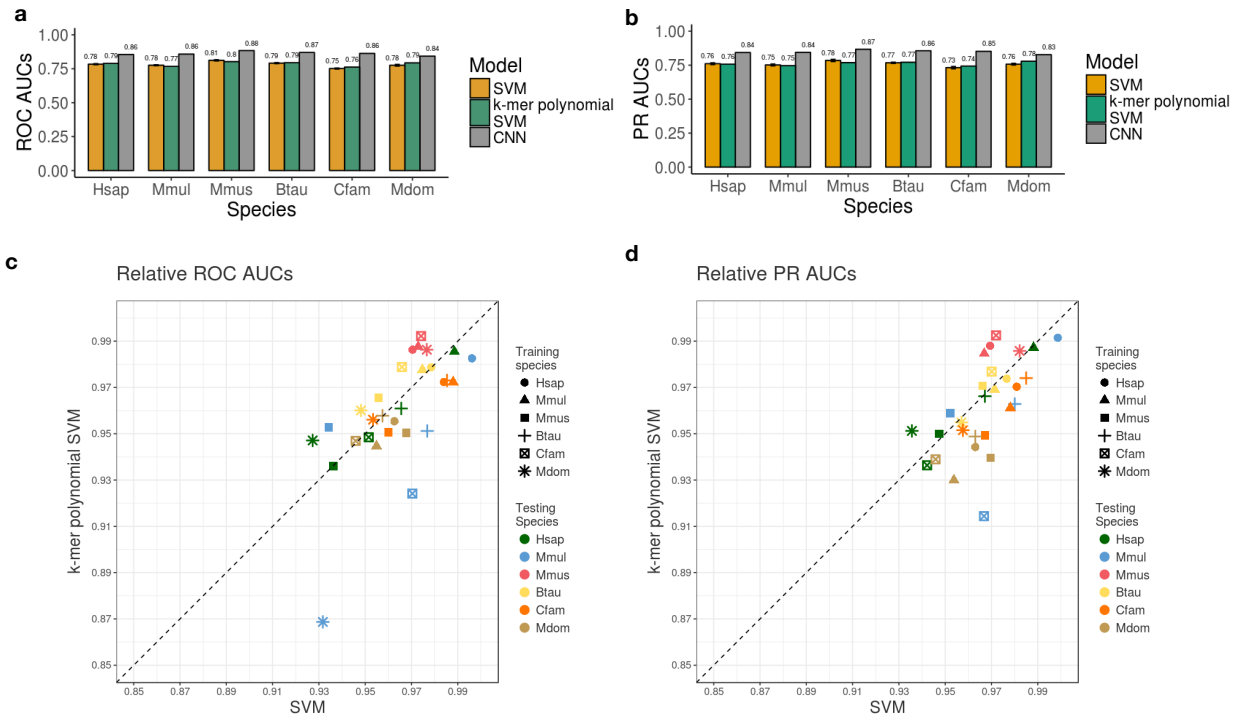

Figure III-22. The 5-mer polynomial kernel SVMs trained on enhancers and random genomic regions (experiments 408–443) performed similarly to 5-mer spectrum SVMs within and across species. The auROCs of 5-mer polynomial kernel SVMs are similar to 5-mer spectrum SVMs within species and are substantially worse than the CNNs in each species. The error bars give the standard error of ten-fold cross-validation for the 5-mer spectrum SVM models. (b) The auPRs of 5-mer polynomial kernel SVMs are similar to 5-mer spectrum SVMs within species and are substantially worse than the CNNs in each species. (c) The relative auROCs of the 5-mer polynomial kernel SVMs applied across species are similar to the 5-mer SVMs applied across the same species. (d) The relative auPRs of the 5-mer polynomial kernel SVMs applied across species are similar to the 5-mer SVMs applied across the same species. This suggests that the 5-mer polynomial kernel SVMs generalized as well across species as the simpler SVM models.

III-4 Discussion

In this study, we trained SVM and CNN classifiers based on DNA sequence patterns to distinguish enhancers from the genomic background in diverse mammalian species. We showed that, in spite of significant changes in the enhancer landscape between species, the SVM models trained using short sequence patterns as features exhibited minimal decreases in performance when applied across species. This indicates that short sequence patterns predictive of enhancer activity captured by these models are largely conserved across mammals. Furthermore, the DNA patterns most predictive of liver enhancer activity across species matched a common set of TF

binding motifs with enrichment for expression in the relevant tissue. The sequence properties predictive of histone-mark defined enhancers were also predictive of enhancers confirmed in transgenic reporter assays. We then showed that CNN models performed better than SVMs at identifying enhancers. They also generalized well across species, but not as well as SVMs. These results suggest that conserved regulatory mechanisms have maintained constraints on short sequence motifs present in enhancers for more than 180 million years.

Confidently identifying and experimentally validating enhancers remains challenging (Benton et al. 2017). We showed that short sequence properties are conserved across species using enhancers identified via two complementary techniques: histone modification profiling and transgenic assays. Each of these approaches has strengths and weaknesses. The histone-modification based enhancer predictions enable genome-wide characterizations across many species, but this approach is prone to false positives (Andersson et al. 2014; Dogan et al. 2015). On the other hand, the transgenic assays clearly demonstrate the competence of a sequence to drive gene expression, but are restricted to a biased set of relatively few sequences from two species that are tested at one developmental stage. By showing the cross-species conservation is maintained in both categories, and that models trained on each set perform similarly, we argue the conservation of enhancer short sequence properties is robust to the methodology used to define enhancers.

The design of this study can serve as a framework for further examining the conservation and divergence of regulatory sequence patterns across species. We trained sequence-based machine learning models within a species, and then applied them to other species; this approach can be applied on a genome-wide scale, is not dependent on knowledge of TF binding motifs, and allows some flexibility in the weights assigned to each feature while directly testing the generalization of overall sequence patterns. Identification of enhancers in more divergent species would enable us to better quantify the depth of enhancer sequence properties conservation. This remains an open question, as more divergent animal species have very little conservation of TF co-associations at putative enhancers despite conservation of TF binding preferences (Boyle et al. 2014b); however, enhancer properties appear to be conserved over greater evolutionary timescales in insects (Stefflova et al. 2013; Kazemian et al. 2014; Villar et al. 2014c) and transcriptional networks seem to evolve at relatively constant rates across animals (Carvunis et al. 2015). Identification of enhancers in the same cellular context for more closely related species

would also enable the investigation of lineage-specific regulatory sequence patterns. Thus, additional comparative studies of regulatory sequence features in more species are needed to better understand both recent and ancient influences on regulatory sequences.

While both the SVM and CNN classifiers correctly distinguished many enhancers from the genomic background, neither performed perfectly. Many factors contribute to this, including: false positives in the training data, noise from the low resolution of the histone modification peaks (i.e., they include non-functional sequence flanking the enhancer), errors in the genome assemblies, and the features considered in our models. As enhancer datasets and prediction methods improve, it will be valuable to continue to evaluate generalization across species. It will also be valuable to train, evaluate, and interpret CNNs on unbalanced data sets. Additionally, the features learned by the enhancer CNNs are difficult to interpret biologically, especially for higher-level neurons. Thus, it is not clear whether the CNN classifiers achieved better performance within species but had worse generalization across species by capturing sophisticated interactions between simpler motifs, by more accurately modeling the sequence-preferences of TFs in each species, via better recognition of the genomic background, or recognition of other unappreciated patterns. The interpretation of sequence features learned by the accurate CNNs could facilitate the understanding of how more complex rules of enhancer sequence architectures change during evolution. The identification and interpretation of conserved and diverged gene regulatory patterns between species is an important area for future work.

Chapter IV

Learning and interpreting regulatory grammar through a deep learning framework

IV-1 Introductions

Enhancers are genomic regions distal to promoters that regulate the dynamic spatiotemporal patterns of gene expression required for the proper differentiation and development of multi-cellular organisms (Shlyueva et al. 2014; Kundaje et al. 2015; Villar et al. 2015b). As a result of their essential role, mutations that disrupt proper enhancer activity can lead to diseases. Indeed, the majority of genetic variants associated with complex disease in genome-wide association studies (GWAS) are non-protein coding, and thought to influence disease by disrupting proper gene expression levels (Maurano et al. 2012; Corradin and Scacheri 2014; Brazel and Vernimmen 2016).

Enhancers function through the coordinated binding of transcription factors (TFs). Recent advances in high-throughput sequencing techniques, such as high-throughput systematic evolution of ligands by exponential enrichment (HT-SELEX), protein binding microarray (PBM), chromatin immunoprecipitation sequencing (ChIP-Seq), have greatly deepened our knowledge of TF binding specifies (Bernstein et al. 2012; Jolma et al. 2013b; Lambert et al. 2018). However, identifying consensus TF binding motifs is not sufficient for inferring TF binding. As shown in many ChIP-seq studies, TFs only bind to a small fraction of all motif occurrences in the genome, and some binding sites do not contain the consensus TF binding motif, indicating a necessity for additional features (Wang et al. 2012). Indeed, many additional features have been suggested to play a role in determining *in vivo* TF binding, such as heterogeneity of a TF's binding motif (Levy and Hannenhalli 2002a), local DNA properties (Dror et al. 2015), broader sequence context and interposition dependence (Mathelier and Wasserman 2013), clusters of homotypic binding sites (Dror et al. 2015), cooperative binding of the TF with its partners (Wang et al. 2006; Yáñez-Cuna et al. 2013; Jolma et al. 2015; Liu et al. 2015), and condition-specific chromatin context (Wang et al. 2006; Heintzman et al. 2009; Kumar and Bucher 2016). While both genomic and epigenomic features are important in determining the *in vivo* occupancy of a TF, recent studies have suggested that the epigenome can be accurately predicted from genomic context (Arvey et al. 2012; Benveniste et al. 2014; Dror et

al. 2015; Whitaker et al. 2015), supporting the fundamental role of sequence in dictating the binding of TFs (Wilson et al. 2008; Ritter et al. 2010; Schmidt et al. 2010; Li and Ovcharenko 2015; Prescott et al. 2015b). Therefore, it is critical to understand the complex mechanisms underlying enhancer regulatory functions and build sufficiently sophisticated models of enhancer sequence architecture.

Combinatorial binding of TFs, i.e., the regulatory grammar, is thought to be essential in determining *in vivo* condition-specific binding (Levy and Hannenhalli 2002b; Arvey et al. 2012; Mathelier and Wasserman 2013; Sharmin et al. 2015). However, how enhancers integrate multiple TF inputs to direct precise patterns of gene expression is not well understood. Most enhancers likely fall on a spectrum represented by two extreme models of enhancer architecture: the enhanceosome model and the billboard model (Slattery et al. 2014; Long et al. 2016). The "enhanceosome model" proposes that enhancer activity is dependent on the cooperative assembly of a set of TFs at enhancers. The cooperative assembly of an enhanceosome is based on physical protein-protein interactions and highly constrained patterns of TF-DNA binding sites. The enhanceosome model does not tolerate shifts in the spacing, orientation, or ordering of the binding site, which can disrupt protein-protein interactions and cooperativity. This model is proposed based on in-depth study of the interferon-β (IFN-β) enhancer (Thanos and Maniatis 1995). Binding sites for the heterodimer ATF-2/c-Jun, interferon response factors IRF-3 and IRF-7, and NFκB (p50:RelA) are tightly clustered in a 55 base pair (bp) stretch of DNA. The individual factors do not activate IFN-β gene expression by themselves, and failure to mobilize any one of the factors abrogates IFN-β transcription entirely (Maniatis et al. 1998). This model likely presents an extreme example because only very few enhancers are found under similarly stringent constraints (Erives and Levine 2004; Papatsenko and Levine 2007; Crocker et al. 2008; Swanson et al. 2010, 2011). However, many examples of less extreme spatial constraints on paired TF-TF co-association and binding-site combinations are found in genome-wide ChIP sequencing studies (Sorge et al. 2012; Cheng et al. 2013; Kazemian et al. 2013) and *in vitro* consecutive affinity-purification systematic evolution of ligands by exponential enrichment (CAP-SELEX) studies. On the other end of the spectrum is the "billboard model", also known as the "information display model" (Kulkarni and Arnosti 2003; Arnosti and Kulkarni 2005), which hypothesizes that instead of functioning as a cooperative unit, enhancers work as an ensemble of separate elements that independently affect gene expression. That is, the positioning of binding

sites within an enhancer is not subject to strict spacing, orientation, or ordering rules. The TFs at billboard enhancers work together to direct precise patterns of gene expression, but their function does not strongly depend on each other. For instance, the loss of a TF binding may lead to change in the target gene expression, but will not cause the complete collapse of enhancer function. The actual mechanisms by which multiple TFs assemble on enhancers are likely a mixture of the two models. Indeed, a massively parallel reporter assay (MPRA) of synthetic regulatory sequences suggested that while certain transcription factors act as direct drivers of gene expression in homotypic clusters of binding sites, independent of spacing between sites, others function only synergistically (Smith et al. 2013).

In recent years, deep neural networks (DNNs) have achieved state-of-art prediction accuracies for many tasks in regulatory genomics, such as predicting splicing activity (Leung et al. 2014; Xiong et al. 2015), specificities of DNA- and RNA-binding proteins (Alipanahi et al. 2015), transcription factor binding sites (TFBS) (Quang and Xie 2015, 2019; Zhou and Troyanskaya 2015), epigenetic marks (Quang and Xie 2015; Zhou and Troyanskaya 2015; Kelley et al. 2016), enhancer activity (Min et al. 2016; Yang et al. 2017) and enhancer-promoter interactions (Singh and Yang 2016). However, in spite of their superior performance, little biological knowledge or mechanistic understanding has been gained from DNN models. In computer vision, the interpretation of DNNs trained on image classification tasks demonstrate that high-level neurons often learn increasingly complex patterns building on those learned by lower level neurons (Zeiler et al. 2010; Springenberg et al. 2014; Zeiler and Fergus 2014; Yosinski et al. 2015; Olah et al. 2017, 2018; Shrikumar et al. 2017a). DNNs trained on DNA sequences might behave similarly, with neurons in low levels learning building blocks of the regulatory grammar, short TF motifs, and those in higher levels learning the regulatory grammar itself, the combinatorial binding rules of TFs (Kelley et al. 2016; Quang and Xie 2016; Chen et al. 2018).

The majority of DNNs trained with genomic sequences use a convolution layer as a first layer and then stack convolution or recurrent layers on top of it. A common approach to interpret the features learned by such DNNs is to convert the first convolution layer neurons to position weight matrices by counting nucleotide occurrences in the set of input sequences that activate the neurons (Alipanahi et al. 2015; Kelley et al. 2016; Chen et al. 2018). With the development of more advanced DNN visualization and interpretation techniques in computer vision, many other

DNN interpretation methods emerged, such as occlusion (Zeiler and Fergus 2014), saliency maps (Simonyan et al. 2013b), guided propagation (Zeiler and Fergus 2014), gradient ascent (Yosinski et al. 2015). Some of these techniques have been applied to visualize features learned by DNNs trained with genomic sequences. For instance, a gradient based approach, DeepLIFT, identified relevant transcription factor motifs in the input sequences learned by a convolutional neural network (Shrikumar et al. 2017a). Saliency maps, gradient ascent and temporal output scores have been used to visualize the sequence features learned by a DNN model for TFBS classification and found informative TF motifs (Lanchantin et al. 2017). These studies demonstrate the power of DNNs in recognizing the TF motifs in the input sequences. However, these studies focused only on the interpretation of the output layer in models for predicting TFBS. Enhancers are much more complex than individual TFBS; they have multiple binding sites in a range of combinations and organizations. It is also unclear whether the intermediate layers of DNNs have the capability of learning rules of combinatorial TF binding from regulatory regions with many TFs, such as enhancers.

Another substantial challenge in the development of methods to interpret DNNs applied to regulatory sequences is our lack of knowledge of the combinatorial rules governing enhancer function in different cell types. Beyond a few foundational examples used to propose possible enhancer architectures, the constraints and interactions that drive enhancer function are largely unknown. Thus, it is difficult to determine if a pattern learned by a neuron is "correct" or biologically relevant. The generation of synthetic DNA sequences that reflect different constraints on regulatory element function has promise to help address these challenges and enable evaluation of the ability of DNNs to learn different regulatory architectures and of algorithms for reconstructing these patterns from the trained networks. Indeed, DeepResolve was recently proposed to interpret the combinatorial logic from intermediate layers of DNNs, and the ability of the neural network to learn the AND, OR, NOT and XOR of two short sequence patterns was demonstrated in a synthetic dataset (Liu and Gifford). However, these simulated combinatorial logics and sequence patterns were not biologically motivated and were simpler than most proposed enhancer architectures.

Here, we develop a biologically motivated framework for simulating enhancer sequences with different regulatory architectures, including homotypic clusters, heterotypic clusters, and enhanceosomes, based on real TF motifs from diverse TF families. We then apply state-of-the-

art residual neural network (ResNet) algorithms to classify these sequences and use this framework to investigate whether the intermediate layers the networks learn the complex combinatorial TF architectures present in the simulated regulatory grammars. In particular, we developed a gradient-based unsupervised clustering approach to interpret the regulatory grammar from the intermediate layers of the neural network. We evaluate the efficiency in retrieving regulatory grammar under a range of scenarios that mimic real-world multi-label regulatory sequence prediction tasks, considering possible heterogeneity in the output enhancer categories and fraction of TFBS not in the regulatory grammar. We demonstrate that ResNet can accurately model simulated regulatory grammar in many multi-label prediction tasks, even when there is heterogeneity in the output categories or a large fraction of TFBS outside of regulatory grammar. We also identified scenarios where the ResNet fails to learn an accurate representation of the regulatory grammar, including using inappropriate control sequences as negative training examples, considering output categories differing in multiple sequence features, and having an overwhelming amount of TFBS outside of the regulatory grammar. In summary, our work makes three main contributions: i) We demonstrate that the ability of DNNs to learn interpretable regulatory grammars is highly dependent on the design of the prediction task. ii) We provide a flexible tool for simulating regulatory sequences based on biologically driven hypotheses about regulatory grammars. iii) We develop and evaluate an algorithm for interpreting the regulatory grammar from the intermediate layers of DNNs trained on enhancer DNA sequences.

IV-2 Materials and Methods

IV-2.1 Data preparation for the simulated sequence analysis

IV.2.1.1 Simulation of regulatory grammar

We used TF binding motifs from the HOCOMOCO v11 database (Kulakovskiy et al. 2017). To make sure that the TF motifs are distinct and diverse, we select one TF from each TF subfamily. This results in a set of 26 TFs. Then the selected TFs are arranged into three types of regulatory grammar representing homotypic clusters, heterotypic clusters, and enhanceosomes.

For the homotypic cluster, we simulated multiple occurrences (3-5) of the same TF in a small window (120 bp) at random locations. For the heterotypic clusters, we simulated a set of four diverse TFs in a small window (120 bp) at random locations. Each TF occurs once in the heterotypic cluster. For the enhanceosome, we simulated a set of four TFs in a small window

with fixed order and spacing. Because it is possible in real enhancers that the same TF factor is used in different regulatory grammars, we allow some of TFs to occur in more than one grammar. We simulated five homotypic TF clusters, five heteotypic clusters and two enhanceosomes.

IV.2.1.2 Simulation of regulatory sequences with different sets of regulatory grammar

To mimic the common enhancer prediction tasks, such as predicting enhancers from different cellular contexts, we designed twelve regulatory sequence classes (Supplementary Table 3) with each regulatory sequence class representing one type of enhancer sequence. Sequences in each class have two different regulatory grammars. Because it is possible that the same regulatory grammar is used in regulatory sequences in different cellular contexts, we allow one regulatory grammar occur in two different regulatory sequence classes. We implemented this design by letting the regulatory sequence classes overlap in one regulatory grammar. For instance, the first regulatory sequence class has homotypic cluster 1 and heterotypic cluster 1, then the second regulatory sequence class has heterotypic cluster 1 and homotypic cluster 2 and then the third regulatory sequence class has homotypic cluster 2 and heterotypic cluster 3, etc. Next, we randomly generated a background DNA sequences of 3000 bp with equal probability of A, C, G, T and inserted 2-4 of each simulated regulatory grammar at random location into the background sequences based on the corresponding regulatory class under the assumption that in the real enhancers, multiple regulatory grammar could occur in one enhancer sequence.

IV.2.1.3 Multiclass classification and heterogenous class classification

We performed two types of classification, including multiclass classification in which each output neuron representing a homogenous set of regulatory sequences and heterogenous class classification in which each output neuron representing a heterogenous set of regulatory sequences. The heterogenous class classification tasks is based on the assumption that in the real enhancer prediction tasks, enhancers in one category, say in one specific cellular context, may have a heterogenous set of sequences harboring different sets of regulatory grammar.

The first one (multiclass classification) has twelve homogeneous output classes, each one corresponding to sequences representing one regulatory sequence class. The second one (heterogenous class classification) has five heterogeneous output classes, each one corresponding to a subset of regulatory sequence classes. More specifically, heterogeneous class 1 has regulatory sequence class 1, 3, and 5; heterogeneous class 2 has regulatory sequence class 2, 4,

and 6; heterogeneous class 3 has regulatory sequence class 7, 9, and 11; heterogeneous class 4 has regulatory sequence class 5, 8, and 10; heterogeneous class 2 has regulatory sequence class 1, 6, and 12.

