# DEVELOPMENT AND APPLICATIONS OF SHAPE-BASED DNA MOTIFS

by

**Peter M. DeFord**

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

Baltimore, Maryland

January, 2021

# Abstract

Transcriptional regulation is imperative for proper development of multicellular organisms, and disregulation of this process can lead to genetic disease. Due to technical limitations, the full human regulome has not been assayed. Computational methods provide resources to fill the gaps in our understanding of these processes. Sequence based representations of transcription factor DNA motifs have long been used for this purpose. We developed a model based on estimates of DNA shape known as Structural Motifs, extending the position weight matrix to accommodate multiple continuous shape parameters at each position. Using expectation maximization, Structural Motifs are discovered *de novo* from transcription factor binding data, and these motifs are specific to their cognate factors. When considered jointly with sequence motifs, Structural Motifs improve classification of transcription factor binding sites. Joint models also provide insight into the readout mechanisms utilized by transcription factors. DNA shape is an important component of the protein-DNA interaction to consider and improves the computational predictions of transcription factor binding, elevating our understanding of the regulatory landscape.

Advisors: James Taylor, Vince Hilser. Readers: Vince Hilser and Michael Schatz.

# Dedication

This work is dedicated to James and all of the values he stood for and inspired others with.  Here's to open science, collaboration, inclusion, and the thoughtful training and encouragement of the next generation.

# Preface

This work was supported by funding from the National Institutes of Health. Specifically the JHU CMDB training grant T32-GM-007231 from the National Institute of General Medical Sciences, as well as R24-DK-106766 granted to the VISION consortium by the National Institute of Diabetes and Digestive and Kidney Diseases.

Chapter 3 (pages 30–47) was posted as a preprint on *bioRxiv* on June 17, 2019. Peter DeFord conceptualized, planned, analyzed, and evaluated the data with support from his advisor James Taylor.

# Acknowledgements

I would like to thank my advisor, James Taylor, for his role in seeing this project through to completion. I joined his lab with no computational experience outside of the week-long bootcamp-style course he taught my first year. Being new to the field, I had very little exposure to many methods, much less well established theories. James helped guide me to the resources I needed, but was also game for diving in and trying things, regardless of the outlandishness of my suggestions. This freedom to explore new ideas stimulated my excitement, and helped carry me through. James also provided me many opportunities to grow, both in the lab and in the classroom.

I also appreciate his support when I was struck with the “two-body problem”, and I began to work remotely part-time. I would like to thank my wife, Sarah Dallas, for her support. There were long nights, travelling, long commutes, more time away from family than we planned, and so much more. She was a trooper through it all, and I appreciate her care and support.

Thanks to all of my friends and colleagues who remained excited and supportive, even when we would go long stretches without talking. You are all appreciated.

# Contents

# List of Tables

# List of Figures

# 1. Introduction

The human body is composed of trillions of cells arranged in complex systems and structures, all working together to carry out the various functions required for life [1]. This feat of coordination is all the more remarkable when considering that all of these cells are working from the same set of instructions: they have identical copies of the DNA code that specifies their jobs [2, 3]. In order to achieve the diversity observed in cellular morphology and function, tight control over the expression of the genes in each cell is required [4, 5, 6]. This is accomplished in large part through the expression and activity of transcriptional regulators.

Sequence specific transcription factors recognize and bind to particular regions of the genome, and through interactions with other proteins serve to modulate the transcription of their target genes[6, 7]. The modulations may be activating, thereby increasing the rate of transcription, or repressive, driving down the rate of transcription of the target gene. This layer of regulation allows for fine tuning of both the temporal and spatial expression of genes across diverse cell types [8, 9].

Variation at the genomic sites that are bound by these transcription factors is a strong contributor to phenotypic diversity across the population [10]. Loss of regulation at a specific locus can contribute not only to heterogeneity of individuals, but also to the development of disease [11]. As such, assaying all of these sites, across all tissue types would reveal mechanisms of disease and candidates for treatment. With current technology this is, however, virtually impossible. Many

tissues and cell types are difficult to sample in high enough quantity and purity to effectively characterize, and antibodies that show high specificity are lacking for many known transcription factors [12].

Computational methods for characterizing the binding specificity of transcription factors seek to fill this gap. By modelling and understanding the binding preferences exhibited by a transcription factor, putative binding sites can be identified in additional cell types by extrapolating based on other experimental data. This avenue of identifying binding sites also allows a larger number of transcription factors to be characterized, as binding preferences can be observed in *in vitro* experiments using purified protein as opposed to relying on specific antibodies for immunoprecipitation.

To date, one of the most widely used models of transcription factor binding specificity has been the position weight matrix [13]. This model is a simple probability matrix that models the probability of seeing any given nucleotide in sequence along the binding site. The abstraction of the DNA molecule to sequences of letters has permitted the use of high throughput technologies for assaying transcription factor binding sites, such as PBMs, ChIP-seq, and SELEX, and has proven to be quite a robust representation [14]. However, considering the three dimensional structure of the interacting molecules (the transcription factor and the DNA) can provide additional valuable information about the interaction, and therefore about the impact of variation in either molecule on that interaction [15].

In this work, we present a method for describing the binding preferences of transcription factors by characterizing the distribution of observed three dimensional DNA shapes at each position across the sequence. Supporting this work we developed computational methods for discovering the shape motifs from high throughput sequencing experiments, and support for making binding predictions for additional sequences. A new Python package provides all of these functions for use in a

scientific setting.

## 1.1 Discovery and Description of Sequence-Specific Transcriptional Regulators

In 1961, eight years after the structure of DNA was discovered [16], but long before large-scale genome sequencing was common, “structural genes” were a known component of cells [17]. These genes were defined as those that provide the template for a RNA or protein product that carries out one of the many cellular processes, for example as enzymes or cytoskeletal factors. Consolidating the data of the time, largely from the *lac* operon in *Escherichia coli*, Jacob and Monod proposed a revolutionary hypothesis. They deduced that there must be another type of element they termed “regulator genes”. They further postulated that these regulators must produce a cytoplasmic product that interacts specifically with certain “operator” sites on the chromosomes to exert their regulatory control on the production of “structural genes” [18].

Following the purification in 1967 of $\lambda$ phage repressor, the cytoplasmic product of one of these regulator genes, M. Ptashne demonstrated that these factors bind preferentially and specifically to particular DNA sequences [19]. It was later hypothesized that the hydrogen-bonding potential in the major groove is sufficient to differentiate between each nucleotide, and that specific amino acid side chains may be able to specifically recognize these differences [20].

As technology advanced, and collections of sequences were curated, methods were proposed to represent the specific sequences recognized by these so-called regulator genes. Since there was some variation in the sequences preferred by a given regulator, a probabilistic model based on a “Perceptron” was used to describe the sites [13]. Encoding the rules dictating the specificity of the interaction in this

way allowed for computational prediction of additional sites where the regulator may bind.

## 1.2 Common Structures of DNA binding domains

As more transcription factors have been purified and studied in detail it has been discovered that, while many factors have a distinct set of sites along the genome to which they bind, most transcription factors can be classified into one of seven main structural families describing the general shape of the DNA binding domain. This classification can be useful, as it has been observed that factors within the same family tend to have similar DNA motifs. Since the structure of the protein domain determines the interaction with DNA, and therefore the specificity of the interaction, the resultant similarity between family members is unsurprising.

### 1.2.1 Helix-Turn-Helix Proteins

As the name implies, a Helix-Turn-Helix protein is composed of a pair of $\alpha$ helices separated by a strand of linker amino acids (the turn) [21]. The two helices interact with each other to maintain a constant orientation and angle, allowing the recognition helix (the more C-terminal of the two) to set along the major groove of the DNA. These factors often function as dimers, with the two recognition helices being separated by approximately one turn of the DNA, such that both may sit in the major groove. Dimerization in this way serves to extend the size of the motif and increase the specificity of the factor [22]. Examples include lambda Cro, and the CAP fragment [23, 24].

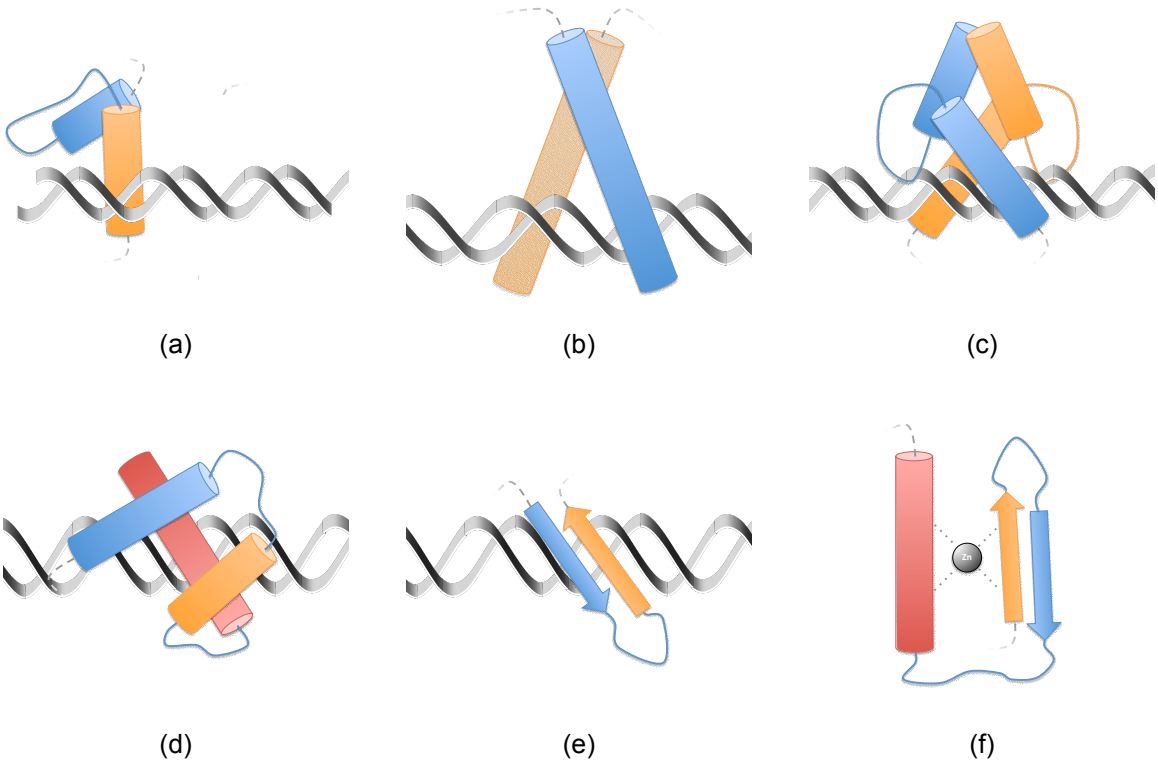



Figure 1.1: Common transcription factor structures. $\alpha$-helices are shown as cylinders, and $\beta$-strands are shown as arrows, with flexible linkers described as lines. (a) Helix-Turn-Helix. The recognition helix (orange) sits in the major groove. (b) Leucine Zipper. The two helices dimerize, forming a Y structure. (c) Helix-Loop-Helix. The C-terminal helices dimerize, positioning the N-terminal helices sit in opposing major grooves. (d) Homeodomain. The recognition helix (red) sizes in the major groove, stabilized by the other helices. (e) $\beta$ Sheet. In this example, the two $\beta$-strands are joined by a flexible linker. (f) Zinc Finger. One module is shown, with a helix and sheet stabilized by the zinc atom (black). This would be one component of a tandem set of fingers.

### 1.2.2 Leucine Zipper Proteins

This type of transcription factor gets its name from the way it dimerizes. Each component of the dimer has an $\alpha$ helix, and they "zip" together to form a coiled coil, primarily through hydrophobic interactions of leucine amino acids along the dimerization interface [25]. Beyond the dimerization region of the helices, they separate and lay in the major groove on opposites of the DNA from one another. Due to the dimeric mode of binding, these factors often bind palendromes, particularly those with a core ACGT motif, including the CACGTG (G box), GACGTC (C box), TACGTA (A box), and AACGTT (T box) motifs [26, 27]. Heterodimerization allows for more diverse binding specificities, often combining the relative specificities of each partner [22]. Common examples include c-Fos and c-Jun [28].

### 1.2.3 Helix-Loop-Helix Proteins

Similar to the leucine zipper class of transcription factors, the Helix-Loop-Helix family also uses $\alpha$ helices to dimerize and to bind the major groove of the DNA [29]. This class is distinguished from the leucine zipper in that there is a flexible linker that loops out between the dimerization and DNA binding domains [30]. The longer N-terminal helix binds the DNA, while the shorter more C-terminal helix folds such that it can interact tightly with a comparable dimerization site. This can be either as a homodimer (usually recognizing a palindromic sequence such as the canonical E-box sequence CACGTG) or a heterodimer (diversifying the range of sequences that can be recognized [31]. This is one of the largest families, and examples include CLOCK, MYC, and MYOD1 [29].

### 1.2.4 Homeodomain Proteins

Homeodomain proteins were discovered in *Drosophila melanogaster*, and are very important for proper development and patterning [32]. They extend the Helix-Turn-Helix motif by adding an additional linker and helix. This additional helix forms additional interactions with the with the nucleotides in the minor groove [33]. The most common example of homeodomain containing factors are the Hox genes, which are found throughout the metazoan kingdom and are important for specifying regions along the developing embryonic anterior-posterior axis [34, 35].

### 1.2.5 $\beta$ Sheet DNA Recognition Proteins

While many of the transcription factor families rely on $\alpha$ helices to interact with the double stranded DNA helix, this class of transcription factor utilizes a $\beta$ sheet with at least two strands, sitting in the major groove [36]. In the cases of dimerization, as is seen with the STAT family of transcription factors, each member of the dimer can contribute an anti-parallel beta strand to the DNA binding beta sheet [37].

### 1.2.6 Zinc Finger Proteins

Zinc fingers are highly specific, modular proteins that have been well studied. Each of the small "finger" modules in a zinc finger protein incorporates a zinc $Zn^{2+}$ ion to stabilize its three dimensional structure (Review [38]). This in turn allows for the finger to be highly specific in recognizing a DNA sequence approximately 3 base pairs long. Arranging a series of zinc fingers in tandem on a single protein permits these proteins to form stable and specific bonds with DNA. Due to this modularity there has been some success in engineering zinc fingers that target specific DNA sequences. This has been based in part on the structure of Zif 268 [39].

## 1.3 Early Representations of Regulator Specificity

Somewhat contemporaneous with the discovery that transcriptional regulators may bind to specific DNA sequences [19], other types of specific protein-DNA interactions were known, including the activity of restriction enzymes [40, 41]. These restriction enzymes were so highly specific, that representing their binding preferences was as straightforward as describing a single sequence. For example, the enzyme EcoRI recognizes the sequence GAATTC [42]. Representing specificities in this way is highly efficient, but limited in that it requires an exact match, and not many interactions follow this rule.

As genetic elements with more variation were discovered, new representations were required to describe the observed differences across the instances of that element. One early such example is the Pribnow box sequence discovered in 1975 in bacteria [43]. While between any two instances of this element there may be up to four sites that differ, if considering the most common nucleotide at each site, most instances differ in nucleotide composition from this 'average' sequence in at most two sites [44]. This 'average' or most common nucleotide sequence is known as a consensus, and only requires sequences to have a high level of identity to be identified as a match to the element the consensus describes. Extra flexibility can be incorporated into this model by utilizing an alphabet that extends beyond the four standard bases, and includes ambiguous nucleotide codes, such as an R for a purine (A or G), or Y for a pyrimidine (C or T) [45].

A limitation of this method was well described by Gary Stormo in 2000: "The concept of the consensus sequence has been widely used to represent the specificity of transcription factors. But exactly how one is defined is somewhat arbitrary. In general it refers to a sequence that matches all of the example sites closely, but not necessarily exactly. There is a trade-off between the number of mismatches al-

lowed, the ambiguity in the consensus sequence, and the sensitivity and precision of the representation" [44].

Upon accumulation of deeper and more diverse datasets, other methods were required to describe the specificity of sites that can be, at times, highly variable. As described earlier, a Perceptron based approach was found to be effective in this case [13]. This and other work from the same time led to the development of the position weight matrix (PWM) and its variants [13, 46]. In short, upon aligning each of the genetic elements, a distribution across the nucleotides is computed at each position in the sequence, as a frequency matrix. The log ratio between these frequencies and the expected background frequencies is then computed. These weights vary for each nucleotide, at each position, giving more weight to positions with more information content, and provide relative rankings for the possible substitutions at each position [47].

**Box 1: PWM Example**

The following example shows how a position weight matrix can be computed from a set of aligned known binding sites for a transcription factor, and how it can be used to score additional sequences. Given the following set of 10 aligned sequences:

TAAGATGATGTAATC

AGGAATGATGTCACG

CAAGGTGAGGTCATC

CTGAGTGACGTCATT

TCCAATGACGGCACA

AGTGATGACGTAATC

GCAAATGATGTCATC

AATGATAATGCCATC

TTAGGTGATGTCATA

GAGAATGAGATGATA

The occurrence of each of the four nucleotides is counted in each position in the motif. This distribution of counts at each position is known as a position frequency matrix. This can be written as follows, where $X_i$ is the $i$th sequence in the data set, $k$ is the nucleotide being counted ($k \in (\texttt{'A'}, \texttt{'C'}, \texttt{'G'}, \texttt{'T'})$), and $j$ is the position in the motif:

$$\text{PFM}_{k,j} = \sum_{i=1}^{N} I(X_{i,j} = k) \tag{1.1}$$

$$\text{PFM} = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 3 & 4 & 4 & 5 & 7 & 0 & 1 & 10 & 0 & 1 & 0 & 2 & 10 & 0 & 3 \\ 2 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 3 & 0 & 1 & 7 & 0 & 2 & 5 \\ 2 & 2 & 3 & 5 & 3 & 0 & 9 & 0 & 2 & 9 & 1 & 1 & 0 & 0 & 1 \\ 3 & 2 & 2 & 0 & 0 & 10 & 0 & 0 & 5 & 0 & 8 & 0 & 0 & 8 & 1 \end{bmatrix}$$

By normalizing each of the columns based on the total number of counts ($N$), this matrix can be converted to a position probability matrix, where each column is now a probability distribution across the nucleotides:

$$\text{PPM}_{k,j} = \frac{1}{N}\text{PFM}_{k,j} \tag{1.2}$$

$$\text{PPM} = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.3 & 0.4 & 0.4 & 0.5 & 0.7 & 0.0 & 0.1 & 1.0 & 0.0 & 0.1 & 0.0 & 0.2 & 1.0 & 0.0 & 0.3 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.3 & 0.0 & 0.1 & 0.7 & 0.0 & 0.2 & 0.5 \\ 0.2 & 0.2 & 0.3 & 0.5 & 0.3 & 0.0 & 0.9 & 0.0 & 0.2 & 0.9 & 0.1 & 0.1 & 0.0 & 0.0 & 0.1 \\ 0.3 & 0.2 & 0.2 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.5 & 0.0 & 0.8 & 0.0 & 0.0 & 0.8 & 0.1 \end{bmatrix}$$

Finally, by taking the log ratio of each value in the PPM relative to some background value for each nucleotide ($b_k$, the most basic representation is

$\frac{1}{4}$ for a uniform distribution of nucleotides). The values in this position weight matrix are now analogous to additive energies corresponding to the favorability of any given nucleotide at that position:

$$\mathrm{PWM}_{k,j} = \log_2(\mathrm{PPM}_{k,j}/b_k) \tag{1.3}$$

$$\mathrm{PWM} = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.26 & 0.68 & 0.68 & 1.00 & 1.49 & ... & -0.32 & 2.00 & -\infty & 0.26 \\ -0.32 & -0.32 & -1.32 & -\infty & -\infty & ... & 1.49 & -\infty & -0.32 & 1.00 \\ -0.32 & -0.32 & 0.26 & 1.00 & 0.26 & ... & -1.32 & -\infty & -\infty & -1.32 \\ 0.26 & -0.32 & -0.32 & -\infty & -\infty & ... & -\infty & -\infty & 1.68 & -1.32 \end{bmatrix}$$

Given this PWM, and some new sequence ($S$), the score con simply be calculated by adding the values across the PWM corresponding to the nucleotides at each position:

$$p(S|\mathrm{PWM}) = \sum_{i=1}^{w} \mathrm{PWM}_{i,S_i}$$

For example, for the sequence $\mathrm{GGAGATGACGTCATT}$:

| | $G$ | $G$ | $A$ | $G$ | $A$ | ... | $C$ | $A$ | $T$ | $T$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $A$ | 0.26 | 0.68 | 0.68 | 1.00 | 1.49 | ... | −0.32 | 2.00 | $-\infty$ | 0.26 |
| $C$ | −0.32 | −0.32 | −1.32 | $-\infty$ | $-\infty$ | ... | 1.49 | $-\infty$ | −0.32 | 1.00 |
| $G$ | −0.32 | −0.32 | 0.26 | 1.00 | 0.26 | ... | −1.32 | $-\infty$ | $-\infty$ | −1.32 |
| $T$ | 0.26 | −0.32 | −0.32 | $-\infty$ | $-\infty$ | ... | $-\infty$ | $-\infty$ | 1.68 | −1.32 |

$p(S_{GGAGATGACGTCATT}|\mathrm{PWM}) = 16.02$

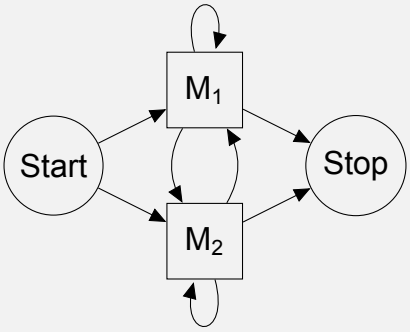
These PWMs can be further optimized to fit quantitative data. This variation no longer describes a pure distribution at each site, but may give a better overall de-

scription of the site, placing emphasis on highly favorable sequences, and selecting against unfavorable substitutions [48].

While the PWM is still the most common representation used today, other forms of motif descriptions have been used for various applications. One such example is a profile hidden markov model [49]. A major advantage of this representation is that it can model gaps/insertions and deletions that may be tolerated by the protein-DNA interaction. It starts with a linear series of states describing each position in the motif. The emissions for each state describe the observed frequencies of each nucleotide at that position, much akin to the PWM. In addition to the transitions between the states corresponding to adjacent positions, there are also insertions and deletion states available. The weights associated with these alternate transitions govern the tolerance of the model for describing sites of varying length.

**Box 2: pHMM Example**

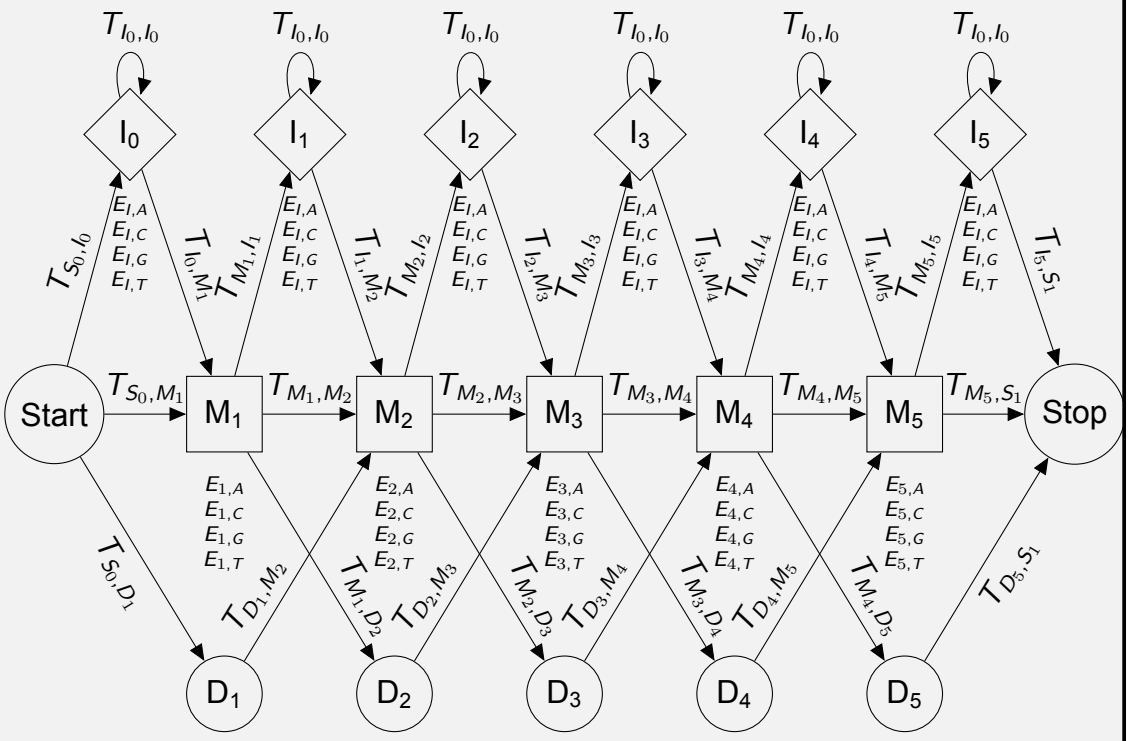
Profile Hidden Markov Models are a unique formulation of a Hidden Markov Model. In a traditional HMM, there are a number of hidden states, with transition probabilities between each. In this example, in addition to the Start and Stop states specifying the beginning and end of the Markov chain, there are two states ($M_1$ and $M_2$).




The transition probabilities along edges departing any state add up to $1.0$.

If there are two states that aren't connected, for example the Start and Stop states in the example, the transition probability between those two states is $0$. In each of the hidden states, there is a distribution of emission probabilities. Again, the emission probabilities from a single state should add up to $1.0$. The emissions are the events that are observed, and are used to infer the order of the hidden states. There are established algorithms for this purpose.

