Systematically Mapping the Epigenetic Context Dependence of Transcription Factor Binding

Judith F. Kribelbauer

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

Columbia University

2018

Systematically Mapping the Epigenetic Context Dependence of Transcription Factor Binding

Judith F. Kribelbauer

Abstract

At the core of gene regulatory networks are transcription factors (TFs) that recognize specific DNA sequences and target distinct gene sets. Characterizing the DNA binding specificity of all TFs is a prerequisite for understanding global gene regulatory logic, which in recent years has resulted in the development of high-throughput methods that probe TF specificity *in vitro* and are now routinely used to inform or interpret *in vivo* studies. Despite the broad success of such methods, several challenges remain, two of which are addressed in this thesis.

Genomic DNA can harbor different epigenetic marks that have the potential to alter TF binding, the most prominent being CpG methylation. Given the vast number of modified CpGs in the human genome and an increasing body of literature suggesting a link between epigenetic changes and genome instability, or the onset of disease such as cancer, methods that can characterize the sensitivity of TFs to DNA methylation are needed to mechanistically interpret its impact on gene expression. We developed a high-throughput *in vitro* method (EpiSELEX-seq) that probes TF binding to unmodified and modified DNA sequences in competition, resulting in high-resolution maps of TF binding preferences. We found that methylation sensitivity can vary between TFs of the the same structural family and is dependent on the position of the ^{5}mCpG within the TF binding site. The importance of our *in vitro* profiling of methylation sensitivity is demonstrated by the preference of human p53 tetramers for ^{5}mCpGs within its binding site core. This previously unknown, stabilizing effect is also detectable in p53 ChIP-seq data when comparing methylated and unmethylated sites genome-wide.

A second impediment to predicting TF binding is our limited understanding of i) how cooperative participation of a TF in different complexes can alter their binding preference, and ii) how the detailed shape of DNA aids in creating a substrate for adaptive multi-TF binding. To address these questions in detail, we studied the *in vitro* binding preferences of three *D. melanogaster* homeodomain TFs: Homothorax (Hth), Extradenticle(Exd) and one of the eight Hox proteins. *In vivo*, Hth occurs in two splice forms: with (Hth^{FL}) and without (Hth^{HM}) the DNA binding domain (DBD). Hth^{HM}-Exd itself is a Hox cofactor that has been shown to induce latent sequence specificity upon complex formation with Hox proteins. There are three possible complexes that can be formed, all potentially having specific target genes: Hth^{HM}-Exd-Hox, Hth^{FL}-Exd-Hox, and Hth^{FL}-Exd. We characterized the *in vitro* binding preferences of each of these by developing new computational approaches to analyze high-throughput SELEX-seq data. We found distinct orientation and spacing preference for Hth^{FL}-Exd-Hox, alternative recognition modes that depend on the affinity class a sequence falls into, and a strong preference for a narrow DNA minor grove near Exd's N-terminal DBD. Strikingly, this shape readout is crucial to stabilize the Hth^{HM}-Exd-Hox complex in the absence of a Hth DBD and can thus be used to distinguish Hth^{HM} from Hth^{FL}-isoform binding. Mutating the amino acids responsible for the shape readout by Exd and reinserting the engineered protein into the fly genome allowed us to classify *in vivo* binding sites based on ChIP-seq signal comparison between "shape-mutant" and wild-type Exd.

In summary, the research presented here has investigated TF binding preferences beyond sequence context by combining novel high-throughput experimental and computational methods. This interdisciplinary approach has enabled us to study binding preferences of TF complexes with respect to the epigenetic landscape of their cognate binding sites. Our novel mechanistic insights into DNA shape readout have provided a new avenue of exploiting guided protein engineering to probe how specific TFs interact with their co-factors in a cellular context, and how flanking genomic sequence helps determine which multi-TF complexes will form and which binding mode a complex adopts.

Contents

List of Figures

List of Tables

Acknowledgements

My journey through the world of science has not only literally taken me across an ocean, but at times, it also made me feel as if I was swimming through one. Getting washed ashore many islands, embarking on scientific adventures, and exploring unknown territories – from chemistry to structural biology, over cancer genomics, microbiology and epigenetics – each experience allowed me to take a look at science from a different angle. Doing so enabled me to pursue a PhD in a truly interdisciplinary setting: in two laboratories – one computational and one experimental, with two advisers, and in an area bringing together chemistry, physics, biology and computer science to address the fundamental question how nature achieves the intricate task of coordinated gene expression.

During this time, I was fortunate to meet many who taught, inspired, and supported me. The list of people is far to long to print, but I would, however, like to take the time to mention a number of individuals integral to my success.

First, I would like to thank Irmgard Sinning, who introduced me to structural biology, and the concept of questioning hypotheses until they stand up to one's own scrutiny. I am also grateful to Kevin Weeks, who gave me the opportunity to conduct my first long-term research project in his lab. My time at UNC allowed me to dive into the fields of high-throughput sequencing and data analysis, which are both important aspects of my research.