IV.2.1.4 Negative set of sequences

We used three approaches to negatives when training the classifiers: no negatives, k-mer shuffled negatives, and TF-shuffled negatives. For the k-mer shuffled negative sequence set, we matched the frequency of k-mers in the negatives to the simulated regulatory (positive) sequences. For the TF-shuffled sequence set, we shuffled the TFBS of the simulated regulatory sequences.

IV-2.2 Model design and training

CNNs have achieved the state-of-art performance on regulatory sequence prediction (Zhou and Troyanskaya 2015; Quang and Xie 2016). The integration of a convolution operation into standard neural networks enables CNNs to learn common patterns that occur at different spatial positions, such as TF motifs in the DNA sequences. Here we use a residual CNN (ResNet), a variant of CNNs that allows connections between non-sequential layers (He et al. 2016) to model the regulatory sequences. Each simulated DNA sequence is represented by a sequence length x 4 matrix with columns representing A, G, C and T.

The basic layers in the network include a convolutional layer, batch normalization layer, pooling layer, and fully connected layer. Every two convolutional layers are grouped into a residual block where an identity shortcut connection adds the input to the residual block to the output of the residual block. This additional identity mapping is an efficient way to deal with vanished gradients that occur in neural networks with large depth and improves performance in many scenarios. The batch normalization layers are added after the activation of each residual block. Batch normalization (Ioffe and Szegedy 2015) helps reduce the covariance shift of the hidden unit and allows each layer of a neural network to learn more independently of other layers. The pooling layers are added after each batch normalization layer. Finally, a dense (fully connected) layer and an output layer are added at the top of the neural network. We used 4 residual blocks, each has two convolutional layers with 32 neurons. The final residual block is connected to a dense layer with 32 neurons and then connected to output layer.

We used rectified (ReLU) activation for all the residual blocks and sigmoid activation for the output fully connected layer activation. We used binary cross-entropy as the loss function and Adam (Kingma and Ba 2014) as the optimizer.

IV-2.3 Model interpretation and grammar reconstruction

IV.2.3.1 Computing saliency values with respect to neuron

We considered two approaches for estimating the importance of each nucleotide in the input sequence with respect to each neuron's activation. The first is guided back-propagation in which we calculated the gradient of the neuron of interest with respect to the input through guided back-propagation and then multiplied the gradient by input sequences. The second is calculating the DeepLIFT score (Shrikumar et al. 2017a) of the neuron of interest with respect to the input using the DeepLIFT algorithm implemented in SHAP (Lundberg and Lee 2017) against the TF-shuffled negatives and then multiplying the DeepLIFT score by input sequences. We refer the resulting values from the above as saliency values and the vector of saliency values for an input sequence as saliency map. We found that the DeepLIFT score performed the better than guided back-propagation. Therefore, for all the main text results we present were calculated with the DeepLIFT approach.

IV-2.4 Box plot, heatmap and hierarchical clustering of TF saliency maps

To analyze which TFs are learned by a specific neuron, we calculate the gradient of a TF binding site with respect to a neuron by averaging a 10 bp window from the start position of the TF binding site. Then, we visualize the distribution of saliency values of the binding sites of each TF in a specific regulatory grammar with respect to a neuron with box plot.

The median saliency values of the binding sites of each TF in a specific regulatory grammar with respect to neurons is stored in a matrix with the shape of number of neurons by the number of TFs. This data matrix is first scaled by column to identify which neurons mostly detect the TF and the scaled matrix is used to generate a heatmap. Then, we performed hierarchical clustering with k=12 (12 is the number of simulated regulatory grammars) or k=13 (when there are non-grammar TFBSs) for both neurons and TFs based on the same data matrix.

IV-2.5 t-SNE and k-means clustering

To reconstruct the regulatory grammar and evaluate how accurately neurons in a layer capture the simulated regulatory grammar, we performed a two dimensional t-SNE and a k-means clustering (k=12) of TFBS using their saliency value profiles across neurons in a layer. To assign

the name of regulatory grammar of a predicted cluster, we used a majority vote, which is the majority of the true labels of regulatory grammar in that cluster. We visualize the k-means clustering by overlaying the predicted regulatory grammar from k-means clustering on top of the t-SNE visualization. We evaluated the accuracy of reconstructing the regulatory grammar by calculating the accuracy (TP+TN/All), the sensitivity (TP/TP+FN), and the precision (TP/TP+FP) of the predicted regulatory grammar.

IV-3 Results

To evaluate the performance of residual neural networks (ResNets) on modeling the regulatory grammar, we performed a simulation analysis (Figure 1a). We designed a set of 12 biologically motivated regulatory grammars consisting of TFs from diverse families. These include five homotypic clusters of the same TF, five heterotypic clusters of different TFs, and two enhanceosomes of different TFs with requirements on the spacing and orientation of their binding sites. Motivated by the fact that enhancers active in a given cellular context likely consist of multiple types with different grammars, we designed twelve "classes" of regulatory sequences. Each class contains a different set of regulatory grammars, but the grammars can occur within multiple classes, and TFs can occur within multiple grammars. Then, using these classes, we simulated 30,000 enhancer sequences which each contain a sequence that matches the pattern defined by one of the classes (Methods).

Our goal is to evaluate the ability of ResNets to learn regulatory grammars and the ability of our proposed framework to reconstruct and visualize these grammars. Using sequences generated from the simulated regulatory grammars, we trained several models corresponding to different classification scenarios found in real-world regulatory sequence prediction tasks (Figure 1b). First, we trained a ResNet on a multi-class classification task using sequences from each of the regulatory classes and TF-shuffled negative sequences. Then, we investigated how well the approach performed when trained in more challenging situations, including no negative training sequences, k-mer matched negatives, heterogeneity in the output categories, and large fractions of TFBSs outside of regulatory grammars in the input sequences.

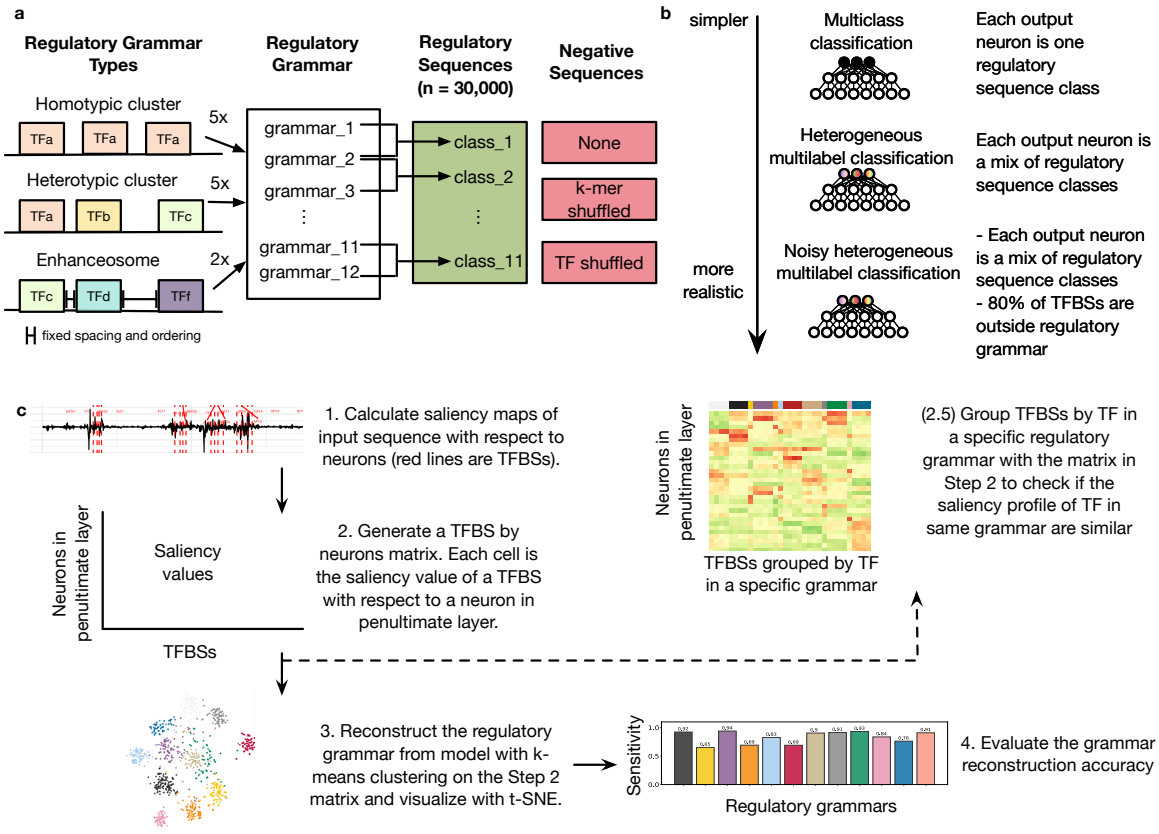

Figure IV-1 Pipeline for analyzing regulatory grammar learned by ResNet models trained on simulated regulatory sequences.Regulatory sequence and negative sequences simulation. Based on the current hypotheses of the types of regulatory grammar, we designed twelve regulatory grammars, including five homotypic clusters, five heterotypic clusters, and two enhanceosomes as prototypes for simulated regulatory sequences. Note that the same TF can be in more than one regulatory grammar. Then, to reflect that regulatory regions active in a cellular context may have multiple grammars, we defined twelve regulatory sequence classes, each with two different grammars. Each regulatory grammar is shared between two classes. Finally, we generated two sets of negative sequences: k-mer shuffled and TF shuffled versions of the simulated positive sequences. (b) Classification tasks. ResNets are trained on simulated regulatory sequences and the negative sets in three increasingly realistic scenarios: 1) multiclass classification in which each output neuron corresponds to one simulated regulatory sequence class, 2) heterogenous multi-label classification in which each output neuron corresponds to a mix of regulatory sequence classes, and 3) noisy heterogenous multi-label classification in which 80% of TFBSs are simulated to occur outside of grammars. (c) Regulatory grammar reconstruction framework. We first calculate the saliency maps of input sequences with respect to the neurons in the penultimate layer. Then we extract the saliency values for each TFBS instance to generate a TFBS by neuron matrix. We then cluster the TFBS by neuron matrix and assign predicted regulatory grammar membership to TFBSs based on the cluster membership. Finally, we quantify the reconstructed grammar by calculating the accuracy, sensitivity, and specificity of the predicted regulatory grammar membership.

We also implemented an intermediate step (2.5) to check if saliency profiles of TFs in the same grammar are similar, which is an indicator whether the model learned the regulatory grammar. In this intermediate step, we group TFBSs by TF into a specific regulatory grammar and then perform hierarchical clustering of the TFs.

IV-3.1 ResNet trained on simulated regulatory sequences and TF-shuffled negatives accurately captures simulated regulatory grammars.

To explore if the ResNet model can learn the regulatory grammar, we started with a multi-class classification task based on simulated regulatory sequences from 12 classes and TF-shuffled negative sequences (Methods; Supplementary Table 1 and 2). We trained a classifier to predict the class of the sequence, either not a regulatory sequence or member of one of the regulatory sequence classes. By constructing the prediction task with TF matched negative sequences, the neural network is forced not only to learn the individual TF motifs, but also learn the combinatorial patterns between the TFs.

The ResNet model accurately predicts the class label of input DNA sequences with near perfect performance: average area under the ROC curve (auROC) of 0.999 and average area under the precision-recall curve (auPR) of 0.982. We then analyzed what features were learned by calculating saliency maps (Methods) of input sequences with respect to each neuron in the penultimate layer (the dense layer immediately before the output layer). We found that neurons in the penultimate layer detect the location of the simulated TFBS. For instance, when we compute the saliency map of a class 6 simulated regulatory sequence with respect to neuron 1 in the penultimate layer, the TFBS have higher saliency value compared to other locations in the sequence, indicating the higher importance of those nucleotides to the activation of neuron 1 (Figure IV-2a).

Next, we visualized the features learned by neuron 1 of the penultimate layer by plotting the mean saliency value of a 10 bp window from the start of each TF binding site using 240 sequences from all simulated regulatory sequence classes (Figure IV-2b). For example, the TFBS from heterotypic cluster 3 have elevated gradients compared to TFBS from other simulated regulatory grammars. This suggests that neuron 1 of the penultimate layer detects TFBS from heterotypic cluster 3. We then took the median gradients of TFBSs in a specific regulatory grammar and generated a matrix with rows of neurons and columns of each TF. We scaled the matrix column-wise. This scaling helps identify which neurons recognize the TF. We plotted the

scaled matrix as a heatmap with hierarchical clustering (Method; Figure IV-2c). We found that: (i) TFBSs from the same regulatory grammar have elevated gradients together and therefore are clustered; (ii) neurons of the penultimate layer can "multi-task", that is, one neuron can detect one or more regulatory grammars. This suggests that the penultimate layer captured the simulated regulatory grammars.

In order to evaluate how well the regulatory grammar can be reconstructed from the penultimate layer, we performed unsupervised clustering of TFBS based on their saliency values with respect to the neurons in the penultimate layer. More specifically, we performed a k-means clustering (k=12) of TFBSs from 240 sequences using their gradients with respect to each neuron of the penultimate layer and visualized it with t-SNE (Figure IV-2d). Each TFBS has a predicted clustering label that is assigned by the k-means clustering algorithm and a true regulatory grammar. We first used majority voting to determine the predicted regulatory grammar for a cluster. For instance, the majority of cluster 1 is from heterotypic cluster 1, so we assign heterotypic cluster 1 as the predicted regulatory grammar for all TFBS in cluster 1. We then calculate the accuracy of the regulatory grammar reconstruction by comparing the predicted regulatory grammar and the true regulatory grammar. On average, 85.1% of TFBS are correctly classified (Figure IV-2e), and homotypic clusters are learned better (sensitivity > 0.97) than heterotypic clusters and enhanceosomes.

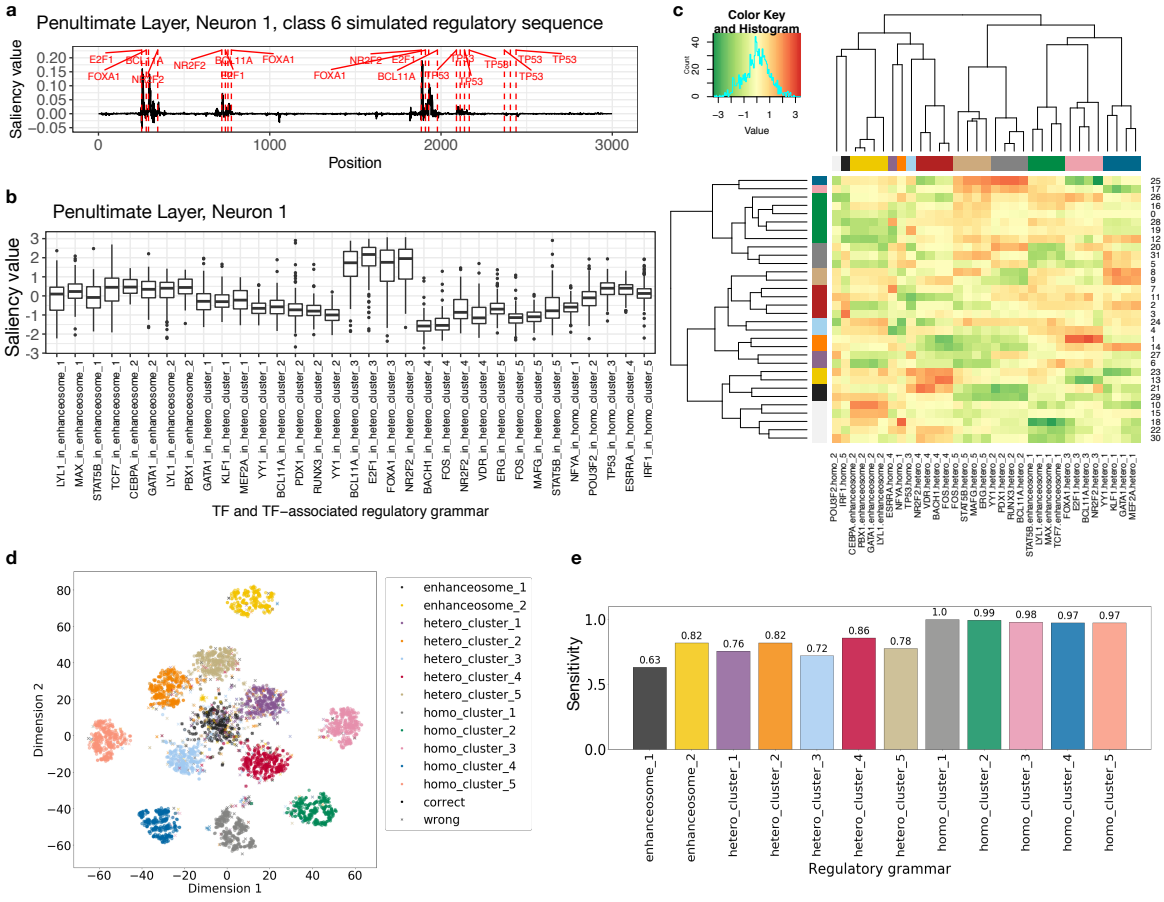

Figure IV-2. ResNet trained on simulated regulatory sequences and TF-shuffled negatives accurately models the regulatory grammar. The saliency map of simulated regulatory sequence from class 6 with respect to neuron 1 in the penultimate layer. (b) The saliency values of the binding sites of each TF in a specific regulatory grammar with respect to neuron 1 in the penultimate layer. (c) Heatmap of the median gradient of the binding sites of each TF in a specific regulatory grammar (x axis) across neurons of the penultimate layer (y axis). The order of x and y axis labels are determined by hierarchical clustering. The color bars on the side indicate the group label assigned by hierarchical clustering. (d) Actual labels of simulated regulatory grammar of the TFBS overlaid on t-SNE visualization of TFBS saliency values across neurons. Correct predictions of the regulatory grammar for a TF is represented by a dot, that is the predicted label agree with the actual label. Incorrect predictions of the regulatory grammar of a TF are indicated by crosses. (e) The sensitivity (TP/TP+FN) of the regulatory grammar predictions.

The same analysis approach can be applied to any layer of the neural network. We found that the neural network built up its representation of the regulatory grammar by first learning the

individual TF motifs in the lower level neurons and gradually grouping TF motifs in the same regulatory grammar together (Figure IV-3).

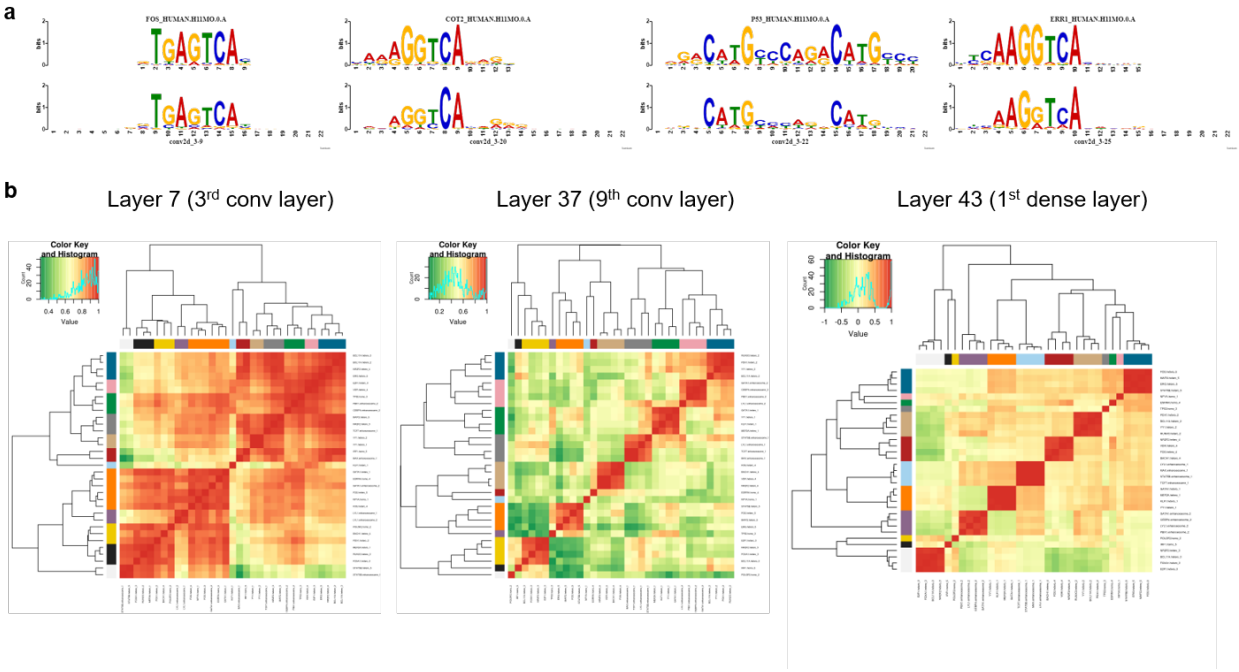

Figure IV-3. ResNet learned individual TF binding motifs in the lower convolutional layer and gradually build up its understanding of regulatory grammar in higher level layers. (a). Simulated TF motifs are learned by neurons in the third convolutional layer. From left to right are four selected examples, neuron 9 learned the FOS motif; neuron 20 learned the COT2 motif; neuron 22 learned the P53 motif; neuron 25 learned ERR1 motif. (b). From layer 7 (third convolutional layer) to Layer 43 (the penultimate dense layer), the ResNet gradually learned the regulatory grammar. The correlation matrix of the saliency value profiles of TFs in a specific regulatory grammar is plotted as the heatmap. In layer 7, TFs from the same regulatory grammar are not clustered. In layer 37, TFs within the same regulatory grammar begin to have a higher correlation. In layer 43, TFs within the same regulatory grammar have near perfect correlation.