In a pHMM, there are a number of match states, one each corresponding to each of the positions in the width of the motif. There are also a set of insert states, and delete states. These in turn correspond to variations in length of a sequence that might match the motif. A unique characteristic of a pHMM relative to the traditional HMM is that these states can only transition from right to left and never returning to an earlier state. This allows for a different set of emission probabilities for each position in the motif. The emission probabilities in a match state are analagous to a column in a PPM.



Above is a diagramatic representation of a profile HMM motif representation corresponding to 5 base pairs wide. Every state has a set of transition probabilities $T$, and the Match ($M_i$) and Insert ($I_i$) states have emission probabilities, $E$, corresponding to the nucleotides present at that position.

Given an ordered sequence ($N$) of nucleotides ($n_i$), there is some alignment ($A$) corresponding to the best path through a profile HMM. This path can be determined using the Viterbi algorithm, for example [49]. A scoring scheme such as the one described below can be used to provide a metric for how well that particular sequence fits the profile HMM.

$$\begin{aligned} N &= (n_1, n_2, n_3, n_4, n_5) \\ &= (\mathsf{C}, \mathsf{G}, \mathsf{A}, \mathsf{T}, \mathsf{A}) \\ A &= (S_0, M_1, ..., M_5, S_1) \end{aligned}$$

$$p(N, A|\mathsf{pHMM}) = \prod_{i=1}^{A} \begin{cases} T_{A_{i-1},A_i} \cdot E_{i,n_i} & \text{if } A_i = M_i \\ T_{A_{i-1},A_i} \cdot E_{I,n_i} & \text{if } A_i = I_i \\ T_{A_{i-1},A_i} & \text{else} \end{cases} \tag{1.4}$$

This path probability can be normalized to all possible paths through the model using the Forward dynamic programming algorithm.

As more complicated problems have been proposed using DNA motifs, additional methods to encode the inherent sequence variation for use in machine learning approaches have been proposed. They have been described as filters in artificial neural nets as early as 1991 [50], and in state-of-the-art methods such as DeepBind [51] . K-mer counts have also been used as the input vectors in other al-

gorithms such as kmer SVMs and gapped kmer SVMs by Michael Beer [52]. Table 1.1 gives some examples of various approaches for encoding sequence representations of DNA motifs. Gary Stormo also provides a good review of these early methods in [44].

| **Method** | **Characteristics** |
|---|---|
| Simple Sequence | Highly specific – One possible matching sequence. |
| Consensus | One sequence representing the variability, sometimes with ambiguous nucleotide characters. Doesn't describe the distribution, just the most likely sequence. |
| PWM [13] | **Weight matrix**. Analagous to $\triangle\triangle G$ contribution of each nucleotide. |
| DAMO [53] | **Optimized Weight Matrix**. Using the same approach as the PWM, each value is then optimized to best fit quantitative binding data, putting more emphasis on important positions. |
| nhmmer [54] | **Profile Hidden Markov Model** framework, allows for searching for sequence matches in a DNA sequence database, including matches with insertions and deletions. |
| Gapped kmer SVM [52] | **Support Vector Machine** approach. Uses kmer counts (including gapped kmers) as inputs. |
| DeepBind [51] | **Convolutional Neural Net**. Convolutional filters simulate PWMs, and there can be interactions between different TFs. |

Table 1.1: Examples of sequence based methods for representing DNA motifs.

## 1.4 Role of DNA Shape in Motif Recognition

It wasn't until 20 years after the discovery of the structure of DNA that a group generated a X-ray crystallography structure with enough resolution to observe and confirm the proposed (and widely accepted) hydrogen bond interactions between DNA nucleotide bases and the amino acid side chains in the transcription factor [55]. Since that time, many crystal structures of DNA have been obtained, allowing for the analysis at atomic resolution of the structure and geometry of the DNA helix.

By 1986, it was recognized that the shape of DNA is variable, and correlates with the underlying sequence of basepairs [56]. Subsequently, several groups identified examples of transcription factors recognizing specific DNA conformations. This ranges from extreme sequence independent structures like SRY and HMG1 each binding four-way junctions [57, 58], to more subtle conformational features such as bending of the promoter sequence to increase the affinity for the TATA binding factor [59].

Over the years, structures of DNA-protein interactions have been solved for an increasing number of transcription factors. A comprehensive study of these structures revealed a number of specific components of DNA shape that participate in interactions with proteins [15]. In this study, Rohs and colleagues found that positively charged arginine residues tend to contact regions along the DNA backbone where the minor groove is narrow, bringing the negatively charged phosphate backbones closer together and increasing the overall negative potential of that position. A common method they observed for creating that narrow minor groove shape was AT reach regions.

Early on, several groups attempted to model the specificity of regulators not by the sequences of the recognized sites, but by the shape. One early such example was Karas, *et al* (1996) in which they used molecular modelling to convert DNA se-

quence into structural parameters, and then created a classifier to identify matches to the TATA sequence motif as a Z-score from a consensus shape profile [60]. Additional approaches included using protein-DNA co structures in the Protein Data Bank to first identify which residues for a given transcription factor interact with the bases of the binding site, determine interaction potentials, and use that to search for other sequences that will be bound by that protein structure [61]. This was followed by other models that relied on known structures to use atom-packing, electrostatics, and other physical energy terms to predict potential binding sites [62].

## 1.5 Significance of Transcriptional Regulators

Within the human body trillions of cells work together to create functioning systems [1]. These cells represent hundreds of distinct cell types that differ both in their physiology and the processes they carry out [63]. While this is amazing in and of itself, it is more remarkable given that each of these cells has an identical copy of the genomic information [2, 3], and yet there is such an incredible diversity.

Throughout the development process as these cells grow and divide, becoming more specialized during the progression from a single cell to a large multicellular organism, they develop distinct profiles of gene expression, contributing to their specialization [4, 5, 6]. Errors in this expression profile can contribute to genetic disease. While in some cases these disease phenotypes are caused by broken malformed proteins, in others the cause is a more subtle difference in the timing or level of expression of that protein [11].

*Cis*-regulatory regions contribute to controlling the timing and level of gene expression in order to allow for proper development and maintaining healthy growth [9]. These are regions of DNA that, when bound by the appropriate factors, serve to promote or inhibit transcription, and can exist in many different states of activ-

ity [6, 7]. The best computational approaches to identifying these modules rely on large amounts of experimental data, primarily targeting transcription factor binding (e.g. [64]). However, the available data is quite limited, with only a handful of experiments available at most for many cell types. Experimentally generating a comprehensive dataset of protein-DNA interactions for all cell types at this point is intractable due to limitations of the number of cells required for some experiments, the quality of reagents, and cost.

## 1.6 Modern Approaches to TFBS Prediction

As described above, the PWM is still one of the most commonly used representation of DNA motifs and transcription factor specificity. This is likely in part due to the widely used resources that have been developed on this framework, including the large number of known motifs available in the JASPAR and TRANSFAC repositories [65, 66, 67], as well as the ease of use of the popular tool suites MEME and HOMER which are also based on PWM motif format [68, 69]. Furthermore, for many applications it appears that using simple sequence based methods is sufficient, particularly if they are optimized based on quantitative binding data [53].

For more generalized tasks such as broad transcription factor binding site prediction, however, these methods fall short. This has led to the continual development of methods to attempt this problem more effectively. As mentioned earlier, machine learning methods have been applied to the this problem. While they are powerful tools, they still only address one part of the problem.

So far, it has been discussed that some transcription factors are able to discriminate between nucleotide bases based on the unique arrangement of hydrogen binding partners in the major groove in a mechanism known as direct readout [70]. Other transcription factors utilize interactions that depend on the shape of

the DNA at the binding site [71]. It can be considered that relying on sequence alone uses sequence as an abstraction of the shape information. In fact, it has been shown that considering dinucleotides can be informative to a similar degree as some shape methods [72, 53]. Alternately, it can be be thought of as DNA shape being a superset of the sequence information, since the arrangement of hydrogen bond partners is arranged by the 3D shape of the DNA.

In this case, shape based methods are required for a more complete view of the protein-DNA interactions of transcription factors with their target sites. To this end, a number of new shape based methods have been developed in recent years, a few of which are summarized here.

Remo Rohs and his coauthors have been at the forefront of developing tools to be used for evaluating DNA shape, including a method to translate DNA pentamers into profiles of several shape parameters [73]. They used these shape values in combination with DNA sequence as input to a gradient boosting decision tree classifier and saw improvement over using just a PWM [74]. They further refined this method by using a more directed approach with multiple linear regression on the shape parameters and various *k*mers using quantitative data from HT-SELEX experiments to optimize the models [75]. Another recent method from another group considers one shape feature at a time, finding regions in variably sized windows that align to generate a shape profile, and uses a hypergeometric distribution to compute scores in order to search for additional instances of that shape motif [76].

One big limitation that all of the methods discussed so far is that they cannot easily distinguish between sites that are bound variably across cell types. As described above, all cells have the same underlying information encoded in their DNA sequence, and yet the TF binding landscape can vary widely. In fact, less that 1% of the available binding motifs are occupied at any given time, in a given cell type [77, 78, 79]. Additionally, many motifs are relatively short, meaning that even due to

chance instances of these motifs will exist relatively frequently, especially when the tolerance for sequence variability is taken into account. In order to make cell-type specific predictions that will carry functional value, additional methods are needed.

Incorporating cell type-specific additional information in the form of chromatin accessibility can increase the specificity of sequence-based methods through approaches such as DNase-footprinting [80, 81, 82, 83, 84, 85]. Traditional footprinting methods largely take a motif-centric approach, placing an emphasis on strong matches to a PWM.

## 1.7 Summary of Goals

**Significance:** Of the many cell types in the human body the majority have the same underlying genetic code, yet they vary widely both in their form and function. Disregulation in the reading of the genetic material can lead to disease. Transcriptional regulation is tightly controlled to promote healthy development, and is driven in part by the activity of transcription factors. To fully understand development and genetic diseases, a comprehensive view of the regulatory landscape is required.

**Problem:** While the amount of available data is constantly increasing, there are still large gaps in our knowledge preventing the generation of comprehensive maps of the regulatory landscape. Some cell types are too rare to make assaying them hundreds of times feasible, others are difficult to grow in culture, and even for immortalized cell lines we have incomplete data. Predictive methods can extrapolate to fill in some of these gaps. In particular, global transcription factor binding data can be used to derive mechanistic insight into gene regulation.

**Goal:** The purpose of this dissertation is to provide tools and insight for the study and prediction of global transcription factor binding site (TFBS) occupancy in a cell

type specific manner from chromatin accessibility. A DNA shape based representation of DNA motifs is described that incorporates relevant features in addition to DNA sequence. Methods are provided for *de novo* discovery and characterization of DNA motifs from high throughput sequencing experiments. These shape motifs are used to predict binding sites and describe the similarity of diverse transcription factor binding preference. Finally, the use of the models developed here to make cell type specific predictions of transcription factor binding in conjunction with DNase accessibility are explored.

# 2. Structural Motifs – StruMs

In the pursuit of accuracy in predicting transcription factor binding sites, increasing complex models have been proposed. Often performance is traded for interpretability. Here, an alternative strategy is proposed that aims to accurately model the protein-DNA interaction while remaining interpretable.

## 2.1 Theoretical formulation

As has been described, the position weight matrix is the most commonly used description of transcription factor specificity as DNA motifs. Additionally the value of considering the shape of the DNA, as opposed to only the sequence of nucleotides, has been discussed. A theoretical framework may be built up from these two principals that results in a model hereafter termed Structural Motifs, or StruMs for short.

As DNA may adopt many possible conformations along a spectrum, and DNA shape contributes to the recognition of binding sites by transcription factors, it may be assumed that for a given transcription factor there is an ideal DNA conformation corresponding to the largest $\triangle G$ of interaction. Given that the inside of the cell is a packed environment, and evidence of the DNA "breathing", e.g. in the case of wrapping/unwrapping histones [86], it can further be assumed that transcription factors can tolerate some variations in the shape of the DNA.

Consider, now, a set of sequences to which the transcription factor is known to bind with high affinity, and for which the values for a shape parameter is known

across the entire site (e.g. the Major Groove Width). The ideal conformation of the binding site most likely lies within the range of observed values, and one maximum likelihood estimation of the ideal conformation would the arithmetic mean of the shape at each position in the site, resulting in a vector of values across the width of the site. Similarly, the tolerance of variability around this ideal is probably encoded in the observed variance, thus the standard deviation at each site provides an estimate of this tolerated variability.

More formally, given training data ($D$) composed of a set of binding sites $S_i$ of the same length and orientation,

$$D = (S_1, S_2, ..., S_n)$$

where, if $k$ is the length of the site, there are $k$ values $v_j$ in each binding site $S_i$ corresponding to the shape value at that position

$$S_i = (v_{i,1}, v_{i,2}, ..., v_{i,k})$$

a set of parameters $\phi$ can be computed that estimate the ideal conformation ($\mu$) and the tolerated variability from ideal ($\sigma$),

$$\phi = \left\{ \begin{matrix} (\mu_1, & ... & \mu_j & ... & \mu_k), \\ (\sigma_1, & ... & \sigma_j & ... & \sigma_k), \end{matrix} \right\} \tag{2.1}$$

where $\mu_j$ and $\sigma_j$ are computed as follows:

$$\mu_j = \frac{1}{n} \sum_{i=1}^{n} v_{i,j} \tag{2.2}$$

$$\sigma_j = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (v_{i,j} - \mu_j)^2} \tag{2.3}$$

If, at each site, there are now $p$ types of shape parameters known, then the length of $S_i$ will now be $k \cdot p = m$.

$$S_i = (v_{i,1}, v_{i,2}, ..., v_{i,m})$$

The estimates are performed the same way, so additional parameters are simply added to the parameter set:

$$\phi = \begin{Bmatrix} (\mu_1, & ... & \mu_j & ... & \mu_m), \\ (\sigma_1, & ... & \sigma_j & ... & \sigma_m), \end{Bmatrix} \tag{2.4}$$

This formulation provides a way to model the ideal conformation and tolerance for the DNA shape within a transcription factor binding site. For the purposes of predicting binding sites, a way to rank or score various new sites is required. With the binding preferences being modelled with the mean and standard deviation, it follows that the values $v_j$ may be distributed according the Normal distribution:

$$v_j \sim \mathcal{N}(\mu_j, \sigma_j^2) \tag{2.5}$$

One of the foundational assumptions in the position weight matrix framework is that each of the positions is independent of the others [13]. While this is not necessarily true in all cases, this assumption has held up well over the years. Using the same assumption here, and given equation 2.5 then calculating a score for a new sequence $S_i$ becomes:

$$P(S_i|\phi) = \prod_{j=1}^{m} P(v_{ij}|\mu_j, \sigma_j^2) \tag{2.6}$$

The result is a model of the DNA shape preferences of transcription factors that is closely analogous to the time tested and still used position weight matrix: At each position in the binding site a distribution is computed, categorical for the PWM, and

continuous for the StruM; assuming each position is independent, the probabilities at each position of observing the value from a new site are multiplied to obtain the score. The similarity of these two methods means that the plethora of tools available for using position weight matrices can be adapted to work with structural motifs, as is done with the *de novo* motif finding algorithm in Chapter 3.

## 2.2 Normality of HMBOX1 shape preferences

The structural motif model described above is an intuitive progression from several base assumptions. The application of this model will therefore depend on how well the assumptions hold. The first assumption to be considered is that, for a given position in a motif, the values for a shape feature are normally distributed across the various observed instances of matches to the motif, and therefore well represented by the mean ($\mu$) and standard deviation ($\sigma$).

The ENCODE project has provided a variety of uniformly processed data files for a whole range of transcription factors [87]. Using the peaks identified from a ChIP-seq experiment mapped to genome build $\mathrm{hg19}$ for the human HMBOX1 transcription factor in K562 cells (accession ENCFF558DSF), a sequence motif was derived using the popular *de novo* motif discovery tool MEME, using the MEME-ChIP variation [88, 89].

Within the output of MEME is a list of the specific sites, and their sequences, that contribute to the motif. Using these sequences as the set of aligned binding sites of the same length and orientation mentioned above, the tool DNAshapeR was used to obtain estimates across the sites for four shape features: minor groove width (MGW), Roll, propeller twist (ProT), and helix twist (HelT) (Figure 2.1) [73]. The DNAshape method using all-atom Monte Carlo simulations of short DNA fragments to derive shape profiles for each of the 512 unique DNA pentamers [90].

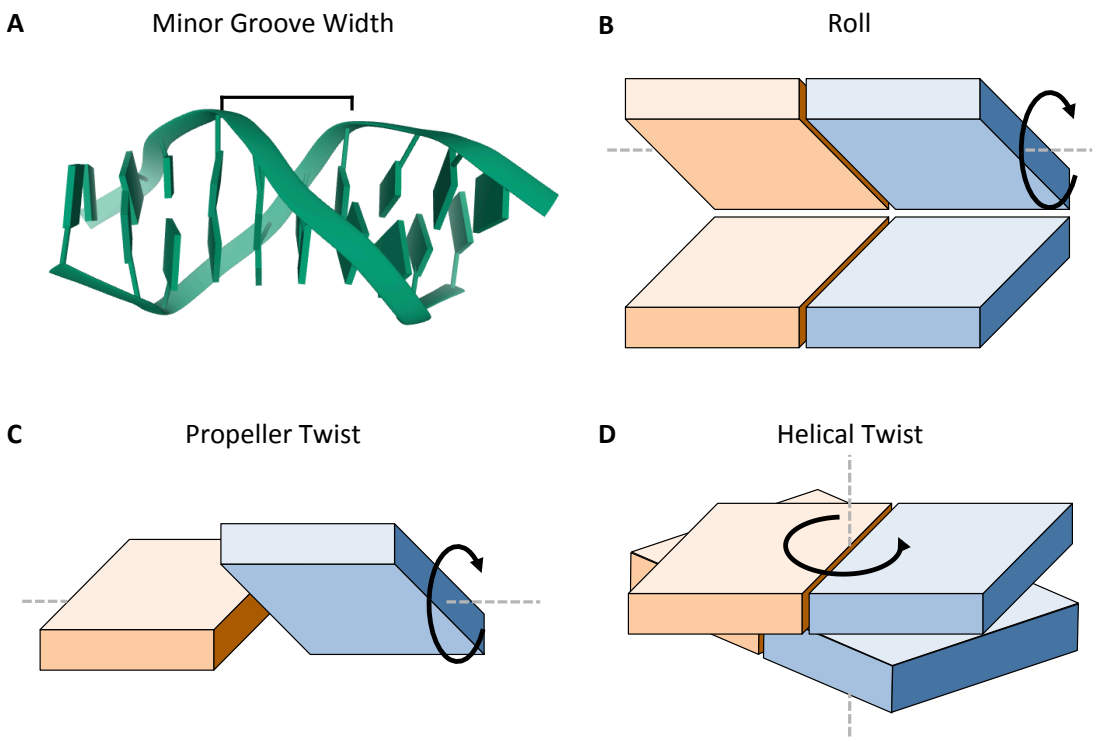



Figure 2.1: Diagrammatic representation of common DNA geometries. For base-pair geometries the base(s) associated with the positive strand is blue, and the negative strand is orange. (a) The minor groove width (MGW) measured as the distance between the phosphate backbones. Shown here relative to a cartoon representation of a DNA decamer (PDB ID: 6JV5). (b) Roll, measured as an angle relative to parallel for two consecutive basepairs. (c) Propeller twist (ProT) measured as the relative angle between to bases in a pair along the long axis. (d) Helical twist (HelT) measured as the angle of rotation along the length of the DNA strand between two consecutive basepairs.

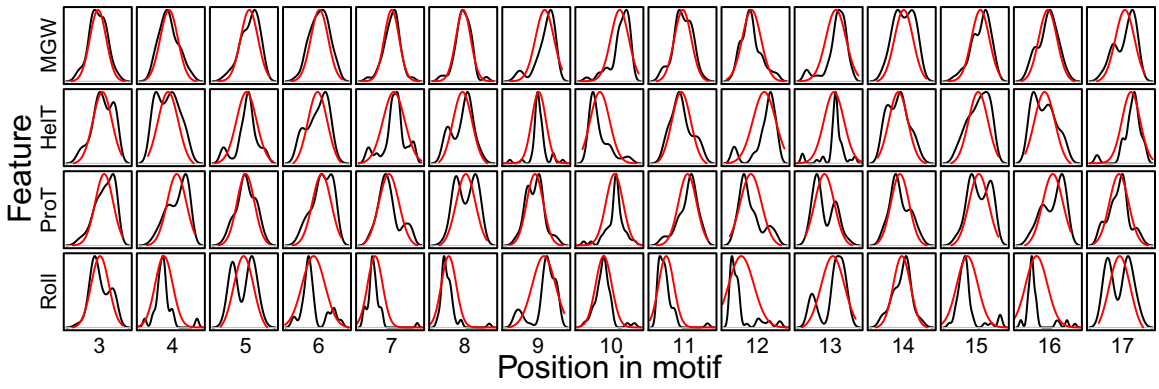



Figure 2.2: Distribution of sequence values for the HMBOX1 motif. The black line represents a kernel density estimation of the true distribution. The red line represents the maximum likelihood estimate of the normal distribution fitting that data. (Roll: Roll; ProT: Propeller Twist; HelT: Helix Twist; MGW: Minor Groove Width)

These profiles are highly correlated with known structures of DNA fragments as determined by Nuclear Magnetic Resonance spectroscopy, and a sliding window approach can be used to predict the overall shape of a DNA sequence. Using the DNAshape method, two distributions were computed for each of these four shape features, at each of the positions across the width of the binding site: a kernel density estimation across the 46 observed sites; and the described maximum likelihood estimation, corresponding to the Normal distribution defined by the mean and standard deviation of the observed values.

Figure 2.2 shows these two distributions, with the observed data in black, and the expected distribution in red. While there are a number of cases where bimodality is observed, the expected Normal distribution is a reasonable approximation of a majority of the observed distributions. The observed discrepancies can be explained, at least in part, by two factors. First, our resolution in structure-space is limited, rendering our values discrete, while we expect these shape features to come from a continuous distribution *in vivo*. Second, this data only encompasses 46 sequences, and sampling error could be contributing to the instances of non-normality being observed, which would be overcome by increased sample sizes at

with high throughput ChIP-seq data, for instance. This supports the parameterization of structural motifs given in equation 2.4 and the assumption from equation 2.5.

## 2.3 Independence of positions

The second assumption that needs investigated is whether the positions in the motif (and the shape values) can be considered independent. This is required to score observed sequences given a structural motif according to equation 2.6, without taking into account covariation of the variables.

Taking each shape feature one at a time, at each position (hereto referred to as a position-specific shape feature) in the same HMBOX1 motif sequences, the pairwise Pearson correlations between each position-specific shape feature were taken.

As can be observed by the striations in Figure 2.3a, there is some amount of correlation between the features at a given position, but limited correlation between adjacent positions. Looking at the distribution of these pairwise correlations in Figure 2.3b, the mean value is 0.01, and the standard deviation is 0.26, meaning that the majority of the features have a correlation close to zero. Along with the evidence of the efficacy of the independence assumption used by the position weight matrix, this provides support for the given scoring scheme (equation 2.6).

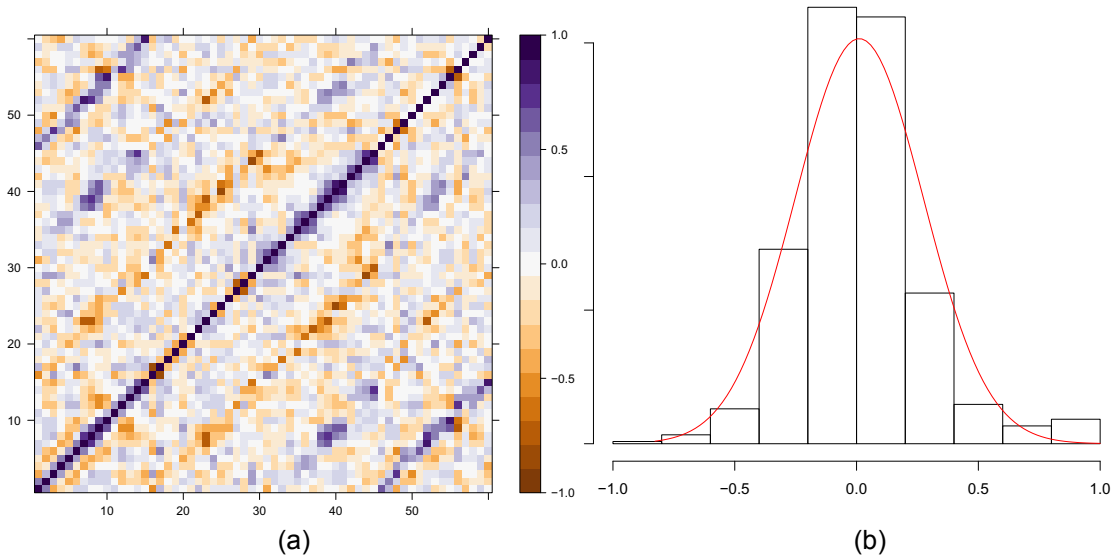



Figure 2.3: Correlation between position-specific shape features in the HMBOX1 structural motif. **(a)** Heat map of the pairwise correlation values for each feature, across all positions. For example, index values 1-15 correspond to MGW at positions 1-15, followed by HelT. The color of the cells corresponds to the correlation, as described by the colorbar. **(b)** Distribution of the values in (a). Black histogram represents the observed values, red line is the best-fit Gaussian.

# 3. DNA shape complements sequence-based representations of transcription factor binding sites

## 3.1 Introduction

The human body has trillions of cells and hundreds of cell types that have many distinct morphologies and functional roles [1, 63]. Yet each of these cells has an identical copy of the underlying developmental program encoded in the DNA [2, 3]. The diversity of form and function that is observed is only possible through tight control of the expression of genes throughout the developmental process [4, 5, 6]. Understanding the mechanisms guiding this control will provide insight into the development of complex multicellular organisms, and associated diseases.