I cannot express enough gratitude to my thesis adviser and co-adviser, who provided me with this unique opportunity to truly combine computational and experimental work. Thank you Harmen, for always devoting the time to passionately discuss science with me, and for teaching me scientific rigor, the importance of statistical significance and the merit

of a well-conceived presentation. I could not have asked for a better mentor; you have fostered a supportive environment and provided me with countless opportunities to grow as a young scientist. Thank you Richard for allowing me to divert your lab away from fly development to the study of mammalian epigenetics, for pushing me out of my comfort zone and encouraging me to explore new ideas. Your guidance has helped me become a better experimental biologist.

Thank you to my committee, Barry Honig, Timothy Bestor, Marcus Noyes, and Chaolin Zhang for taking an interest in my work, providing advice, and helping me navigate research areas entirely new to me. An extended thank you to the CMBS program, the directors Ron Liem and Donna Farber, and in particular Zaia Sivo for their continuous support during my PhD.

I have also been fortunate to have had great collaborators. Many thanks to Remo Rohs, Carol Prives and Stavros Lomvardas and their labs.

I am indebted to the members of the Bussemaker and Mann labs, for the many fruitful discussions, provision of resources, and for their help in trouble-shooting experiments. My time in lab has truly been a joyful experience. A special thank you to Roumen Voutev, without whom I would have never been able to distinguish male from female flies, let alone the more complicated fly genetics.

Whether it was during my time in Heidelberg, in Chapel Hill or in New York, I have met a number of truly inspiring people, and made relationships that I am sure will last for a lifetime. Thank you to all my friends who supported me along the way, especially Mariana, Kathryn, Bryan and Petar – without you, Grad School would not have not been the same.

Thank you Adam for your patience, encouragement and making sure I always wind up with a smile on my face, even during those past tiring weeks. Life is a lot easier when you are around.

Finally, I would like to thank my family, without whom I would not be where I am today. Thank you for your endless support and love, making me the person I am today.

Für meine Eltern

Danke für eure Unterstützung und Liebe in all den Jahren. Ihr habt mir gezeigt, wie wichtig es ist an sich selbst zu glauben und seine Ziele nicht aus den Augen zu verlieren. Ihr habt mir die Kraft und das Selbstvertrauen gegeben meinen Weg unbeirrt fortzuführen. Ohne euch wäre ich nicht der Mensch, der ich heute bin. Danke.

Preface

The work described in this thesis was conceived by Judith F. Kribelbauer, Richard S. Mann and Harmen J. Bussemaker. The majority of the experiments, data analysis and validation was done by Judith F. Kribelbauer.

Namiko Abe, Siying Chen, Oleg Laptenko and Siqian Feng contributed to the experimental work – Namiko performed the SELEX-seq experiments for the wild-type Hth-Exd-Hox 21-mer libraries, Siying performed the SELEX-seq for Pbx-HoxA1, Oleg provided purified Δ30-p53, and Siqian generated the Antp-GFP fly line. The work benefited from input by the members of the Mann lab, providing guidance through the misleading nomenclature and puzzling details of fly genetics.

Chaitanya Rastogi generated the NRLB models used in Chapter 3. He and past and current members of the Bussemaker Lab – Tomas Rube, Vince FitzPatrick, Gabriella D. Martini, Ron Tepper, and Harmen J. Bussemaker – improved the data analysis through many invaluable discussions.

Oleg Laptenko, Will Freed-Pastor, Carol Prives, Richard S. Mann, and Harmen J. Bussemaker contributed to interpreting the findings of Chapter 2. Satyanarayan Rao and Remo Rohs have been essential for extending the work described in Chapter 2.

Detailed author contributions (designated by initials) are provided for each chapter.

Chapter 1

Introduction

1.1 General Overview

For a cell to maintain homeostasis as well as respond to different stimuli, it needs to coordinate the expression of a variety of different genes. Proper timing, tuning of transcript levels, and inter-cellular communication to restrict expression to distinct, spatially separated domains are all crucial, yet challenging tasks, for which cells have adopted a complex set of regulatory mechanisms. Understanding key aspects of gene regulation has thus long been an integral part of biological research. Its study recently gained momentum with the emergence of high-throughput technologies that enable the interrogation of virtually every layer of the regulatory network. Despite this progress, we are still only at the dawn of understanding how the many components involved in it work together. To give a brief overview, we can divide the regulatory network into four subcategories: i) regulation that occurs at the chromatin level (e.g. chromatin accessibility and 3D architecture), ii) regulation that acts on specific DNA signatures associated with their cognate genes, iii) post-transcriptional, and iv) post-translational regulation. For all these, well-studied examples exist and misregulation at any layer can have fatal consequences. For the purpose of this thesis however, only regulation under ii) that occurs at the DNA level and specifically the interplay between transcription factors (TFs) and their DNA target sites will be considered.