Taken together, these results demonstrate that ResNet models can largely capture simulated regulatory grammars if trained to perform a multi-class prediction with TF-shuffled negatives, and that our unsupervised clustering method based on saliency maps is able to reconstruct the regulatory grammar.

IV-3.2 Regulatory grammar can be learned by the ResNet model without TF-shuffled negatives

Although the ResNet model demonstrated the ability to capture the simulated regulatory grammars when trained against TF-shuffled negatives, we cannot construct perfect TF-shuffled

negatives in the real-world, because the true TFs are not known. Indeed, in many applications, only the positive regulatory sequences (Zhou and Troyanskaya 2015; Quang and Xie 2016; Zhou et al. 2018) or k-mer shuffled negatives are used for training machine learning models. Therefore, we tested whether the ResNet model can learn the simulated regulatory grammar if trained with no negatives or k-mer shuffled negatives.

We trained five models for multi-class classification against: no negatives, 1-mer shuffled negatives, 4-mer shuffled negatives, 8-mer shuffled negatives, and 12-mer shuffled negatives. Then, we evaluate their performance at predicting simulated regulatory sequences. The model trained with 8-mer shuffled negatives achieved the highest accuracy at distinguishing TF-shuffled negatives from simulated regulatory sequences (average auROC 0.998, auPR 0.957, Figure IV-4a).

To further explore the regulatory grammar learned by the ResNet model trained against 8-mer shuffled negatives, we calculated saliency maps over a set of input sequences (n=240) from each class of simulated regulatory sequences with respect to neurons in the penultimate layer. We performed hierarchical clustering on the median gradients for the binding sites for each TF in a specific regulatory grammar as we did in the previous results section. We found that TFBS from the same regulatory grammar were grouped together. Next, we performed k-means clustering (k=12) of the TFBS from the 240 sequences and overlaid the clustering label on the tSNE visualization (Figure IV-4b). We calculated the accuracy of predicted regulatory grammar for each TF. The average grammar reconstruction accuracy of this model is on par with the model trained against TF-shuffled negatives (85.3% vs. 85.1%).

These results suggest that the model trained against 8-mer shuffled negatives can learn a good representation of the regulatory grammar and therefore 8-mer shuffled negatives can be used as a substitute for TF-shuffled negatives in practice.

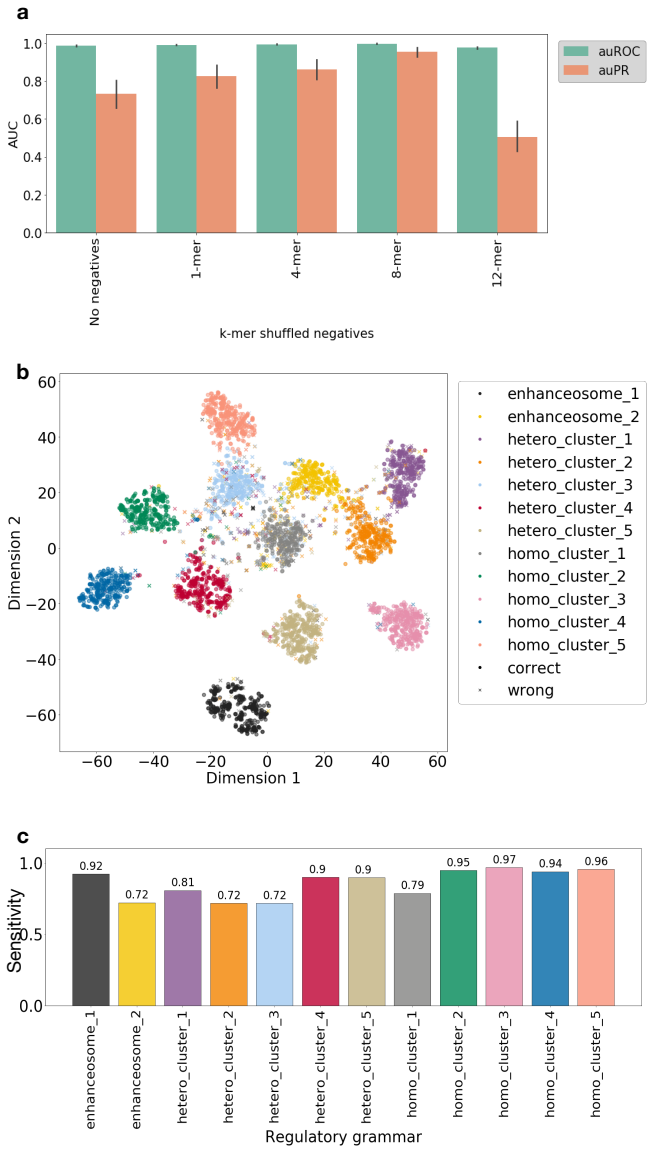

Figure IV-4. ResNet trained on simulated regulatory sequences against 8-mer shuffled negatives accurately models the regulatory grammar. (a) The performance of five different ResNet models trained on simulated regulatory sequences against different k-mer shuffled negatives at predicting the regulatory class of the simulated regulatory sequences vs. TFs-shuffled negatives test dataset. (b) Actual labels of simulated regulatory grammar of the TFBS overlaid on t-SNE visualization of TFBS saliency values across neurons. (c) The sensitivity of predicted labels in (b) of the ResNet model trained on the simulated regulatory sequences against 8-mer shuffled negatives.

IV-3.3 Regulatory grammar can be learned by the ResNet model in the presence of heterogeneity in the regulatory sequences

A common task in regulatory sequence prediction is to predict regulatory sequences that exert a certain set of functions, e.g., sequences active in different cellular contexts. For instance, in the DeepSEA model, some of the predicted epigenetic marks are H3K27ac peaks—a marker for active regulatory regions—from different cellular contexts. It is likely that sequences with a heterogeneous set of grammars are active in each cellular context.

To mimic this type of heterogeneity, we performed a heterogenous multi-label classification by pooling a number of simulated regulatory classes together as one heterogeneous class to generate five heterogeneous classes (Method; Figure IV-1b). We also allowed one regulatory class to be used in several heterogeneous classes. For example, in our simulation, regulatory sequences in heterogenous class 1 consist of regulatory class 1, 3, and 5. Regulatory class 1 sequences also belong to heterogenous class 5, and regulatory class 5 sequences also belong to heterogenous class 4. This multi-function of a regulatory sequence class is often observed in real-word regulatory sequences as many enhancers are active in more than one cellular context.

We trained the ResNet model against k-mer shuffled negatives (k=1, 4, 8, 12). Again, the model trained against 8-mer shuffled negatives performed the best when evaluated against the TF-shuffled negatives (average auROC 0.99, auPR 0.93). We performed hierarchical clustering and unsupervised clustering (Figure IV-5a, b) as we did in the previous sections. The model trained to predict the heterogenous classes can still learn the majority of the regulatory grammars. The average accuracy of reconstructing regulatory grammar in this setting is 89.2%, which is similar to that of the multi-class classifications against TF-shuffled negatives (85.1%) and against k-mer shuffled negatives (85.3%).

These results suggest that the model trained on regulatory sequences with heterogenous output categories can still largely capture the regulatory grammars that are essential for the heterogenous multi-label classification.

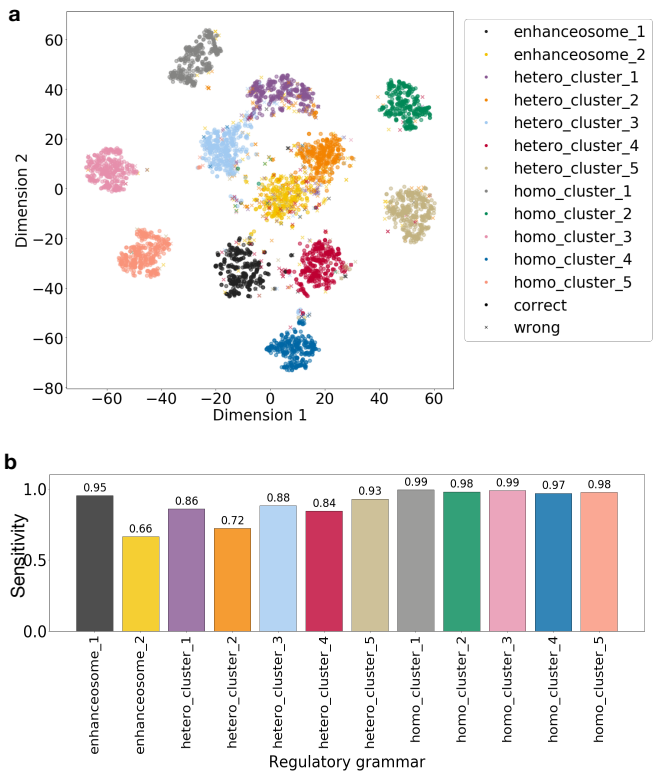

Figure IV-5. Regulatory grammar can be learned by ResNet despite heterogeneity in the regulatory sequences. Actual labels of simulated regulatory grammar of the TFBSs overlaid on t-SNE visualization of TFBS saliency values across neurons. (b) The sensitivity of predicted labels in (b) across regulatory grammars.

IV-3.4 Regulatory grammar can be learned by ResNet when a large fraction of TFBSs are not in grammars and there is heterogeneity in the regulatory sequences

In all previous prediction tasks, the simulated TFBSs in the input sequences are always in a regulatory grammar. However, in the real regulatory sequences, it is likely that only a fraction of TFBS are in regulatory grammars, while others are individual motifs scattered along the sequence. To mimic this scenario, we simulated a set of regulatory sequences with 80% of TFBSs randomly scattered in the sequence outside of any regulatory grammar and 20% of TFBSs in regulatory grammar.

We trained a ResNet model on this 80% non-grammar TFBSs dataset with the five heterogenous classes as output categories against 8-mer shuffled negatives. We found that the TFBSs outside

of the regulatory grammars (single TFBS) have lower saliency values compared to the TFs in simulated regulatory grammars (Figure IV-6a) except for those in enhanceosome 2.

Next, we performed unsupervised clustering analysis as in the previous sections (Figure IV-6b). Although the TFBSs in regulatory grammars still cluster, many of the TFBSs outside of regulatory grammar overlap the TFBSs in regulatory grammars in t-SNE space. This makes identifying the regulatory grammars challenging. To better reconstruct the regulatory grammar from the unsupervised clustering analysis, we took advantage of the fact that the non-grammar TFBSs have lower saliency values and only kept the TFBSs with top 10% sum of saliency values across neurons in the penultimate layer. Intuitively, this filtering helps improve the reconstruction of regulatory grammar by only focusing on TFBSs with high influence on the prediction. We repeated the unsupervised clustering analysis on these filtered TFBSs (Figure IV-6c). We found that nearly all TFBSs outside of regulatory grammars are filtered out (97.7%) and a smaller fraction of TFBSs in regulatory grammars are filtered (59.3%). After filtering, the remaining TFBSs are sufficient to reconstruct 11 of the 12 simulated regulatory grammars. The regulatory grammar that we failed to reconstruct, enhanceosome 2, has the lowest sum of saliency values across neurons in the penultimate layer (Figure IV-6a), suggesting their lower importance for accurate predictions.

These results suggest that even with only a small fraction of TFBSs in regulatory grammars and heterogeneity in the output categories, we can still reconstruct most of the simulated regulatory grammars.

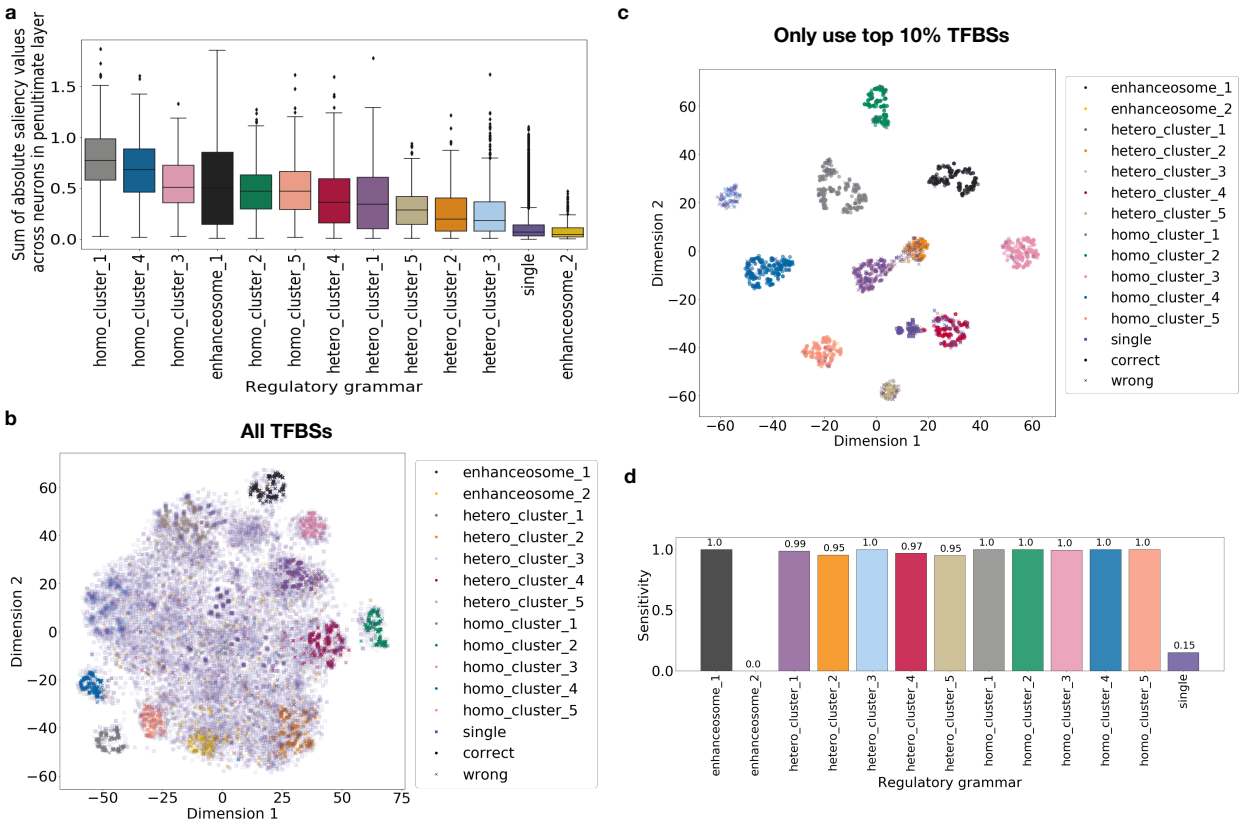

Figure IV-6. Regulatory grammar can be learned by ResNet when TFBSs are outside of regulatory grammars and there is heterogeneity in the regulatory sequence categories. Sum of saliency values for TFBSs in each regulatory grammar across neurons in penultimate layer (b) Actual labels of simulated regulatory grammar of the TFBS overlaid on t-SNE visualization of TFBS saliency values across neurons. (c) Actual labels of simulated regulatory grammar of the TFBS filtered to only those in the top 10% sum of saliency values across neurons in penultimate layer overlaid on the t-SNE visualization. (d) The sensitivity of predicted labels in (c) across regulatory grammars.

IV-3.5 Regulatory grammar cannot be learned if multiple grammars are able to distinguish one regulatory sequence class from another

As shown in Figure IV-4 and Figure IV-5, some regulatory grammars, especially enhanceosome 2, are reconstructed from ResNet model with limited accuracy. This suggests that the essentiality of a regulatory grammar may influence the ability to reconstruct regulatory grammars from the model. To further investigate this hypothesis, we simulated three heterogenous regulatory classes (Table 1) with non-overlapping subsets of regulatory grammars, so that multiple regulatory grammars could distinguish one heterogenous regulatory class from another. Then we trained the

model against TF-shuffled negative sequences. By setting up the training this way, the model will have to distinguish sequences with TFBSs in regulatory grammars from those with TFBSs not in regulatory grammars. However, the model does not need to learn all the regulatory grammars or distinguish one regulatory grammar from the other to make accurate predictions.

Table 1. Simulated heterogenous regulatory sequence classes with multiple regulatory grammars that can distinguish one class from another.

	Regulatory grammar used in the first type of sequence	*Regulatory grammar used in the second type of sequence*
Heterogeneous Regulatory Sequence Class 1	homotypic cluster 1, homotypic cluster 2	homotypic cluster 4, heterotypic cluster 4
Heterogeneous Regulatory Sequence Class 2	heterotypic cluster 1, heterotypic cluster 2	homotypic cluster 5, enhanceosome 1
Heterogeneous Regulatory Sequence Class 3	homotypic cluster 3, heterotypic cluster 3	heterotypic cluster 5, enhanceosome 2

As expected, the model performed well at distinguishing positives and negatives (average auROC 0.995, auPR 0.978). However, when visualizing the saliency values of TFBSs of the neurons in the penultimate layer, there is limited resolution to recover individual regulatory grammars; multiple regulatory grammars have similar saliency profiles and overlap in the t-SNE space (Figure IV-7a). More specifically, the grammars that co-occur in the same regulatory sequence classes tend to cluster together. For example, in regulatory sequence class 1, homotypic cluster 4 clustered with homotypic cluster 1; in regulatory class 2, homotypic cluster 5 clustered with heterotypic cluster 2 and heterotypic cluster 1; in regulatory sequence class 3, homotypic cluster 3 clustered with heterotypic cluster 5. However, the remaining regulatory grammars, including homotypic cluster 2, heterotypic cluster 3, heterotypic cluster 4, enhanceosome 1, and enhanceosome 2, are scattered in the t-SNE visualization (Figure IV-7a). The regulatory grammars that are scattered show lower sum of saliency values across neurons, suggesting lower attention they received from the neural network (Figure IV-7b). This observation is consistent with our hypothesis that if there are multiple regulatory grammars that can distinguish one class of sequences from another, the neural network will not learn to distinguish one regulatory grammar

from another nor learn all the distinct regulatory grammars. This scenario is likely to happen in many real enhancer classification tasks and would make reconstruction of individual regulatory grammars challenging.

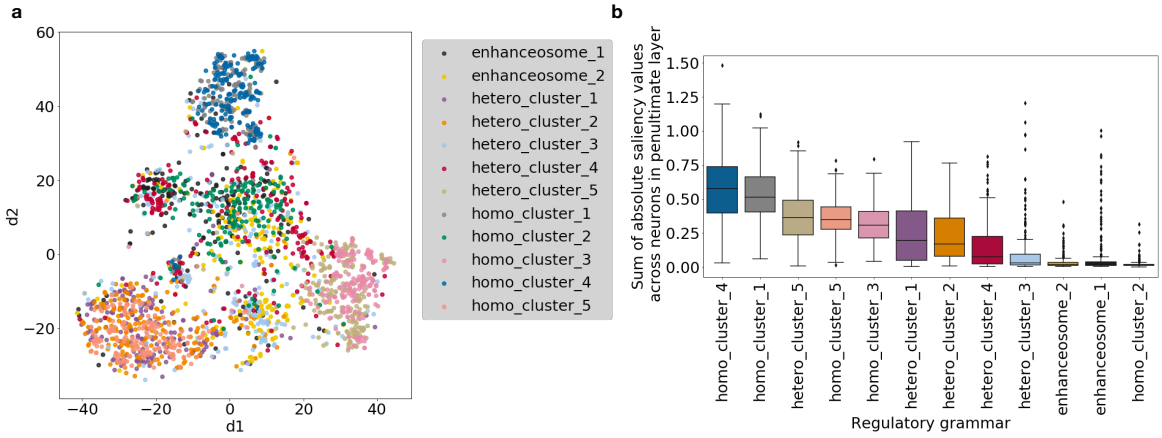

Figure IV-7. The ResNet model fails to learn the correct representation of individual grammars when there are multiple regulatory grammars that can distinguish one heterogenous regulatory class from another. a) Actual labels of simulated regulatory grammar of the TF binding sites overlaid on t-SNE visualization of TFBS saliency values across neurons. b) Sum of saliency values for TFBSs in each regulatory grammar across neurons in the penultimate layer.

IV-3.6 ResNet trained on developmental heart enhancers failed to capture the known heart heterotypic clusters

Transcription factors TBX5, NKX2-5, and GATA4 function as a heterotypic cluster and coordinately control cardiac gene expression and differentiation (Luna-Zurita et al. 2016). To test if the neural network can learn the heterotypic cluster from known heart enhancers, I trained a residual neural network with enhancers from three stages of mouse heart development, including embryonic stem cells (ESC, N=6359), mesoderm (MES, N=4775), cardiac precursors (CP, N=5549), and cardiomyocytes (CM, N=6894), against 8-mer shuffled negatives (Wamstad et al. 2012). I used a Bayesian hyperparameter search approach to select the best number of neurons in each layer and the number of layers. The neural network achieved only moderate accuracy for developmental stage prediction (ROC AUCs of 0.79, 0.72, 0.62, and 0.64, PR AUCs of 0.70, 0.42, 0.36, and 0.47 for ESC, MES, CP, and CM enhancers). The lower accuracy at predicting CM and CP stage enhancers reflects the similarity between those two stages.