Transcription is controlled by the activity of *cis*-regulatory modules; regions of DNA that, when bound by the appropriate factors, serve to promote or inhibit transcription, and can exist in many different states of activity [6, 7]. The best computational approaches to identifying these modules rely on large amounts of experimental data, primarily targeting transcription factor binding (e.g. [64]). However, the available data is quite limited, with only a handful of experiments available at most for many cell types. Experimentally generating a comprehensive dataset of

protein-DNA interactions for all cell types at this point is intractable due to limitations of the number of cells required for some experiments, the quality of reagents, and cost.

Computational methods that predict transcription factor (TF) binding complement the available data, and allow for extrapolation across rare or otherwise unwieldy cell types. Historically this is done by representing the chains of nucleotides that make up DNA as sequences composed of an alphabet of 4 letters. Patterns within these sequences can be identified, for example using letter frequencies at each position within a set of aligned binding sites as in the position weight matrix (PWM) [13]. This can be quite effective, and has been shown to perform well for many transcription factors [14]. Some transcription factors have been shown to be especially well represented in this manner, displaying extreme sequences preferences enforced through base pair-specific contacts along the major groove. This mode of binding site recognition by transcription factors is known as direct- or sequence-readout.

In reality, DNA is a complex three-dimensional macromolecule that is tightly packed into the nucleus. Other transcription factors have been shown to take advantage of the three dimensional shape of DNA molecule to recognize their binding sites in a mode known as indirect- or shape-readout [15, 91, 71]. For example, it has been shown that the narrowing of the minor groove and corresponding increase in electrostatic potential drives interactions with positively charged arginine side chains in Oct-1/PORE complex binding [15].

The sequence representation of DNA and the associated TF binding site (TFBS) models are in reality an abstraction of the chemical and physical interactions of the protein molecules with the DNA. As indicated above, even sequence-readout relies on the proper 3D positioning and electrostatic compatibility of hydrogen bond donors and acceptors between the two molecules.

We hypothesized that TF binding preferences could be modeled using estimates of DNA shape parameters across the binding site, and this would provide increased discrimination over sequence-based approaches. To this end we designed a set of methods adapting the time-tested position weight matrix to incorporate DNA shape instead of sequence, known as Structural Motifs (StruMs). StruMs specifically model TFBSs and are complementary to sequence-based methods.

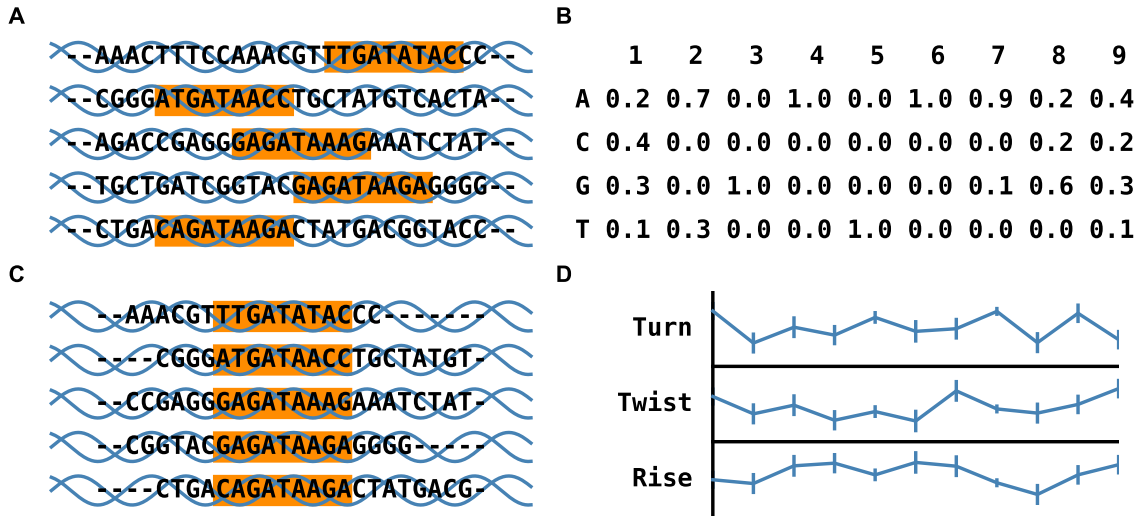



Figure 3.1: Graphical overview of structural motifs. **(a)** A series of DNA regions with a known binding site for GATA1 (orange). **(c)** These same regions are aligned by their binding site. **(b)** The alignment is used to calculate a distribution of base frequencies at each position, represented as a traditional PWM. **(d)** For the StruM paradigm, a distribution (mean, standard deviation) is computed at each position of the binding site for several shape features.

## 3.2 Materials and Methods

### 3.2.1 Structural Motifs

The StruM is an extension of the PWM [13, 92]. Each position-specific feature is assumed to be independent allowing for log-probabilities to be combined additively, and the model finds a simple distribution across each of these features. The

construction of a motif requires two things: 1) aligned sequences corresponding to binding sites; and 2) a method for estimating shape features. It has been shown that, as with proteins, the primary sequence of the DNA plays a large role in determining the local shape of the DNA [15]. Here the Dinucleotide Property Database (DiProDB) [93] is used to estimate shape parameters for each dinucleotide in the sequence.

**Definition**

Given a set of $n$ training sequences, each sequence is converted to a structural representation. In this case each consecutive dinucleotide is looked up in DiProDB, and that column is appended to the feature vector. The length of this vector is $k$ (the length of the binding site) times $p$ (the number of shape features being considered); this value $k \cdot p$ is noted hereafter as $m$. This set of training structures ($D$) is used to compute the parameters ($\phi$). These are represented as a mean ($\mu$) and standard deviation ($\sigma$) for each feature at each position.

$$D = \begin{Bmatrix} (v_{11}, & v_{12}, & \dots & v_{1m}), \\ (v_{21}, & v_{22}, & \dots & v_{2m}), \\ \vdots & \vdots & \vdots & \vdots \\ (v_{n1}, & v_{n2}, & \dots & v_{nm}), \end{Bmatrix} \qquad \phi = \begin{Bmatrix} (\mu_1, & \mu_2, & \dots & \mu_m), \\ (\sigma_1, & \sigma_2, & \dots & \sigma_m), \end{Bmatrix}$$

The model begins with the assumption that the DNA shapes preferred for interaction with a given transcription factor at any given position specific feature ($v_j$) have an optimum shape, and sample adjacent shapes according to a normal distribution.

$$v_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$$

Assuming that each feature and each position is independent, then calculating

the score ($s$) for the $i$-th sequence becomes:

$$s_i = \prod_{j=1}^{m} P(v_{ij}|\mu_j, \sigma_j^2)$$

In order to avoid underflow issues during computation, all calculations are done in log space.

### 3.2.2 Data

All experimental data is obtained from ENCODE [87]. The ChIP-seq data used Transcription Factor ChIP-seq data from K562 cells mapped to *hg19*. This data was filtered for targets that were annotated as being sequence specific transcription factors in [94]. Specifically the conservative IDR thresholded peaks were downloaded in the ENCODE narrowPeak BED format. TF family assignments were done using the assignments in [95]. Accession numbers for all datasets used are available in the Supplementary Information.

### 3.2.3 ChIP peak classification

#### *De novo* motif finding

For each TF analyzed, the sequences for the top 500 peaks were retrieved based on signal enrichment. The sequence corresponding to 100 bp around the peak identified in the BED file was then extracted. A PWM was learned using `meme-chip` [89] with the parameters `-norand -meme-nmotifs 1 -dreme-m 0 -spamo-skip -dna -nmeme 500 -seed`. The model was left as the default (`zoops`, and the random seed was set by hashing each TF's accession number. The sites reported by MEME as being used to train the final motif were extracted from the output. These aligned binding sites were used to compute position specific frequencies for mononucleotides (PWM) [13] and dinucleotides (din-

ucleotide weight matrix, DWM) [96]. After translation into structural space, the distribution of shape values at each position was calculated for the StruM. This variation will be referred to as the Maximum Likelihood StruM (ML-StruM). Additionally, the 100 bp centered sequences were used to fit an additional structural motif by expectation-maximization, matching its length to that of the PWM learned by MEME (StruM). Similarly to the incorporation of pseudocounts in nucleotide frequency estimations, a minimum threshold of 0.1 was imposed on $\sigma$ for the StruM.

**Motif performance**

For the next 500 sequences in the ChIP-seq experiment, the maximum score for each motif type was calculated. To generate scores for a matched set of negative sequences, two strategies were employed: shuffling these testing sequences, or taking 100 bp flanking sequences from the top 500 peaks. Upon generation of the negative set, the scoring process was repeated. A simple threshold was varied across the scores to generate a receiver operating characteristic (ROC) curve and a precision-recall curve (PRC) and the area under the curve (AUC) was calculated. This process was repeated for 355 TF ChIP-seq experiments in K562 cells.

**Specificity of StruM by TF family**

Using TF family assignments in TFClass ([95]), the second 500 sequences for each TF was pooled by TF family. For each TF, the second 500 sequences for that TF were compared to 500 sequences randomly sampled from the other TF families. As a control, 500 sequences randomly selected from that TF's family pool were compared to the subset from the other families.

#### Complementarity of methods

To assess the complementarity of StruMs with sequence-based methods, the scores for each sequence were passed as a vector of length two (PWM score, StruM score) to a logistic regression classifier as implemented in Python's `sklearn`[97, 98]. Ten-fold cross validation was performed with `sklearn.cross_validation.cross_val_score`, and the average AUC was retrieved.

### 3.2.4 Proximity of PWMs and StruMs

Given the motif derived by MEME for the ChIP-seq experiments described above, FIMO was used to scan the input sequences for statistically significant matches, as executed by `meme-chip` [88, 89, 99]. These significant matches were merged if they were within 100 bp of each other.

The sequences within 100 bp of the center of the clusters of significant matches were extracted, and scored with the EM derived StruM. The highest scoring position for each sequence was recorded. The average distance between the genomic locations of these best StruM matches to the nearest significant match to the PWM was then computed.

### 3.2.5 Program Versions

All programs and packages used in this analysis were downloaded and installed on a system running Ubuntu 16.04.6 LTS (GNU/Linux 4.4.0-150-generic x86_64) using Conda (4.6.2). The following packages and versions were used: MEME-suite (meme-chip, fimo) (4.12.0), Bedtools (2.27.1), Sci-kit learn (0.20.1), numpy (1.15.4), matplotlib (2.2.3), python (2.7.15), scipy (1.1.0).

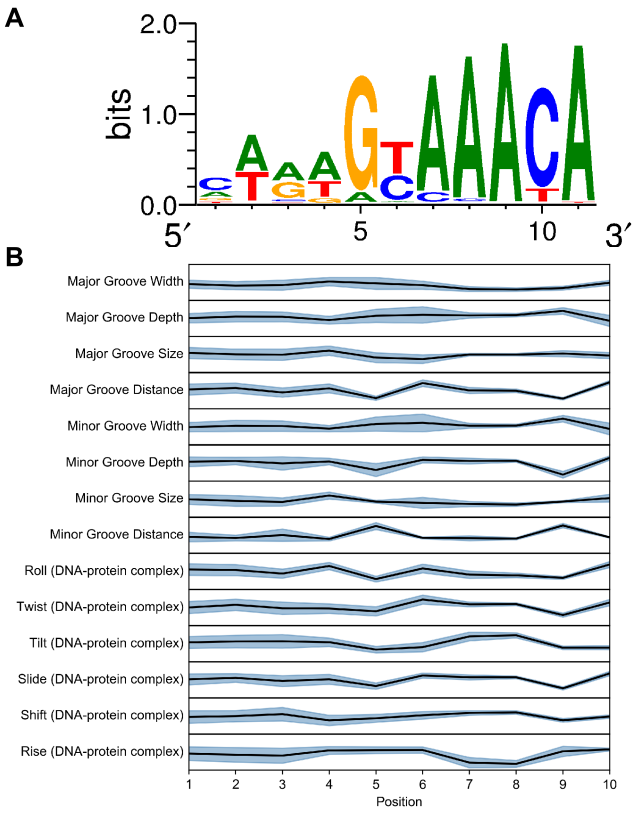



Figure 3.2: Motif for FOXA1. **(a)** Standard web logo representation of PWM for FOXA1. **(b)** Graphical representation of a StruM based on the same sequences as (a). The line plot represents the average value, and the shaded region is one standard deviation above and below.

## 3.3 Results

### 3.3.1 Overview of the StruM model

The traditional representation of binding preferences for transcription factors, or their motifs, is the PWM. Given a set of sequences that are known to be bound by a TF, e.g. GATA1 (Figure 3.1a), these sequences can be aligned by the binding site (Figure 3.1c). Assuming that each of the positions in the binding site are independent, they can each be represented by a distribution of nucleotide frequencies; one distribution per position in the motif (Figure 3.1b).

While most binding site representations are sequence-based, the physical interaction between the TF and TFBS must be compatible in both terms of electrostatics and sterics [100]. It has been shown that some transcription factors prefer specific shape configurations of the DNA [91, 70, 71]. It may be that the PWM and other sequence-based representations of TF binding motifs are abstracting these shape preferences, as sequence and DNA shape are tightly linked. We hypothesized that binding motifs could be modeled directly by DNA shape parameters.

The StruM model operates under the same basic assumptions as the PWM, but extends the model to correspond to shape values. If quantitative values can be obtained for characteristics of the DNA such as the Rise, Twist, and Turn, a distribution can be computed at each position of the binding site for these features. The StruM model parameterizes these distributions with the mean value and the observed standard deviation (Figure 3.1d).

Figure 3.2 shows the motif for FOXA1, learned using MEME [88]. Figure 3.2a is the traditional web logo representation of the PWM. A StruM trained on the same binding site regions identified by MEME show several interesting features (Figure 3.2b). The first thing to note is that there are indeed clear patterns. If there was no

preference for a given shape, the average values would all be near zero, and the standard deviations would consistently be near one, given the scaled shape values used in the model. Rather, clear patterns are observed in both the average values and the variance. For example, the variance trends from high to low from left to right, corresponding to the information content in the PWM across the positions.

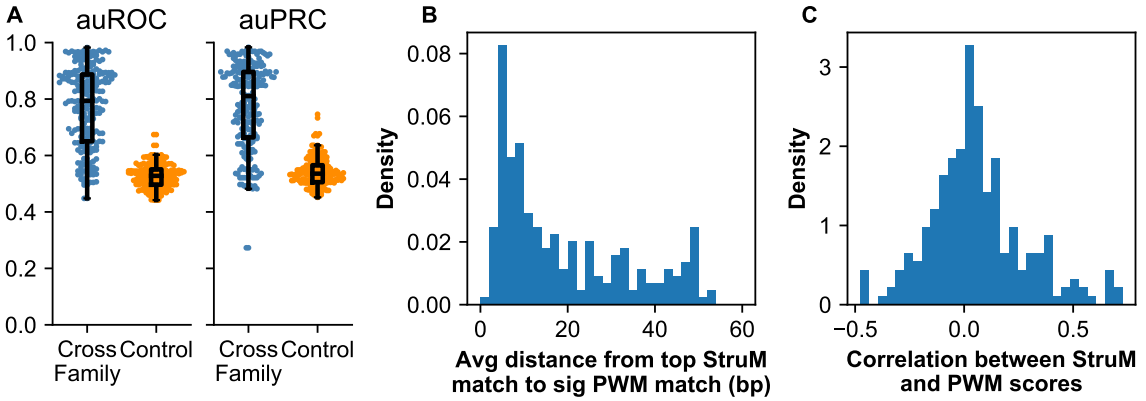



Figure 3.3: Specificity of StruMs. **(a)** The area under the ROC curve or PRC for TFs that could be unambiguously assigned to a TF family. Blue points correspond to the ability of the StruM to discriminate between ChIP-seq peaks for that TF *vs.* peaks for a different TF family (Cross Family). The orange colored points are the AUCs for the StruMs ability to discriminate between peaks from other TFs in the same family *vs.* peaks for a different TF family (Control). **(b)** Distribution of average distances from each StruM match to nearest significant PWM match. **(c)** The majority of correlations between scores assigned to kmers by PWMs and StruMs are near zero.

### 3.3.2 StruMs specifically model TF binding sites

There is evidence to suggest that promoters and other similar genomic elements may share certain general shape features [101]. A simple example of conserved promoter structure would be the prevalence of the TATA-box [102, 103]. One possibility in evaluating the performance of shape based models to recognize TF binding sites is that they may be instead modeling general features of the type of genomic element that TF may target.

To determine whether a StruM is specific to the binding sites of the TF targeted in the ChIP experiment used to train the TFBS model the StruM was presented with a simple classification task to evaluate its ability to discriminate between ChIP peaks for its cognate TF from ChIP peaks deriving from other TF families. As a control, it was assessed whether StruM could distinguish between ChIP peaks from the same TF structural family versus the set of peaks from other TF families.

As shown in Figure 3.3a, StruMs were well able to distinguish between cognate TF binding sites and those belonging to other TF families (Avg. auROC = 0.77, Avg. auPRC = 0.77). In contrast, the control sequences appeared indistinguishable from the extra familial sequences (Avg. auROC = 0.53, Avg. auPRC = 0.54). Using a two-sided paired t-test this was a statistically significant difference (auROC: p-value = $4.33 \times 10^{-78}$, auPRC: p-value = $1.62 \times 10^{-71}$) indicating that the StruMs are specific to the TF on which they were trained.

Once peaks are called from a ChIP-seq experiment and a motif is identified, the next step is frequently to identify the probable binding locations of the target factor at a higher resolution. One approach is to look for significant matches to the identified motif near the peak summit. If StruMs are accurately describing the binding sites of their cognate factors, high scoring positions within the peak should correspond with significant matches to the PWM. In order to evaluate this, FIMO [99] was used to identify significant matches to the PWM in the original ChIP sequences. The 200 bp surrounding each significant match was scored using the StruM, and the top scoring position for each sequence was retained. For each ChIP experiment, the average distance between each StruM match site and the nearest site identified by FIMO was calculated. As observed in the distribution in Figure 3.3b, the majority of StruMs recognized sites on average within 15 bp of the sites identified by FIMO. In many of these instances this offset seems to be the result of the motifs not aligning perfectly on the TFBS (Figure B.1, Figure B.3). The other peak in the distribution

near 50 bp corresponds with StruMs that failed to accurately model the binding site (e.g. Figure B.1), as there is a strong negative correlation between the average distance and the auROC of the StruM model (Figure B.2, $R = -0.69$).

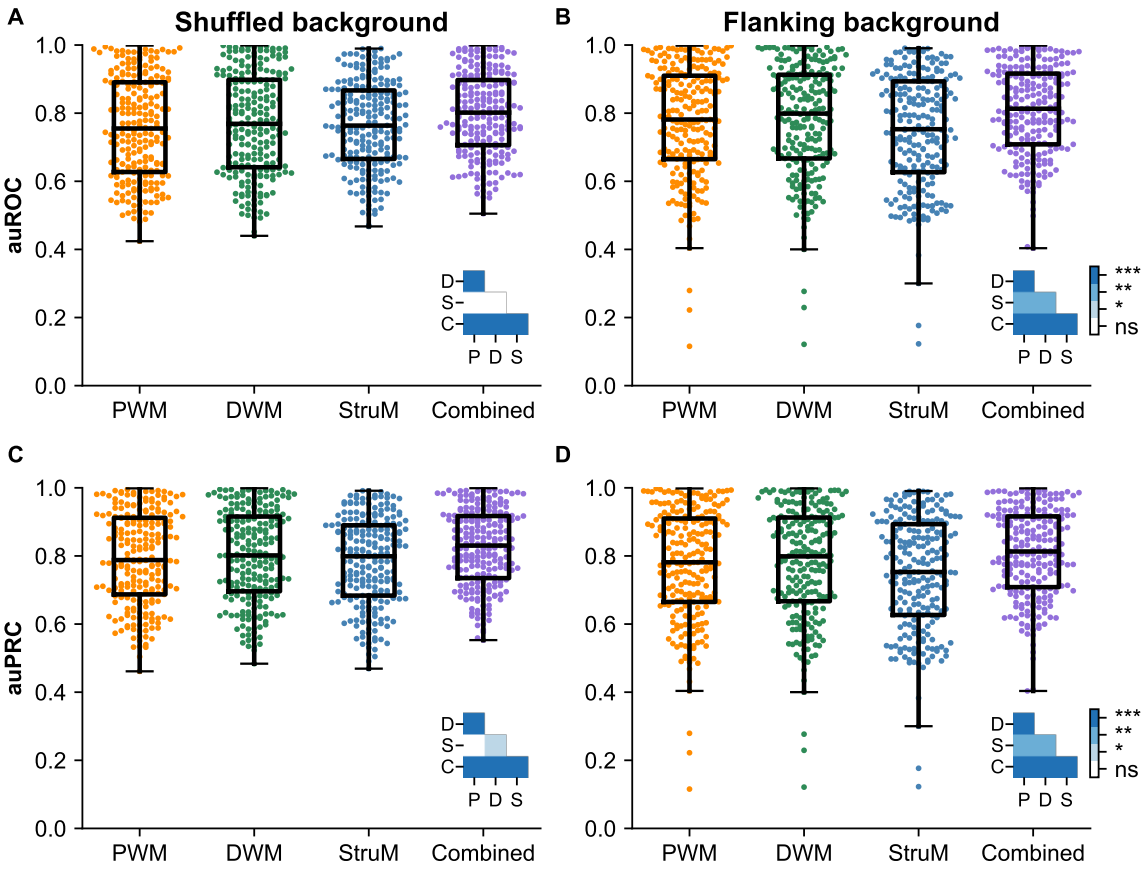



Figure 3.4: Classification of ChIP-seq peaks *vs* non-peak sequences. **(a-d)** Area under the curve for each TFBS model trained on the top 500 peaks, and tested using the next 500 sequences as the positive examples. The combined model used 10-fold cross validation to generate the score. The negative sequences came either from (a,c) shuffled peak sequences or (b,d) flanking sequences. Panels (a-b) represent the area under the ROC curve, and panels (c-d) represent the area under the PRC. Significance for each pairwise comparison is shown in the inset, with the following abbreviations: P–PWM, D–DWM, S–StruM, C–Combined model.

In the previous experiment, the PWM was used as the baseline for confirming the specificity of the StruM. To extend this analysis, a comparison was made of the ability of three models to discriminate between regions identified via ChIP-seq ex-

periments and a negative control: either randomized peak sequences, or flanking sequences. The three models compared for each factor were: Position weight matrix (PWM), Dinucleotide weight matrix (DWM), and a Structural Motif (StruM). The area under the receiver operating characteristic curve (auROC) and the area under the precision-recall curve (auPRC) were calculated for each ChIP-seq experiment, using each of the models (Figure 3.4).

After evaluating 229 ChIP-seq experiments in K562 cells from the ENCODE consortium, StruMs perform at a similar level with the sequence based methods. It is interesting to note that using shuffled sequences as the negative set resulted in performance that was quite comparable across the three models. When using flanking sequences, the sequence-based approaches showed a slight but statistically significant edge over the StruM.

### 3.3.3 StruMs encode motifs differently than sequence-based methods

It is quite interesting to note that models constructed in this way (a probability distribution at each position of the motif, assuming independence between positions) perform relatively similarly, regardless of the feature being considered, be it mono- or dinucleotides, or shape features estimated from dinucleotides. In addition to similar performance in a simple classification problem, the different representations identify similar positions as being the putative targets for a given TF.

One might therefore expect there to be a strong correlation between the scores produced by each model for a given sequence. However when scoring a set of 1000 randomly generated sequences the models have near zero correlation (Figure 3.3c. Average correlation = 0.059, standard deviation = 0.22). This indicates that while the separate motif representations model the same site, they are encoding the information at that site very differently.

### 3.3.4 Shape and sequence are complementary

Given that both sequence- and shape-based methods can model the same sites with a similar level of accuracy, we investigated whether these methods were in fact redundant, despite showing very little relatedness between the ordering of random sequences. If these models are in fact parameterizing distinct features of the binding site, one would expect a combined model to outperform either model alone.

Towards this end a simple logistic regression model was trained for each TF, passing the maximum score from the PWM and the StruM as a length 2 vector for each sequence as input. Using 10-fold cross validation the combined model significantly outperformed each motif alone. (Figure 3.4a-d. Paired t-test; p-values $< 4 \times 10^{-12}$).

As further evidence of the complementarity of the models, the increase in auROC vs. shuffled sequences was plotted by the combined model over the StruM against the increase over the PWM performance (Figure 3.5a). In the event where the combined model simply agreed with the best performing model one would expect the points to fall along the x- and y-axes. Rather it was found that the majority of the points (68%) fall in the first quadrant off of the axes again indicating the complementarity of sequence and structural motif representations. Most of the remainder, accounting for 22% of all experiments, performed best with the StruMs alone.