TFs are proteins that contain a DNA binding domain (DBD), are capable of recognizing and binding specific DNA sequences (transcription factor binding sites or TFBS) and can

modulate the expression levels of their target genes (which in many cases is the gene closest to the TFBS). The basic aspects of those regulatory units were worked out more than 30 years ago, starting with the discovery that proteins can recognize double-helical nucleotides in a sequence-specific manner (Seeman et al., 1976), such as seen for the cro-repressor DNA complex (Steitz et al., 1982). Soon thereafter, recurring DNA sequences, motifs for short, were first found in sets of genes with similar function. One example are the heat-shock genes, controlled by heat-shock proteins that recognize the cognate heat-shock response element (Gene and Pelham, 1982) (Davidson et al., 1983). New insights at the time predominantly came from atomic-resolution structures of protein-DNA complexes. With a few examples at hand, it was soon established that certain factors used the same secondary structural features to recognize DNA (Sauer et al., 1982) (Pabo and Sauer, 1992). The idea emerged that a set of simple rules for protein-DNA binding could be inferred, termed "DNA recognition code", and that soon it would be feasible to engineer proteins to achieve desirable DNA-sequence preferences (Pabo and Sauer, 1984) (Suzuki et al., 1995).

However, with increasing numbers of structures solved, the idea of a simple code did not solidify. Different TFs, despite sharing structural homology between their DBDs, were found to span a much broader repertoire of DNA binding mechanisms than initially thought. Nevertheless, with new biochemical techniques such as footprinting – an alternative approach to study TF-DNA recognition – some TFs within the same structural family were found to indeed share a similar DNA motif, whereas others, most prominently the zinc-finger-nucleases (ZFN), did not. At the beginning of the 1990s, it had become obvious that no simple code that generalizes TF-DNA interactions would be found, but rather that several aspects, including direct contacts with bases and the backbone, but also the overall structure and flexibility of the DNA molecule, would have to be taken into account (Pabo and Sauer, 1992). Carl Pabo appropriately described the challenges ahead in his 1992 review article (Pabo and Sauer, 1992):

"It is difficult to dissect these interactions in a way that assigns specific energetic

contributions to individual contacts."

This statement still holds true 25 years later, despite major technological improvements that allow us to probe TF binding specificity in high-throughput for a large set of TF families. In addition, many more structures have been solved that allow us to study both the structural properties of DNA molecules, as well as the binding mechanisms TFs use to recognize their cognate binding sites.

With the advent of high-throughput sequencing and the resulting genome assemblies for many species, combined with the structural and experimental insight, we learned that TF-DBDs are highly conserved and a typical DBD hovers around 60 amino acids independent of the animal kingdom a species belongs to (Charoensawan et al., 2010). This suggests that specific DNA sequence recognition is best achieved by a few secondary structure elements that define the individual TF families. As a consequence, TFs from the same structural family recognize similar sequence motifs, leaving us with the question, how more complex, multicellular organisms orchestrate regulation of highly specialized multicellular compartments?

A famous example of such a "specificity paradox" was provided by the homeotic transformation experiments done in fruit flies (Lewis, 1978) (Morata et al., 1983). Different body segments could be transformed into one another by simply swapping out homeobox TFs (Hox TFs) that all recognize similar DNA sequences *in vitro*, yet control the development of distinct body patterns, such as formation of the leg or head. Possible explanations for this paradox can be found when looking at i) the fraction TFs make up of the total protein pool and ii) the make-up of the protein sequences not coding for the DBD. As a rule of thumb, we can conclude that more complex organisms i) tend to devote a larger fraction of their genome to TFs (e.g. 5% in animals versus half that in fungi or plants) and that ii) their TFs are longer than the average protein in the genome (Charoensawan et al., 2010). Moreover, the average number of DBDs per TF is larger than one, suggesting that both combinations of DBD-specific motifs and amino acids outside the DBD contribute to the regulatory circuit.

Additional specificity could thus be achieved either through contacts with other TFs which refine target sites by cooperative binding, or by amino acids outside the DBD, sensing DNA geometry of flanking sequences. The former is supported by many examples of specific genes regulated by TF pairs, whereas the latter is less well-documented. This is partially a result of the transient nature and the high degree of flexibility of flanking DNA-TF contacts, which makes it difficult to capture them in crystal structures or with current TF binding assays.

With the emergence of new techniques that enable probing of TF sequence specificity in high-throughput, it is now becoming possible to go beyond the simple picture of a one-to-one mapping of TF-sequence recognition. Not only can we start mapping cooperative binding, but we can also pick up subtle differences in TF sequence preference that can contribute to the high degree of *in vivo* specificity seen for individual TFs. Last but not least, we now have the ability to also take into consideration how different chemical modifications of DNA bases (that exist throughout the genome with varying degrees) affect TF-specific readout.