Next, we calculated saliency maps for the top 150 sequences for the CM enhancer category using the 8-mer shuffled negative sequences as reference with respect to the neurons in the penultimate layer. We chose CM stage because it has largest overlap with TF binding sites of TBX-5, NKX2-5, and GATA4 among the four stages (Wamstad et al. 2012). We then used FIMO to identify transcription factor binding site in those sequences using HOCOMOCO mouse motif database (Grant et al. 2011; Kulakovskiy et al. 2016). We then visualized the saliency profile of the representative TF from each TF family (we select the TF with median motif counts within a TF family) and also TBX5, NKX2-5, GATA4, and MEIS1 with t-SNE. We include MEIS1 because it has been suggested to be similar to the in vivo binding motifs for TBX5 in hear developmental enhancers (Luna-Zurita et al. 2016). However, TBX5, NKX2-5 and GATA4 are not clustered, suggesting that the neural network did not learn this heterotypic cluster (Figure IV-8). Several factors could have prevented the neural network from learning the heterotypic cluster. First, the model has only moderate accuracy at predicting enhancers from different stages. This suggests that the representation learned by the neural network is not very accurate. Second, as in the previous simulation analysis, it is likely that many sequence features can be used to distinguish enhancers from one stage to another so that it is not necessary for the neural network to learn the full heterotypic cluster. These results illustrate the difficulties in learning the correct representation of regulatory grammar through neural networks in common real enhancer prediction tasks.

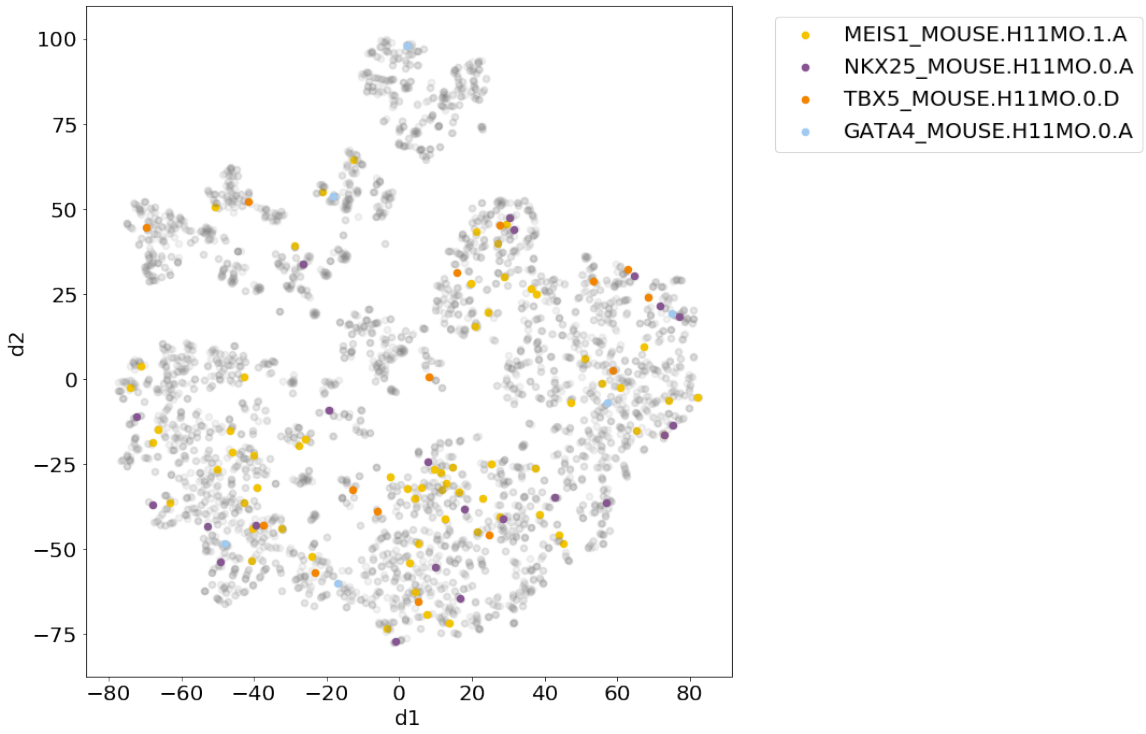

Figure IV-8. ResNet trained on developmental heart enhancers did not learn the heterotypic cluster of TBX-5, NKX2-5, and GATA4. TF binding sites overlaid on t-SNE visualization of TFBS saliency values across neurons. The colored dots are motifs that either belong to TBX5, NKX2-5, and GATA4 or motifs that have been suggested to correlate with TBX5 binding (MEIS1).

IV-4 Conclusion and discussion

We trained a variant of CNNs, ResNets, to model sequences with simulated regulatory grammars (combinatorial binding of TFs). Then we developed a gradient-based unsupervised clustering approach to interpret the features learned by neurons in the intermediate layers of the neural network. We found that ResNets can model the simulated regulatory grammars even when there is heterogeneity in the regulatory sequences and a large fraction of TFBSs outside of regulatory grammars.

We also identified scenarios when the ResNet model failed to learn the regulatory grammar. The networks strive to learn simple representations of the training data. As a result, the ResNet models in our studies failed to learn the simulated regulatory grammar when there is a lack of constraint in negative training samples or between the positive output categories. For instance, we found that the choice of negative training samples influences the ability of the neural network to learn regulatory grammar. The model trained against no negatives or short k-mer shuffled

negatives (k=1-6) or very long k-mer shuffled negatives (k=12) did not learn accurate representation of the regulatory grammar and often misclassified TF-shuffled negatives as positives. The model trained against 8-mer shuffled negatives performed the best when evaluated on the TF-shuffled negatives. This is because when shorter k-mers (k=1-6) are used to generate the negative training samples, the neural network can distinguish the positives from negatives by learning the individual TF motifs, many of which are longer than 6 bp, rather than learning the regulatory grammar of the TFs. With longer k-mers (k=12), the reason is likely that k-mers are not well shuffled in the negatives and very similar to the positives. Indeed, the ResNet model trained against 12-mer shuffled negatives has a low accuracy (auPR 0.506). The 8-mer shuffled negative provides a sweet spot where the negatives are well shuffled and the network is forced to learn the TF motifs and regulatory grammars. Another challenging situation occurs when there are multiple sequence features that can distinguish one output category from another. Under this scenario, it is not necessary for the neural network to accurately learn all the features nor distinguish one feature from another.

In addition to these scenarios, there is also another situation in which the ResNet model failed to learn the regulatory grammar. When the majority of the TFBSs are not in a regulatory grammar, the non-grammar TFBSs overlap those in regulatory grammars in the unsupervised clustering analysis and make it impossible to recover the grammars. Fortunately, we could use the observation that many of the TFBSs outside of regulatory grammars have low saliency values to filter out those TFBSs, and focus the unsupervised clustering analysis on TFBS with high saliency values to improve the accuracy of grammar reconstruction. This gradient magnitude-based filtering method may be less efficient when there is an overwhelmingly large number of TFBSs outside of regulatory grammar and larger sample sizes might be needed to train the neural network to better retrieve the regulatory grammars.

While we demonstrate potential to interpret biologically relevant patterns learned by deep neural network models in some realistic scenarios, our work has several caveats. First, the synthetic dataset and proposed methods assume that combinatorial binding of TFs does not change their motifs. However, this assumption may not be always true. In vitro analyses of the combinatorial binding of pairs of TFs indicate that many pairs of TFs have different binding motifs when they bind together compared to their consensus motifs (Jolma et al. 2015). Although there is nothing preventing the neural network from learning such altered motifs, the

unsupervised clustering methods based on individual TFBS may have limited accuracy in identifying such altered motifs. Second, we did not simulate noisy labels in the synthetic dataset which could occur in the real regulatory sequence prediction tasks. The common methods of experimentally finding enhancers, such as ChIP-seq on histone modifications, DNase-Seq, CAGE-seq, and MPRAs, often produce mislabeled regulatory regions and vague region boundaries. This could be improved in the future by integrating methods for learning from noisy labeled data.

In summary, we demonstrated the power and limitation of convolutional neural network at modeling the regulatory grammar and provided a backpropagation gradient based unsupervised learning approach to retrieve the learned regulatory grammar from inner layers of the neural network.

Chapter V

Conclusions and future directions

In this dissertation, I investigated the mechanism underlying the rapidly evolving regulatory landscape of enhancers and evaluated the capability of state-of-the-art deep neural networks at modeling enhancer sequence architectures. In Chapter II, I demonstrated differences in the functional activity across cells between highly conserved enhancers and enhancer with species-specific regulatory activity. Defining the conservation of enhancers that are alignable across ten mammalian species based on active histone modification marks (H3K27ac and H3K4me1) in primary liver tissue, I found that the conserved-activity enhancers had higher density and diversity of TF binding sites, more target genes and more broadly active target genes and were active in more cellular contexts comparing to species-specific enhancers. These pieces of evidence suggest that highly conserved enhancers are more pleiotropic and under more evolutionary constraint. In Chapter III, I investigated the evolution of enhancer sequence properties across mammalian species. I demonstrated that enhancer sequence k-mer spectrum SVM models trained on enhancer sequences from one species could be applied to predict enhancers in another species with great accuracy in adult liver, developing limb, and developing brain tissues. Furthermore, the top predictive k-mers in species-specific enhancer SVM models matched a common set of binding motifs for TFs enriched for expression in relevant tissues. This suggests the overall conservation of the enhancer sequence properties over 180 million years of mammalian evolution even though the enhancer activity of specific loci often changes between closely related species. I also applied convolutional neural networks (CNN) to the cross-species prediction framework. I found that the CNN predicted enhancers with better accuracy, but worse generalization across species. It has been hypothesized that the higher layer neurons of CNNs capture the combinatorial effects of transcription factor binding. The better within-species performance and worse cross-species performance of CNNs compared to SVM models motivated me to investigate whether the CNNs have the capability to learn such "regulatory grammar" and whether we can extract such information from the neural networks. Therefore, in Chapter IV, I created synthetic dataset of enhancers with simulated regulatory grammars based on previous hypothesis of combinatorial TF binding rules, such as homotypic clusters,

heterotypic clusters, and enhanceosomes, with real TF motifs from all TF families to test the capability of CNNs to learn regulatory grammars. I also created scenarios mimicking the current common tasks of enhancer prediction, such as predicting heterogenous set of enhancers active in a specific cellular context, having the majority of TF binding sites outside of regulatory grammar, and training against differently created negative sets. As a proof of concept, I demonstrated that the CNNs are capable of learning the correct association of TFs in the regulatory grammar when trained against TF-switched negatives, that is the positions of TFs are maintained but the specific motifs simulated at a position is randomized, using the backpropagation gradient based feature importance score. Moreover, I found that it is common for neurons in CNNs to multitask and learn multiple regulatory grammars and only when considering the activation pattern of the whole layer can I recover the individual regulatory grammar. Next, I demonstrated that the CNNs were capable of capturing regulatory grammar under more realistic situations—output categories containing a heterogenous set of enhancers and input sequences, and/or input sequences have large amount of non-grammar TFs. Finally, I identified situations where CNNs failed to capture accurate representation of individual regulatory grammars. This happens if the model is not trained against constrained negative samples like TF-switched negatives or k-mer shuffled negatives, or if the output categories are not constrained, e.g., when multiple sequence features can distinguish one output category of enhancers from another. Neural network models strive to learn the simplest representation for accurate predictions. This result suggests that in most current trained regulatory sequence models, it is likely that the neural network did not learn accurate, biologically relevant representations of individual higher-order TF interactions, because of lack of constraint in the output labels of enhancers or lack of a stringent set of negative sequences.

V-1 Mechanisms of enhancer turnover and redundancy

In Chapter II and III, I investigated the underlying mechanisms of rapid enhancer evolution. I mainly focused on the function and sequence properties of individual enhancer elements. However, studies suggest that the nearby regulatory landscape play an important role on the turnover of the regulatory elements. For example, genes with complex regulatory landscapes exhibit high expression levels that remain evolutionarily stable and conserved regulatory activity associates with high and evolutionarily stable gene expression (Berthelot et al. 2017). Their

observation of correlation between evolutionally stable gene expression and the presence of many regulatory elements regardless of their conservation is consistent with the idea of redundancy among regulatory elements. Another recent study used genome editing to create 23 mouse deletion lines and inter-crosses, including both single and combinatorial enhancer deletions at seven distinct loci required for limb development (Osterwalder et al. 2018). They found that none of that the ten deletions of individual enhancers caused noticeable changes in limb morphology but removal of pairs of enhancers showed effects, also suggesting the redundancy of enhancer functions.

These pieces of evidence suggest that the regulatory elements in the same regulatory neighborhood function in synergy. It will be interesting to investigate the interplay between the function of gene, the density of regulatory elements, the TF composition of the regulatory elements, and the evolutionary conservation. More specifically, one can ask the following questions: If one enhancer is lost during evolution, would there be replacement enhancers? Would this answer be different based on the density of existing regulatory element and the function or expression conservation level of target genes? If there were replacement enhancers, would the new enhancers have the similar TF composition as the lost one or does only the number of enhancers matter? Is there a location preference for the new enhancers? And also questions about enhancer gain. This question is partially addressed by an analysis of recently evolved enhancers that suggested that recently evolved enhancers contributed only weakly to gene expression (Berthelot et al. 2017). However, it hasn't been analyzed in close examination of TF composition similarities with other nearby regulatory elements. For example, if one enhancer is gained, does it compensate for the loss of others with similar TF composition? Would the answer change based on the density of existing regulatory elements and the function or expression conservation level of target genes? Can we find cases of recently evolved enhancers contributed greatly to gene expression? If so, what genes they are affecting? What sequence characteristics do they have?

V-2 Towards more accurate and interpretable machine learning model of regulatory sequences

In this Chapter IV, I analyzed the capability of deep neural networks to learn representations of the regulatory grammar—the combinatorial binding rules of transcription factors—under different scenarios with both simulated and real enhancer data. Deep neural networks showed

limited power at learning individual regulatory grammars when there is not enough constraint in the training sequences and output enhancer categories. This observation is different from the neural network trained with millions of images where the neurons in the inner layers of the network have been demonstrated to learn human-recognizable features. This could due to several reasons. First, the range of values of one hot encoded DNA sequence matrix for neural network training is binary, either 0 or 1, while pixels can take a wider range of values in image classification. This could make some interpretation methods that are efficient for image based neural network not efficient for DNA sequences. For example, when I used gradient ascent to optimize the input sequence that maximally activate a neuron, the input sequence was often trapped in local minimum and different initialization generated different final optimal input sequence with limited information content. Moreover, the binary nature of the input DNA sequence matrices could make the training process different too. For example, it may be easier for the neurons of sequence-based neural network to multitask. Second, the output categories of regulatory sequences are often not as homogenous as those in the image classification tasks. In image classification, the output categories are usually very distinct classes of objects that share same set of properties and very different from other classes. For example, cat images will usually have ears, eyes, noses, whiskers, legs, tails, and a body. However, this is not true with most of the regulatory sequence classifications. For example, the enhancers active in heart may have different sets of TF binding sites and some of the heart enhancers may also be enhancers in another tissue, say, brain. Third, the data are noisy. The current labels of regulatory sequences are usually from genome-wide functional genomics experiments, which likely have many false positives and false negatives. Finally, the size of data for neural network training is important. ImageNet has 14,197,122 well-defined images in 21841 synsets. While in DNA sequences, we usually have limited noisy annotations (currently the maximum number of annotation used is 2002, (Zhou et al. 2018)). More data could give more constraint to the neural network and encourages it to learn an accurate representation of features.

There are improvements we can implement to solve some of the above-mentioned problems. The first is improving the neural network to learn with noisy labels. Deep neural networks are so powerful that they can “memorize” noisy data in the training set and overfit. Therefore, it is important to train neural networks that are robust to noisy labels. Goldberger et al. used an additional softmax layer that connects the correct labels to the noisy labels to estimate the noise

transition matrix. Under the intuition that as the labels become less noisy, the performance of the neural network would improve, Jiang et al. proposed MentorNet, which pre-trained an neural network for selecting clean instances to guide the actual training (Jiang et al. 2018). Arplt et al. demonstrated that neural networks start with learning easy samples and gradually adapt to hard instances in the later stages of training (Arplt et al. 2017). This observation could also be used to stop the training early to focus on clean training examples. Han et al. proposed a co-teaching paradigm to solve this problem, which trained two peer neural networks to teach each other the errors in the data (Han et al. 2018). Applying these techniques to the deep learning of regulatory sequences would improve the accuracy of the network as well as identify false positives in enhancer datasets.

Another way to possibly improve performance is to use a complex but interpretable learning algorithm. Poon et al. proposed a fully traceable deep leaning network, the Sum-Product network (SPN), which has full probabilistic semantics and tractable inference over many layers and is better at stating rules expressively than convolutional neural networks (CNNs) or recurrent neural networks (RNNs) (Poon and Domingos 2011). SPNs have not been used in modeling regulatory sequences, but they outperform CNNs and RNNs in many tasks (Poon and Domingos 2011; Gens and Domingos 2012). The following is a potential structure of SPNs that could be used to train enhancer sequence predictors. There are two classes of sequences in this task: S_1 are enhancers and S_2 are non-enhancers. The regulatory sequences may consist of parts (multiple regulatory units), and then each part is a mixture of subtype of parts. For each subtype of parts, it may consist of smaller parts, which is a mixture of subtypes of smaller parts, and at the end would be a 4 x W filter to learn the individual transcription factor motifs. In the SPNs, the decomposition of parts is performed by the product nodes and the clustering of subtypes is performed by the sum nodes. Given the deep but fully traceable structure, SPNs have the potential of being accurate and interpretable at same time for enhancer sequence modeling.

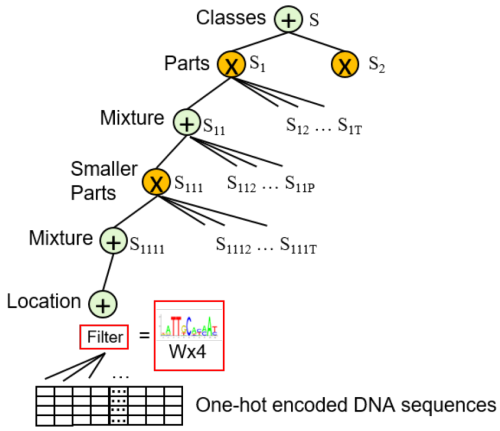

Figure V-1. The structure of a sum-product network trained with enhancer sequences

V-3 Towards disease prediction from genetic data for individuals

The investigation of enhancer evolution and enhancer sequence architecture performed in this dissertation are efforts towards understanding the non-coding genome to facilitate the interpretation of non-coding variants' effects on phenotypic traits. Researchers have tried to uncover the effect of non-coding variants on human health at different levels, from directly testing the association between non-coding variants and diseases, to analyzing influence on gene expression, to measuring the effect on regulatory activity.

Genome-wide association studies (GWAS) are a popular method of linking genomic variants to human disease. In GWAS, the association between the variants is tested using either generalized linear model (GLM) if the phenotypic is a quantitative trait or contingency table/logistic regression if the phenotypic is a dichotomous trait. GWAS has generated over 100,000 variant-disease associations to date and greatly deepened our understanding for the genetic etiology of diseases. However, this method suffers from the following weaknesses: 1) It cannot identify the causative single nucleotide polymorphsims (SNPs) because of linkage disequilibrium (LD). 2) It cannot determine the effect of rare variants because power in GWAS studies is strongly influenced by variant frequency. Many pieces of evidence suggest that rare variants play an important role in human diseases. For example, a recent study demonstrated that ultrarare variants drive substantial *cis* heritability of human gene expression in lymphoblastoid cell lines (Hernandez et al. 2019). 4) The power for identifying risk variants is dependent on the sample size. Variants with small effects or with low minor allele frequency may not be identified. 3) It only produces the association between a single variant and the disease. For complex diseases, there is no direct way for obtaining

the overall disease risk for an individual, who has abundant variants across the genome. Genome-wide polygenic scores (GPS) have been proposed as a way to summarize the risk scores from GWAS studies across variants for a disease. Khera et al. demonstrate that with larger studies and improved algorithms, they were able to use GPS to identify 8.0, 6.1, 3.5, 3.2, and 1.5% of the population at greater than threefold increased risk for coronary artery disease, atrial fibrillation, type 2 diabetes, inflammatory bowel disease, and breast cancer, respectively (Khera et al. 2018). 4) GWASs has population biases. The vast majority of the GWAS studies are done with the individuals of European descent. This limit the generalization of the genetic risk prediction from GWASs to other populations (Need and Goldstein 2009; Popejoy and Fullerton 2016; Hindorff et al. 2018; Martin et al. 2019).

Analyzing influence of variants on lower level molecular measurements, such as gene expression and regulatory activity, may mitigate the issue of power due to small variant effect on organism-level phenotypic traits. Because even though the variant does not cause a significant change on the phenotype, it may show detectable effect on the level of gene expression or regulatory activity. The GTEx project aims to identify expression quantitative trait loci (eQTL), variants that have a significant influence on gene expression in various tissues. However, eQTL analyses share some caveats with GWAS, such as the difficulty of identifying the casual variants and its limitation on detecting effects of rare variants.