### 3.3.5 Towards distinguishing between direct- and indirect-readout mechanisms

Despite the similarity in performance on average between the sequence- and shape-based TFBS representations, there were several factors that showed a stronger than average preference for one method over another. We sought to understand

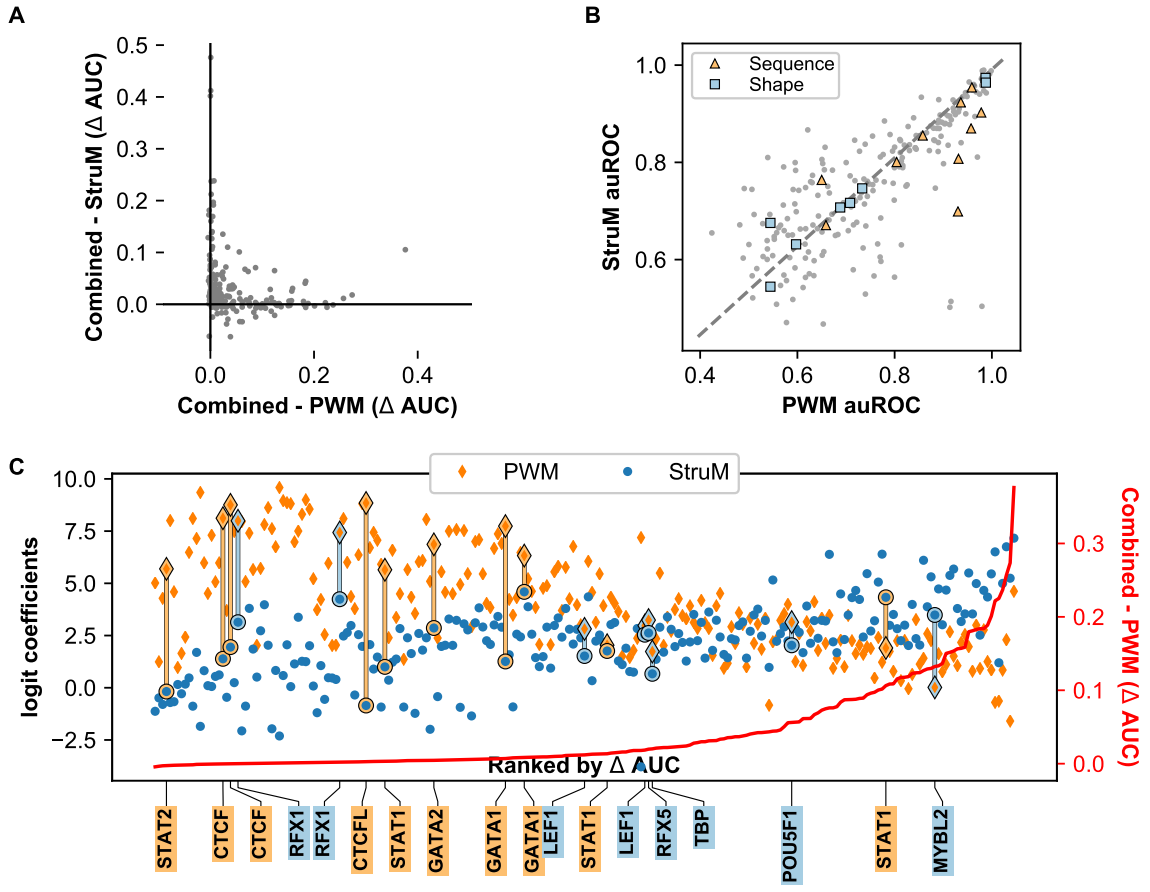



Figure 3.5: Combined PWM-StruM model. **(a)** The improvement of the combined model over StruMs (y-axis) was plotted against the improvement over PWMs (x-axis) for each TF. Points falling into the first quadrant show a positive impact on performance by joining multiple motif representations. **(b)** TFs that are known to utilize the base-readout mechanism (orange diamonds) fall on both sides of the line, while known shape-readers (blue circles) are generally better predicted by StruMs. **(c)** Coefficients for PWM and StruM from the combined logistic regression model were plotted against their rank improvement of the combined model over PWM alone. The absolute value of that improvement is indicated by the red line and corresponds to the values on the right-hand axis. The known sequence- and shape-readers are identified by gold and blue lines, respectively.

whether these differences were limited to a subset of factors. First, the relationship between the AUC values for PWMs and StruMs for specific factors was examined (Figure 3.5b). Points falling along the dashed line representing $y = x$ represent experiments where both representations perform equally well in predicting whether a sequence is likely to be bound by the factor.

We hypothesized that factors falling above the line employ shape- or indirect-readout mechanisms, whereas points below the line represent factors employing primarily base- or direct-readout. Several examples of known base- and shape-reading TFs are highlighted. In line with this hypothesis, those factors with a larger than average residuals tend to segregate by their readout mechanism.

Next, the coefficients of the combined logistic regression model were considered. One would expect that shape-readers would give more weight to the PWM than to the StruM. In direct contrast to the previous observations, the logistic regression model prefers the PWM a majority of the time, regardless of the annotated readout mechanism of the factor (Figure 3.5c). In fact only 1 out of 8 shape-readers weighted the StruM score more strongly, and the model preferred the StruM for only 1 out of 10 experiments for the known sequence-readers.

## 3.4 Discussion

In this work we have presented a novel representation of transcription factor binding site preferences termed **Stru**ctural **M**otifs, or **StruM**s. This model is an extension of the formulation behind the time-tested PWM that accommodates distributions of shape features. The flexibility of this model allows for variations of the StruM that can be tailored to specific tasks, and the incorporation and integration of additional data types. The DiProDB is used as a simple and fast system for converting sequences to a structural representation [93]. Other methods for shape estimation

such as the DNAshapeR package would fit well within this representation [73]. Arbitrary other data types such as DNase hypersensitivity may likewise be incorporated, as long as quantitative values can be generated for each position in the sequence. The StruM representation shows an ability to specifically model TF binding sites, as well as differentiate between 'peak' and 'non-peak' sequences in a ChIP-seq experiment.

Despite an average similarity to sequence methods in performance for these tasks, representing motifs using shape features is not universally appropriate. Some transcription factors employ a direct-readout mechanism whereby they recognize their binding site via interactions with specific base pairs. For these factors, the shape is an abstraction of the sequence information, rather than the other way around, and the PWM (or other sequence-based representation) is preferable for predicting binding sites.

A number of methods have been developed recently which seek to discover TF binding motifs in local DNA shape [76, 74]. As with StruMs, the values used for DNA shape are estimated directly from the sequence using a table like DiProDB [93] or from simulations like DNAshape [73]. The local structure of naked DNA is likewise determined by the sequence composition. This raises the question of whether consider higher order nucleotide features could fully capture the information contained in shape features. Indeed recent work has shown that given appropriate training data, dinucleotide features alone are sufficient to model most shape features [72, 14].

While it is possible to model the contribution of shape features indirectly through dinucleotide models, and indeed the dinucleotide model (DWM) displayed a strong performance, the StruM parameterizes the same information in a very different manner from these sequence-based methods. This disparity between the scores generated by the different model types turned out to be useful; combining the scores

into a single model performed even better than either method alone. Thus the different mechanisms of encoding the same information are complementary to one another.

In summary, StruMs provide a novel way of considering transcription factor binding sites, which is complementary to sequence-based approaches. In fact many transcription factors may utilize a blend of direct- and indirect-readout mechanisms, agreeing with recent evidence (Reviewed in [91]). In this context, and given the observed complementarity, representing motifs using StruMs provides valuable information about the binding site preferences of transcription factors, and that when used in conjunction with sequence based methods can produce high confidence cell type-specific predictions of TFBSs. This will be valuable especially for studying the TF binding landscape of rare cell types for which carrying out extensive TF ChIP-seq experiments would be prohibited either by cost or the ability to collect enough cells for those experiments.

## 3.5 Availability

Source code for StruMs and related tools as well documentation are available in the GitHub repository (https://github.com/pdeford/StructuralMotifs). All code required to replicate the analyses in this paper are available at (https://github.com/pdeford/strum_paper).

## 3.6 Funding

This work was supported by the National Institutes of Health [T32 GM 007231, R24 DK 106766].

# 4. Intrafamilial Discrimination

One challenge in predicting transcription factor binding sites is that members of the same transcription factor family have very similar motifs [104]. Intuitively this makes sense, as the transcription factor families are defined based on the structural similarity of their DNA binding domains. High structural conservation and similarities in these functional domains result in sequence preferences that are quite similar.

We hypothesized that, compared to sequence based approaches, Structural Motifs might be better able to discriminate between binding sites within a transcription factor family. Having more parameters associated with each position in the binding site may allow for increased sensitivity to variations from the preferred shape of a given transcription factor.

To assess this we used the motifs discovered in Chapter 3, and evaluated several key characteristics. First we determined the similarity of the discovered motifs for replicate experiments for the same TF. Secondly we assessed the within cluster similarity when TFs were clustered based on their PWM or StruM representation. Finally, we consider next steps for utilizing the observed differences in motif similarity between PWMs and StruMs to improve the discriminatory power of predictive methods when presented with similar motifs.

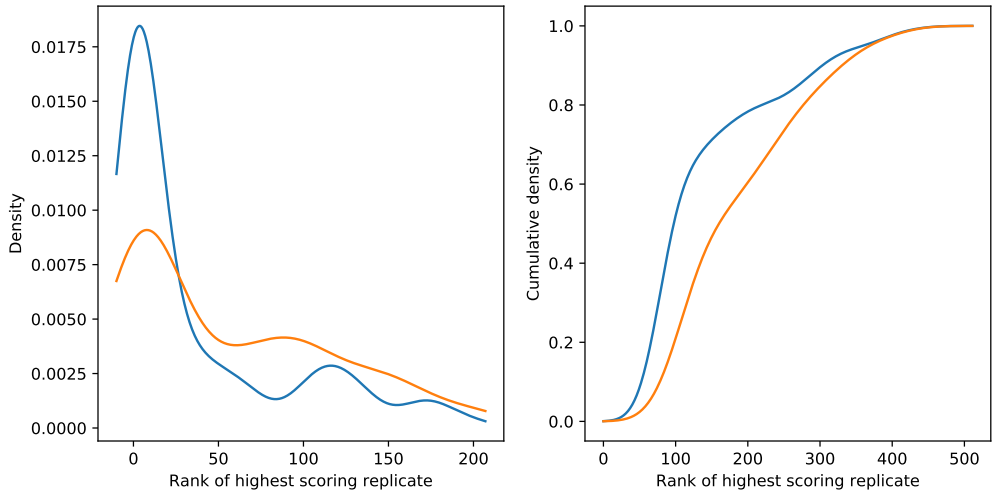



Figure 4.1: Similarity of motifs from replicate experiments. X-axis is the rank of replicate motifs given that all motifs are ranked by their similarity to the reference motif. The blue line corresponds to position weight matrices, and the orange line corresponds to Structural Motifs. **Left:** The density distribution of replicate ranks. **Right:** The cumulative density distributions of replicate ranks, used to calculate the KS statistic.

## 4.1 Replicate experiments yield the most similar motifs

Firstly, we assessed the similarity between motifs from several transcription factor families. The main ones considered were bZIP, C2H2 Zinc Finger (C2H2 ZF), basic Helix-Loop-Helix (bHLH), and Ets transcription factors. As the position weight matrix is the most widely used representation of sequence motifs, it was used as the standard to which the Structural Motif was compared.

There are several methods for aligning position weight matrices and scoring the similarity between any two models (e.g. TOMTOM [105]). The extensions to Structural Motifs were complicated, and it was unclear if the two sets of similarity scores would be comparable. Another method of scoring similarity between motifs

that is agnostic to alignment was recently introduced, known as the MoSBAT energy score [106]. In short, many sequences are assigned a score by each of the motifs being considered. Then, for any two motifs, the relatedness is the Pearson correlation between the scores assigned to all of those sequences by the motifs.

Before comparing position weight matrices to structural motifs at a broad scale, the MoSBAT energy score was applied to the Structural Motifs for many ChIP datasets. Some transcription factors had replicate datasets. For those with replicates, the Pearson Dissimilarity was computed for all other datasets, the motifs were sorted by that distance, and the ranks of the replicate motifs were extracted. As shown in Figure 4.1, the distributions of replicate ranks for position weight matrices and structural motifs were quite similar. A Kolmogorov-Smirnov test between the distributions had a p-value of 0.0018. This indicated that while the structure of the distributions is quite similar with the mode near zero, the Structural Motifs do show a heavier right tail, indicating that the replicate motifs are consistent, but less similar overall compared to the position weight matrices.

Overall this demonstrates that the MoSBAT energy score is an appropriate distance measure for Structural Motifs, as the replicates are generally the most similar motifs for a given reference. Additionally, this provides a model agnostic way of comparing motif similarity, that can be applied in the exact same way for both sequence- and shape-based motifs, allowing for better comparisons.

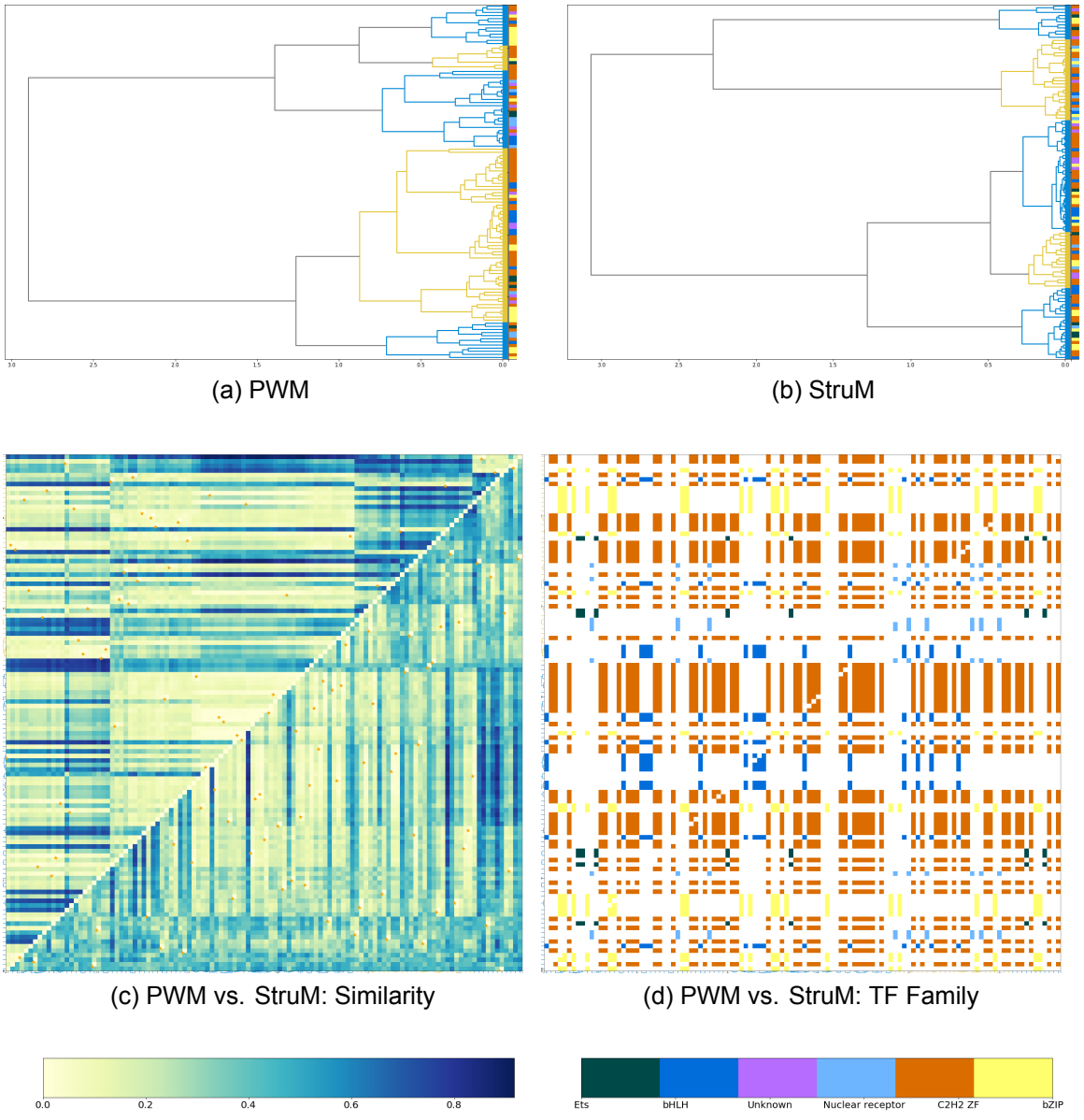



Figure 4.2: Clustering of motifs. **(a-b)** Motifs are clustered by Pearson Dissimilarity on the MoSBAT energy scores. Leaves are colored by what Structural family the TF belongs to. **(a)** PWM Motifs. **(b)** StruMs Motifs. **(c-d)** Joint clustering by PWM and StruM. The heatmap is clustered in one direction by StruM similarity (bottom) and the other by PWM similarity (left). Orange dots mark where the same TF appears in both clustering schemes. **(c)** Similarity heatmap. Upper triangle corresponds to the similarity of StruM motifs. Upper triangle corresponds to the similarity of StruM motifs. **(d)** Colormap indicates which leaves in the opposite dendrogram correspond to the same TF family.

## 4.2 Structural Motifs are less similar within transcription factor families than position weight matrices.

Based on the Pearson dissimilarity metric computed above on the MoSBAT energy scores for position weight matrices and Structural Motifs, the motifs can be clustered hierarchically. The identified clusters reveal which motifs are the most similar. We expect, based on the level of conservation observed among members of the same family, that the motifs will cluster predominantly by their family membership.

Clustering the position weight matrix motifs, there are large blocks of C2H2 zinc fingers that cluster together (Figure 4.2a). Additionally, blocks of bZIP, Nuclear receptor, and bHLH motifs can be observed. In contrast, the structural motifs appear to cluster more randomly, and the single family blocks of motifs are generally smaller (Figure 4.2b).

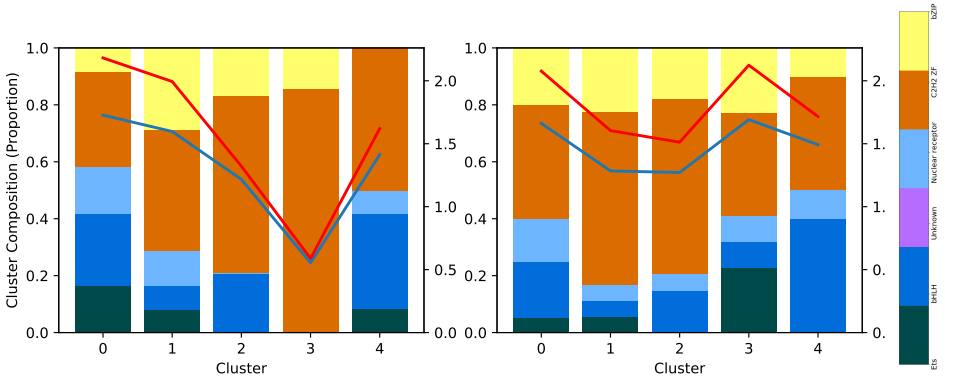
Assessing both sets of clustering together reveals the differences between the similarities identified in both motif types. In Figure 4.2c, the similarity matrix was clustered and sorted according to the same dendrograms from panels (a) and (b). The upper triangle represents the similarities between StruMs, while the lower triangle is the similarities between PWMs. If the cluster memberships were similar, there would be blocks of color corresponding to each of the clusters. Additionally, orange points indicate where each single TF was positioned in both dendrograms. Again, if there was consistency between the two clustering schemes, there would be groups of points that co-occur in close proximity. Rather, there is a fairly even distribution of the orange points, and the striations observed in the heatmap indicate that the two methods result in very different clusters. This is further confirmend by Figure 4.2d, where instead of the similarity heatmap, for each TF the position of

the other TFs of the same family are indicated. The number of blocks for each color is different depending on the reference axis being considered, and again striations appear, preferentially positioned up and down indicating that the StruM clusters are less pure.

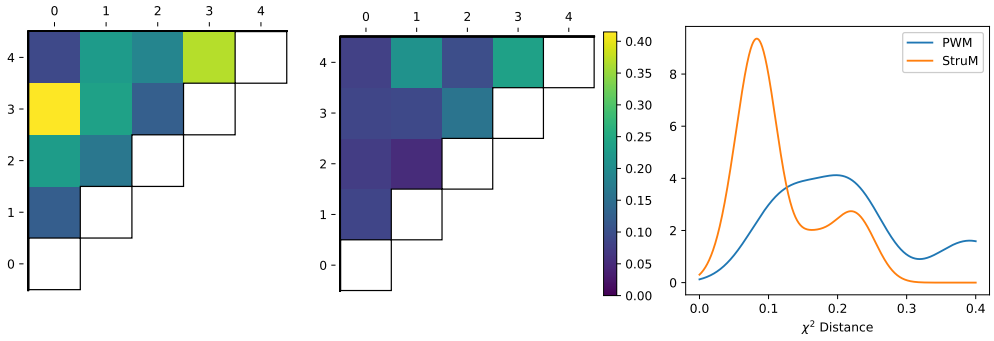
To examine the purity of the clusters more rigorously, the composition of TF families present in each cluster was evaluated using two related measures: the Entropy and the Gini Impurity. If the clustering is totally random, the composition of the clusters will be fairly uniform, maximizing the Entropy and Impurity of each cluster. In contrast, clusters with a bias towards a single transcription factor family will have lower values. In Figure 4.3 it is fairly clear that the composition of the StruM clusters (b) are more uniform in general than the PWM clusters (a). In fact, the $\chi^2$ Distance among the StruM clusters is less on average than the Distance among the PWM clusters (Figure 4.3c–e).

We can directly compare each of the clusters from the PWM similarity tree to each of the clusters in the StruM similarity tree. Using the Jaccard index to assess the similarity of each pairwise combination of clusters between the two motif types by considering the membership of each. Figure 4.4a is a heatmap showing each of these pairwise comparisons. These values are summarized in the distribution in Figure 4.4b, and show that majority of the scores are fairly close to zero (average=0.1, standard devation=0.06), with the maximum being an outlier, and still only 0.26. This suggests that the clusters are very different from one another. In fact, even if only the maximum score for each cluster is considered, the average Jaccard score is still only 0.15 and 0.17 for either the PWMs and StruMs, respectively (Figure 4.4c).

Thus far, it has been established that the similarities identified by both motif types are different, resulting in clusters with low overall levels of identity between the two methods. In addition, the Structural Motif clusters are generally less pure

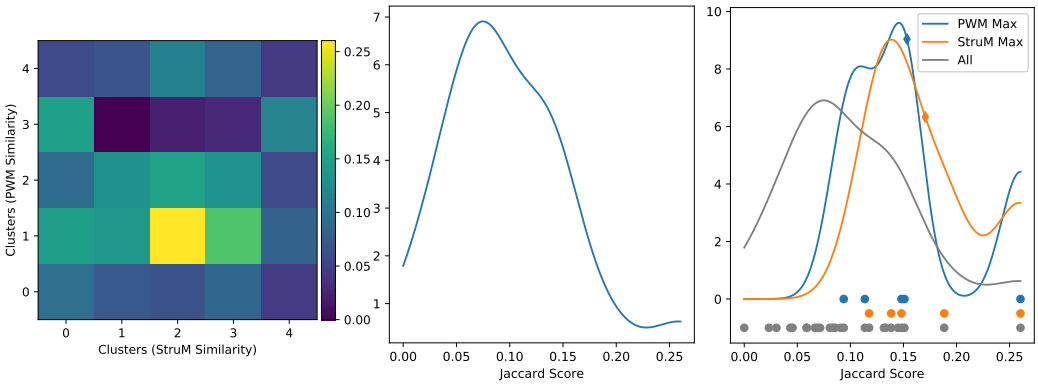



(a) PWM Similarity (b) StruM Similarity




(c) PWM (d) StruM (e) $\chi^2$ Distance

Figure 4.3: Motif cluster composition. **a–b** The composition of TF families present in each of the motif clusters for PWMs (**a**) or StruMs (**b**). The right-hand axis corresponds to the Entropy (Red line) and Gini Impurity (Blue line). **c–e** $\chi^2$ Distance between each of the PWM (**c**) or StruM (**d**) clusters, or the distribution of values from those heatmaps (**e**).




(a) (b) (c)

Figure 4.4: Similarity between motif clusters. **(a)** Heatmap of the Jaccard score between each of the PWM clusters (left) with the StruM clusters (bottom) from Figure 4.2a-b. **(b)** Distribution of Jaccard scores from panel ***a***. **(c)** Distribution of Jaccard scores from panel ***a*** (grey), the maximum score for each PWM cluster (blue), or StruM cluster (orange). The points along the bottom indicate the scores being included in each density plot. The diamond is the average of the distribution.

than the position weight matrix clusters. This implies that, if considered together, transcription factors should become more separable from the other members in their structural family.

## 4.3 Future applications

A more useful extension of separating motifs from within the same transcription factor family is the ability to identify to which transcription factor a given sequence is a binding site for, from within the structural family.

Due to the extreme degree of similarity observed between motifs for members of the same structural transcription factor family, simple sequence or shape based models may be insufficient, as the most preferred series of DNA bases (the consensus sequence for sequence-based methods) is often identical for multiple members of that family.

As will be discussed in the next chapter, additional features might be able to

be incorporated into the structural motif to improve this resolution, though surprisingly we found that DNase sensitivity is insufficient for this task. The more likely solution is to include more of the putative binding site's genomic context as input to a higher order model. For example by considering the motifs of potential binding partners for transcription factors that are known to dimerize. This type of interaction is easily captured through convolutional layers in artificial neural network methods, as has been observed with models like DeepBind for transcription factor binding, and Basset for predicting the impact of sequence variation on DNase accessibility [51, 107].

# 5. Areas Needing Further Development

## 5.1 Cell Type-specific Predictions of Binding Sites

### 5.1.1 DNase signatures

One important feature of the way the structural motif model was constructed, is that it can incorporate an arbitrary number of shape features, as long as a quantitative measure can be produced for each position in the binding site. The literature has long described DNase signatures that are present at TFBSs due to unequal protection of the nucleotides by the bound transcription factor [108]. As a result we extended the StruM to include DNase-seq signals to promote the discrimination between bound and unbound sites that are otherwise similar. In theory, this would allow for the distinction between bound and unbound instances of the motif, even if they are identical at the sequence level.

However, recent work by Sung, *et al*. (2014) has shown that these apparent DNase signatures are directly a result of the sequence bias of the nuclease, and can be predicted from DNA tetramers [109]. Due to this fact, directly incorporating DNase accessibility merely contributes an abstraction of the overall level of accessibility in the model. As such, this avenue of cell type specific predictions proved

unfruitful.