Another lower level measurement of non-coding variant effect is at the level of regulatory activity. This is usually done with reporter assay, especially the high throughput massively parallel reporter assay (MPRA). In this type of assay, the regulatory activity of DNA sequence is tested by how well they drive the expression of reporter gene in when constructed in the same plasmid. MPRA has been done at different scales. The largest MPRA experiment to date surveyed the effect of 5.9 million SNPs in K562 and HepG2 cells, including 57% of the known common SNPs, on enhancer and promoter activity and identified more than 30,000 SNPs that alter the activity of putative regulatory elements. The majority of DNA elements tested in this study only contain a single SNP, so they can assign the alteration of regulatory activity to the SNP. Because the effect on regulatory activity is directly read from the experiment, this approach largely avoids power and rare variant problem. The caveats of this study in terms of linking variants to diseases is that the effect on the regulatory activity cannot be directly translated to the effect on the phenotypic trait.

Some other measurements of non-coding variant effects are evolution-based metrics, like constraint calculated based on the allele frequency in the population (Di Iulio et al. 2018) or computational approaches integrating evolutionary conservation, genomic, chromatin, and gene expression information (Fu et al. 2014; Kircher et al. 2014; Ritchie et al. 2014; Zhou and Troyanskaya 2015; Zhou et al. 2018). These computational methods achieve state-of-art prediction performance at many disease variant prioritization tasks and are a powerful way to integrate various data sources. However, there is still much room for improvement in the computational methods. As shown in the work of Kircher and colleagues (Kircher et al. 2019), experimental saturation mutagenesis on 20 disease-associated gene promoters and enhancers for over 30000 single nucleotides showed low correlation with the predicted variant effect prediction made by many of the computation methods, such as DeepSEA (Zhou et al. 2018), CADD (Kircher et al. 2014), Eigen (Ionita-Laza et al. 2016), and FATHMM-MKL (Shihab et al. 2015), suggesting poor performance of these computational methods at predicting variants effect on gene expression.

As I discussed above, there are several problems to solve to achieve an accurate estimation of an individual's risk of disease, including fine mapping the causal variant, measuring effects of rare variants, identifying variants with small effects, summarizing the effects of variants across genome, and population biases. Modern experimental techniques suffer from different sets of these problems. A future method for better linking the non-coding variants to diseases and predicting the disease risk for individuals should address all these problems. In my opinion, machine learning models trained from DNA sequences with integrated data and better ensemble algorithms are a promising path to achieving this goal. There is likely a finite set of rules governing the genetic effect of non-coding variants on phenotypic traits. With the help with larger and more diverse data and better algorithms for integrating them, machine learning models may learn the principle biological rules that act upon non-coding variants, such as how they affect regulatory activity, gene expression, and ultimately phenotypic traits. Such a model could detect causal variants, be agnostic to rare variants because of learning the principle rules, be sensitive enough to capture even changes in the regulatory activity, and be able to summarize the variant effects across an individual's genome because the variants would be modeled together.

Several DNA sequence based machine learning methods have been developed to realize some of the steps (non-coding variants to regulatory activity, to gene expression, to phenotypic traits). Movva et al. developed a deep learning model predicting regulatory activity of MPRA experiments

from DNA sequences (Movva et al. 2019). Zhou et al. chained the first two steps, modeling the regulatory activity from DNA-sequences, and then modeling the gene expression, through training a two-step model to predict tissue specific gene expression from DNA sequences (Zhou et al. 2018). They first trained deep learning models predicting ~2000 epigenetic marks from DNA sequences and then use the representation learned from the model to encode the sequences nearby a gene as feature to train a regression model for gene expression. A logical next step would be chaining all steps, from variants in an individual's genome, to models of regulatory activity, to models of gene expression, and finally to phenotypic traits. To get to this step, we might need more regulatory activity data, gene expression, and phenotypic traits to ensure that the model at each step is accurate and reliable for next step. Even today, all available data have not yet been used in training such models. For example, the newly generated data by van Arensbergen et al (van Arensbergen et al. 2019) has not been included in any machine leaning models. This dataset of 5.9 million SNPs is >100 times larger than previous MPRA dataset and would likely be powerful for learning a better representation of genomic sequences for regulatory activity.

Another under-utilized resource is biobanks linked to electronic health records (EHRs). There have been many studies, especially phenome-wide association studies (Denny et al. 2010), that use billing codes in EHRs and DNA samples from biobanks to find associations between genetic variants and phenotypes. However, this kind of approach cannot be used to directly predict disease risk for individuals and have not integrated other function annotations of variants.

As the size of such data continue to grow, I see the potential to integrate EHR and biobanks with existing evolutionary, genomic, and population genetic information with complex, non-linear machine learning algorithms to aggregate variants in an individual for better disease risk prediction. This would be even better if not only the genotyping information, but also exome sequencing or whole genome sequencing (WGS) data were available. Similarly, natural language processing models could be used to generate better representation of phenotypic traits than provided by billing codes. An example of the effort of integrating health record and genetic information for individual disease risk prediction is done by Guturu et al (Guturu et al. 2016). They identified ancestral transcription factor binding sites disrupted by an individual's variants and then look for the affected target genes. They compared the function of potential affected genes with the individuals' health record and found some concordance. This is only done in five people with

simple quantification of genetic information and limited heath data. With the biobank, EHR, the abundant genomic data, and population genetics data, there is potential for a much better model.

In this dissertation, I demonstrated two uses of machine learning methods at improving our understanding of regulatory genome. In Chapter I and II, I demonstrated the difference in pleiotropic function between conserved and species-specific enhancers and the largely conserved underlying sequence properties of enhancers sequence elements through support vector machine algorithms. In Chapter IV, I demonstrated the power and limitation of deep neural networks at learning the accurate representation of complex regulatory grammar in enhancer sequences. These results contribute to our understanding of regulatory genome and the machine learning model of regulatory genome. Ultimately, there is still much room for machine learning methods to improve our understanding of the regulatory genome, learn biologically interpretable features from the resulting models, and dissect noncoding variant effects.

Appendix

A. **Summary of performance of all classification tasks**

Columns	*Description*
classifiers	The type of classifier. This can be SVM, CNN, gkm-SVM, gappy k-mer SVM or mistach k-mer SVM.
C	C parameter for SVM training
enhancer_length	Whether the original length is used or center 3000 bp is used
training_tissue	Enhancers in which tissue is used for training
testing_tissue	Enhancers in which tissue is used for testing
training_species	Enhancers and genomic background in which species is used for training
testing_species	Enhancers and genomic background in which species is used for testing
negative_set_size	Whether unbalanced negatives (10x) or balanced negatives(1x) are used
gc_controlled	Whether the negatives are GC-controlled or not
repeats_controlled	Whether the negatives are repeats-controlled or not
repeats_removed	Whether this experiment is done with enhancers and negatives that are both devoid of repeats or not
shared_removed	If the training and testing species are different, "shared_removed=TRUE" means orthologous enhancers are removed from both training and testing species; If the training and testing species are the same, "shared_removed=TRUE" means overlapped enhancers are removed from both training and testing set.
tenfold_cross_validation	Whether the experiment is done by tenfold cross validation or not
auROC	The Area under ROC curves for this prediction tasks. If the prediction is done by tenfold cross-validation within the same data, the mean auROC is shown.
auROC_std	The standard deviation of auROC if the prediction is done by ten-fold cross-validation within the same data.
relative_auROC	The relative auROC is calculated by dividing the auROC of this prediction task by the mean auROC of ten fold cross-validation of the testing data.
auPR	The Area under PR curves for this prediction tasks. If the prediction is done by ten fold cross-validation within the same data, the mean auPR is shown.
auPR_std	The standard deviation of auPR if the prediction is done by ten fold cross-validation within the same data.
relative_auPR	The relative auPR is calculated by dividing the auROC of this prediction task by the mean auPR of ten fold cross-validation of the testing data.

9	8	7	6	5	4	3	2	1	Experiment number
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	**classifiers**
15	15	15	15	15	15	15	15	15	**C**
original	original	original	original	original	original	original	original	original	**enhancer_length**
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	**training_tissue**
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	**testing_tissue**
Mmul	Mmul	Mmul	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	**training_species**
Mmus	Mmul	Hsap	Mdom	Cfam	Btau	Mmus	Mmul	Hsap	**testing_species**
10x	10x	10x	10x	10x	10x	10x	10x	10x	**negative_set_size**
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**gc_controlled**
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**repeats_controlled**
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**repeats_removed**
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	**shared_removed**
FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	**tenfold_cross_validatio n**
0.823	0.812	0.810	0.758	0.759	0.803	0.811	0.803	0.821	**auROC**
NA	0.004	NA	NA	NA	NA	NA	NA	0.003	**auROC_std**
0.979	1.000	0.987	0.955	0.974	0.964	0.964	0.989	1.000	**relative_auROC**
0.305	0.325	0.307	0.252	0.237	0.306	0.306	0.322	0.332	**auPR**
NA	0.008	NA	NA	NA	NA	NA	NA	0.010	**auPR_std**
0.882	1.000	0.925	0.869	0.888	0.913	0.884	0.991	1.000	**relative_auPR**

20	19	18	17	16	15	14	13	12	11	10
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Btau	Btau	Mmus	Mmus	Mmus	Mmus	Mmus	Mmus	Mmul	Mmul	Mmul
Mmul	Hsap	Mdom	Cfam	Btau	Mmus	Mmul	Hsap	Mdom	Cfam	Btau
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
0.781	0.785	0.761	0.755	0.795	0.841	0.770	0.768	0.762	0.766	0.800
NA	NA	NA	NA	NA	0.004	NA	NA	NA	NA	NA
0.962	0.956	0.958	0.969	0.954	1.000	0.948	0.935	0.960	0.983	0.960
0.286	0.274	0.201	0.231	0.296	0.346	0.270	0.250	0.250	0.239	0.300
NA	NA	NA	NA	NA	0.012	NA	NA	NA	NA	NA
0.880	0.825	0.693	0.865	0.884	1.000	0.831	0.753	0.862	0.895	0.896

31	30	29	28	27	26	25	24	23	22	21
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Mdom	Cfam	Cfam	Cfam	Cfam	Cfam	Cfam	Btau	Btau	Btau	Btau
Hsap	Mdom	Cfam	Btau	Mmus	Mmul	Hsap	Mdom	Cfam	Btau	Mmus
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
0.759	0.754	0.779	0.789	0.820	0.774	0.773	0.762	0.752	0.833	0.832
NA	NA	0.004	NA	NA	NA	NA	NA	NA	0.003	NA
0.924	0.950	1.000	0.947	0.975	0.953	0.942	0.960	0.965	1.000	0.989
0.245	0.236	0.267	0.283	0.317	0.257	0.248	0.257	0.236	0.335	0.336
NA	NA	0.010	NA	NA	NA	NA	NA	NA	0.005	NA
0.738	0.814	1.000	0.845	0.916	0.791	0.747	0.886	0.884	1.000	0.971

42	41	40	39	38	37	36	35	34	33	32
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Mdom	Mdom	Mdom	Mdom	Mdom
Mdom	Cfam	Btau	Mmus	Mmul	Hsap	Mdom	Cfam	Btau	Mmus	Mmul
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
0.705	0.705	0.717	0.731	0.720	0.751	0.794	0.740	0.777	0.815	0.757
NA	NA	NA	NA	NA	0.004	0.003	NA	NA	NA	NA
0.941	0.967	0.960	0.954	0.982	1.000	1.000	0.950	0.933	0.969	0.932
0.193	0.195	0.202	0.223	0.214	0.233	0.290	0.222	0.270	0.309	0.252
NA	NA	NA	NA	NA	0.005	0.007	NA	NA	NA	NA
0.843	0.911	0.898	0.851	0.964	1.000	1.000	0.831	0.806	0.893	0.775

53	52	51	50	49	48	47	46	45	44	43
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Mmus	Mmus	Mmus	Mmus	Mmus	Mmul	Mmul	Mmul	Mmul	Mmul	Mmul
Cfam	Btau	Mmus	Mmul	Hsap	Mdom	Cfam	Btau	Mmus	Mmul	Hsap
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE
0.699	0.705	0.766	0.694	0.696	0.694	0.715	0.713	0.737	0.733	0.733
NA	NA	0.005	NA	NA	NA	NA	NA	NA	0.005	NA
0.959	0.944	1.000	0.947	0.927	0.927	0.981	0.954	0.962	1.000	0.976
0.198	0.204	0.262	0.208	0.194	0.183	0.199	0.194	0.229	0.222	0.209
NA	NA	0.010	NA	NA	NA	NA	NA	NA	0.006	NA
0.925	0.907	1.000	0.937	0.833	0.799	0.930	0.862	0.874	1.000	0.897

64	63	62	61	60	59	58	57	56	55	54
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Cfam	Cfam	Cfam	Cfam	Btau	Btau	Btau	Btau	Btau	Btau	Mmus
Btau	Mmus	Mmul	Hsap	Mdom	Cfam	Btau	Mmus	Mmul	Hsap	Mdom
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
0.711	0.736	0.702	0.685	0.695	0.710	0.747	0.744	0.701	0.715	0.697
NA	NA	NA	NA	NA	NA	0.004	NA	NA	NA	NA
0.952	0.961	0.958	0.912	0.928	0.974	1.000	0.971	0.956	0.952	0.931
0.197	0.228	0.195	0.169	0.182	0.196	0.225	0.237	0.203	0.201	0.189
NA	NA	NA	NA	NA	NA	0.006	NA	NA	NA	NA
0.876	0.870	0.878	0.725	0.795	0.916	1.000	0.905	0.914	0.863	0.825

75	74	73	72	71	70	69	68	67	66	65
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Hsap	Hsap	Hsap	Mdom	Mdom	Mdom	Mdom	Mdom	Mdom	Cfam	Cfam
Mmus	Mmul	Hsap	Mdom	Cfam	Btau	Mmus	Mmul	Hsap	Mdom	Cfam
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
0.790	0.834	0.877	0.749	0.682	0.696	0.725	0.670	0.686	0.697	0.729
NA	NA	NA	0.004	NA	NA	NA	NA	NA	NA	0.004
0.954	0.986	1.000	1.000	0.936	0.932	0.946	0.914	0.913	0.931	1.000
0.227	0.270	0.344	0.229	0.185	0.184	0.220	0.176	0.181	0.182	0.214
NA	NA	NA	0.008	NA	NA	NA	NA	NA	NA	0.007
0.819	0.922	1.000	1.000	0.864	0.818	0.840	0.793	0.777	0.795	1.000

86	85	84	83	82	81	80	79	78	77	76
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Mmus	Mmus	Mmul	Mmul	Mmul	Mmul	Mmul	Mmul	Hsap	Hsap	Hsap
Mmul	Hsap	Mdom	Cfam	Btau	Mmus	Mmul	Hsap	Mdom	Cfam	Btau
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
0.815	0.846	0.802	0.669	0.715	0.807	0.846	0.887	0.782	0.663	0.685
NA	NA	NA	NA	NA	NA	0.015	NA	NA	NA	NA
0.963	0.965	1.096	1.081	0.915	0.975	1.000	1.011	1.068	1.071	0.877
0.254	0.302	0.283	0.156	0.179	0.241	0.293	0.357	0.238	0.153	0.166
NA	NA	NA	NA	NA	NA	0.035	NA	NA	NA	NA
0.867	0.878	1.408	1.106	0.734	0.870	1.000	1.038	1.184	1.085	0.680

97	96	95	94	93	92	91	90	89	88	87
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Cfam	Btau	Btau	Btau	Btau	Btau	Btau	Mmus	Mmus	Mmus	Mmus
Hsap	Mdom	Cfam	Btau	Mmus	Mmul	Hsap	Mdom	Cfam	Btau	Mmus
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
0.854	0.808	0.691	0.781	0.845	0.859	0.893	0.789	0.644	0.700	0.828
NA	NA	NA	0.009	NA	NA	NA	NA	NA	NA	0.017
0.974	1.104	1.116	1.000	1.021	1.015	1.018	1.078	1.040	0.896	1.000
0.295	0.306	0.178	0.244	0.307	0.321	0.384	0.256	0.151	0.179	0.277
NA	NA	NA	0.020	NA	NA	NA	NA	NA	NA	0.042
0.858	1.522	1.262	1.000	1.108	1.096	1.116	1.274	1.071	0.734	1.000

108	107	106	105	104	103	102	101	100	99	98
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Mdom	Mdom	Mdom	Mdom	Mdom	Mdom	Cfam	Cfam	Cfam	Cfam	Cfam
Mdom	Cfam	Btau	Mmus	Mmul	Hsap	Mdom	Cfam	Btau	Mmus	Mmul
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
0.732	0.617	0.645	0.757	0.788	0.825	0.747	0.619	0.647	0.762	0.813
0.030	NA	NA	NA	NA	NA	NA	0.041	NA	NA	NA
1.000	0.997	0.826	0.914	0.931	0.941	1.020	1.000	0.828	0.920	0.961
0.201	0.134	0.148	0.200	0.225	0.268	0.210	0.141	0.149	0.206	0.253
0.040	NA	NA	NA	NA	NA	NA	0.030	NA	NA	NA
1.000	0.950	0.607	0.722	0.768	0.779	1.045	1.000	0.611	0.744	0.863

119	118	117	116	115	114	113	112	111	110	109
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Mmul	Mmul	Mmul	Mmul	Mmul	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap
Cfam	Btau	Mmus	Mmul	Hsap	Mdom	Cfam	Btau	Mmus	Mmul	Hsap
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
0.699	0.676	0.725	0.709	0.696	0.705	0.694	0.681	0.725	0.698	0.709
NA	NA	NA	0.005	NA	NA	NA	NA	NA	NA	0.004
0.983	0.960	0.948	1.000	0.982	0.896	0.976	0.967	0.948	0.984	1.000
0.184	0.165	0.201	0.193	0.181	0.197	0.184	0.173	0.207	0.191	0.191
NA	NA	NA	0.004	NA	NA	NA	NA	NA	NA	0.003
0.948	0.882	0.807	1.000	0.948	0.732	0.948	0.925	0.831	0.990	1.000

130	129	128	127	126	125	124	123	122	121	120
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Btau	Btau	Btau	Btau	Mmus	Mmus	Mmus	Mmus	Mmus	Mmus	Mmul
Btau	Mmus	Mmul	Hsap	Mdom	Cfam	Btau	Mmus	Mmul	Hsap	Mdom
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
0.704	0.732	0.685	0.684	0.699	0.676	0.675	0.765	0.667	0.668	0.687
0.005	NA	NA	NA	NA	NA	NA	0.006	NA	NA	NA
1.000	0.957	0.966	0.965	0.888	0.951	0.959	1.000	0.941	0.942	0.873
0.187	0.211	0.181	0.177	0.197	0.181	0.176	0.249	0.180	0.174	0.182
0.006	NA	NA	NA	NA	NA	NA	0.006	NA	NA	NA
1.000	0.847	0.938	0.927	0.732	0.933	0.941	1.000	0.933	0.911	0.677

141	140	139	138	137	136	135	134	133	132	131
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Mdom	Mdom	Mdom	Cfam	Cfam	Cfam	Cfam	Cfam	Cfam	Btau	Btau
Mmus	Mmul	Hsap	Mdom	Cfam	Btau	Mmus	Mmul	Hsap	Mdom	Cfam
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
0.702	0.623	0.632	0.697	0.711	0.684	0.729	0.687	0.676	0.706	0.700
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.918	0.879	0.891	0.886	1.000	0.972	0.953	0.969	0.953	0.897	0.985
0.187	0.144	0.144	0.186	0.194	0.173	0.209	0.179	0.164	0.191	0.184
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.751	0.746	0.754	0.691	1.000	0.925	0.839	0.927	0.859	0.710	0.948

152	151	150	149	148	147	146	145	144	143	142
SVM	SVM	SVM	SVM	SVM	SVM	mismatch kmer SVM	gappy kmer SVM (k=2	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Mmul	Mmul	Mmul	Hsap	Hsap	Hsap	Hsap	Hsap	Mdom	Mdom	Mdom
Mmus	Mmul	Hsap	Mmus	Mmul	Hsap	Hsap	Hsap	Mdom	Cfam	Btau
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE
0.870	0.898	0.862	0.796	0.864	0.889	0.818	0.818	0.787	0.641	0.651
NA	0.001	NA	NA	NA	0.003	NA	NA	0.004	NA	NA
0.973	1.000	0.970	0.890	0.962	1.000	1.000	1.000	1.000	0.902	0.925
0.377	0.460	0.352	0.294	0.409	0.443	0.313	0.319	0.269	0.150	0.155
NA	0.004	NA	NA	NA	0.007	0.007	0.009	0.009	NA	NA
0.885	1.000	0.795	0.690	0.889	1.000	1.000	1.000	1.000	0.773	0.829

163	162	161	160	159	158	157	156	155	154	153
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb
Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb	Cotney et al. limb
Mmus	Mmus	Mmul	Mmul	Mmul	Hsap	Hsap	Hsap	Mmus	Mmus	Mmus
Mmul	Hsap	Mmus	Mmul	Hsap	Mmus	Mmul	Hsap	Mmus	Mmul	Hsap
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE
0.735	0.752	0.722	0.773	0.806	0.664	0.746	0.829	0.894	0.874	0.823
NA	NA	NA	0.002	NA	NA	NA	0.004	0.001	NA	NA
0.951	0.907	0.955	1.000	0.972	0.878	0.965	1.000	1.000	0.973	0.926
0.232	0.219	0.201	0.270	0.291	0.169	0.262	0.334	0.426	0.421	0.292
NA	NA	NA	0.004	NA	NA	NA	0.009	0.005	NA	NA
0.859	0.656	0.863	1.000	0.871	0.725	0.970	1.000	1.000	0.915	0.659

174	173	172	171	170	169	168	167	166	165	164
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Cotney et al. limb
Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Cotney et al. limb
Hsap	Mmus	Mmus	Mmus	Mmul	Mmul	Mmul	Hsap	Hsap	Hsap	Mmus
Hsap	Mmus	Mmul	Hsap	Mmus	Mmul	Hsap	Mmus	Mmul	Hsap	Mmus
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE
0.811	0.918	0.884	0.918	0.901	0.902	0.925	0.902	0.896	0.929	0.756
0.003	0.001	NA	NA	NA	0.002	NA	NA	NA	0.002	0.004
1.000	1.000	0.980	0.988	0.981	1.000	0.996	0.983	0.993	1.000	1.000
0.317	0.545	0.521	0.529	0.489	0.542	0.545	0.493	0.527	0.560	0.233
0.005	0.006	NA	NA	NA	0.005	NA	NA	NA	0.009	0.004
1.000	1.000	0.961	0.945	0.897	1.000	0.973	0.905	0.972	1.000	1.000

185	184	183	182	181	180	179	178	177	176	175
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain
Villar et al. liver	Villar et al. liver	Villar et al. liver	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain	Reilly et al. brain
Hsap	Hsap	Hsap	Mmus	Mmus	Mmus	Mmul	Mmul	Mmul	Hsap	Hsap
Cfam	Btau	Mmus	Mmus	Mmul	Hsap	Mmus	Mmul	Hsap	Mmus	Mmul
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
0.731	0.780	0.804	0.828	0.734	0.777	0.773	0.792	0.779	0.790	0.756
NA	NA	NA	0.002	NA	NA	NA	0.003	NA	NA	NA
0.968	0.957	0.963	1.000	0.927	0.958	0.934	1.000	0.961	0.954	0.955
0.208	0.268	0.289	0.336	0.298	0.247	0.256	0.360	0.266	0.295	0.347
NA	NA	NA	0.008	NA	NA	NA	0.004	NA	NA	NA
0.885	0.899	0.868	1.000	0.828	0.779	0.762	1.000	0.839	0.878	0.964

196	195	194	193	192	191	190	189	188	187	186
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
VISTA forebrain	VISTA forebrain	VISTA forebrain	VISTA forebrain	VISTA branchialarch	VISTA branchialarch	VISTA branchialarch	VISTA branchialarch	Villar et al. liver	Villar et al. liver	Villar et al. liver
VISTA forebrain	VISTA forebrain	VISTA forebrain	VISTA forebrain	VISTA branchialarch	VISTA branchialarch	VISTA branchialarch	VISTA branchialarch	Villar et al. liver	Villar et al. liver	Villar et al. liver
Mmus	Mmus	Hsap	Hsap	Mmus	Mmus	Hsap	Hsap	Cfam	Btau	Mmus
Mmus	Hsap	Mmus	Hsap	Mmus	Hsap	Mmus	Hsap	Cfam	Btau	Mmus
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
0.714	0.801	0.762	0.919	0.793	0.816	0.791	0.822	0.755	0.815	0.835
0.099	NA	NA	NA	0.095	NA	NA	NA	0.003	0.003	0.006
1.000	0.872	1.067	1.000	1.000	0.993	0.997	1.000	1.000	1.000	1.000
0.212	0.305	0.242	0.638	0.341	0.377	0.296	0.389	0.235	0.298	0.333
0.169	NA	NA	NA	0.160	NA	NA	NA	0.008	0.009	0.015
1.000	0.478	1.142	1.000	1.000	0.969	0.868	1.000	1.000	1.000	1.000

207	206	205	204	203	202	201	200	199	198	197
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
VISTA heart	VISTA heart	VISTA heart	VISTA hindbrain	VISTA hindbrain	VISTA hindbrain	VISTA hindbrain	VISTA midbrain	VISTA midbrain	VISTA midbrain	VISTA midbrain
VISTA heart	VISTA heart	VISTA heart	VISTA hindbrain	VISTA hindbrain	VISTA hindbrain	VISTA hindbrain	VISTA midbrain	VISTA midbrain	VISTA midbrain	VISTA midbrain
Mmus	Hsap	Hsap	Mmus	Mmus	Hsap	Hsap	Mmus	Mmus	Hsap	Hsap
Hsap	Mmus	Hsap	Mmus	Hsap	Mmus	Hsap	Mmus	Hsap	Mmus	Hsap
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE
0.825	0.743	0.824	0.803	0.753	0.821	0.895	0.806	0.785	0.760	0.898
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
1.001	0.964	1.000	1.000	0.841	1.022	1.000	1.000	0.874	0.943	1.000
0.328	0.226	0.366	0.346	0.281	0.361	0.556	0.399	0.346	0.282	0.562
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.896	0.739	1.000	1.000	0.505	1.043	1.000	1.000	0.616	0.707	1.000

218	217	216	215	214	213	212	211	210	209	208
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
VISTA forebrain	VISTA forebrain	VISTA branchialarch	VISTA branchialarch	VISTA branchialarch	VISTA branchialarch	VISTA vistalimb	VISTA vistalimb	VISTA vistalimb	VISTA vistalimb	VISTA heart
VISTA forebrain	VISTA forebrain	VISTA branchialarch	VISTA branchialarch	VISTA branchialarch	VISTA branchialarch	VISTA vistalimb	VISTA vistalimb	VISTA vistalimb	VISTA vistalimb	VISTA heart
Hsap	Hsap	Mmus	Mmus	Hsap	Hsap	Mmus	Mmus	Hsap	Hsap	Mmus
Mmus	Hsap	Mmus	Hsap	Mmus	Hsap	Mmus	Hsap	Mmus	Hsap	Mmus
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE
0.766	0.908	0.711	0.807	0.787	0.808	0.779	0.790	0.786	0.811	0.771
NA	NA	0.102	NA	NA	NA	NA	NA	NA	NA	NA
1.076	1.000	1.000	0.999	1.107	1.000	1.000	0.974	1.009	1.000	1.000
0.256	0.634	0.190	0.348	0.371	0.361	0.275	0.312	0.304	0.432	0.306
NA	NA	0.069	NA	NA	NA	NA	NA	NA	NA	NA
0.945	1.000	1.000	0.964	1.953	1.000	1.000	0.722	1.105	1.000	1.000

229	228	227	226	225	224	223	222	221	220	219
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
VISTA heart	VISTA hindbrain	VISTA hindbrain	VISTA hindbrain	VISTA hindbrain	VISTA midbrain	VISTA midbrain	VISTA midbrain	VISTA midbrain	VISTA forebrain	VISTA forebrain
VISTA heart	VISTA hindbrain	VISTA hindbrain	VISTA hindbrain	VISTA hindbrain	VISTA midbrain	VISTA midbrain	VISTA midbrain	VISTA midbrain	VISTA forebrain	VISTA forebrain
Hsap	Mmus	Mmus	Hsap	Hsap	Mmus	Mmus	Hsap	Hsap	Mmus	Mmus
Hsap	Mmus	Hsap	Mmus	Hsap	Mmus	Hsap	Mmus	Hsap	Mmus	Hsap
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE
0.717	0.783	0.744	0.769	0.883	0.668	0.818	0.780	0.907	0.712	0.839
NA	NA	NA	NA	NA	0.089	NA	NA	NA	NA	NA
1.000	1.000	0.843	0.982	1.000	1.000	0.902	1.168	1.000	1.000	0.924
0.204	0.287	0.290	0.266	0.554	0.179	0.436	0.317	0.654	0.271	0.402
NA	NA	NA	NA	NA	0.093	NA	NA	NA	NA	NA
1.000	1.000	0.523	0.927	1.000	1.000	0.667	1.771	1.000	1.000	0.634

240	239	238	237	236	235	234	233	232	231	230
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	orginal	original	original	original	original	original	original	original	original
Villar et al. liver	Villar et al. liver	non-H3K27ac VISTA limb	non-VISTA Cotney et al. limb	VISTA vistalimb	VISTA vistalimb	VISTA vistalimb	VISTA vistalimb	VISTA heart	VISTA heart	VISTA heart
Roadmap Pancreas	Roadmap Brain Hippocampus	non-H3K27ac VISTA limb	non-H3K27ac VISTA limb	VISTA vistalimb	VISTA vistalimb	VISTA vistalimb	VISTA vistalimb	VISTA heart	VISTA heart	VISTA heart
Hsap	Hsap	Hsap	Hsap	Mmus	Mmus	Hsap	Hsap	Mmus	Mmus	Hsap
Hsap	Hsap	Hsap	Hsap	Mmus	Hsap	Mmus	Hsap	Mmus	Hsap	Mmus
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE
0.672	0.754	0.803	0.823	0.766	0.797	0.768	0.828	0.743	0.697	0.667
NA	NA	0.050	NA	NA	NA	NA	NA	0.052	NA	NA
0.883	0.920	1.000	1.025	1.000	0.963	1.003	1.000	1.000	0.972	0.898
0.175	0.210	0.389	0.352	0.268	0.346	0.306	0.466	0.205	0.142	0.178
NA	NA	0.123	NA	NA	NA	NA	NA	0.098	NA	NA
0.845	0.673	1.000	0.905	1.000	0.742	1.142	1.000	1.000	0.696	0.868

251	250	249	248	247	246	245	244	243	242	241
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Roadmap Gastric	Roadmap Liver, Adult	Roadmap Pancreas	Roadmap Brain Hippocampus	Roadmap CD14, Primary Cells	Roadmap Bone Marrow Derived	Roadmap Ovary	Roadmap Lung	Roadmap Left Ventricle	Roadmap Gastric	Roadmap Liver, Adult
Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap
Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
0.688	0.755	0.617	0.683	0.744	0.697	0.685	0.739	0.779	0.733	0.788
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.962	0.981	0.939	0.909	0.880	0.857	0.825	0.847	0.928	0.931	0.968
0.186	0.238	0.142	0.172	0.217	0.168	0.189	0.225	0.245	0.209	0.251
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.935	0.922	0.934	0.758	0.650	0.535	0.668	0.623	0.698	0.816	0.797

262	261	260	259	258	257	256	255	254	253	252
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Roadmap Lung	Roadmap Left Ventricle	Roadmap Gastric	Roadmap Liver, Adult	Roadmap Pancreas	Roadmap Brain Hippocampus	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Roadmap Lung	Roadmap Left Ventricle	Roadmap Gastric	Roadmap Liver, Adult	Roadmap Pancreas	Roadmap Brain Hippocampus	Roadmap CD14, Primary Cells	Roadmap Bone Marrow Derived	Roadmap Ovary	Roadmap Lung	Roadmap Left Ventricle
Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap
Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
0.873	0.839	0.787	0.814	0.761	0.820	0.645	0.713	0.608	0.642	0.726
0.002	0.003	0.005	0.004	0.005	0.004	NA	NA	NA	NA	NA
1.000	1.000	1.000	1.000	1.000	1.000	0.880	0.893	0.877	0.892	0.933
0.361	0.351	0.256	0.315	0.207	0.312	0.154	0.197	0.145	0.168	0.210
0.004	0.008	0.008	0.009	0.004	0.008	NA	NA	NA	NA	NA
1.000	1.000	1.000	1.000	1.000	1.000	0.703	0.675	0.797	0.792	0.795

273	272	271	270	269	268	267	266	265	264	263
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Roadmap Bone Marrow Derived	Roadmap Ovary	Roadmap Lung	Roadmap Left Ventricle	Roadmap Gastric	Roadmap Liver, Adult	Roadmap Pancreas	Roadmap Brain Hippocampus	Roadmap CD14, Primary Cells	Roadmap Bone Marrow Derived	Roadmap Ovary
Roadmap Bone Marrow Derived	Roadmap Ovary	Roadmap Lung	Roadmap Left Ventricle	Roadmap Gastric	Roadmap Liver, Adult	Roadmap Pancreas	Roadmap Brain Hippocampus	Roadmap CD14, Primary Cells	Roadmap Bone Marrow Derived	Roadmap Ovary
Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap
Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
0.798	0.693	0.720	0.778	0.715	0.770	0.657	0.751	0.845	0.813	0.830
0.004	0.005	0.003	0.003	0.004	0.004	0.004	0.004	0.003	0.003	0.002
1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.292	0.182	0.212	0.264	0.199	0.258	0.152	0.227	0.334	0.314	0.283
0.007	0.006	0.004	0.004	0.006	0.008	0.004	0.009	0.009	0.009	0.004
1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

284	283	282	281	280	279	278	277	276	275	274
CNN	CNN	CNN	CNN	CNN	CNN	CNN	CNN	CNN	CNN	SVM
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	15
center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Roadmap CD14, Primary Cells
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Roadmap CD14, Primary Cells
Mmul	Mmul	Mmul	Mmul	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap
Btau	Mmus	Mmul	Hsap	Mdom	Cfam	Btau	Mmus	Mmul	Hsap	Hsap
1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	10x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE
0.825	0.847	0.858	0.822	0.794	0.804	0.820	0.841	0.821	0.855	0.733
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.006
0.948	0.958	1.000	0.961	0.942	0.932	0.943	0.951	0.957	1.000	1.000
0.802	0.826	0.844	0.800	0.778	0.791	0.801	0.820	0.818	0.843	0.219
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	0.004
0.937	0.953	1.000	0.949	0.941	0.928	0.936	0.946	0.969	1.000	1.000

295	294	293	292	291	290	289	288	287	286	285
CNN	CNN	CNN	CNN	CNN	CNN	CNN	CNN	CNN	CNN	CNN
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Btau	Btau	Btau	Mmus	Mmus	Mmus	Mmus	Mmus	Mmus	Mmul	Mmul
Mmus	Mmul	Hsap	Mdom	Cfam	Btau	Mmus	Mmul	Hsap	Mdom	Cfam
1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
0.857	0.799	0.763	0.795	0.774	0.803	0.884	0.767	0.749	0.798	0.824
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.969	0.931	0.892	0.943	0.897	0.923	1.000	0.894	0.876	0.947	0.955
0.838	0.788	0.739	0.779	0.777	0.797	0.867	0.773	0.740	0.781	0.810
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.967	0.934	0.877	0.942	0.912	0.931	1.000	0.916	0.878	0.944	0.951

306	305	304	303	302	301	300	299	298	297	296
CNN	CNN	CNN	CNN	CNN	CNN	CNN	CNN	CNN	CNN	CNN
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Mdom	Mdom	Cfam	Cfam	Cfam	Cfam	Cfam	Cfam	Btau	Btau	Btau
Mmul	Hsap	Mdom	Cfam	Btau	Mmus	Mmul	Hsap	Mdom	Cfam	Btau
1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
0.681	0.760	0.816	0.863	0.830	0.857	0.713	0.758	0.793	0.808	0.870
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.794	0.889	0.968	1.000	0.954	0.969	0.831	0.887	0.941	0.936	1.000
0.582	0.739	0.803	0.852	0.811	0.840	0.607	0.729	0.777	0.802	0.856
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.690	0.877	0.971	1.000	0.947	0.969	0.719	0.865	0.940	0.941	1.000

317	316	315	314	313	312	311	310	309	308	307
SVM	SVM	SVM	SVM	SVM	SVM	SVM	CNN	CNN	CNN	CNN
15	15	15	15	15	15	15	NA	NA	NA	NA
center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Cfam	Btau	Btau	Btau	Btau	Btau	Btau	Mdom	Mdom	Mdom	Mdom
Btau	Mmus	Mmul	Mdom	Hsap	Cfam	Btau	Mdom	Cfam	Btau	Mmus
1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE
0.764	0.799	0.758	0.743	0.757	0.740	0.791	0.843	0.787	0.782	0.837
NA	NA	NA	NA	NA	NA	0.005	NA	NA	NA	NA
0.966	0.984	0.977	0.957	0.966	0.985	1.000	1.000	0.912	0.899	0.947
0.745	0.777	0.738	0.729	0.736	0.721	0.768	0.827	0.772	0.759	0.817
NA	NA	NA	NA	NA	NA	0.006	NA	NA	NA	NA
0.970	0.990	0.980	0.963	0.967	0.985	1.000	1.000	0.906	0.887	0.942

328	327	326	325	324	323	322	321	320	319	318
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Cfam	Cfam	Cfam	Cfam	Cfam
Mmus	Mmul	Mdom	Hsap	Cfam	Btau	Mmus	Mmul	Mdom	Hsap	Cfam
1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE
0.788	0.773	0.747	0.784	0.739	0.774	0.791	0.753	0.734	0.746	0.751
NA	NA	NA	0.005	NA	NA	NA	NA	NA	NA	0.005
0.970	0.996	0.963	1.000	0.984	0.979	0.974	0.970	0.946	0.952	1.000
0.761	0.751	0.729	0.761	0.718	0.750	0.763	0.727	0.716	0.717	0.732
NA	NA	NA	0.009	NA	NA	NA	NA	NA	NA	0.011
0.969	0.999	0.963	1.000	0.981	0.977	0.972	0.967	0.946	0.942	1.000

339	338	337	336	335	334	333	332	331	330	329
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Mmul	Mmul	Mmul	Mmul	Mmul	Mdom	Mdom	Mdom	Mdom	Mdom	Mdom
Mmul	Mdom	Hsap	Cfam	Btau	Mmus	Mmul	Mdom	Hsap	Cfam	Btau
1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE
0.776	0.741	0.775	0.742	0.771	0.793	0.723	0.776	0.727	0.716	0.750
0.004	NA	NA	NA	NA	NA	NA	0.008	NA	NA	NA
1.000	0.955	0.989	0.988	0.975	0.977	0.932	1.000	0.927	0.953	0.948
0.752	0.722	0.752	0.716	0.746	0.771	0.710	0.757	0.712	0.701	0.735
0.009	NA	NA	NA	NA	NA	NA	0.009	NA	NA	NA
1.000	0.954	0.988	0.978	0.971	0.982	0.944	1.000	0.936	0.958	0.957

350	349	348	347	346	345	344	343	342	341	340
kmer polynomial	kmer polynomial	gkm SVM (L=11, k=7	gkm SVM (L=11, k=7	SVM	SVM	SVM	SVM	SVM	SVM	SVM
0.1	1	0.1	1	15	15	15	15	15	15	15
center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Hsap	Hsap	Hsap	Hsap	Mmus	Mmus	Mmus	Mmus	Mmus	Mmus	Mmul
Hsap	Hsap	Hsap	Hsap	Mmus	Mmul	Mdom	Hsap	Cfam	Btau	Mmus
1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
0.755	0.709	0.751	0.763	0.812	0.725	0.751	0.734	0.721	0.756	0.790
NA	NA	NA	NA	0.006	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	1.000	0.934	0.968	0.936	0.960	0.956	0.973
0.723	0.676	0.722	0.745	0.785	0.716	0.734	0.721	0.708	0.742	0.759
NA	NA	NA	NA	0.011	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	1.000	0.952	0.970	0.947	0.967	0.966	0.967

361	360	359	358	357	356	355	354	353	352	351
CNN	CNN	SVM	SVM	SVM	SVM	SVM	SVM	kmer	kmer	kmer
NA	NA	15	15	15	15	15	15	15	0.001	0.01
center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Hsap	Hsap	Cfam	Btau	Mmus	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap
Btau	Mmus	Hsap	Hsap	Hsap	Cfam	Btau	Mmus	Hsap	Hsap	Hsap
1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	90% train and 10% test	90% train and 10% test	90% train and 10% test
0.675	0.670	0.658	0.657	0.696	0.647	0.658	0.644	0.709	0.789	0.789
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.890	0.876	0.977	0.970	0.955	0.954	0.974	0.947	NA	NA	NA
0.661	0.668	0.643	0.639	0.673	0.631	0.643	0.639	0.676	0.757	0.759
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.904	0.891	0.981	0.973	0.958	0.953	0.975	0.963	NA	NA	NA

372	371	370	369	368	367	366	365	364	363	362
SVM	SVM	SVM	SVM	SVM	SVM	SVM	CNN	CNN	CNN	CNN
15	15	10	0.01	0.1	1	0.001	NA	NA	NA	NA
original	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Cfam	Btau	Mmus	Hsap
Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Cfam
10x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x
Flanking regions	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
Flanking regions	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
Flanking regions	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
Flanking regions	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE
0.751	0.782	0.783	0.776	0.785	0.784	0.726	0.717	0.722	0.764	0.688
NA	0.005	0.004	0.005	0.003	0.005	0.007	NA	NA	NA	NA
1.000	NA	NA	NA	NA	NA	NA	0.896	0.938	0.957	0.909
0.226	0.760	0.760	0.752	0.763	0.761	0.698	0.707	0.707	0.746	0.675
NA	0.008	0.008	0.007	0.007	0.010	0.009	NA	NA	NA	NA
1.000	NA	NA	NA	NA	NA	NA	0.896	0.932	0.946	0.911