### 5.1.2 Modulated StruMs

Many of the available measurements for DNA shape are derived from essentially naked DNA. This includes those found in the Dinucleotide Property Database [93] as well as those produced by simulations in DNAshapeR [73]. Naked DNA would represent DNA in its most accessible form. At any given time, however, the majority of the genome is inaccessible.

We hypothesize that in inaccessible regions, the DNA undergoes distortions that make individual nucleotides appear more similar, sterically. Effectively, the differences between nucleotides are minimized relative to the distortion that may be present. We propose that in cases where the DNA is not sensitive to DNase, and therefore inaccessible, the shape values across distinct sequences should appear more similar. In practice we will make these sequences appear more "average", scaled according to the sensitivity to DNase.

**Box 3: DNA accessibility model**

**Inaccessible DNA** $\xleftrightarrow{f(v_{i,j}|A_i)}$ **Accessible DNA**

- Shape distortions outweigh differences between bases.
- Shape features are forced towards the mean value, here normalized to $0$.

- Naked DNA, comparable to results of shape estimators.
- Observe shape features as their true values ($v_{i,j}$).

In the accessible state, the differences in shape between nucleotides will be overshadowed by the larger scale torsion on the helix. Bases that are completely inaccessible will appear completely identical, in that sense, to a transcription factor scanning for its binding site. On the other end of the spectrum, completely accessible DNA is largely equivalent to naked DNA, with much fewer stresses on the helix. In this state, the shape features should agree with the values produced by shape prediction tools. Intermediate values of accessibility should therefore correspond to an intermediate value of the shape value, between the predicted value and the average.

This can be expressed in terms of a scaling factor that is a function of accessibility, $f(A)$. In the inaccessible state, the scaling factor should be zero. In the perfectly accessible state, the scaling factor should be one.

$$f(0\%) = 0$$

$$f(100\%) = 1$$

The particular function used to scale based on the accessibility should be monotonically increasing, but may not be linear. In fact, as will be discussed, a logistic function may be appropriate.

**Theoretical Framework**

The standard method for calculating scores is shown in equation 2.6, reproduced here:

$$P(S_i|\phi) = \prod_{j=1}^{m} P(v_{ij}|\mu_j, \sigma_j^2) \qquad \text{(2.6 revisited)}$$

The structural motif is trained on ”naked” DNA and so all of the parameters will reflect that. If a region of DNA is inaccessible, we would expect the the sequence's shape values would diverge from the transcription factor's preferred conformation. In short we would have a modulated form of the scoring function:

$$P(S_i|\phi, A_i) = \prod_{j=1}^{m} P\Big(f(v_{ij}|A_i) \;\Big|\; \mu_j, \sigma_j^2\Big) \tag{5.1}$$

An appropriate function for $f(v)$ may be the logistic function. It will buffer many low accessibilities as being in accessible. At the high end it will eventually reach a level that will be considered essentially naked. In the middle it will have a transition region, with a slope specific to each TF/StruM model. The form may look something like:

$$f(v_{ij}|A_i;\theta) = \frac{v_{ij}}{1 + e^{-\theta(A_i - A_0)}} \tag{5.2}$$

where $A_i$ is the accessibility of the region, $A_0$ is a parameterized factor to shift the distribution appropriately, and $\theta$ is the transcription factor-specific scaling factor that determines the slope of the middle region.

### 5.1.3 Methods.

**Training Data** All data was obtained from the ENCODE Uniform Processing Pipeline. For a given transcription factor several datasets are required. First are the peaks from ChIP-seq experiments in at least three cell types. These provide sufficient numbers of regions that are known to be bound by the transcription factor, that are unbound in the cell type of interest. The cell type of interest is then determined by the second dataset required: the DNase signal for at least one of the three (or more) cell types represented in the ChIP experiments. Finally the sequences for those ChIP peaks are required.

The initial structural motif is computed *de novo* from the top 500 peaks from the union of all available cell types. For the rest of the peaks, the sequence $\pm 50$ bp from the peak of each region was scored with the motif, and the highest scoring position was considered the binding site. The positive examples will be those peaks that are present in the cell type of interest, and the negative examples are those peaks from the other cell types that share no overlap with any of the positive examples. These positive and negative sets were then split into a training and testing set of data.

**Parameter Initialization.** Parameter fitting by gradient descent (described in the next section) is computationally expensive in this case. It has the additional risk that, given the dimensionality of the data, the loss function may have local minima which may trap optimization algorithm. To accommodate these two factors a good initial guess at the parameters is important. This is achieved by fitting a simple logistic regression classifier to the accessibility data, with the classes being defined as whether the peaks come from the positive or negative training sets. In this case, the intercept of the classifier will correspond to $A_0$, and the coefficient will approximate $\theta$.

**Parameter optimization.** Each transcription factor will have a different characteristic response to DNA accessibility. Pioneer transcription factors, for example, should be relatively invariant to the accessibility of the region, while others may be completely dependent on the state of the chromatin. Thus each structural motif will need modulated individually, in a way that represents that particular transcription factor's activity. This is done through fitting the parameters $A_0$ and $\theta$ by gradient descent, using a Cross Entropy loss function.

The general form of the Cross Entropy loss function [110] is:

$$J(\theta, A_0) = -\frac{1}{m}\sum_{i=1}^{m}\left[y_i \ln\Big(P(x_i)\Big) + (1-y_i)\ln\Big(1-P(x_i)\Big)\right] \tag{5.3}$$

Given the similarity in construction of the structural motif to the position weight matrix, similar methods can be employed. In fact, the principals for defining $P(x_i)$ were derived from those used in the PWM tool GOMER [111]. In short, $P$ is determined as a biochemical value depending on concentrations and energies.

$$P = \frac{[x]}{K_d + [x]} \tag{5.4}$$

where $[x]$ is the concentration of the factor $x$, and the dissociation constant $K_d$ is defined as

$$K_d = e^{-\Delta G/RT} \tag{5.5}$$

Based on the the Gibbs free energy of reaction ($\Delta G$), the ideal Gas Constant ($R$), and temperature ($T$).

$$\Delta G = \sum_i RT \ln \frac{f_{bi}}{p_b} \tag{5.6}$$

$$\Delta\Delta G = RT \ln \frac{f_{bi}}{p_b} \tag{5.7}$$

While this method was conceptualized for the discrete basepair frequencies of the PWM, it can be adapted for use in the structural motif case by changing $f_j$ and $b_j$ to derive from a continuous distribution (assuming that the data is normalized such that, under a uniform distribution of nucleotides, the mean value for any shape would be zero, and the standard deviation would be one):

$$f_j = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(f(v_{ij})-\mu)^2}{2\sigma^2}} \tag{5.8}$$

$$b_j = \frac{1}{\sqrt{2\pi}} e^{\frac{-(f(v_{ij}))^2}{2}} \tag{5.9}$$

With one final simplifying assumption that the temperature is the same for training and testing, this produces a final value for $P$ of:

$$P = \frac{[x]}{[x] + e^{-\sum_{j=1}^{p} \frac{f(v_{ij})^2}{2} - \frac{(f(v_{ij})-\mu_j)^2}{2\sigma_j^2} - \ln \sigma_j}} \tag{5.10}$$

Due to the risk of local minima, a grid of parameters were explored centered on the initialized parameters. This grid encompassed 4 values evenly spaced $\pm 2$ from each parameter ($A_0$ and $\theta$), for a total of 16+1 parameter sets to be optimized.

### 5.1.4 Results

#### Optimization

Three transcription factors were selected as representative for this analysis: GATA2, MNT, and SUZ12. Following the initialization step of using a logistic regression classifier on the average accessibility, gradient descent was used with the cross entropy loss function to further optimize the model parameters. A grid of parameters centered on the initialized values was selected to optimize. As can be observed in Figure 5.1, this grid was important. Some initializations were in local minima and could not be further optimized, others were partially optimized before being likewise caught, but the majority led to a more optimum solution, minimizing $J(\theta)$ for example.

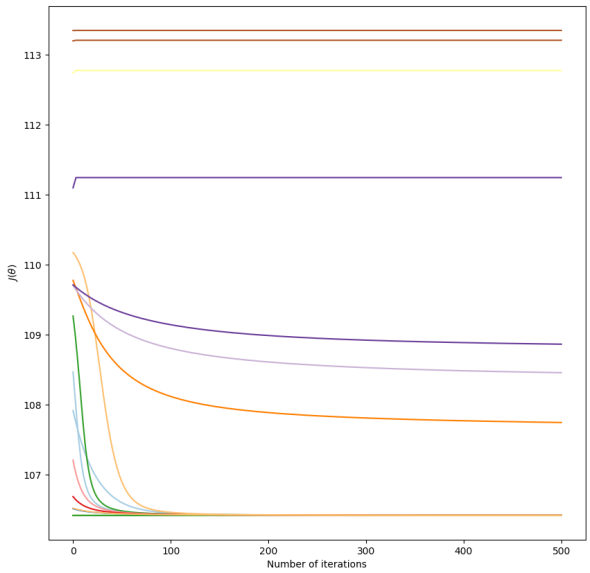


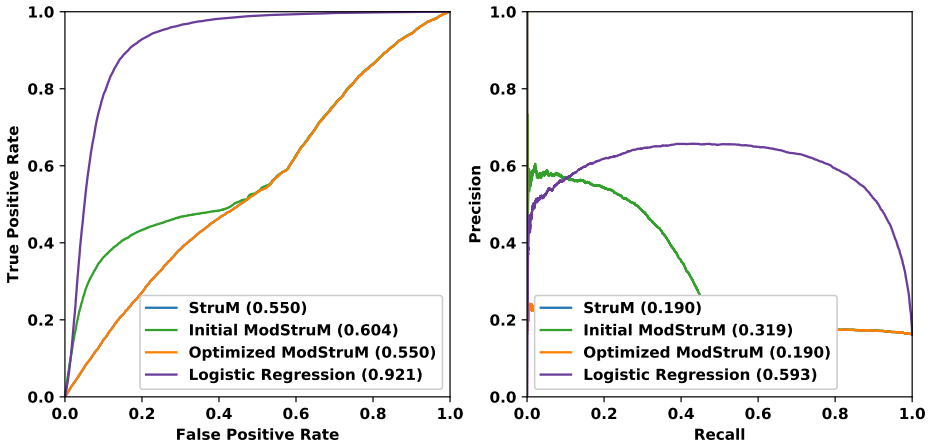

Figure 5.1: Optimization of parameters for the MNT model. Shown here are the cost function values at each step in the gradient descent optimization of $\theta$. Each line represents one parameter set used as the starting place for the optimization.

### Cell type specific predictions

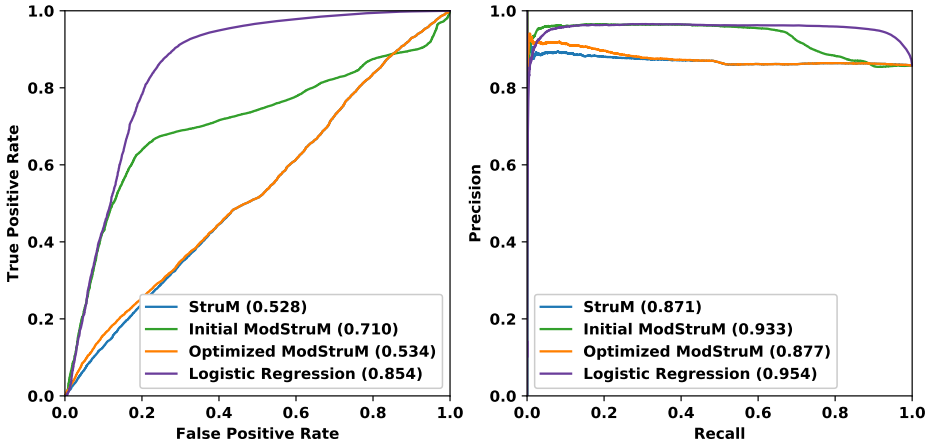
To evaluate the utility of the DNase modulated structural motif, it was applied to a simple classification problem. Non-overlapping ChIP-seq peaks from two or more cell types were used. Using DNase accessibility from one of the cell types, the model was used to predict which of the peaks originated from the same cell type as the DNase data by scoring each peak. An ROC curve and a precision recall curve was generated to evaluate the performance.

Four models were evaluated in this way. First was a naïve structural motif that is unmodulated. Second is a modulated structural motif using the logistic regression initialized values of $\theta$ and $A_0$. The third model is the fully optimized modulated structural motif. Finally is a simple logistic regression classifier on the accessibility of the region (given that a motif has been identified).

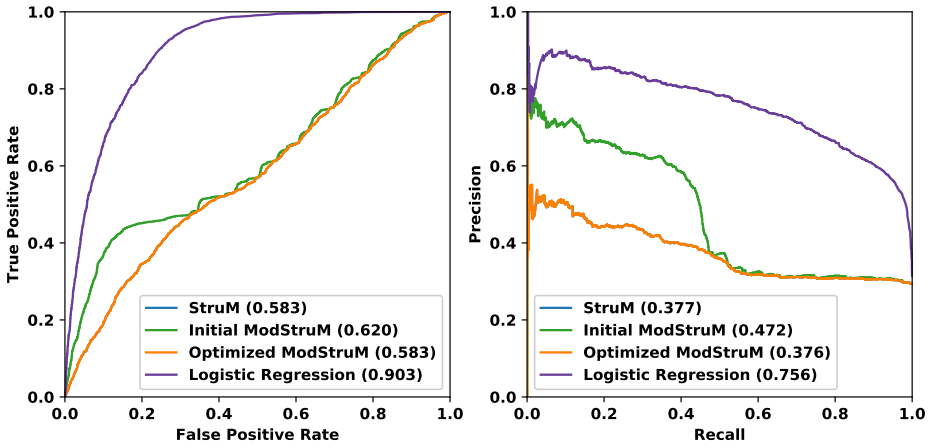
Figure 5.2 shows the performance curves for these four models applied to each




(a) GATA2




(b) MNT




(c) SUZ12

Figure 5.2: Performance of the Modulated StruM. The left panels are the ROC curve, the right are precision-recall curves. The curves shown are for a naïve StruM (blue), a ModStruM with the initialized values (green), the optimized ModStruM (orange), and logistic regression (purple). **(a)** GATA2, **(b)** MNT, **(c)** SUZ12.

of the three representative transcription factors. As both the positive and negative training sets include regions that are bound by the factors at some point, it is expected that a model without cell type specific information would be unable to distinguish between the two, and the area under the ROC curve would be close to 0.5. This is indeed what is observed as the performance of the unmodulated structural motif. In all three cases the area under the ROC curve is less than 0.6. Surprisingly, this poor performance is mirrored almost exactly by that of the fully optimized modulated structural motif. The modulated structural motif utilizing the logistic regression intialized parameters gains some improvements over the naïve version, but is vastly overshadowed by the performance of the logistic regression control.

### 5.1.5 Discussion

Based on the Figure 5.1 it initially appeared that the optimization strategy of the modulated structural motif parameters was successful. However the poor performance of the optimized model displayed in Figure 5.2 contradicts that assessment. This may suggest that the cross entropy loss function may not have been appropriate for this application. Interestingly the logistic regression initialized modulated strum *did* show an improvement, indicating that this form of parameter selection may be sufficient. This has the benefit of being much faster, both due to the reduced complexity of the problem, and the obviation of the need to sample multiple possible parameter sets.

Despite the success of the modulated structural motif in improving the cell type specific predictive power of the structural motif model, its performance was still overshadowed by a simple logistic regression model. This model still requires the identification of putative binding sites through the use of a structural motif or similar. We conclude that traditional DNase footprinting methods used in conjunction with

structural motifs is more effective than incorporating the DNase information directly into the structural motif itself.

# 6. Discussion

Sequence based DNA motifs have long been used to represent the genomic sites that can be bound by TFs. These abstractions of transcription factor DNA binding preferences are intuitive, computationally lightweight, and have been shown to be quite robust. Shape based models have the potential to better capture the binding preferences by more directly modelling the protein-DNA interaction.

The Structural Motif uses estimates of of DNA shape features derived from sequence-to-shape methods, such as the Dinucleotide Property Database. This method models the shape preferences at each position in a motif with a Normal distribution. By assuming the independence of positions across the binding site, novel sequences can be scored in a straightforward manner, by computing the probability at each position of observing the DNA as or more distorted from the transcription factors preferred configuration as the given shape, and multiplying those probabilities across the sites and shapes.

Using expectation maximization, we demonstrate that *de novo* Structural Motifs are discovered within the output from ChIP-seq experiments. Aligning with our hypothesis, these *de novo* motifs are highly specific to their cognate TFs, accurately distinguishing true binding sites from a mixed pool of transcription factor binding sites of other transcription factors (Average auROC = 0.77, SD = 0.14). This is reinforced by the finding that for a given sequence the high quality matches to a Structural Motif are generally quite close to the best match to the position weight

matrix for the same transcription factor (Average Distance = 20 bp, SD = 16.6). In a simple classification experiment, identifying sequences containing a true transcription factor binding site versus background sequences, Structural Motifs show a similar degree of specificity as not only the position weight matrix, but also the higher order dinucleotide weight matrix (Average auROC: PWM=0.76, DWM=0.77, StruM=0.76).

In spite of the similar level of performance in these classification experiments, overall the scores produced by the sequence and shape models were very poorly correlated. For the 229 transcription factor ChIP-seq experiments considered, the average correlation between StruM and PWM scores was only 0.06. Because of this poor correspondence, we hypothesized that these different representations of binding sites were encoding the transcription factor preferences in fundamentally different ways. This hypothesis was borne out by observing that the performance in these classification experiments could be further improved by integrating the two motif representations into a joint logistic regression model. The joint model outperformed both single models in 68% of the experiments.

Taking advantage of the differences in the types of information encoded by sequence- and shape-based models, we utilized the model agnostic MoSBAT energy score to investigate whether shape models might increase the ability to discriminate between binding sites for highly similar motifs. After first confirming that motifs derived replicate experiments for a given transcription factor were generally more similar than motifs from other transcription factors, we investigated the similarity of motifs across a variety of transcription factor families. Upon clustering of the transcription factor based on either the similarity of the position weight matrices or Structural Motifs, we determined that the Structural Motifs had a higher level of entropy (avg PWM = 1.5, avg StruM = 1.8) and lower overall $\chi^2$ distance (avg PWM = 0.22, avg StruM = 0.12) than the position weight matrix, indicating that the

clusters had a more uniform composition in terms of the transcription factor families represented there. The difference in these cluster compositions is highlighted by considering the Jaccard score between each of the position weight matrix clusters and each of the Structural Motif clusters. The maximum Jaccard score between any pair of clusters was only 0.26, indicating that the clustering was significantly different.

In the initial design of the Structural Motif model, one of the main objectives was to create a model that was not only simple and intuitive, but also extensible. The Structural Motif framework allows for inclusion of arbitrary features in addition to the shape estimates from DiProDB. We hypothesized that the inclusion of experimental data from other assays would provide the requisite context for making cell type specific predictions. We successfully incorporated DNase-seq data into the Structural Motif as a proof of concept. Contrary to our hypothesis, we did not achieve making cell type specific predictions, which is explained by nucleotide resolution DNase signals showing an extreme sequence bias, as opposed the the previously hypothesized transcription factor specific protection of bases within a binding site.

Sequence based methods have remained popular in part due not only to the robustness of the models, but also due to their relatively lightweight computational requirements. With modern high performance computing resources available even on personal laptops, more complex models are now much more feasible even for simple applications. Even so, Structural Motifs are fairly computationally lightweight, being constructed as an extension of the position weight matrix. The main increase in computational complexity comes from two sources: the increased number of parameters required to model each site in the motif, and the computation time for calculating p-values. Using the provided Python package, a Structural Motif requires a maximum of up to 48 times the number of parameters as the position weight matrix, depending on the shape features included in the model. Addition-

ally, since the p-values are computed from a Normal distribution, there are fast implementations for calculating the probabilities, and depending on the application precalculated probabilities tables can be used to provide estimated p-values. In our experience, a *de novo* motif can be calculated (with 5 random restarts) from the top 500 peaks in a high quality ChIP-seq experiment in a matter of minutes on a personal laptop. This design makes these models not only intuitive, but accessible for general use due to their simplicity and speed.

Although at present these models are intrinsically linked to the underlying DNA sequence as the shape parameters are estimated from sequence, the structural models capture a unique set of information relative to the sequence models. This difference turns out to be useful beyond just a new perspective on the interaction, as it allows for a more complete representation of the transcription factor binding preferences when sequence and shape are considered together, resulting in overall higher quality predictions from joint models. Our results imply that this increased specificity may even extend to discriminating between binding sites within the same transcription factor families, though additional work is required in this area. Higher order models may be required that can consider multiple motifs simultaneously to best address this problem.

While estimating the shape parameters from the underlying sequence does not restrict the structural models to parallel the sequence models, it does force some strict assumptions on the shape models. Namely, this method of defining shape parameters operates under the assumption that the sequences being analyzed are from naked DNA, not taking into account local distortions of DNA shape, for example as a result even of histone wrapping or tension/DNA bending introduced by the binding of other proteins nearby. For this reason we expect that Structural Motifs will perform poorly for binding sites located in regions of more complex three dimensional DNA structure. Particularly, these models are poorly suited for predicting the

binding sites of pioneer transcription factors that bind primarily to packaged DNA.

An additional limitation of Structural Motifs is that chromatin state is not natively included. This prevents “vanilla” Structural Motifs from making cell type specific predictions, as has been described. And while the model allows for integration of additional relevant features, DNase sensitivity, for example, doesn’t provide information at high enough resolution to be useful directly incorporated into the model. Other feature types may, in the future, be used in this manner, but we have not yet explored that possibility.

Finally, we have found that not all transcription factors are well represented by shape motifs. Transcription factors with known sequence preferences that operate via the direct base readout method of binding site recognition tend to be well represented with sequence based methods. In these situations models based on indirect-readout, such as Structural Motifs, tend to provide a less complete representation of the binding site than sequence methods.

Ultimately, based on our findings we recommend considering both sequence- and shape-based methods for making predictions of transcription factor binding sites. While the ease of use and widespread acceptance of the position weight matrix is great for many applications, for broad scale predictions Ensemble methods will likely produce better results. Combining shape and sequence methods allow for more confident predictions by considering a broader view of the binding preferences. Additionally, integrating both types of methods into higher order models that enable the consideration of coordination or competition between transcription factors as well as cooperative binding will further refine these predictions. Models such as DeepBind [51] and Basset [107] have demonstrated the utility of applying convolutional neural networks for this purpose.

While incorporating DNase directly into structural motifs was insufficient for making cell type predictions, it still has significant utility in identifying potential bind-

ing sites. Incorporating other feature types informing the model of the chromatin state context of a given transcription factor binding site (such as methylation or available ChIP-seq data from other factors) will lead to the most confident predictions. As better or additional sequence-to-shape estimating methods become available, such as DNAshapeR [73], these can also be included in the structural motif model to improve its overall performance by increasing the accuracy of the shape representation of the binding sites.

This work demonstrates the utility of shape based methods for representing DNA motifs. Importantly, these methods don't stand alone but are better suited to be evaluated alongside sequence methods. With additional work incorporating chromatin state information and integrating multiple models cell type specific predictions will become yet more confident, expanding our overall understanding of the regulatory landscape that contributes to healthy development, as well as providing insight into the mechanisms and treatment of genetic diseases.

Consistent with the goal of providing a model that is intuitive and accessible for general use, the Python package developed for working with Structural Motifs is freely available on GitHub, and was designed for use with Python 2.7. It can be found at https://github.com/pdeford/StructuralMotifs and installation instructions and documented is also provided there. The documentation is included in the accompanying website https://pdeford.github.io/StructuralMotifs/, and provides a description of all of the available functions, as well as examples of how to use the package.

# A. Supplementary Methods

## A.1 Training Structural Motifs

Two methods were applied to training the Structural Motifs used in this paper. These approaches differ in whether they take a sequence-centric approach to identifying the binding site. In the case where the training is directed by sequence, the kmers identified by MEME as contributing to the `meme-chip` PWM are taken as being representative of the binding site. Each one of these aligned binding site sequences is translated to 'structural space'. This is done by iteratively considering each dinucleotide in the sequence, and looking up the corresponding values in DiProDB for the features associated with the '`full`' filter mode in the StruM package (see Table A.1). If the binding site identified by the PWM is of width $k$, and there are $p$ features, the size of the feature vector for this sequence is of length $(k-1)\cdot p$. The 'Maximum Likelihood' StruM, is then computed by taking the arithmetic mean and standard deviation at each of these $(k-1)\cdot p$ position-specific features across all of the aligned binding sites.