373	SVM	15	original	Villar et al. liver	Villar et al. liver	Hsap	Mmul	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.733	NA	0.987	0.219	NA	0.982
374	SVM	15	original	Villar et al. liver	Villar et al. liver	Hsap	Mmus	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.706	NA	0.938	0.197	NA	0.831
375	SVM	15	original	Villar et al. liver	Villar et al. liver	Hsap	Btau	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.713	NA	0.952	0.199	NA	0.896
376	SVM	15	original	Villar et al. liver	Villar et al. liver	Hsap	Cfam	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.695	NA	1.112	0.381	NA	1.241
377	SVM	15	original	Villar et al. liver	Villar et al. liver	Hsap	Mdom	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.681	NA	0.950	0.171	NA	0.847
378	SVM	15	original	Villar et al. liver	Villar et al. liver	Mmul	Hsap	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.738	NA	0.983	0.209	NA	0.925
379	SVM	15	original	Villar et al. liver	Villar et al. liver	Mmul	Mmul	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.743	NA	1.000	0.223	NA	1.000
380	SVM	15	original	Villar et al. liver	Villar et al. liver	Mmul	Mmus	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.712	NA	0.946	0.193	NA	0.814
381	SVM	15	original	Villar et al. liver	Villar et al. liver	Mmul	Btau	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.705	NA	0.941	0.190	NA	0.856
382	SVM	15	original	Villar et al. liver	Villar et al. liver	Mmul	Cfam	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.705	NA	1.128	0.389	NA	1.267
383	SVM	15	original	Villar et al. liver	Villar et al. liver	Mmul	Mdom	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.681	NA	0.950	0.169	NA	0.837

384	SVM	15	original	Villar et al. liver	Villar et al. liver	Mmus	Hsap	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.692	NA	0.921	0.188	NA	0.832
385	SVM	15	original	Villar et al. liver	Villar et al. liver	Mmus	Mmul	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.688	NA	0.926	0.193	NA	0.865
386	SVM	15	original	Villar et al. liver	Villar et al. liver	Mmus	Mmus	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.753	NA	1.000	0.237	NA	1.000
387	SVM	15	original	Villar et al. liver	Villar et al. liver	Mmus	Btau	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.694	NA	0.927	0.192	NA	0.865
388	SVM	15	original	Villar et al. liver	Villar et al. liver	Mmus	Cfam	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.682	NA	1.091	0.376	NA	1.225
389	SVM	15	original	Villar et al. liver	Villar et al. liver	Mmus	Mdom	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.678	NA	0.946	0.167	NA	0.827
390	SVM	15	original	Villar et al. liver	Villar et al. liver	Btau	Hsap	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.692	NA	0.921	0.183	NA	0.810
391	SVM	15	original	Villar et al. liver	Villar et al. liver	Btau	Mmul	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.690	NA	0.929	0.186	NA	0.834
392	SVM	15	original	Villar et al. liver	Villar et al. liver	Btau	Mmus	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.723	NA	0.960	0.209	NA	0.882
393	SVM	15	original	Villar et al. liver	Villar et al. liver	Btau	Btau	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.749	NA	1.000	0.222	NA	1.000
394	SVM	15	original	Villar et al. liver	Villar et al. liver	Btau	Cfam	10x	Flanking regions	Flanking regions	Flanking regions	Flanking regions	TRUE	0.691	NA	1.106	0.373	NA	1.215

405	404	403	402	401	400	399	398	397	396	395
SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM	SVM
15	15	15	15	15	15	15	15	15	15	15
original	original	original	original	original	original	original	original	original	original	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Mdom	Mdom	Mdom	Mdom	Cfam	Cfam	Cfam	Cfam	Cfam	Cfam	Btau
Btau	Mmus	Mmul	Hsap	Mdom	Cfam	Btau	Mmus	Mmul	Hsap	Mdom
10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions
Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions
Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions
Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions	Flanking regions
TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
0.689	0.715	0.690	0.693	0.604	0.625	0.610	0.616	0.606	0.608	0.671
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.920	0.950	0.929	0.923	0.842	1.000	0.814	0.818	0.816	0.810	0.936
0.183	0.203	0.193	0.191	0.125	0.307	0.128	0.131	0.122	0.125	0.161
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.824	0.857	0.865	0.845	0.619	1.000	0.577	0.553	0.547	0.553	0.797

416	415	414	413	412	411	410	409	408	407	406
kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	SVM	SVM
0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	15	15
center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	original	original
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Mdom	Mdom
Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Mdom	Cfam
1x	1x	1x	1x	1x	1x	1x	1x	1x	10x	10x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Flanking regions	Flanking regions
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Flanking regions	Flanking regions
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Flanking regions	Flanking regions
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Flanking regions	Flanking regions
90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	TRUE	TRUE
0.763	0.803	0.731	0.739	0.767	0.730	0.793	0.751	0.725	0.717	0.683
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	1.000	1.093
0.737	0.770	0.716	0.719	0.747	0.719	0.757	0.731	0.706	0.202	0.380
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	1.000	1.238

427	426	425	424	423	422	421	420	419	418	417
kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial
0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap
Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap
1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test
0.760	0.667	0.797	0.741	0.758	0.749	0.709	0.762	0.741	0.767	0.754
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.739	0.582	0.763	0.714	0.732	0.709	0.683	0.744	0.721	0.749	0.740
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

438	437	436	435	434	433	432	431	430	429	428
kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial	kmer polynomial
0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp	center 3000bp
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap
Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap	Hsap
1x	1x	1x	1x	1x	1x	1x	1x	1x	1x	1x
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE
90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test
0.778	0.778	0.789	0.805	0.729	0.793	0.758	0.795	0.792	0.742	0.777
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0.754	0.751	0.757	0.777	0.707	0.779	0.735	0.772	0.760	0.724	0.748
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

443	442	441	440	439
kmer	kmer	kmer	kmer	kmer
polynomial	polynomial	polynomial	polynomial	polynomial
0.001	0.001	0.001	0.001	0.001
center	center	center	center	center
3000bp	3000bp	3000bp	3000bp	3000bp
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver	Villar et al. liver
Hsap	Hsap	Hsap	Hsap	Hsap
Hsap	Hsap	Hsap	Hsap	Hsap
1x	1x	1x	1x	1x
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
FALSE	FALSE	FALSE	FALSE	FALSE
90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test	90% train and 10% test
0.754	0.747	0.778	0.792	0.750
NA	NA	NA	NA	NA
NA	NA	NA	NA	NA
0.731	0.720	0.748	0.758	0.724
NA	NA	NA	NA	NA
NA	NA	NA	NA	NA

B. Liver expression of the shared TF motifs in the liver GC-controlled analysis

The combined TF motifs are considered liver expressed when all of the components are expressed in liver. NA means that the TF or at least one of the TFs in the compound TF motifs is not found.

TF motifs matched by all species' top 5-mers	Liver expression in human	Other TFs in the same subfamily that are liver expressed in human	Liver expression in macaque	Liver expression in mouse	Liver expression in cow	Liver expression in dog	Liver expression in opossum
BACH1::MAFK	Yes[a]		Yes[a]	Yes[a]	Yes[a]	Yes[a]	Yes[a]
NFE2L1::MAFG	Yes[a]		Yes[a]	Yes[a]	Yes[a]	Yes[a]	Yes[a]
NR1H2::RXRA	Yes[a]		Yes[a]	Yes[a]	Yes[a]	Yes[a]	NA
PPARG::RXRA	Yes[a]		Yes[a]	Yes[a]	Yes[a]	Yes[a]	Yes[a]
SMAD2::SMAD3::SMAD4	Yes[a]		Yes[a]	Yes[a]	Yes[a]	Yes[a]	NA
STAT5A::STAT5B	Yes[a]		Yes[a]	Yes[a]	Yes[a]	Yes[a]	Yes[a]
E2F1	Yes		Yes	Yes	Yes	Yes	Yes
E2F4	Yes		Yes	Yes	Yes	Yes	Yes
E2F6	Yes		Yes	Yes	Yes	Yes	Yes
ESR1	Yes		Yes	Yes	Yes	Yes	Yes
FOS	Yes		Yes	Yes	Yes	Yes	Yes
FOSL2	Yes		Yes	Yes	Yes	Yes	Yes
GABPA	Yes		Yes	Yes	Yes	Yes	Yes
JUN	Yes		Yes	Yes	Yes	Yes	Yes
JUNB	Yes		Yes	Yes	Yes	Yes	NA
JUND	Yes		Yes	Yes	NA	Yes	Yes
MAFB	Yes		Yes	Yes	Yes	Yes	Yes
MAFF	Yes		Yes	Yes	Yes	Yes	Yes
MAFK	Yes		Yes	Yes	Yes	Yes	Yes
NFE2L2	Yes		Yes	Yes	Yes	Yes	Yes
NR5A2	Yes		Yes	Yes	Yes	Yes	Yes
RXRA	Yes		Yes	Yes	Yes	Yes	Yes
SREBF1	Yes		Yes	Yes	Yes	Yes	Yes
SREBF2	Yes		Yes	Yes	Yes	Yes	Yes
THAP1	Yes		Yes	Yes	Yes	Yes	Yes
USF2	Yes		Yes	Yes	Yes	Yes	Yes
NFE2::MAF	No[a]	NFE2L1, NFE2L2 (NF-E2-like factors, 1.1.1.2)	Yes	Yes	Yes	Yes	NA
RXRA::VDR	No[a]	NR1I2 (Vitamin D receptor 2.1.2.4)	Yes	Yes	Yes	Yes	Yes

ESR2	No	ESRRA, ESR1 (ER-like receptors 2.1.1.2)	Yes	Yes	No	No	Yes
FEV	No	ETS1, ETS2, GABPA, FLI1, ETV2, ETV3, ERF(Ets-like factors 3.5.2.1)	Yes	NA	No	NA	No
FOSL1	No	FOSL2 (Fos factors 1.1.2.1)	Yes	Yes	No	No	NA
PAX2	No	NA(PAX-2-like factors 3.2.2.2)	No	Yes	No	No	Yes
SOX2	No	SOX5, SOX6, SOX7, SOX12, SOX13(SOX-related factors, 4.1.1)	No	No	No	No	No
Data source	Gene Expression Atlas (https://expressionatlas.org/hg19/adult/)	Gene Expression Atlas (https://expressionatlas.org/hg19/adult/)	Berthelot et al 2017	Rudolph et al. 2016	Berthelot et al 2017	Berthelot et al 2017	Berthelot et al 2017

REFERENCES

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33**: 831–838.

Amano T, Sagai T, Tanabe H, Mizushina Y, Nakazawa H, Shiroishi T. 2009. Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. *Dev Cell*.

Amoutzias GD, Veron a. S, Weiner J, Robinson-Rechavi M, Bornberg-Bauer E, Oliver SG, Robertson DL. 2007. One billion years of bZIP transcription factor evolution: Conservation and change in dimerization and DNA-binding site specificity. *Mol Biol Evol* **24**: 827–835.

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al. 2014. An atlas of active enhancers across human cell types and tissues. *Nature* **507**: 455–61.

Arnosti DN, Kulkarni MM. 2005. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J Cell Biochem* **94**: 890–898.

Arplt D, Jastrzębskl S, Bailas N, Krueger D, Bengio E, Kanwal MS, Maharaj T, Fischer A, Courville A, Benglo Y, et al. 2017. A closer look at memorization in deep networks. In *34th International Conference on Machine Learning, ICML 2017*.

Arunachalam M, Jayasurya K, Tomancak P, Ohler U. 2010. An alignment-free method to identify candidate orthologous enhancers in multiple Drosophila genomes. *Bioinformatics* **26**: 2109–2115.

Arvey A, Agius P, Noble WS, Leslie C. 2012. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res* **22**: 1723–1734.

Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*.

Banerji J, Rusconi S, Schaffner W. 1981a. Expression of a ??-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**: 299–308.

Banerji J, Rusconi S, Schaffner W. 1981b. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell*.

Benton ML, Talipineni SC, Kostka D, Capra JA. 2017. Genome-wide Enhancer Maps Differ

Significantly in Genomic Distribution, Evolution, and Function. *bioRxiv* 1–23.

Benveniste D, Sonntag H-J, Sanguinetti G, Sproul D. 2014. Transcription factor binding predicts histone modifications in human cell lines. *Proc Natl Acad Sci U S A* **111**: 13367–13372.

Bergstra J, Yamins D, Cox DD. 2013. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. *12th PYTHON Sci CONF (SCIPY 2013)* 13–20.

Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**: 57–74.

Berthelot C, Villar D, Horvath J, Odom DT, Flicek P. 2017. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *BioRxiv* 1–31.

Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette J, Jiang L, et al. 2014a. Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**: 453–456.

Boyle AP, Araya CL, Brdlik C, Cayting P, Cheng C, Cheng Y, Gardner K, Hillier LW, Janette J, Jiang L. 2014b. Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**: 453–456.

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell*.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343–348.

Brazel AJ, Vernimmen D. 2016. The complexity of epigenetic diseases. *J Pathol* **238**: 333–344.

Burzynski GM, Reed X, Taher L, Stine ZE, Matsui T, Ovcharenko I, McCallion AS. 2012. Systematic elucidation and in vivo validation of sequences enriched in hindbrain transcriptional control. *Genome Res* **22**: 2278–2289.

C. elegans sequencing consortium. 1998. Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium [published erratum appears in Science 1999 Jan 1;283(5398):35]. *Science (80-)*.

Carvunis AR, Wang T, Skola D, Yu A, Chen J, Kreisberg JF, Ideker T. 2015. Evidence for a common evolutionary rate in metazoan transcriptional networks. *Elife* **4**.

Chan ET, Quon GT, Chua G, Babak T, Trochesset M, Zirngibl RA, Aubin J, Ratcliffe MJHH, Wilde A, Brudno M, et al. 2009. Conservation of core gene expression in vertebrate tissues. *J Biol* **8**: 33.

Chen L, Fish AE, Capra JA. 2018. Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. *PLoS Comput Biol*.

Cheng Q, Kazemian M, Pham H, Blatti C, Celniker SE, Wolfe SA, Brodsky MH, Sinha S. 2013. Computational Identification of Diverse Mechanisms Underlying Transcription Factor-DNA Occupancy. *PLoS Genet* **9**.

Cheng Y, Ma Z, Kim B-HH, Wu W, Cayting P, Boyle AP, Sundaram V, Xing X, Dogan N, Li J, et al. 2014. Principles of regulatory information conservation between mouse and human. *Nature* **515**: 371–375.

Chesmore KN, Bartlett J, Cheng C, Williams SM. 2016. Complex patterns of association between pleiotropy and transcription factor evolution. *Genome Biol Evol* **8**: 3159–3170.

Chollet F, others. 2015. Keras. *GitHub Repos*.

Chuong EB, Rumi MAK, Soares MJ, Baker JC. 2013. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat Genet* **45**: 325–329.

Consortium RE, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–330.

Conway JR, Lex A, Gehlenborg N. 2017. UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**: 2938–2940.

Corradin O, Scacheri PC. 2014. Enhancer variants: Evaluating functions in common disease. *Genome Med* **6**: 85.

Cotney J, Leng J, Oh S, DeMare LE, Reilly SK, Gerstein MB, Noonan JP. 2012. Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. *Genome Res* **22**: 1069–1080.

Cotney J, Leng J, Yin J, Reilly SK, DeMare LE, Emera D, Ayoub AE, Rakic P, Noonan JP. 2013. The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell* **154**: 185–196.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA, et al. 2010. Histone H3K27ac separates active from poised

enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**: 21931–21936.

Crocker J, Tamori Y, Erives A. 2008. Evolution acts on enhancer organization to fine-tune gradient threshold readouts. *PLoS Biol* **6**: 2576–2587.

Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, Wang D, Masys DR, Roden DM, Crawford DC. 2010. PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*.

Di Iulio J, Bartha I, Wong EHM, Yu HC, Lavrenko V, Yang D, Jung I, Hicks MA, Shah N, Kirkness EF, et al. 2018. The human noncoding genome defined by genetic diversity. *Nat Genet*.

Dogan N, Wu W, Morrissey CS, Chen KB, Stonestrom A, Long M, Keller CA, Cheng Y, Jain D, Visel A, et al. 2015. Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics and Chromatin* **8**.

Dooley S, ten Dijke P. 2012. TGF-β in progression of liver disease. *Cell Tissue Res* **347**: 245–56.

Dror I, Golan T, Levy C, Rohs R, Mandel-Gutfreund Y. 2015. A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Res* **25**: 1268–1280.

Eddy SR. 2012. The C-value paradox, junk DNA and ENCODE. *Curr Biol*.

Eddy SR. 2013. The ENCODE project: Missteps overshadowing a success. *Curr Biol*.

Eferl R, Sibilia M, Hilberg F, Fuchsbichler A, Kufferath I, Guertl B, Zenz R, Wagner EF, Zatloukal K. 1999. Functions of c-Jun in liver and heart development. *J Cell Biol* **145**: 1049–1061.

Erceg J, Saunders TE, Girardot C, Devos DP, Hufnagel L, Furlong EEMM. 2014. Subtle Changes in Motif Positioning Cause Tissue-Specific Effects on Robustness of an Enhancer's Activity. *PLoS Genet* **10**: e1004060.

Erhan D, Bengio Y, Courville A, Vincent P. 2009. Visualizing higher-layer features of a deep network. *Bernoulli*.

Erives A, Levine M. 2004. Coordinate enhancers share common organizational features in the Drosophila genome. *Proc Natl Acad Sci U S A* **101**: 3851–6.

Ernst J, Kellis M. 2012. ChromHMM: Automating chromatin-state discovery and characterization. *Nat Methods* **9**: 215–216.

Erwin GD, Oksenberg N, Truty RM, Kostka D, Murphy KK, Ahituv N, Pollard KS, Capra JA.

2014. Integrating Diverse Datasets Improves Developmental Enhancer Prediction. *PLoS Comput Biol* **10**: e1003677.

Fish A, Chen L, Capra JA. 2017. Gene regulatory enhancers with evolutionarily conserved activity aremore pleiotropic than those with species-specific activity. *Genome Biol Evol* **9**: 2615–2625.

Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science (80-)* **312**: 276–279.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res* **42**: 749–755.

Fu Y, Liu Z, Lou S, Bedford J, Mu XJ asmin., Yip KY, Khurana E, Gerstein M. 2014. FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol*.

Galis F, Van Dooren TJMM, Metz JAJJ. 2002. *Conservation of the segmented germband stage: Robustness or pleiotropy?*

Gens R, Domingos P. 2012. Discriminative Learning of Sum-Product Networks. *Nips* 1–9.

Ghandi M, Lee D, Mohammad-Noori M, Beer MA. 2014. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* **10**: e1003711.

Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res*.

Gorkin DU, Lee D, Reed X, Fletez-Brant C, Bessling SL, Loftus SK, Beer M a., 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.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017–1018.

Guillaume F, Otto SP. 2012. Gene functional trade-offs and the evolution of pleiotropy. *Genetics* **192**: 1389 LP – 1409.

Guillon N, Tirode F, Boeva V, Zynovyev A, Barillot E, Delattre O. 2009. The oncogenic EWS-FLI1 protein binds in vivo GGAA microsatellite sequences with potential transcriptional activation function. *PLoS One* **4**: e4932.

Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between

motifs. *Genome Biol* **8**: R24.

Guturu H, Chinchali S, Clarke SL, Bejerano G. 2016. Erosion of Conserved Binding Sites in Personal Genomes Points to Medical Histories. *PLOS Comput Biol* **12**: e1004711.

Hahn MW, Wray GA. 2002. The g-value paradox. *Evol Dev*.

Han B, Yao Q, Yu X, Niu G, Xu M, Hu W, Tsang IW, Sugiyama M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*.

He K, Zhang X, Ren S, Sun J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

He X, Zhang J. 2006. Toward a molecular understanding of pleiotropy. *Genetics* **173**: 1885–1891.

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

Hernandez RD, Uricchio LH, Hartman K, Ye C, Dahl A, Zaitlen N. 2019. Ultrarare variants drive substantial cis heritability of human gene expression. *Nat Genet*.

Hindorff LA, Bonham VL, Brody LC, Ginoza MEC, Hutter CM, Manolio TA, Green ED. 2018. Prioritizing diversity in human genomics research. *Nat Rev Genet*.

Hornik K. 1991. Approximation capabilities of multilayer feedforward networks. *Neural Networks*.

Hsu C-H, Ovcharenko I. 2013. Effects of gene regulatory reprogramming on gene expression in human and mouse developing hearts. *Philos Trans R Soc Lond B Biol Sci* **368**: 20120366.

Huang Y-FF, Gulko B, Siepel A. 2017. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet* 069682.

Iguyon I, Elisseeff A. 2003. An introduction to variable and feature selection. *J Mach Learn Res*.

Ioffe S, Szegedy C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv Prepr arXiv150203167*.

Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. 2016. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* **advance on**: 214–220.

Jiang L, Zhou Z, Leung T, Li LJ, Fei-Fei L. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *35th International Conference on Machine Learning, ICML 2018*.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science (80-)*.

Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al. 2013a. DNA-binding specificities of human transcription factors. *Cell* **152**: 327–339.

Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G. 2013b. DNA-binding specificities of human transcription factors. *Cell* **152**: 327–339.

Jolma A, Yin Y, Nitta KR, Dave K, Popov A, Taipale M, Enge M, Kivioja T, Morgunova E, Taipale J. 2015. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **527**: 384–8.

Kapur K, Schüpbach T, Xenarios I, Kutalik Z, Bergmann S. 2011. Comparison of strategies to detect epistasis from eQTL data. *PLoS One* **6**: e28415.

Kazemian M, Pham H, Wolfe SA, Brodsky MH, Sinha S. 2013. Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development. *Nucleic Acids Res* **41**: 8237–8252.

Kazemian M, Suryamohan K, Chen JY, Zhang Y, Samee MAH, Halfon MS, Sinha S. 2014. Evidence for deep regulatory similarities in early developmental programs across highly diverged insects. *Genome Biol Evol* **6**: 2301–2320.

Kelley DR, Snoek J, Rinn JL. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* **26**: 990–999.

Khera A V., Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, et al. 2018. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science (80-)*.

Kingma DPP, Ba J. 2014. Adam: A method for stochastic optimization. *arXiv Prepr arXiv14126980*.

Kircher M, Witten DM, Jain P, O'roak BJ, Cooper GM, Shendure J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*.

Kircher M, Xiong C, Martin B, Schubach M, Inoue F, Bell RJA, Costello JF, Shendure J, Ahituv N. 2019. Saturation mutagenesis of twenty disease-associated regulatory elements at single

base-pair resolution. *Nat Commun*.

Kleftogiannis D, Kalnis P, Bajic VB. 2016. Progress and challenges in bioinformatics approaches for enhancer identification. *Brief Bioinform* **17**: 967–979.

Kulakovskiy I V., Vorontsov IE, Yevshin IS, Soboleva A V., Kasianov AS, Ashoor H, Ba-Alawi W, Bajic VB, Medvedeva YA, Kolpakov FA, et al. 2016. HOCOMOCO: Expansion and enhancement of the collection of transcription factor binding sites models. *Nucleic Acids Res* **44**: D116–D125.

Kulakovskiy IV V, Vorontsov IEE, Yevshin ISS, Sharipov RNN, Fedorova ADD, Rumynskiy EII, Medvedeva YAA, Magana-Mora A, Bajic VBB, Papatsenko DAA, et al. 2017. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res* **46**: D252--D259.

Kulkarni MM, Arnosti DN. 2003. Information display by transcriptional enhancers. *Development* **130**: 6569–75.

Kumar S, Bucher P. 2016. Predicting transcription factor site occupancy using DNA sequence intrinsic and cell-type specific chromatin features. *BMC Bioinformatics* **17**: 4.

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* **518**: 317–329.

Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The Human Transcription Factors. *Cell*.

Lanchantin J, Singh R, Wang B, Qi Y. 2017. Deep Motif Dashboard : Visualizing and Understanding Genomic Sequences Using Deep Neural Networks. *bioRxiv*.

Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, Beer MA. 2015. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* **47**: 955–61.

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

Lee PH, Lee C, Li X, Wee B, Dwivedi T, Daly M. 2018. Principles and methods of in-silico prioritization of non-coding regulatory variants. *Hum Genet*.

Leslie C, Eskin E, Noble WS. 2002. The spectrum kernel: a string kernel for SVM protein classification. *Pac Symp Biocomput* **575**: 564–575.

Lettice LA, Heaney SJH, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de

Graaff E. 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet*.

Leung MKK, Xiong HY, Lee LJ, Frey BJ. 2014. Deep learning of the tissue-regulated splicing code. *Bioinformatics* **30**.

Levy S, Hannenhalli S. 2002a. Identification of transcription factor binding sites in the human genome sequence. *Mamm Genome* **13**: 510–514.

Levy S, Hannenhalli S. 2002b. Identification of transcription factor binding sites in the human genome sequence. *Mamm Genome* **13**: 510–514.

Li S, Ovcharenko I. 2015. Human enhancers are fragile and prone to deactivating mutations. *Mol Biol Evol* **32**: 2161–2180.

Li Y, Chen C yu, Kaye AM, Wasserman WW. 2015. The identification of cis-regulatory elements: A review from a machine learning perspective. *BioSystems*.

Liu G, Gifford D. Visualizing Feature Maps in Deep Neural Networks using DeepResolve A Genomics Case Study.

Liu L, Zhao W, Zhou X. 2015. Modeling co-occupancy of transcription factors using chromatin features. *Nucleic Acids Res* **44**.

Liu Y, Hauser M. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Chemtracts*.

Long HK, Prescott SL, Wysocka J. 2016. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell* **167**.

Lu Z, Pu H, Wang F, Hu Z, Wang L. 2017. The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems*.

Ludwig MZ, Manu, Kittler R, White KP, Kreitman M. 2011. Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. *PLoS Genet* **7**: e1002364.

Luna-Zurita L, Stirnimann CU, Glatt S, Kaynak BL, Thomas S, Baudin F, Samee MAH, He D, Small EM, Mileikovsky M, et al. 2016. Complex Interdependence Regulates Heterotypic Transcription Factor Distribution and Coordinates Cardiogenesis. *Cell*.

Lundberg SMM, Lee S-I. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30* (eds. I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett), pp. 4765–4774, Curran

Associates, Inc.

Mahendran A, Vedaldi A. 2015a. Understanding deep image representations by inverting them. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Mahendran A, Vedaldi A. 2015b. Visualizing Deep Convolutional Neural Networks Using Natural Pre-Images. *arXiv Prepr arXiv151202017*.

Mahrenholz CC, Abfalter IG, Bodenhofer U, Volkmer R, Hochreiter S. 2011. Complex Networks Govern Coiled-Coil Oligomerization – Predicting and Profiling by Means of a Machine Learning Approach. *Mol Cell Proteomics* **10**: M110.004994.

Maniatis T, Falvo J V., Kim TH, Kim TK, Lin CH, Parekh BS, Wathelet MG. 1998. Structure and function of the interferon-beta enhanceosome. *Cold Spring Harb Symp Quant Biol* **63**: 609–620.

Martin AR, Daly MJ, Robinson EB, Hyman SE, Neale BM. 2019. Predicting Polygenic Risk of Psychiatric Disorders. *Biol Psychiatry*.

Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, Shi W, Shyr C, Tan G, Worsley-Hunt R, et al. 2016. JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **44**: D110–D115.

Mathelier A, Wasserman WW. 2013. The next generation of transcription factor binding site prediction. *PLoS Comput Biol* **9**: e1003214.

Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen C, Chou A, Ienasescu H, et al. 2014. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* **42**: D142--7.

Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science (80-)* **337**: 1190–1195.

Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Kinney JB, et al. 2012. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol*.

Mendenhall EM, Williamson KE, Reyon D, Zou JY, Ram O, Joung JK, Bernstein BE. 2013. Locus-specific editing of histone modifications at endogenous enhancers. *Nat Biotechnol*.

Merkin J, Russell C, Chen P, Burge CB. 2012. Evolutionary dynamics of gene and isoform

regulation in Mammalian tissues. *Science (80-)* **338**: 1593–1599.

Miguel-Escalada I, Pasquali L, Ferrer J. 2015. Transcriptional enhancers: Functional insights and role in human disease. *Curr Opin Genet Dev*.

Min X, Chen N, Chen T. 2016. DeepEnhancer : Predicting Enhancers by Convolutional Neural Networks. 637–644.

Min X, Chen N, Chen T, Jiang R. 2017. DeepEnhancer: Predicting enhancers by convolutional neural networks. In *Proceedings - 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016*, pp. 637–644.

Mirsky A., Hans R. 1951. The DNA content of animal cells and its evolutionary significance. *J Gen Physiol*.

Mordvintsev A, Olah C, Tyka M. 2015. Inceptionism: Going Deeper into Neural Networks. *Res Blog*.

Movva R, Greenside P, Marinov GK, Nair S, Shrikumar A, Kundaje A. 2019. Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. *PLoS One*.

Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML. 2004. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet*.

Need AC, Goldstein DB. 2009. Next generation disparities in human genomics: concerns and remedies. *Trends Genet*.

Nguyen A, Yosinski J, Clune J. 2016. Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks. *Arxiv* 23.

Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EEM, et al. 2015a. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife* **2015**.

Nitta KR, Jolma A, Yin Y, Morgunova E, Kivioja T, Akhtar J, Hens K, Toivonen J, Deplancke B, Furlong EEM, et al. 2015b. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife* **4**: 1–20.

Nord AS, Blow MJ, Attanasio C, Akiyama JA, Holt A, Hosseini R, Phouanenavong S, Plajzer-Frick I, Shoukry M, Afzal V, et al. 2013. Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155**: 1521–1531.

Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, MacIsaac KD, Rolfe PA, Conboy CM, Gifford DK, Fraenkel E. 2007. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39**: 730–732.

Ohno S. 1972. So much "junk" DNA in our genome. *Brookhaven Symp Biol*.

Olah C, Mordvintsev A, Schubert L. 2017. Feature Visualization. *Distill*.

Olah C, Satyanarayan A, Johnson I, Carter S, Schubert L, Ye K, Mordvintsev A. 2018. The Building Blocks of Interpretability. *Distill*.

Orgel LE, Crick FHC. 1980. Selfish DNA: The ultimate parasite. *Nature*.

Osterwalder M, Barozzi I, Tissiéres V, Fukuda-Yuzawa Y, Mannion BJ, Afzal SY, Lee EA, Zhu Y, Plajzer-Frick I, Pickle CS, et al. 2018. Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature*.

Paaby AB, Rockman M V. 2013. *The many faces of pleiotropy*.

Palme J, Hochreiter S, Bodenhofer U. 2015. KeBABS: an R package for kernel-based analysis of biological sequences. *Bioinformatics* 1–3.

Papakostas S, Vøllestad LA, Bruneaux M, Aykanat T, Vanoverbeke J, Ning M, Primmer CR, Leder EH. 2014. Gene pleiotropy constrains gene expression changes in fish adapted to different thermal conditions. *Nat Commun* **5**: 4071.

Papatsenko D, Levine M. 2007. A rationale for the enhanceosome and other evolutionarily constrained enhancers. *Curr Biol* **17**.

Philipsen S, Suske G. 1999. A tale of three fingers: the family of mammalian Sp/XKLF transcription factors. *Nucleic Acids Res* **27**: 2991–3000.

Polychronopoulos D, King JWD, Nash AJ, Tan G, Lenhard B. 2017. Conserved non-coding elements: Developmental gene regulation meets genome organization. *Nucleic Acids Res*.

Poon H, Domingos P. 2011. Sum-product networks: A new deep architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 689–690.

Popejoy AB, Fullerton SM. 2016. Genomics is failing on diversity. *Nature*.

Prescott SL, Srinivasan R, Marchetto MC, Gage FH, Swigut T, Selleri L, Gage FH, Swigut T, Wysocka J. 2015a. Enhancer Divergence and cis -Regulatory Evolution in the Human and Chimp Neural Crest Article Enhancer Divergence and cis -Regulatory Evolution in the Human and Chimp Neural Crest. 68–83.

Prescott SL, Srinivasan R, Marchetto MC, Grishina I, Narvaiza I, Selleri L, Gage FH, Swigut T,

Wysocka J. 2015b. Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell* **163**: 68–83.

Quang D, Xie X. 2015. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *bioRxiv* **44**: 032821.

Quang D, Xie X. 2016. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *bioRxiv* **44**: 032821.

Quang D, Xie X. 2019. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods* 1–28.

Rajagopal N, Xie W, Li Y, Wagner U, Wang W, Stamatoyannopoulos J, Ernst J, Kellis M, Ren B. 2013. RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. *PLoS Comput Biol*.

Rebollo R, Romanish MT, Mager DL. 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet* **46**: 21–42.

Reilly SK, Yin J, Ayoub AE, Emera D, Leng J, Cotney J, Sarro R, Rakic P, Noonan JP. 2015. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science (80-)* **347**: 1155–1159.

Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*.

Ritchie GRS, Dunham I, Zeggini E, Flicek P. 2014. Functional annotation of noncoding sequence variants. *Nat Methods*.

Ritter DI, Li Q, Kostka D, Pollard KS, Guo S, Chuang JH. 2010. The importance of Being Cis: Evolution of Orthologous Fish and Mammalian enhancer activity. *Mol Biol Evol* **27**: 2322–2332.

Romero IG, Ruvinsky I, Gilad Y. 2012. *Comparative studies of gene expression and the evolution of gene regulation*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.

Romiguier J, Ranwez V, Douzery EJPP, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Res* **20**: 1001–1009.

Rudolph KLM, Schmitt BM, Villar D, White RJ, Marioni JC, Kutter C, Odom DT. 2016. Codon-Driven Translational Efficiency Is Stable across Diverse Mammalian Cell States. *PLoS Genet*

12.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-jimenez CP, Mackay S, et al. 2010. Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science (80-)* **328**: 1036–1041.

Schoenfelder S, Fraser P. 2019. Long-range enhancer–promoter contacts in gene expression control. *Nat Rev Genet*.

Sharmin M, Corrada Bravo H, Hannenhalli SS. 2015. Heterogeneity of Transcription Factor binding specificity models within and across cell lines. *bioRxiv* **8219**: 028787.

Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, Gaunt TR, Campbell C. 2015. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*.

Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: from properties to genome-wide predictions. *Nat Rev Genet* **15**: 272–286.

Shrikumar A, Greenside P, Kundaje A. 2017a. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3145–3153.

Shrikumar A, Greenside P, Shcherbina A, Kundaje A, Shcherbina A, Kundaje A, Shcherbina A, Kundaje A. 2017b. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3145–3153.

Simonti CN, Pavličev M, Capra JA. 2017. Transposable Element Exaptation into Regulatory Regions is Rare, Influenced by Evolutionary Age, and Subject to Pleiotropic Constraints. *Mol Biol Evol*.

Simonyan K, Vedaldi A, Zisserman A. 2013a. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv Prepr arXiv13126034*.

Simonyan K, Vedaldi A, Zisserman A. 2013b. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps . *arXiv.org* **cs.CV**.

Singh S, Yang Y. 2016. Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep Neural Networks. 1–12.

Slattery M, Zhou T, Yang L, Dantas Machado AC, Gordân R, Rohs R, Gordan R, Rohs R. 2014.

Absence of a simple code: how transcription factors read the genome. *Trends Biochem Sci* **39**: 381–399.

Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, Lee BT, et al. 2016. ENCODE data at the ENCODE portal. *Nucleic Acids Res* **44**: D726–D732.

Smemo S, Campos LC, Moskowitz IP, Krieger JE, Pereira AC, Nobrega MA. 2012. Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum Mol Genet*.

Smit A, Hubley R, Green P. 2013. RepeatMasker Open-4.0. 2013-2015 . *http://www.repeatmasker.org*.

Smith RP, Taher L, Patwardhan RP, Kim MJ, Inoue F, Shendure J, Ovcharenko I, Ahituv N. 2013. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet* **45**: 1021–8.

Sorge S, Ha N, Polychronidou M, Friedrich J, Bezdan D, Kaspar P, Schaefer MH, Ossowski S, Henz SR, Mundorf J, et al. 2012. The cis-regulatory code of Hox function in Drosophila. *EMBO J* **31**: 3323–33.

Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. 2014. Striving for Simplicity: The All Convolutional Net. *ICLR 2015*.

Stampfel G, Kazmar T, Frank O, Wienerroither S, Reiter F, Stark A. 2015. Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature* **528**: 147–51.

Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, Karagianni P, Brazma A, Adams DJ, Talianidis I, Marioni JC, et al. 2013. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell* **154**: 530–540.

Stergachis AB, Neph S, Sandstrom R, Haugen E, Reynolds AP, Zhang M, Byron R, Canfield T, Stelhing-Sun S, Lee K, et al. 2014. Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature* **515**: 365–370.

Su M, Han D, Boyd-Kirkup J, Yu X, Han J-DJ. 2014. *Evolution of Alu Elements toward Enhancers*.

Swanson CI, Evans NC, Barolo S. 2010. Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. *Dev Cell* **18**: 359–70.

Swanson CI, Schwimmer DB, Barolo S. 2011. Rapid evolutionary rewiring of a structurally

constrained eye enhancer. *Curr Biol* **21**: 1186–1196.

Taher L, McGaughey DM, Maragh S, Aneas I, Bessling SL, Miller W, Nobrega MA, McCallion AS, Ovcharenko I. 2011. Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Res* **21**: 1139–1149.

Taher L, Narlikar L, Ovcharenko I. 2012. Clare: Cracking the LAnguage of regulatory elements. *Bioinformatics* **28**: 581–583.

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

The Gene Ontology Consortium. 2000. Gene Ontology: tool for the unification of biology. *Nat Genet* **25**: 25–29.

The Gene Ontology Consortium. 2015. Gene Ontology Consortium: going forward. *Nucleic Acids Res* **43**: D1049–D1056.

The GTEx Consortium, GTEx Consortium Gte, Ardlie KG, DeLuca DS, Segrè A V., Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, et al. 2015. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-)* **348**: 648–60.

Thomas CAA. 1971. The Genetic Organization of Chromosomes. *Annu Rev Genet*.

Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, et al. 2015. Tissue-based map of the human proteome. *Science (80-)* **347**: 1260419–1260419.

van Arensbergen J, Pagie L, FitzPatrick VD, de Haas M, Baltissen MP, Comoglio F, van der Weide RH, Teunissen H, Võsa U, Franke L, et al. 2019. High-throughput identification of human SNPs affecting regulatory element activity. *Nat Genet*.

Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. 2009. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **10**: 252–263.

Varki A, Altheide TK. 2005. Comparing the human and chimpanzee genomes: Searching for needles in a haystack. *Genome Res*.

Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ. 2015a. Enhancer evolution across 20 mammalian species. *Cell* **160**: 554–566.

Villar D, Berthelot C, Flicek P, Odom DT, Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M,

Pignatelli M. 2015b. Enhancer Evolution across 20 Mammalian Species. *Cell* **160**: 554–566.

Villar D, Berthelot C, Flicek P, Odom DTT, Villar D, Berthelot C, Aldridge S, Rayner TFF, Lukk M, Pignatelli M, et al. 2015c. Enhancer Evolution across 20 Mammalian Species. *Cell* **160**: 554–566.

Villar D, Flicek P, Odom DT. 2014a. Evolution of transcription factor binding in metazoans-mechanisms and functional implications. *Nat Rev Genet*.

Villar D, Flicek P, Odom DT. 2014b. Evolution of transcription factor binding in metazoans—mechanisms and functional implications. *Nat Rev Genet* **15**: 221–233.

Villar D, Flicek P, Odom DT. 2014c. Evolution of transcription factor binding in metazoans - mechanisms and functional implications. *Nat Rev Genet* **15**: 221–233.

Visel A, Minovitsky S, Dubchak I, Pennacchio LA. 2007. VISTA Enhancer Browser — a database of tissue-specific human enhancers. **35**: 88–92.

Wamstad JA, Alexander JM, Truty RM, Shrikumar A, Li F, Eilertson KE, Ding H, Wylie JN, Pico AR, Capra JA, et al. 2012. Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell*.

Wang J, Zhuang J, Iyer S, Lin XY, Whitfield TW, Greven MC, Pierce BG, Dong X, Kundaje A, Cheng Y, et al. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res*.

Wang L, Jensen S, Hannenhalli S. 2006. An interaction-dependent model for transcription factor binding. *Syst Biol Regul Genomics* 225–234.

Wei G-H, Badis G, Berger MF, Kivioja T, Palin K, Enge M, Bonke M, Jolma A, Varjosalo M, Gehrke AR, et al. 2010a. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J* **29**: 2147–60.

Wei G-HH, Badis G, Berger MF, Kivioja T, Palin K, Enge M, Bonke M, Jolma A, Varjosalo M, Gehrke AR, et al. 2010b. Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo. *EMBO J* **29**: 2147–60.

Whitaker JW, Chen Z, Wang W. 2015. Predicting the human epigenome from DNA motifs. *Nat Methods* **12**: 265–272.

Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VLJJ, Fisher EMCC, Tavaré S, Odom DT. 2008. Species-specific transcription in mice carrying human chromosome 21. *Science* **322**: 434–8.

Woo YH, Li WH. 2012. Evolutionary conservation of histone modifications in mammals. *Mol Biol Evol* **29**: 1757–1767.

Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, et al. 2015. The human splicing code reveals new insights into the genetic determinants of disease. *Science (80-)* **347**: 1254806–1254806.

Yamazaki H, Katsuoka F, Motohashi H, Engel JD, Yamamoto M. 2012. Embryonic lethality and fetal liver apoptosis in mice lacking all three small Maf proteins. *Mol Cell Biol* **32**: 808–16.

Yáñez-Cuna JO, Kvon EZ, Stark A. 2013. Deciphering the transcriptional cis-regulatory code. *Trends Genet* **29**: 11–22.

Yang B, Liu F, Ren C, Ouyang Z, Xie Z, Bo X, Shu W. 2017. BiRen: Predicting enhancers with a deep-learning-based model using the DNA sequence alone. *Bioinformatics* **33**: 1930–1936.

Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. 2015. Understanding Neural Networks Through Deep Visualization. *Int Conf Mach Learn - Deep Learn Work 2015* 12.

Zeiler MD, Fergus R. 2014. Visualizing and Understanding Convolutional Networks arXiv:1311.2901v3 [cs.CV] 28 Nov 2013. *Comput Vision–ECCV 2014* **8689**: 818–833.

Zeiler MDD, Krishnan D, Taylor GWW, Fergus R. 2010. Deconvolutional networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. 2018. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* **50**: 1171.

Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* **12**: 931–4.