The second method used to train Structural Motifs was an Expectation Maximization approach (algorithm described more fully below) directly on the structural representation of the training sequences. This version of the StruM was trained on the same set of sequences passed to MEME, i.e. the 100 bp surrounding the peak summit of the top 500 most enriched peaks in the ChIP experiment. In order

to speed up the training time, the Expectation Maximization algorithm was done using the ‘proteingroove’ filter mode of the StruM package (see Table A.2). This filter mode only uses 14 features related to the major and minor grooves and values derived from protein-DNA complexes. 10 random restarts were used to initialize the motif. After convergence of the models, the one with the highest likelihood was retained. This intermediate ‘proteingroove‘ StruM was used to score the training sequences, and the best scoring kmer from each sequence was extracted. These were then translated to structural space with full 96 features available in the ‘full’ mode, and the parameters derived in a maximum likelihood fashion from this set of kmers.

| Count | Feature |
|---|---|
| 1 | Bend |
| 1 | Clash Strength |
| 2 | Enthalpy |
| 2 | Entropy |
| 1 | Flexibility_shift |
| 1 | Flexibility_slide |
| 9 | Free energy |
| 1 | Major Groove Depth |
| 1 | Major Groove Distance |
| 1 | Major Groove Size |
| 1 | Major Groove Width |
| 2 | Melting Temperature |
| 1 | Minor Groove Depth |
| 1 | Minor Groove Distance |
| 1 | Minor Groove Size |
| 1 | Minor Groove Width |

| | |
|---|---|
| 1 | Persistance Length |
| 1 | Probability contacting nucleosome core |
| 1 | Propeller Twist |
| 3 | Rise |
| 2 | Rise (DNA-protein complex) |
| 1 | Rise stiffness |
| 1 | Rise_rise |
| 4 | Roll |
| 2 | Roll (DNA-protein complex) |
| 1 | Roll stiffness |
| 1 | Roll_rise |
| 1 | Roll_roll |
| 1 | Roll_shift |
| 1 | Roll_slide |
| 2 | Shift |
| 2 | Shift (DNA-protein complex) |
| 1 | Shift stiffness |
| 1 | Shift_rise |
| 1 | Shift_shift |
| 1 | Shift_slide |
| 3 | Slide |
| 2 | Slide (DNA-protein complex) |
| 1 | Slide stiffness |
| 1 | Slide_rise |
| 1 | Slide_slide |
| 4 | Stacking energy |
| 3 | Tilt |

| | |
|---|---|
| 2 | Tilt (DNA-protein complex) |
| 1 | Tilt stiffness |
| 1 | Tilt_rise |
| 1 | Tilt_roll |
| 1 | Tilt_shift |
| 1 | Tilt_slide |
| 1 | Tilt_tilt |
| 1 | Tip |
| 6 | Twist |
| 2 | Twist (DNA-protein complex) |
| 1 | Twist stiffness |
| 1 | Twist_rise |
| 1 | Twist_roll |
| 1 | Twist_shift |
| 1 | Twist_slide |
| 1 | Twist_tilt |
| 1 | Twist_twist |
| 1 | Wedge |

Table A.1: Features from the Dinucleotide Property Database used in the ‘full’ filtering mode of the StruM package. The ‘Count’ column represents how many times that feature appears, as the DiProDB table references some features multiple times, from different source in the literature.

| Feature |
|---|
| Major Groove Depth |
| Major Groove Distance |
| Major Groove Size |
| Major Groove Width |
| Minor Groove Depth |
| Minor Groove Distance |
| Minor Groove Size |
| Minor Groove Width |
| Rise (DNA-protein complex) |
| Roll (DNA-protein complex) |
| Shift (DNA-protein complex) |
| Slide (DNA-protein complex) |
| Tilt (DNA-protein complex) |
| Twist (DNA-protein complex) |

Table A.2: Features from the Dinucleotide Property Database used in the ‘proteingroove’ filtering mode of the StruM package.

## A.2 Expectation Maximization training of StruMs

Our approach to expectation maximization was modeled after the OOPS model (only one per sequence) used by MEME [88]. Due to the formulation of StruMs as a combination of normal distributions, the parameters can be estimated using a variation of a weighted average.

### A.2.1 E-step

The likelihood ($l_{ij}$) of the $j$-th position in the $i$-th sequence being the start of the binding site is taken to be the score of the StruM at that position multiplied by the likelihood of the flanking regions matching the background model ($\phi_B$):

$$l_{ij} = \prod_{n=1}^{j-1} P(v_{ij}|\phi_B) \prod_{n=j}^{j+k-1} P(v_{ij}|\phi_{i-j+1}) \prod_{n=j+k}^{N} P(v_{ij}|\phi_B)$$

The likelihoods are then normalized on a by-sequence basis to produce $M$, the matrix of expected start positions:

$$M_{ij} = \frac{l_{ij}}{\sum_{j'=1}^{m} l_{ij'}}$$

### A.2.2 M-step

The maximization step takes these likelihoods and calculates maximum likelihood values for $\mu$ and $\sigma$ for each of the $m$ position-specific features:

$$\mu_j = \sum_{i=1}^{n} \sum_{\mathsf{v}} \frac{v_{ij} \cdot M_{ij}}{\sum_i \sum_j M_{ij}}$$

$$\sigma_j = \sum_{i=1}^{n} \sum_{\mathsf{v}} \frac{(v_{ij} - \mu_j)^2 \cdot M_{ij}}{\sum_i \sum_j M_{ij} - \frac{\sum_i \sum_j M_{ij}^2}{\sum_i \sum_j M_{ij}}}$$

## A.3 Filtering Position-Specific Features

As is the case with PWMs, not all position-specific features will contribute equally to the specificity of the motif. Analagously to positions with very low information content in a PWM, position-specific features in a StruM with large values for $\sigma$ don't reveal much information about the binding site, and can tolerate high amounts of variability at that site.

These non-specific features may contribute to the noisiness of the signal without appreciably contributing to the specificity of the motif. By filtering out non-specific features, not only might the signal to noise ration be improved, but also reduce the time required to score a kmer with the motif.

Two methods of identifying non-specific features in the StruM were explored. The first was simply based on the value of $\sigma$. Large values of $\sigma$ by definition correspond to large amounts of variation in the training data for that feature. Excluding position-specific features with a value for $\sigma$ greater than some threshold would limit the score for a sequence to only derive from specific features.

The second method considered was to use a Fisher score, based on the Fisher Linear Discriminant. This strategy requires a negative set, and for each feature compares the difference in mean values for the positive and negative sets, to the difference in variability observed for the positive and negative sets for that feature. More specifically, the Fisher score $V$ for the $i$-th position-specific feature can be computed as:

$$V_i = \frac{(\mu_{i+} - \mu_{i-})^2}{\sigma_{i+}^2 + \sigma_{i-}^2} \tag{A.1}$$

A larger value for $V_i$ corresponds to a larger difference between the two sets, after accounting for their variability. After training the StruM, the positive set was constructed by taking the best scoring kmer from each training sequence. The

negative set was then generated by randomly shuffling each of the kmers in the positive set.

In order to automatically identify a threshold for these two methods, the position specific features were rank-ordered by each of the metrics. A univariate spline was fit to these rank-ordered values. The point of inflection in this spline was selected as the threshold. Features with values above the threshold were retained for the $\log_{10}$Fisher score (Figure A.1) and features with values below the threshold were retained for $\sigma$ (Figure A.2).

The performance of using the full StruM and each of the filtered versions was evaluated using three metrics: The Fisher score, auROC, and auPRC (Figure A.3). In general, the the filtered versions performed as well as or slightly better than the original full version of the motif (data not shown). We therefore elected to use filtered-StruMs for this analysis. Specifically, filtering on the variance threshold was used as it does not require any sort of negative set to compute.

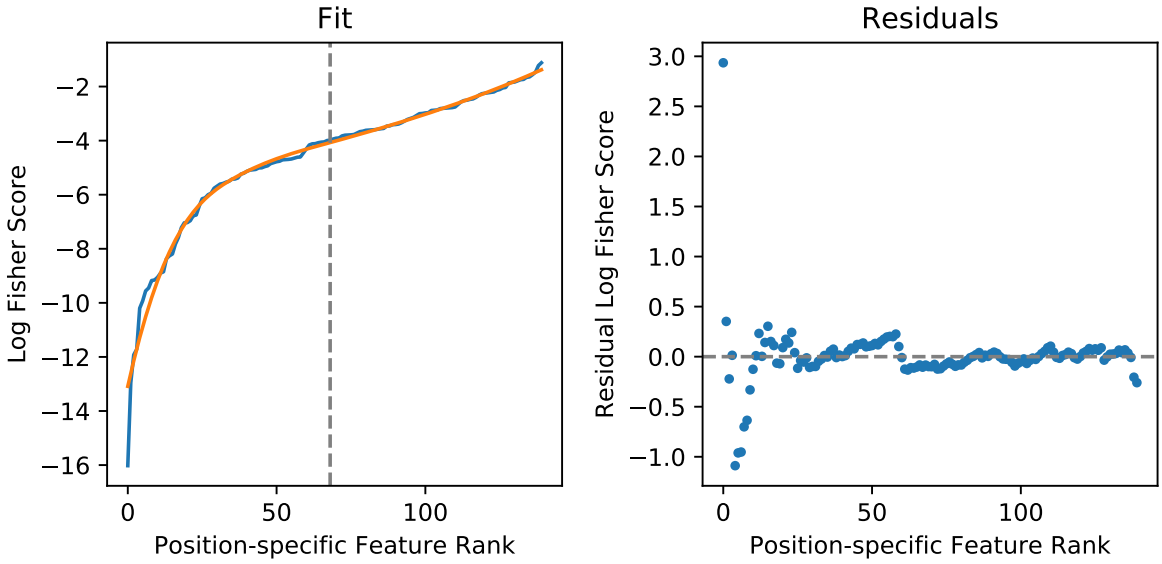



Figure A.1: Position-specific StruM features were rank ordered by their log Fisher score determined from the binding sites, and a shuffled set of sequences. A univariate spline was fit (orange line) and the point of inflection determined as the threshold (vertical grey line). The residuals from the fit are on the right hand side.

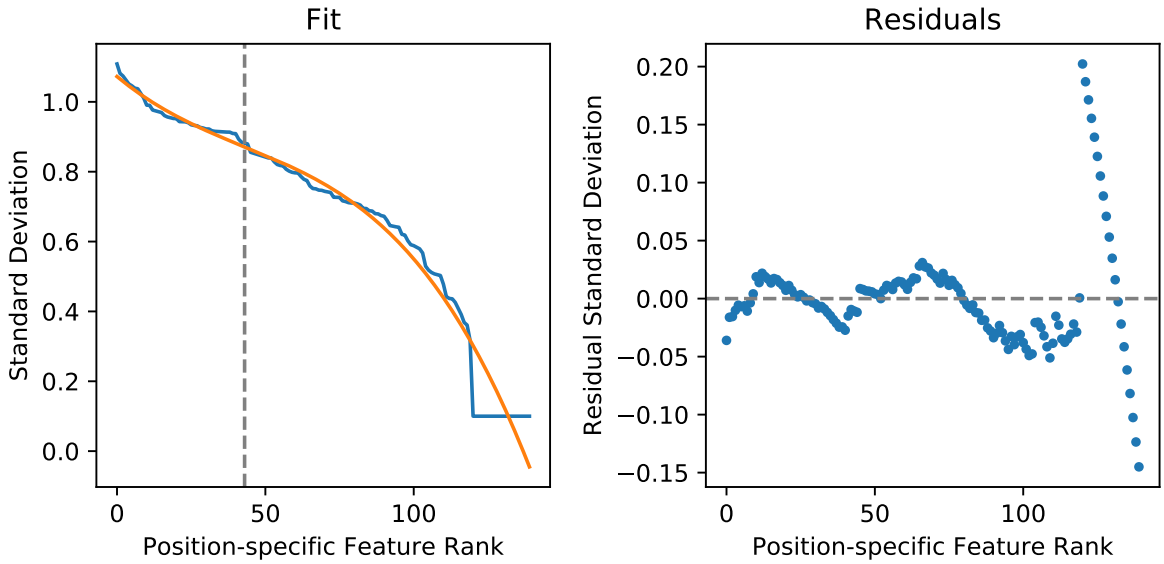



Figure A.2: Position-specific StruM features were rank ordered by their value for $\sigma$. A univariate spline was fit (orange line) and the point of inflection determined as the threshold (vertical grey line). The residuals from the fit are on the right hand side.

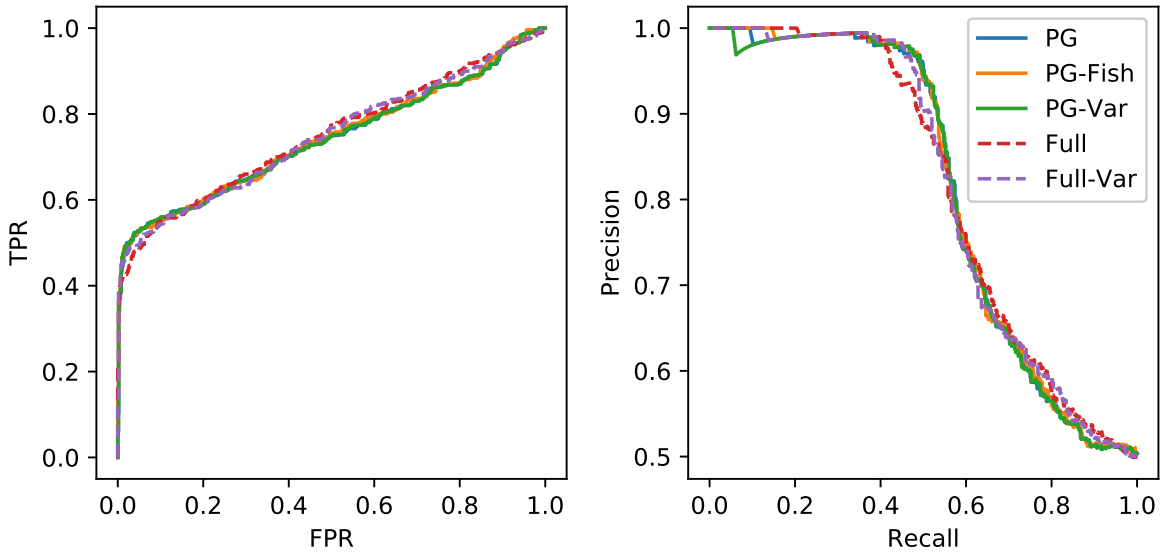



Figure A.3: The performance of the the full StruM compared to the filtered version.

## A.4 Dataset accessions

The following table describes the accessions for the datasets used in Chapter 3. The columns specify which of the analyses each dataset was included in, as follows:

1. Target: The target of the ChIP-seq experiment.

2. Accession: The ENCODE accession number for the dataset used.

3. coefficients: Whether the experiment was included in the calculations for assessing the relative size of the coefficients in the logistic regression model.

4. alignments: Whether the experiment was included in the calculations for aligning the PWM and StruM motifs.

5. classification: Whether the experiment was included in the calculations for the performance of the motif representations in classifying sequences as coming from the TF vs. background.

6. correlation: Whether the experiment was included in the calculations for correlations for scores between the PWMs and StruMs.

7. positions: Whether the experiment was included in the calculations for distances between best PWM and StruM matches in a sequence.

8. specificities: Whether the experiment was included in the calculations for the performance of motifs in classifying sequences as coming from the cognate TF or other sequences.

| Target | Accession | coefficients | alignments | classification | correlation | positions | specificities |
|---|---|---|---|---|---|---|---|
| 3xFLAG-ATF1 | ENCFF948LKX | Y | Y | Y | Y | Y | N |
| 3xFLAG-PBX2 | ENCFF003PSH | Y | Y | Y | Y | Y | N |
| AFF1 | ENCFF466GTG | Y | Y | Y | Y | Y | N |
| AFF1 | ENCFF579KSC | Y | Y | Y | Y | Y | N |
| ARHGAP35 | ENCFF991AML | Y | Y | Y | Y | Y | N |
| ARID1B | ENCFF676SJQ | Y | Y | Y | Y | Y | Y |
| ARID2 | ENCFF056IYP | Y | Y | Y | Y | Y | Y |
| ARID3A | ENCFF974QZG | Y | Y | Y | Y | Y | Y |
| ARNT | ENCFF032MDN | Y | Y | Y | Y | Y | Y |
| ARNT | ENCFF521JOW | Y | Y | Y | Y | Y | Y |
| ARNT | ENCFF934MUQ | Y | Y | Y | Y | Y | Y |
| ASH1L | ENCFF701QFG | Y | Y | Y | Y | Y | N |
| ATF2 | ENCFF525YRJ | Y | Y | Y | Y | Y | Y |
| ATF3 | ENCFF737ZGG | Y | Y | Y | Y | Y | Y |
| ATF7 | ENCFF280UZU | Y | Y | Y | Y | Y | Y |
| BCLAF1 | ENCFF145KNX | Y | Y | Y | Y | Y | N |
| BCLAF1 | ENCFF401FXT | Y | Y | Y | Y | Y | N |
| BCOR | ENCFF467YYR | Y | Y | Y | Y | Y | N |
| BHLHE40 | ENCFF432TSX | Y | Y | Y | Y | Y | N |
| BMI1 | ENCFF259NPX | Y | Y | Y | Y | Y | N |
| BRD4 | ENCFF868NFS | Y | Y | Y | Y | N | N |
| BRD9 | ENCFF148RRJ | Y | Y | Y | Y | Y | N |
| CBFA2T2 | ENCFF339YXQ | Y | Y | Y | Y | Y | N |
| CBFA2T3 | ENCFF202JYJ | Y | Y | Y | Y | Y | N |
| CC2D1A | ENCFF209YTE | Y | Y | Y | Y | Y | N |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| CDC5L | ENCFF820ZHR | Y | Y | Y | Y | Y | Y |
| CEBPB | ENCFF819GCB | Y | Y | Y | Y | Y | Y |
| CEBPZ | ENCFF803JOE | Y | Y | Y | Y | Y | N |
| CHAMP1 | ENCFF131TQV | Y | Y | Y | Y | Y | N |
| CHAMP1 | ENCFF201PDX | Y | Y | Y | Y | Y | N |
| CREB3L1 | ENCFF719NIX | Y | Y | Y | Y | Y | Y |
| CREM | ENCFF744NXL | Y | Y | Y | Y | Y | Y |
| CTBP1 | ENCFF125XVE | Y | Y | Y | Y | Y | N |
| CTCF | ENCFF559HEE | Y | Y | Y | Y | Y | Y |
| CTCF | ENCFF681OMH | Y | Y | Y | Y | Y | Y |
| CTCFL | ENCFF985NLY | Y | Y | Y | Y | Y | Y |
| CUX1 | ENCFF010SPN | Y | Y | Y | Y | Y | Y |
| DACH1 | ENCFF318UQW | Y | Y | Y | Y | Y | N |
| DEAF1 | ENCFF633OQQ | Y | Y | Y | Y | Y | N |
| DNMT1 | ENCFF997JHM | Y | Y | Y | Y | Y | Y |
| DPF2 | ENCFF206HJJ | Y | Y | Y | Y | Y | N |
| DPF2 | ENCFF663KFV | Y | Y | Y | Y | Y | N |
| E2F1 | ENCFF414ZZX | Y | Y | Y | Y | Y | Y |
| E2F1 | ENCFF666FOX | Y | Y | Y | Y | Y | Y |
| E2F6 | ENCFF142VQA | Y | Y | Y | Y | Y | Y |
| E2F7 | ENCFF987GXS | Y | Y | Y | Y | Y | Y |
| E2F8 | ENCFF749OIM | Y | Y | Y | Y | Y | Y |
| E4F1 | ENCFF211YKN | Y | Y | Y | Y | Y | Y |
| eGFP-ADNP | ENCFF219ZRU | Y | Y | Y | Y | Y | N |
| eGFP-ATF1 | ENCFF884CMO | Y | Y | Y | Y | Y | N |
| eGFP-ATF3 | ENCFF820BLL | Y | Y | Y | Y | Y | N |
| eGFP-BACH1 | ENCFF476YGX | Y | Y | Y | Y | Y | N |
| eGFP-CEBPB | ENCFF231HJU | Y | Y | Y | Y | Y | N |
| eGFP-CEBPG | ENCFF674WGB | Y | Y | Y | Y | Y | N |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| eGFP-CREB3 | ENCFF459GME | Y | Y | Y | Y | Y | N |
| eGFP-CUX1 | ENCFF193JMA | Y | Y | Y | Y | Y | N |
| eGFP-ELF1 | ENCFF140VIF | Y | Y | Y | Y | Y | N |
| eGFP-ETS2 | ENCFF392QMP | Y | Y | Y | Y | Y | N |
| eGFP-ETV1 | ENCFF332WRL | Y | Y | Y | Y | Y | N |
| eGFP-FOXJ2 | ENCFF797AAQ | Y | Y | Y | Y | Y | N |
| eGFP-GATA2 | ENCFF879REQ | Y | Y | Y | Y | Y | N |
| eGFP-GTF2A2 | ENCFF915SSM | Y | Y | Y | Y | Y | N |
| eGFP-GTF2E2 | ENCFF394WVZ | Y | Y | Y | Y | Y | N |
| eGFP-HDAC8 | ENCFF531QQR | Y | Y | Y | Y | Y | N |
| eGFP-HDAC8 | ENCFF621HTW | Y | Y | Y | Y | N | N |
| eGFP-HINFP | ENCFF413OIG | Y | Y | Y | Y | Y | N |
| eGFP-ID3 | ENCFF792DSI | Y | Y | Y | Y | Y | N |
| eGFP-IRF1 | ENCFF372XLP | Y | Y | Y | Y | Y | N |
| eGFP-IRF9 | ENCFF850QMO | Y | Y | Y | Y | Y | N |
| eGFP-KLF13 | ENCFF387ZED | Y | Y | Y | Y | Y | N |
| eGFP-KLF1 | ENCFF581HPR | Y | Y | Y | Y | Y | N |
| eGFP-MAFG | ENCFF921SVE | Y | Y | Y | Y | Y | N |
| eGFP-MEF2D | ENCFF776UTN | Y | Y | Y | Y | Y | N |
| eGFP-NFE2 | ENCFF759YKT | Y | Y | Y | Y | Y | N |
| eGFP-NFE2L1 | ENCFF573QWP | Y | Y | Y | Y | Y | N |
| eGFP-NFE2L1 | ENCFF625IYZ | Y | Y | Y | Y | Y | N |
| eGFP-NR2C1 | ENCFF633PGJ | Y | Y | Y | Y | Y | N |
| eGFP-NR2C2 | ENCFF978GHB | Y | Y | Y | Y | Y | N |
| eGFP-NR4A1 | ENCFF837QOK | Y | Y | Y | Y | Y | N |
| eGFP-PTTG1 | ENCFF320ZVD | Y | Y | Y | Y | Y | N |
| eGFP-RELA | ENCFF939ETO | Y | Y | Y | Y | Y | N |
| eGFP-TAF7 | ENCFF511GKD | Y | Y | Y | Y | Y | N |
| eGFP-TEAD2 | ENCFF006PAK | Y | Y | Y | Y | Y | N |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| eGFP-TFDP1 | ENCFF179TYN | Y | Y | Y | Y | Y | N |
| eGFP-TSC22D4 | ENCFF494TFV | Y | Y | Y | Y | Y | N |
| eGFP-VEZF1 | ENCFF053OWV | Y | Y | Y | Y | Y | N |
| eGFP-ZBTB11 | ENCFF624CZP | Y | Y | Y | Y | Y | N |
| eGFP-ZBTB40 | ENCFF413TRG | Y | Y | Y | Y | Y | N |
| eGFP-ZFX | ENCFF575JXW | Y | Y | Y | Y | Y | N |
| eGFP-ZKSCAN8 | ENCFF115ZHW | Y | Y | Y | Y | Y | N |
| eGFP-ZNF175 | ENCFF197LEN | Y | Y | Y | Y | Y | N |
| eGFP-ZNF24 | ENCFF048DTR | Y | Y | Y | Y | Y | N |
| eGFP-ZNF354B | ENCFF504AID | Y | Y | Y | Y | Y | N |
| eGFP-ZNF395 | ENCFF632AQL | Y | Y | Y | Y | Y | N |
| eGFP-ZNF507 | ENCFF610XQH | Y | Y | Y | Y | Y | N |
| eGFP-ZNF512 | ENCFF617CTX | Y | Y | Y | Y | Y | N |
| eGFP-ZNF584 | ENCFF498DWU | Y | Y | Y | Y | Y | N |
| eGFP-ZNF589 | ENCFF433UZU | Y | Y | Y | Y | Y | N |
| eGFP-ZNF639 | ENCFF644EZR | Y | Y | Y | Y | Y | N |
| eGFP-ZNF644 | ENCFF377ZLR | Y | Y | Y | Y | Y | N |
| eGFP-ZNF740 | ENCFF583QKD | Y | Y | Y | Y | Y | N |
| eGFP-ZNF740 | ENCFF669BPX | Y | Y | Y | Y | Y | N |
| eGFP-ZNF83 | ENCFF755MMS | Y | Y | Y | Y | Y | N |
| EGR1 | ENCFF529GVQ | Y | Y | Y | Y | Y | Y |
| EGR1 | ENCFF630ANY | Y | Y | Y | Y | Y | Y |
| ELF1 | ENCFF067ZUO | Y | Y | Y | Y | Y | Y |
| ELF1 | ENCFF368HEW | Y | Y | Y | Y | Y | Y |
| ELF4 | ENCFF976WHF | Y | Y | Y | Y | Y | Y |
| ELK1 | ENCFF519LUE | Y | Y | Y | Y | Y | Y |
| EP300 | ENCFF821GNB | Y | Y | Y | Y | Y | N |
| ESRRA | ENCFF200PMR | Y | Y | Y | Y | Y | N |
| ETV6 | ENCFF095DJV | Y | Y | Y | Y | Y | Y |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ETV6 | ENCFF808CFK | Y | Y | Y | Y | Y | Y |
| EWSR1 | ENCFF414CVQ | Y | Y | Y | Y | N | N |
| FOXA1 | ENCFF547KFA | Y | Y | Y | Y | Y | Y |
| FOXK2 | ENCFF540IYS | Y | Y | Y | Y | Y | Y |
| FOXK2 | ENCFF960WIT | Y | Y | Y | Y | Y | Y |
| FOXM1 | ENCFF869BQV | Y | Y | Y | Y | Y | Y |
| FUS | ENCFF137NMV | Y | Y | Y | Y | Y | N |
| GABPA | ENCFF678KYI | Y | Y | Y | Y | Y | Y |
| GATA1 | ENCFF178NBS | Y | Y | Y | Y | Y | Y |
| GATA1 | ENCFF715NLX | Y | Y | Y | Y | Y | Y |
| GATA2 | ENCFF727SRR | Y | Y | Y | Y | Y | Y |
| GATAD2A | ENCFF987JFX | Y | Y | Y | Y | Y | Y |
| GATAD2B | ENCFF540MNX | Y | Y | Y | Y | Y | Y |
| GMEB1 | ENCFF815MOP | Y | Y | Y | Y | Y | N |
| GTF2F1 | ENCFF334QGA | Y | Y | Y | Y | Y | N |
| GTF2F1 | ENCFF449AHL | Y | Y | Y | Y | Y | N |
| GTF2F1 | ENCFF988FFD | Y | Y | Y | Y | Y | N |
| HCFC1 | ENCFF780MBM | Y | Y | Y | Y | Y | N |
| HDAC1 | ENCFF130EPK | Y | Y | Y | Y | Y | N |
| HDAC1 | ENCFF652JEE | Y | Y | Y | Y | Y | N |
| HDAC1 | ENCFF925QJY | Y | Y | Y | Y | Y | N |
| HDAC1 | ENCFF951GMF | Y | Y | Y | Y | Y | N |
| HDAC2 | ENCFF458TCO | Y | Y | Y | Y | Y | N |
| HDAC2 | ENCFF631IAC | Y | Y | Y | Y | Y | N |
| HDAC2 | ENCFF713HRG | Y | Y | Y | Y | Y | N |
| HDAC2 | ENCFF809VBW | Y | Y | Y | Y | Y | N |
| HDAC3 | ENCFF765UDL | Y | Y | Y | Y | Y | N |
| HDAC6 | ENCFF025HHM | Y | Y | Y | Y | Y | N |
| HES1 | ENCFF419FNX | Y | Y | Y | Y | Y | Y |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| HMBOX1 | ENCFF829KBM | Y | Y | Y | Y | Y | Y |
| HNRNPK | ENCFF936IJD | Y | Y | Y | Y | Y | N |
| HNRNPL | ENCFF650VQU | Y | Y | Y | Y | Y | N |
| IKZF1 | ENCFF570TDY | Y | Y | Y | Y | Y | Y |
| IKZF1 | ENCFF766NVP | Y | Y | Y | Y | Y | Y |
| IRF1 | ENCFF256GYM | Y | Y | Y | Y | Y | Y |
| IRF1 | ENCFF546ZGF | Y | Y | Y | Y | Y | Y |
| IRF1 | ENCFF999SXR | Y | Y | Y | Y | Y | Y |
| IRF2 | ENCFF324USK | Y | Y | Y | Y | Y | Y |
| JUNB | ENCFF209OUI | Y | Y | Y | Y | Y | Y |
| JUN | ENCFF934OCU | Y | Y | Y | Y | Y | N |
| KAT8 | ENCFF567GFO | Y | Y | Y | Y | Y | N |
| KDM1A | ENCFF526UKS | Y | Y | Y | Y | Y | N |
| KDM1A | ENCFF569IJM | Y | Y | Y | Y | Y | N |
| KLF16 | ENCFF063BGI | Y | Y | Y | Y | Y | Y |
| LEF1 | ENCFF595DQM | Y | Y | Y | Y | Y | N |
| LEF1 | ENCFF866QXR | Y | Y | Y | Y | Y | N |
| MAFF | ENCFF483XQO | Y | Y | Y | Y | Y | Y |
| MAFK | ENCFF715WON | Y | Y | Y | Y | Y | Y |
| MAX | ENCFF221GAR | Y | Y | Y | Y | Y | Y |
| MAX | ENCFF679HVZ | Y | Y | Y | Y | Y | Y |
| MBD2 | ENCFF496CSN | Y | Y | Y | Y | Y | N |
| MEF2A | ENCFF883WDT | Y | Y | Y | Y | Y | N |
| MEIS2 | ENCFF678QHT | Y | Y | Y | Y | Y | Y |
| MGA | ENCFF582HCK | Y | Y | Y | Y | Y | Y |
| MIER1 | ENCFF945LXQ | Y | Y | Y | Y | Y | Y |
| MITF | ENCFF522LWR | Y | Y | Y | Y | Y | Y |
| MLLT1 | ENCFF148VLB | Y | Y | Y | Y | Y | N |
| MLLT1 | ENCFF460XQR | Y | Y | Y | Y | Y | N |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| MNT | ENCFF228QNJ | Y | Y | Y | Y | Y | Y |
| MNT | ENCFF578SWL | Y | Y | Y | Y | Y | Y |
| MTA1 | ENCFF407VAS | Y | Y | Y | Y | Y | Y |
| MTA2 | ENCFF752DRC | Y | Y | Y | Y | Y | Y |
| MTA2 | ENCFF778TXT | Y | Y | Y | Y | Y | Y |
| MTA3 | ENCFF120WQF | Y | Y | Y | Y | Y | Y |
| MXI1 | ENCFF413YYC | Y | Y | Y | Y | Y | Y |
| MYBL2 | ENCFF865IDL | Y | Y | Y | Y | Y | N |
| MYC | ENCFF390IMT | Y | Y | Y | Y | Y | N |
| MYC | ENCFF596JXE | Y | Y | Y | Y | Y | N |
| MYC | ENCFF623EST | Y | Y | Y | Y | Y | N |
| MYC | ENCFF836PBB | Y | Y | Y | Y | Y | N |
| MYC | ENCFF884VNW | Y | Y | Y | Y | Y | N |
| MYNN | ENCFF795LBM | Y | Y | Y | Y | Y | Y |
| NCOA1 | ENCFF152WSS | Y | Y | Y | Y | Y | Y |
| NCOA1 | ENCFF271LVS | Y | Y | Y | Y | Y | Y |
| NCOA1 | ENCFF939JEH | Y | Y | Y | Y | Y | Y |
| NCOA2 | ENCFF082PCH | Y | Y | Y | Y | Y | Y |
| NCOA2 | ENCFF550NJW | Y | Y | Y | Y | Y | Y |
| NCOA4 | ENCFF074DBO | Y | Y | Y | Y | Y | N |
| NCOA6 | ENCFF758SJG | Y | Y | Y | Y | Y | N |
| NCOR1 | ENCFF062BMU | Y | Y | Y | Y | Y | Y |
| NCOR1 | ENCFF098XOZ | Y | Y | Y | Y | Y | Y |
| NCOR1 | ENCFF691VAI | Y | Y | Y | Y | Y | Y |
| NEUROD1 | ENCFF852XSD | Y | Y | Y | Y | Y | Y |
| NFATC3 | ENCFF185YRG | Y | Y | Y | Y | Y | Y |
| NFATC3 | ENCFF358IAH | Y | Y | Y | Y | Y | Y |
| NFE2 | ENCFF496KKT | Y | Y | Y | Y | Y | Y |
| NFIC | ENCFF072WWL | Y | Y | Y | Y | Y | N |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| NFRKB | ENCFF220IME | Y | Y | Y | Y | Y | N |
| NFRKB | ENCFF739UMP | Y | Y | Y | Y | Y | N |
| NFXL1 | ENCFF015PTS | Y | Y | Y | Y | Y | N |
| NFYB | ENCFF561PGE | Y | Y | Y | Y | Y | N |
| NONO | ENCFF219YST | Y | Y | Y | Y | Y | N |
| NONO | ENCFF591UOR | Y | Y | Y | Y | Y | N |
| NR2C1 | ENCFF297OMH | Y | Y | Y | Y | Y | Y |
| NR2C2 | ENCFF541SHT | Y | Y | Y | Y | Y | Y |
| NR2F1 | ENCFF078SJN | Y | Y | Y | Y | Y | Y |
| NR2F2 | ENCFF823SRC | Y | Y | Y | Y | Y | Y |
| NR2F6 | ENCFF510ZUJ | Y | Y | Y | Y | Y | Y |
| NR3C1 | ENCFF091PDT | Y | Y | Y | Y | Y | Y |
| NR3C1 | ENCFF571LPJ | Y | Y | Y | Y | Y | Y |
| NRF1 | ENCFF450JCL | Y | Y | Y | Y | Y | N |
| NRF1 | ENCFF836FKF | Y | Y | Y | Y | Y | N |
| NUFIP1 | ENCFF517GZV | Y | Y | Y | Y | Y | N |
| PCBP1 | ENCFF632GIY | Y | Y | Y | Y | Y | N |
| PCBP2 | ENCFF970WJK | Y | Y | Y | Y | Y | N |
| PHB2 | ENCFF505XYY | Y | Y | Y | Y | Y | N |
| PHF20 | ENCFF128DRH | Y | Y | Y | Y | Y | N |
| PHF21A | ENCFF894CCA | Y | Y | Y | Y | Y | N |
| PKNOX1 | ENCFF853VOT | Y | Y | Y | Y | Y | Y |
| PML | ENCFF070CZW | Y | Y | Y | Y | Y | N |
| POLR2A | ENCFF231OJM | Y | Y | Y | Y | Y | N |
| POLR2A | ENCFF275TFD | Y | Y | Y | Y | Y | N |
| POLR2A | ENCFF278KVO | Y | Y | Y | Y | Y | N |
| POLR2A | ENCFF286HZT | Y | Y | Y | Y | Y | N |
| POLR2A | ENCFF772UWM | Y | Y | Y | Y | Y | N |
| POLR2A | ENCFF845UBO | Y | Y | Y | Y | Y | N |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| POLR2A | ENCFF998LIN | Y | Y | Y | Y | Y | N |
| POLR2B | ENCFF810SHF | Y | Y | Y | Y | Y | N |
| POU5F1 | ENCFF577BMW | Y | Y | Y | Y | Y | Y |
| PRDM10 | ENCFF548RXT | Y | Y | Y | Y | Y | Y |
| RB1 | ENCFF400NKF | Y | Y | Y | Y | Y | N |
| RBM22 | ENCFF301USB | Y | Y | Y | Y | Y | N |
| RCOR1 | ENCFF796XFQ | Y | Y | Y | Y | Y | Y |
| REST | ENCFF120MVT | Y | Y | Y | Y | Y | Y |
| REST | ENCFF603SNP | Y | Y | Y | Y | Y | Y |
| RFX1 | ENCFF611GXL | Y | Y | Y | Y | Y | Y |
| RFX1 | ENCFF934JXG | Y | Y | Y | Y | Y | Y |
| RFX5 | ENCFF581CEN | Y | Y | Y | Y | Y | Y |
| RLF | ENCFF569QYK | Y | Y | Y | Y | Y | Y |
| RNF2 | ENCFF321RTR | Y | Y | Y | Y | Y | N |
| RNF2 | ENCFF513CNT | Y | Y | Y | Y | Y | N |
| RNF2 | ENCFF697DZC | Y | Y | Y | Y | Y | N |
| RNF2 | ENCFF885TLC | Y | Y | Y | Y | Y | N |
| RUNX1 | ENCFF168KBY | Y | Y | Y | Y | Y | Y |
| RUNX1 | ENCFF259VDF | Y | Y | Y | Y | Y | Y |
| SAFB | ENCFF189QCJ | Y | Y | Y | Y | Y | N |
| SETDB1 | ENCFF865APC | Y | Y | Y | Y | Y | N |
| SIN3A | ENCFF210HYN | Y | Y | Y | Y | Y | N |
| SIN3A | ENCFF435GRA | Y | Y | Y | Y | Y | N |
| SIN3B | ENCFF004IKJ | Y | Y | Y | Y | Y | N |
| SIX5 | ENCFF615QOD | Y | Y | Y | Y | Y | Y |
| SKIL | ENCFF048GFY | Y | Y | Y | Y | Y | N |
| SMAD1 | ENCFF388JWW | Y | Y | Y | Y | Y | N |
| SMAD5 | ENCFF410JJC | Y | Y | Y | Y | Y | N |
| SMARCA4 | ENCFF171AYO | Y | Y | Y | Y | Y | N |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| SMARCA4 | ENCFF251WLL | Y | Y | Y | Y | Y | N |
| SMARCA4 | ENCFF670PUR | Y | Y | Y | Y | Y | N |
| SMARCA5 | ENCFF989TYW | Y | Y | Y | Y | Y | Y |
| SMARCB1 | ENCFF056FNQ | Y | Y | Y | Y | Y | N |
| SMARCC2 | ENCFF287WLC | Y | Y | Y | Y | Y | Y |
| SMARCE1 | ENCFF811SJN | Y | Y | Y | Y | Y | N |
| SMC3 | ENCFF483CZB | Y | Y | Y | Y | Y | N |
| SP1 | ENCFF321TMN | Y | Y | Y | Y | Y | Y |
| SPI1 | ENCFF387FGQ | Y | Y | Y | Y | Y | Y |
| SREBF1 | ENCFF318KNL | Y | Y | Y | Y | Y | Y |
| STAT1 | ENCFF270FXG | Y | Y | Y | Y | Y | Y |
| STAT1 | ENCFF595BKT | Y | Y | Y | Y | Y | Y |
| STAT1 | ENCFF964OWK | Y | Y | Y | Y | Y | Y |
| STAT2 | ENCFF273MGI | Y | Y | Y | Y | Y | Y |
| STAT5A | ENCFF126MOY | Y | Y | Y | Y | Y | Y |
| SUZ12 | ENCFF019XTP | Y | Y | Y | Y | Y | N |
| SUZ12 | ENCFF319JHV | Y | Y | Y | Y | Y | N |
| TAF1 | ENCFF052NLP | Y | Y | Y | Y | Y | N |
| TAL1 | ENCFF519DOC | Y | Y | Y | Y | Y | Y |
| TAL1 | ENCFF661CJD | Y | Y | Y | Y | Y | Y |
| TARDBP | ENCFF059VEO | Y | Y | Y | Y | Y | N |
| TARDBP | ENCFF063ILV | Y | Y | Y | Y | Y | N |
| TARDBP | ENCFF235UZG | Y | Y | Y | Y | Y | N |
| TBP | ENCFF159SBO | Y | Y | Y | Y | Y | N |
| TCF12 | ENCFF139SRI | Y | Y | Y | Y | Y | N |
| TCF12 | ENCFF766PSK | Y | Y | Y | Y | Y | N |
| TCF7 | ENCFF850HGO | Y | Y | Y | Y | Y | N |
| TCF7L2 | ENCFF205MMX | Y | Y | Y | Y | Y | N |
| TEAD4 | ENCFF857RDC | Y | Y | Y | Y | Y | Y |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| THAP1 | ENCFF981YOM | Y | Y | Y | Y | Y | Y |
| THRA | ENCFF752OUI | Y | Y | Y | Y | Y | N |
| TRIM24 | ENCFF608XIU | Y | Y | Y | Y | Y | N |
| TRIM24 | ENCFF647EPQ | Y | Y | Y | Y | Y | N |
| TRIM28 | ENCFF016PZU | Y | Y | Y | Y | Y | N |
| TRIM28 | ENCFF600QVD | Y | Y | Y | Y | Y | N |
| TRIM28 | ENCFF906BNB | Y | Y | Y | Y | Y | N |
| U2AF1 | ENCFF342ZVG | Y | Y | Y | Y | Y | N |
| UBTF | ENCFF005KSQ | Y | Y | Y | Y | Y | N |
| UBTF | ENCFF613NZO | Y | Y | Y | Y | Y | N |
| YBX3 | ENCFF300XAN | Y | Y | Y | Y | Y | N |
| YY1 | ENCFF049NZU | Y | Y | Y | Y | Y | Y |
| YY1 | ENCFF100JMN | Y | Y | Y | Y | Y | Y |
| YY1 | ENCFF291MOZ | Y | Y | Y | Y | Y | Y |
| ZBED1 | ENCFF373FBG | Y | Y | Y | Y | Y | Y |
| ZBTB2 | ENCFF593XRH | Y | Y | Y | Y | Y | Y |
| ZBTB33 | ENCFF681IOP | Y | Y | Y | Y | Y | Y |
| ZBTB40 | ENCFF593VVO | Y | Y | Y | Y | Y | Y |
| ZBTB5 | ENCFF414CUZ | Y | Y | Y | Y | Y | Y |
| ZBTB5 | ENCFF654HDK | Y | Y | Y | Y | Y | Y |
| ZBTB7A | ENCFF345YFV | Y | Y | Y | Y | Y | Y |
| ZBTB8A | ENCFF383KZY | Y | Y | Y | Y | Y | N |
| ZC3H11A | ENCFF804JPD | Y | Y | Y | Y | Y | N |
| ZEB2 | ENCFF132ZXL | Y | Y | Y | Y | Y | Y |
| ZEB2 | ENCFF568FLE | Y | Y | Y | Y | Y | Y |
| ZFP36 | ENCFF262ZUB | Y | Y | Y | Y | Y | N |
| ZFP91 | ENCFF277XFF | Y | Y | Y | Y | Y | Y |
| ZHX1 | ENCFF908BDF | Y | Y | Y | Y | Y | Y |
| ZKSCAN1 | ENCFF090QJO | Y | Y | Y | Y | Y | Y |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ZMYM3 | ENCFF988WAF | Y | Y | Y | Y | Y | N |
| ZNF143 | ENCFF480IMY | Y | Y | Y | Y | Y | Y |
| ZNF184 | ENCFF077KPD | Y | Y | Y | Y | Y | Y |
| ZNF184 | ENCFF951UWN | Y | Y | Y | Y | Y | Y |
| ZNF24 | ENCFF015AWN | Y | Y | Y | Y | Y | Y |
| ZNF24 | ENCFF016XTQ | Y | Y | Y | Y | Y | Y |
| ZNF24 | ENCFF933ZIS | Y | Y | Y | Y | Y | Y |
| ZNF274 | ENCFF521HBG | Y | Y | Y | Y | Y | Y |
| ZNF280A | ENCFF346SGM | Y | Y | Y | Y | Y | Y |
| ZNF282 | ENCFF188ZMQ | Y | Y | Y | Y | Y | Y |
| ZNF316 | ENCFF663HVK | Y | Y | Y | Y | Y | Y |
| ZNF318 | ENCFF737LSF | Y | Y | Y | Y | Y | N |
| ZNF384 | ENCFF641YZX | Y | Y | Y | Y | Y | Y |
| ZNF407 | ENCFF478DQU | Y | Y | Y | Y | Y | Y |
| ZNF407 | ENCFF571OOG | Y | Y | Y | Y | Y | Y |
| ZNF592 | ENCFF575STA | Y | Y | Y | Y | Y | Y |
| ZNF639 | ENCFF360FDG | Y | Y | Y | Y | Y | Y |
| ZNF639 | ENCFF774SSI | Y | Y | Y | Y | Y | Y |
| ZNF830 | ENCFF413PJP | Y | Y | Y | Y | Y | N |
| ZSCAN29 | ENCFF496FNH | Y | Y | Y | Y | Y | Y |
| ZSCAN29 | ENCFF609XKW | Y | Y | Y | Y | Y | Y |
| ZZZ3 | ENCFF056MQX | Y | Y | Y | Y | Y | N |

# B. Supplemental Figures

Figure B.1: *MEME-trained PWMs aligned to motifs generated from StruMs.* MEME was used to find a PWM for each ChIP experiment, and expectation maximization was concurrently used to generate a StruM. The highest scoring kmer in each training sequence was used to generate a StruM-PWM. The MEME-PWM and StruM-PWM were aligned by retaining the maximum scoring alignment of either the forward or reverse complemented MEME-PWM to the StruM-PWM. The alignment score was calculated by summing across each aligning column in the motifs, where a column score is the Pearson correlation coefficient of the aligned columns. This figure shows the motif alignments sorted by their alignment score, with the highest scoring alignments at the top left.

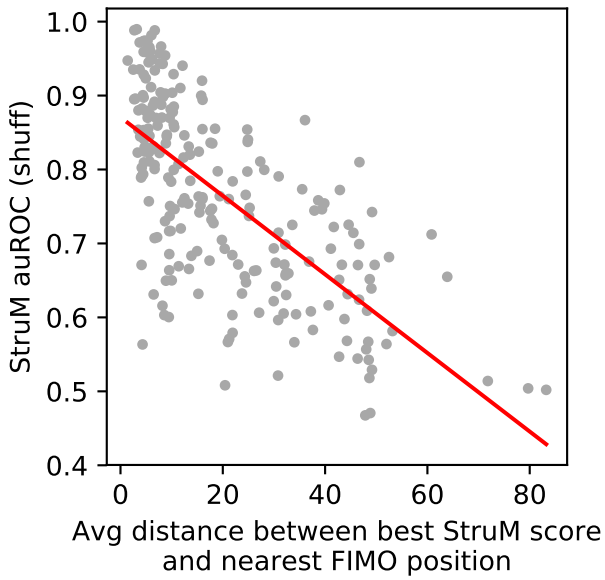



Figure B.2: The performance of the StruM is inversely related to the distance between the top StruM matches and the nearest binding site identified by FIMO using the PWM.

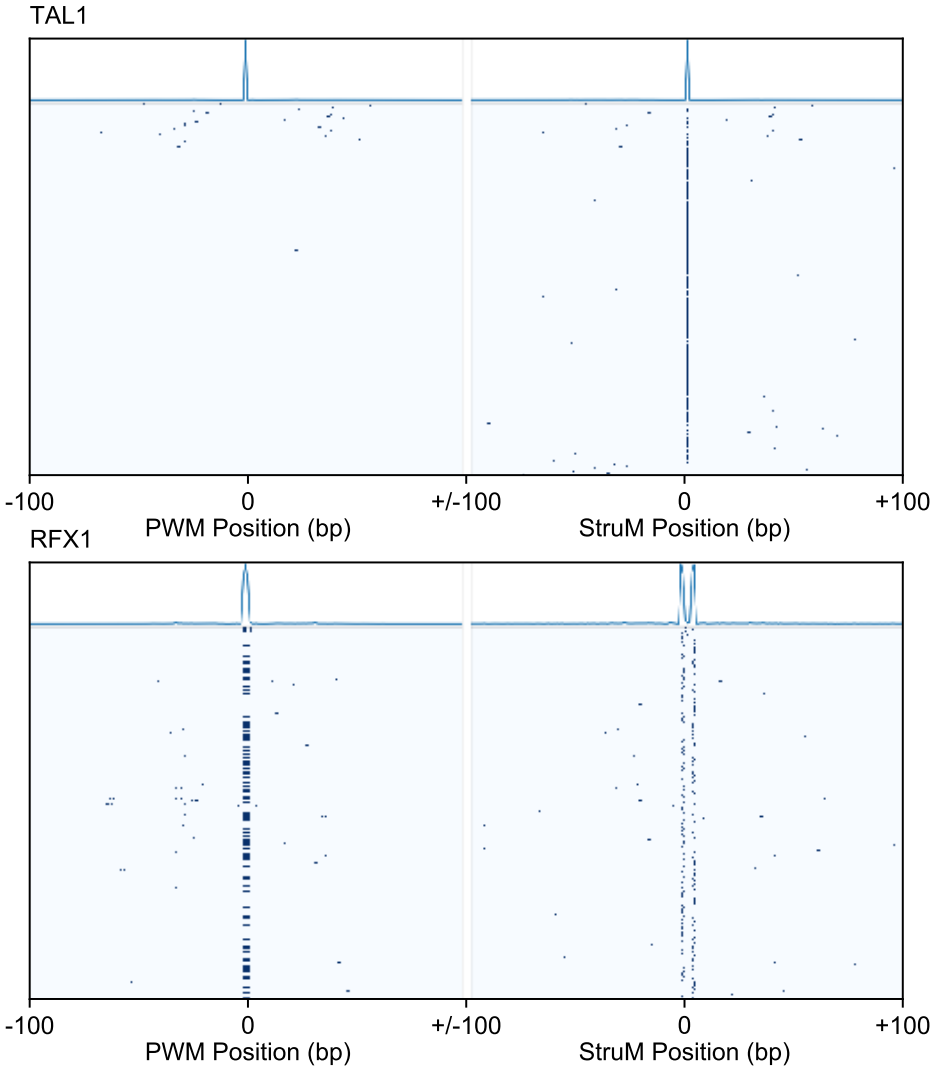



Figure B.3: Distribution of top scoring StruM positions relative to PWM matches identified by FIMO. **(top)** Example of good correlation, small average distance with a single peak. **(bottom)** Example of good correlation, at a consistent small flanking distance..

# C. References


[1] E. Bianconi, A. Piovesan, F. Facchin, A. Beraudi, R. Casadei, F. Frabetti, L. Vitale, M.C. Pelleri, S. Tassani, F. Piva, and *et al*. An estimation of the number of cells in the human body. *Ann Hum Biol.*, 40:463–471, 2013.

[2] J.B. Gurdon, R.A. Laskey, and O.R. Reeves. The developmental capacity of nuclei transplanted from keratinized skin cells of adult frogs. *J of Embryology and Exp Morphology*, 34:93–112, 1975.

[3] I. Wilmut, A.E. Schnieke, J. McWhir, A.J. Kind, and K.H.S. Campbell. Viable offspring derived from fetal and adult mammalian cells. *Nature*, 385:810–813, 1997.

[4] G. Keller. Embryonic stem cell differentiation: emergence of a new era in biology and medicine. *Genes and Dev*, 19:1129–1155, 2005.

[5] S. Ben-Tabou de Leon and E.H. Davidson. Gene regulation: gene control network in development. *Annu Rev Biophys Biomol Struct.*, 36:191–212, 2007.

[6] M. Levine and R. Tjian. Transcription regulation and animal diversity. *Nature*, 424:147–151, 2003.

[7] D.S. Latchman. Transcription factors: an overview. *The International J of Biochem & Cell Biol*, 29:1305–1312, 1997.

[8] T. Isshiki, B. Pearson, Holbrook S., and Doe C.Q. *Drosophila* neuroblasts sequentially express transcription factors which specify the temporal identify of their neuronal progeny. *Cell*, 106:511–521, 2001.

[9] J. Lewis. Autoinhibition with transcriptional delay: a simple mechanism for the zebrafish somitogenesis oscillator. *Curr Biol*, 13:1398–1408, 2003.

[10] J.E.G. Gallagher, W. Zhen, X. Rong, N. Miranda, Z. Lin, B. Dunn, H. Zhao, and M.P. Snyder. Divergence in a master variator generates distinct phenotypes and transcriptional responses. *Genes and Dev*, 28:409–421, 2014.

[11] A. Mathelier, W. Shi, and W.W. Wasserman. Identification of altered *cis*-regulatory elements in human disease. *Trends in Genetics*, 31:P67–76, 2015.

[12] M. Baker. Reproducibility crisis: blame it on the antibodies. *Nature*, 521:274–276, 2015.

[13] G.D. Stormo, T.D. Schneider, L. Gold, and A. Ehrenfeucht. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Res*, 10:2997–3011, 1982.

[14] S. Ruan and G.D. Stormo. Comparison of discriminative motif optimization using matrix and DNA shape-based models. *BMC Bioinformatics*, 19:1–8, 2018.

[15] R. Rohs, S.M. West, A.L. Sosinsky, R.S. Mann, and B. Honig. The role of DNA shape in protein-DNA recognition. *Nature*, 461:1248–1253, 2009.

[16] J.D. Watson and F.H.C. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[17] A.B. Pardee, F. Jacob, and J. Monod. The genetic control and cytoplasmic expression of "inducibility" in the synthesis of $\beta$-galactosidase by *e. coli*. *JMB*, 1:165–178, 1959.

[18] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3:318–356, 1961.

[19] M. Ptashne. Specific binding of the $\lambda$ Phage Repressor to $\lambda$ DNA. *Nature*, 214:232–234, 1967.

[20] N.C. Seeman, J.M. Rosenberg, and A. Rich. Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Nat. Acad. Sci. USA*, 73:804–808, 1976.

[21] R. Wintjens and M. Rooman. Structural classification of HTH DNA-binding domains and protein-DNA interaction modes. *JMB*, 262:294–313, 1996.

[22] J.D. Klemm, S.L. Schreiber, and G.R. Crabtree. Dimerization as a regulatory mechanism in signal transduction. *Annu Rev Immunol.*, 16:569–592, 1998.

[23] W.F. Anderson, D.H. Ohlendor, Y. Takeda, and Matthews B.W. Structure of the cro repressor from bacteriophage lambda and its interaction with dna. *Nature*, 290:754–758, 1981.

[24] D.B. McKay and T.A. Steitz. Structure of catabolite gene activator protein at 2.9 A resolution suggests binding to left-handed B-DNA. *Nature*, 290:744–749, 1981.

[25] W.H. Landschulz, P.F. Johnson, and S.L. McKnight. The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. *Science*, 240:1759–1764, 1988.

[26] Z.G. E, Y.P. Zhang, J.H. Zhou, and L. Wang. Mini review roles of the *bzip* gene family in rice. *Genet Mol Res.*, 13:3025–3036, 2014.

[27] A. Nijhawan, M. Jain, A.K. Tyagi, and J.P. Khurana. Genomic survey and gene expression analysis of the basic leucine zipper transcription factor family in rice. *Plant Phys*, 146:333–350, 2008.

[28] J. Hess, P. Angel, and M. Schorpp-Kistner. AP-1 subunits: quarrel and harmony among siblings. *J Cell Sci.*, 117:5965–5973, 2004.

[29] C. Murre, McCaw P.S., and D. Baltimore. A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. *Cell*, 56:777–783, 1989.

[30] Davis R.L., Cheng. P.F., Lassar A.B., and H. Weintraub. The myod dna binding domain contains a recognition code for muscle-specific gene activation. *Cell*, 60:733–746, 1990.

[31] J. Chaudhary and M.K. Skinner. Basic helix-loop-helix proteins can act at the E-box within the serum response element of the c-fos promoter to influence hormone-induced promoter activation in Sertoli cells. *Mol. Endocrinol.*, 13:774–786, 1999.

[32] E.B. Lewis. A gene complex controlling segmentation in *drosophila*. *Nature*, 276:565–570, 1978.

[33] W.J. Gehring, M. Affolter, and T.R. Bürglin. Homeodomain proteins. *Annu Rev Biochem*, 63:487–526, 1994.

[34] W. McGinnis and R. Krumlauf. Homeobox genes and axial patterning. *Cell*, 68:283–302, 1992.

[35] T.R. Bürglin and M. Affolter. Homeodomain proteins: an update. *Chromosoma*, 125:497–521, 2015.

[36] S.E.V Phillips. Specific $\beta$-sheet interactions. *Curr Opin Struc Biol*, 1:89–98, 1991.

[37] E. Wingender. Criteria for an updated classification of human transcription factor DNA-binding domains. *J Bioinfo Comp Bio*, 11:1340007, 2013.

[38] A. Klug. The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Ann Rev Biochem*, 79:213–231, 2010.

[39] N.P. Pavletich and C.O. Pabo. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science*, 252:809–817, 1991.

[40] U. Kühnlein, S. Linn, and W. Arber. Host specificity of DNA produced by *escherichia coli*, XI. *in vitro* modifiction of phage fd replicative form. *PNAS*, 63:556–562, 1969.

[41] H.O. Smith and K.W. Welcox. A restriction enzyme from *Hemophilus influenzae*: I. purification and general properties. *J Mol Bio*, 51:379–391, 1970.

[42] J. Hedgpeth, H.M. Goodman, and H.W. Boyer. DNA nucleotide sequence restricted by the RI endonuclease. *PNAS*, 69:3448–3452, 1972.

[43] D. Pribnow. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *PNAS*, 72:784–788, 1975.

[44] Gary D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16:16–23, 2000.

[45] Nomenclature Committee of the International Union of Biochemistry (NC-IUB). Nomenclature for incompletely specified bases in nucleic acid sequences: Recommendations 1984. *Eur J Biochem*, 150:1–5, 1985.

[46] R. Staden. Computer methods to locate signals in nucleic acid sequences. *Nuc Acids Res*, 12:505–519, 1984.

[47] P. D'haeseleer. What are DNA sequence motifs? *Nat Biotech*, 24:423–425, 2006.

[48] G.D. Storm, T.D. Schneider, and L. Gold. Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucl Acids Res*, 14:6661–6679, 1986.

[49] S.R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755–763, 1998.

[50] M.C. O'Neill. Training back-propagation neural networks to define and detect DNA-binding sites. *Nucl Acids Res*, 19:313–318, 1991.

[51] B. Alipanahi, A. Delong, M. Weirauch, and B. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33:831–838, 2015.

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

[53] S. Ruan and G.D. Stormo. Comparison of discriminative motif optimization using matrix and DNA shape-based models. *BMC Bioinformatics*, 19:1–8, 2018.

[54] T.J. Wheeler and S.R. Eddy. nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, 29:2487–2489, 2013.

[55] J.M. Rosenberg, N.C. Seeman, and *et al*. Double helix at atomic resolution. *Nature*, 243:150–154, 1973.

[56] Z. Shakked and D. Rabinovich. The effect of the base sequence on the fine structure of the DNA double helix. *Prog Biophys Mol Biol*, 47:159—195, 1986.

[57] S. Ferrari, V.R. Harley, A. Pontiggia, P.N. Goodfellow, R. Lovell-Badge, and M.E. Bianchi. SRY, like HMG1, recognizes sharp angles in DNA. *EMBO J.*, 11:4497–4506, 1992.

[58] P.M. Pil and S.J. Lippard. Specific binding of chromosomal protein HMG1 to DNA damaged by the anticancer drug cisplatin. *Science*, 256:234–237, 1992.

[59] J.D. Parvin, R.J. McCormick, P.A. Sharp, and D.E. Fisher. Pre-bending of a promoter sequence enhances affinity for the TATA-binding factor. *Nature*, 373:724–727, 1995.

[60] H. Karas, R. Knuppel, W. Schulz, H. Sklenar, and E. Wingender. Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements. *CABIOS*, 12:441–446, 1996.

[61] H. Kono and A. Sarai. Structure-based predictions of DNA target sites by regulatory proteins. *Proteins*, 35:114–131, 1999.

[62] A.V. Morozov, J.J. Havranek, D. Baker, and E.D. Siggia. Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res.*, 33:5781–5798, 2005.

[63] J.W. Valentine, A.G. Collins, and C.P. Meyer. Morphological complexity increase in metazoans. *Paleobiology*, 20:131–142, 1994.

[64] G.D. Erwin, N. Oksenberg, R.M. Truty, D. Kostka, K.K. Murphy, N. Ahituv, K.S. Pollard, and J.A. Capra. Integrating diverse datasets improves developmental enhancer prediction. *PLoS Comput Biol*, 10:e1003677, 2014.

[65] A. Sandelin, W. Alkema, P. Engstrom, W.W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, 32:D91–4., 2004.

[66] O. Fornes, J.A. Castro-Mondragon, Khan A., and *et al*. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 2019.

[67] V. Matys, O.V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A.E. Kel, and E. Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nuc Ac Res*, 34:D108–110, 2006.

[68] Timothy L. Bailey, Mikael Bodén, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37:W202–W208, 2009.

[69] S. Heinz, C. Benner, N. Spann, E. Bertolino, and *et al*. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for Macrophage and B Cell identities. *Mol Cell*, 38:576–589, 2010.

[70] M. Kitayner, H. Rozenberg, Rohs R., O. Suad, D. Rabinovich, B. Honig, and Z. Shakked. Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat. Struct. Mol. Biol.*, 17:423–429, 2010.

[71] S. Stella, D. Cascio, and R.C. Johnson. The shape of the DNA minor groove directs binding by the DNA-bending protein Fis. *Genes Dev.*, 24:814–826, 2010.

[72] H.T. Rube, C. Rastogi, J.F. Kribelbauer, and H.J. Bussemaker. A unified approach for quantifying and interpreting DNA shape readout by transcription factors. *Mol. Sys. Biol.*, 14:e7902, 2018.

[73] T. Chiu, F. Comoglio, T. Zhou, L. Yang, R. Paro, and R. Rohs. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinfo.*, 32:1211–1213, 2016.

[74] A. Mathelier, B. Xin, TP. Chiu, L. Yang, R. Rohs, and W.W. Wasserman. DNA shape features improve transcription factor binding site predictions in vivo. *Cell Sys.*, 3:278–286, 2016.

[75] Remo Rohs, Arttu Jolma, Lin Yang, Jussi Taipale, Yimeng Yin, Yaron Orenstein, and Ron Shamir. Transcription factor family□specific DNA shape readout revealed by quantitative specificity models. *Mol Syst Biol*, 13:910, 2017.

[76] Md Abul Hassan Samee, Benoit G. Bruneau, and Katherine S. Pollard. A *De Novo* shape motif discovery algorithm reveals preferences of transcription factors for DNA shape beyond sequence motifs. *Cell Systems*, 8:27–42.e6, 2019.

[77] X.-y. Li, S. MacArthur, R Bourgon, D Nix, D.A Pollard, V.N Iyer, H Aaron, L Simirenko, M Stapleton, C.L Luengo-Hendriks, and *et al*. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol.*, 6:e27, 2008.

[78] Y Zhang, W Wu, Y Cheng, D.C King, R.S Harris, J Taylor, F Chiaromonte, and R.C. Hardison. Primary sequence and epigenetic determinants of in vivo

occupancy of genomic DNA by GATA1. *Nuc. Acids Res.*, 37:7024–7038, 2009.

[79] M.J. Guertin and J.T. List. Chromatin landscape dictates HSF binding to target DNA elements. *PLoS Genet.*, 6:e1001114, 2010.

[80] G.G Yardimci, C.L Frank, G.E Crawford, and C.U. Ohler. Explicit DNAse sequence bias modeling enables high-resolution transcription factor footprint detection. *Nuc. Acids Res.*, 42:11865–11878, 2014.

[81] K. Luo and A.J. Hartemink. Using DNAse digestion data to accurately identify transcription factor binding sites. *Pac Symp Biocomput.*, pages 80–91, 2013.

[82] R.P Pique-Regi, J.F Degner, A.A Pai, D.G Gaffney, Y Gilad, and J.K. Pritchard. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, 21:447–455, 2011.

[83] S Neph, J Vierstra, A.B Stergachis, A.P Reynolds, E Haugen, B Vernot, R.E Thurman, R Sandstrom, A.K Johnson, M.T Maurano, and *et al*. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489:83–90, 2013.

[84] J. Kähärä and H. Lädesmäki. BinDNase: a discriminatory approach for transcription factor binding prediction using DNAse I hypersensitivity data. *Bioinformatics*, 31:2852–2859, 2015.

[85] J Piper, M.C Elze, P Cauchy, P.N Cockerill, C Bonifer, and S. Ott. Wellington: a novel method for the accurate identification of digital genomic footprints from DNAse-seq data. *Nuc. Acids Res.*, 41:e201, 2013.

[86] 2016. Coupling between histone conformations and DNA geometry in nucleosomes on a microscopic timescale: Atomistic insights into nucleosome functions. *J Mol Biol*, 428:221–237, Shaytan, A.K. and Armeev G.A. and Goncearenco, A. and Zhurkin, V.B. and Landsman, D. and Panchenko, A.R.

[87] ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Bio*, 9:e1001046, 2011.

[88] T.L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *ISMB*, pages 28–36, 1994.

[89] P. Machanick and T.L. Bailey. MEME-ChIP: motif analysis of large DNA datasets. *Bioinfo.*, 27:1696–1697, 2011.

[90] T. Zhou, L. Yang, Y. Lu, I. Dror, A.C.D. Machado, T. Ghane, R. Di Felice, and R. Rohs. DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nuecleic Acids Res.*, 41:W56–W62, 2013.

[91] M. Slattery, T. Zhou, L. Yang, C.D. Machad, R. Gordan, and R. Rohs. Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sci.*, 39:381–399, 2014.

[92] X. Xia. Position weight matrix, Gibbs sampler, and the associated significance tests in motif characterization and prediction. *Scientifica*, 2012:917540, 2012.

[93] M. Friedel, S. Nikolajewa, J. Suehnel, and T. Wilhelm. DiProDB: a database for dinucleotide properties. *Nucleic Acids Res.*, 37:D37–40, 2009.

[94] S.A. Lambert, A. Jolma, L.F. Campitelli, P.K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T.R. Hughes, and M.T. Weirauch. The human transcription factors. *Cell*, 172:650–665, 2018.

[95] E. Wingender, T. Schoeps, and J. Donitz. TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res*, 41:D165–170, 2013.

[96] R. Siddharthan. Dinucleotide weight matrices for predicting transcription factor binding sites: Generalizing the position weight matrix. *PLoS One*, 5:e9722, 2010.

[97] D.R. Cox. The regression analysis of binary sequences (with discussion). *J Roy Stat Soc B.*, 20:215–242., 1958.

[98] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and *et al*. Scikit-learn: Machine learning in Python. *J. of Machine Learning Res.*, 12:2825–2830, 2011.

[99] C.E. Grant, T.L. Bailey, and W.S. Noble. FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, 27:1017–1018, 2011.

[100] Z. Otwinowski, R.W. Schevitz, R.G. Zhang, C.L. Lawson, A. Joachimiak, R.Q. Marmorstein, B.F. Luisi, and P.B. Sigler. Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*, 335:321–329, 1988.

[101] A. Kumar and M. Bansal. Unveiling DNA structural features of promoters associated with various types of TSSs in prokaryotic transcriptomes and their role in gene expression. *DNA Res*, 24:25–35, 2017.

[102] D.B. Nikolov and S.K. Burley. RNA polymerase II transcription initiation: A structural view. *PNAS*, 94:15–22, 1997.

[103] Y. Suzuki, T. Tsunoda, J. Sese, H. Taira, J. Mizushima-Sugano, H. Hata, T. Ota, T. Isogai, T. Tanaka, Y. Nakamura, A. Suyama, Y. Sakaki, S. Morishita, K. Okubo, and S. Sugano. Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res*, 11:677–684, 2001.

[104] M. Suzuki and N. Yagi. DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *PNAS*, 91:12357–12361, 1994.

[105] S. Gupta, J.A. Stamatoyannopolous, T. Bailey, and W.S. Noble. Quantifying similarity between motifs. *Genome Biology*, 8:R24, 2007.

[106] S.A. Lambert, M. Albu, and T.R. Hughes. Motif comparison based on similarity of binding affinity profiles. *Bioinformatics*, 32:3504–3506, 2016.

[107] D.R. Kelley, J. Snoek, and J. Rinn. Basset: learning the regulatory code of the accessible genome with deep convolution neural networks. *Genome Res*, 26:990–999, 2016.

[108] P.D. Jackson and G. Felsenfeld. A method for mapping intranuclear protein-dna interactions and its application to a nuclease hypersensitive site. *PNAS*, 82:2296–2300, 1985.

[109] M.H. Sung, M.J. Guertin, S. Baek, and G.L. Hager. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol Cell*, 56:275–285, 2014.

[110] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[111] J.A. Granek and N.D. Clarke. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol*, 6:R87, 2005.

# D. Other Collaborative Projects

In addition to the work described in this document, I was actively engaged in open and collaborative science. This took the form of supporting colleagues in the analysis of their data, supervising computational projects, and providing both technical and domain specific expertise. A number of these projects are summarized below.

## Published Collaborations

**Natural variation in stochastic photoreceptor specification and color preference in Drosophila.**

> Anderson, C., Reiss, I., Zhou, C., Cho, A., Siddiqi, H., Mormann, B., Avelis, C.M., **DeFord, P.**, Bergland, A., Roberts, E., Taylor, J., Vasiliauskas, D., Johnston, R.J. (2017) *Elife*, doi: 10.7554/eLife.29593.

The Johnston lab investigates the mechanisms determining how cells make fate decisions, particularly through their study of the stochastic expression of the *Drosophila melanogaster* transcription factor, Spineless (Ss). Caitlin Andersen identified a genetic variant affecting the ratio of $Ss^{ON}$ to $Ss^{OFF}$ cells, due to altering the affinity for the transcription factor Klumpfuss (Klu). To provide quantitative insight on the impact of the variant on Klu binding, we analyzed high throughput SELEX-seq (Systematic evolution of ligands by exponential enrichment) to determine relative enrichments of sequence either with or without the variant of interest. The variant, known as *sin* was demonstrated to increase the affinity of the site for Klu binding.

**The insulator protein BEAF-32 is required for Hippo pathway activity in the terminal differentiation of neuronal subtypes.**

Jukam, D., Viets, K., Anderson, C., Zhou, C., **DeFord, P.**, Yan, J., Cao, J., Johnston, R.J. (2016) *Development*, doi: 10.1242/dev.134700.

In an analysis of the role of BEAF-32 in regulating Hippo signaling in determining R8 photoreceptor neuron fates, ChIP data was employed to show direct interactions between BEAF-32 and a subset of genes in the R8 regulatory network. We analyzed ChIP-chip and ChIP-seq data to identify both putative direct ChIP peaks (enrichment > 2.5 fold) and putative indirect ChIP peaks (enrichment < 2.5 fold). The peaks were found to be specific to genes related to Hippo signaling.

**A cloud-based learning environment for comparing RNA-seq aligners.**

Baskin, E., **DeFord, P.M.**, Dennis, A., Misner, I., Tan, F.J., Busby, B. (2016) *F1000R*, doi: 10.12688/f1000research.8684.1.

As part of a 2015 hackathon hosted by the National Center for Biotechnology Information, our team investigated educational environments for teaching novices about the process of RNA-sequencing, in particular the alignment step. To facilitate this, we developed a tutorial guiding users through RNA-seq alignment, and provided tools for comparing the output from different aligners. This provides users the tools and background they need to make educated decisions about which tools to use for their particular project, and encourages them to think deeply about each step of their analysis.

## Unpublished Data

### Sisterless A activity in *Drosophila melanogaster*

Collaboration with Raghav Goyal, Mark Van Doren.

Sisterless A (sisA) is a transcription factor being studied by the Van Doren lab, due to its role in activating the sex-specific promoter of the *sex lethal (Sxl)* gene in somatic tissues. We analyzed novel ChIP-seq data for sisA binding to describe the first motif identified for sisA as well as the genes sisA targets in different contexts.

### Analysis of 21U-RNA regulation by GEI-11 in *C. elegans* via ChIP-exo

Collaboration with Rebecca Tay, Charlotte Choi, John Kim.

GEI-11 has been implicated in the regulation of piRNAs in *C. elegans*. To hone in on its role, particularly via the use of alternate promoters, a ChIP-exo experiment was run to increase the resolution of the identified peaks, given the small size of targets. The results of the analysis suggested refinements to the experimental design to increase the power to detect these interactions.

### Investigating the role of DNA methylation in social insects.

Supervision of student project by Sage Corzine and Claudia Perez.

Some insects are highly social while others are notably individualized. The level of cooperativity also depends somewhat on each individuals role (picture the queen vs. the workers). DNA methylation data for queens and workers in *Apis mellifera* and *Solenopsis invicta* were compared to identify patterns corresponding with social behavior.

# E. Curriculum Vitae

Peter M. DeFord

Johns Hopkins University

3400 N Charles St

Baltimore, MD 21218

## RESEARCH INTERESTS

Mechanisms regulating gene expression. Qualities and characteristics of *cis*-regulatory elements. Transcription factor binding. Applying statistical models to high-throughput next generation sequencing datasets.

## EDUCATION

**Johns Hopkins University** Baltimore, Maryland

Ph.D. Biology 2021

Program in Cell, Molecular, Developmental Biology and Biophysics

**Boise State University** Boise, Idaho

B.S. Biology, Cell and Molecular Biology emphasis, Chemistry minor 2014

Honors Scholar, *Cum Laude*

## AWARDS AND HONORS

**Victor Corces Teaching Award** (Cell Biology, Spring) 2016

Department of Biology, Johns Hopkins University

**Owen Scholars Fellowship** 2014–2017

Department of Biology, Johns Hopkins University

**Student Research Initiative Fellow** 2014

Boise State University

**Idaho INBRE Fellow** 2013

Idaho IDeA Network of Biomedical Research Excellence

## RESEARCH EXPERIENCE AND TRAINING

**Graduate Research** Advisor: James Taylor 2015-2020

Johns Hopkins University Baltimore, Maryland

- Developed statistical models to describe transcription factor motifs with DNA shape.
- Used machine learning to model and predict features of transcriptional regulation.
- Analyzed mechanisms of protein-DNA interactions through feature reduction of statistical models.

**Undergraduate Honors Research** Advisor: Allan R. Albig 2012–2014

Boise State University Boise, Idaho

- Investigated initiation events of angiogenesis.
- Investigated integrin and Notch signaling pathway interactions through luciferase reporters in an over expression cell culture model.
- Cloned proteins of interest into a strong promoter context to investigate extracellular matrix molecule mediated control of angiogenesis.

- Analyzed the relative strengths of pro- and anti-angiogenic signals contributing to the angiogenic switch by cell migration assays through purified matrix proteins.

## TEACHING EXPERIENCE

**Johns Hopkins University** Baltimore, Maryland

- Certificate of Completion, *Teaching Academy* 2019
- Guest Lecturer, *Introduction to Scientific Computing in BME using Python, Matlab, and R* 2018
- Teaching Assistant, *Undergraduate Cell Biology Lab* 2016
- Teaching Assistant, *Undergraduate Biochemistry Lab* 2015
- Teaching Assistant, *Graduate Quantitative Biology Bootcamp* 2015–2019
- Teaching Assistant, *Graduate Quantitative Biology Lab* 2015–2019

**Cold Spring Harbor Laboratory** Cold Spring Harbor, New York

- Teaching Assistant, *Computational Genomics* 2016–2019

**Boise State University** Boise, Idaho

- Teaching Assistant, *Applied Statistics with Computers* 2012–2014

## PUBLICATIONS

- **DeFord, P.M.**, & Taylor, J. (2019) DNA shape complements sequence-based representations of transcription factor binding sites. *bioRxiv* [Preprint], doi: 10.1101/666735
- Anderson, C., Reiss, I., Zhou, C., Cho, A., Siddiqi, H., Mormann, B., Avelis, C.M., **DeFord, P.**, Bergland, A., Roberts, E., Taylor, J., Vasiliauskas, D., Johnston, R.J. (2017) Natural variation in stochastic photoreceptor specification and color preference in Drosophila. *Elife*, doi: 10.7554/eLife.29593.
- Jukam, D., Viets, K., Anderson, C., Zhou, C., **DeFord, P.**, Yan, J., Cao, J., Johnston, R.J. (2016) The insulator protein BEAF-32 is required for Hippo

pathway activity in the terminal differentiation of neuronal subtypes. *Development*, doi: 10.1242/dev.134700.

- **DeFord, P.M.**, Brown, K., Richards, R.L., King, A., Newburn, K., Westover, K., Albig, A.R. (2016) MAGP2 controls Notch via interactions with RGD binding integrins: Identification of a novel ECM-integrin-Notch signaling axis. *Exp. Cell Res.*,
  doi: 10.1016/j.yexcr.2016.01.011.
- Baskin, E., **DeFord, P.M.**, Dennis, A., Misner, I., Tan, F.J., Busby, B. (2016) A cloud-based learning environment for comparing RNA-seq aligners. *F1000R*, doi: 10.12688/f1000research.8684.1.

## PRESENTATIONS

- **DeFord, P.M.**, Taylor, J. *StruMs—A flexible and information-rich representation of DNA motifs.* Invited talk at the 2016 Biological Data Sciences Meeting, Cold Spring Harbor Laboratory, NY.
- **DeFord, P.M.** Invited talk at the Waksman Student Scholars event, Johns Hopkins University. Invited by Dr. Forrest Spencer. May 22, 2016.
- **DeFord, P.M.**, Albig, A. *Characterization of basement membrane induced endothelial cell quiescence in the presence of growth factors*. Poster presented at the 2014 Boise State University Undergraduate Research and Scholarship Conference, Boise, ID.
- **DeFord, P.M.**, Westover, K., Albig, A. *The Influence of Integrin Binding RGD domains on Notch Signaling and Angiogenesis*. Poster presented at the 2014 National Conference on Undergraduate Research, Lexington, KY.