In what follows, an overview about existing methods that characterize and quantify TF binding specificity *in vitro* and *in vivo* will be given, including their advantages and drawbacks. It is followed by a section on the different computational methods available to analyze such binding data. Lastly, it will conclude with a section about aspects of TF binding, briefly touched upon above, that are context-dependent and go beyond the simple picture of direct DNA-sequence recognition. Recent advances made in those areas will be discussed, including TF-binding to epigenetic DNA modifications and to low-affinity binding sites, DNA shape recognition, and latent specificity as a result of cooperative binding with other TFs.

1.2 Methods for Quantifying Transcription Factor Binding and Specificity

To dissect the regulatory logic of gene expression and to ultimately predict functional and phenotypic outcomes, we need to understand the binding mechanisms that make the strengths of TF-DNA interaction vary with sequence. Over the past few decades, many different techniques have evolved whose aim is to elucidate the biophysical and -chemical properties specifying such interactions. Two alternative approaches, focusing on two distinct aspects of TF-DNA binding can be distinguished: i) the use of structural biology to obtain high-resolution maps of the interaction surface of a TF bound to a specific DNA ligand, and ii) the use of genomic and biochemical assays to infer the range of binding preferences and energies of a TF to many different sequences. The former set of methods allows inference of highly-detailed structural readout mechanisms, but is limited to one sequence at a time, whereas the latter characterizes the relative ranking of a wider range of sequences, but generally provides no insight into the underlying binding mechanisms. In the sections below, select methods addressing either aspect of protein-DNA interaction will be discussed, along with their strengths and weaknesses:

1.2.1 Low- & Medium-Throughput Methods

Atomic Resolution Structures

Perhaps, the gold standard in understanding how a specific protein interacts with DNA is to solve the atomic structure of a TF bound to a DNA molecule. Generally, this has been done using X-ray crystallography. One of the first examples, was the study of the cro repressor of bacteriophage λ to a cro repressor DNA site (Anderson et al., 1981; Steitz et al., 1982), which gave valuable insight into the mechanisms proteins use to bind to DNA. For instance, the authors found that the DNA retained its B-conformation and that most contacts occurred between protein side chains and bases within the major groove of the DNA without the need

to further expose those bases. In addition, they also found mostly positively charged amino acids lying along the minor groove of the DNA and hypothesized that amino acids within the C-terminal arm of the cro protein could serve as "feelers" to readout flanking DNA. Since then, many more structures have been solved, and structural building principles of TF DNA readout have been identified, with the inferred similarities driving the classification of TFs into structural families (Garvie and Wolberger, 2001) (Luscombe, 2001). While these structures shared similar folds and overall binding mechanisms, they also revealed the great extent to which TF-DNA interactions are specific to a given complex, making it obvious that a simple family code, capable of accounting for all facets of binding, did not exist (Pabo and Sauer, 1992). In many ways, crystal structures have helped us investigate TF-DNA interactions by revealing in detail the complex structures of hydrogen bonding between DNA bases or backbone and protein amino acids, and by identifying a recurring set of protein folds used to interact with DNA. Yet, at the same time, they are highly limited, only providing us with a glimpse at a single bound state of a TF with a particular DNA ligand, but giving little insight into what it takes for the TF to recognize a variety of DNA target sequences. Some of these short-comings were already pointed out in earlier papers describing the first TF-DNA structures (Pabo and Sauer, 1984): More flexible regions thought to contact well defined positions when in solution, were found to be disordered in the crystal structure, perhaps a consequence of subtly different configurations, thus limiting the resolution (Anderson et al., 1981). Although hydrogen bonding between protein alpha-helices with bases in the major groove were thought to predominantly determine sequence selectivity (Pabo and Sauer, 1984), many van der Waals interactions were also observed (Pabo and Sauer, 1984). Later on, it was demonstrated that the latter were equally important for TF-sequence recognition by comparing protein-DNA contacts across many structures. Van der Waals interactions made up more than half of all contacts (Luscombe, 2001). New approaches, such as time-resolved cryo-Electron Microscopy (cryo-EM) (Frank, 2017), might thus be needed to probe more subtle, yet important aspects of TF-sequence recognition in

terms of structural biology.

In vitro Footprinting

A second, complementary methodology for studying how TFs interact with DNA that does not require detailed knowledge about the three-dimensional structure and detailed interaction maps is to ask what sequences TFs recognize and what common features those sequences share. One of the earliest such techniques is DNA footprinting, which relies on the protection of DNA to exonuclease cutting when a TF is bound. The method was first decribed as an *in vitro* procedure in 1978 (Galas and Schmitz, 1978) but was soon after modified to probe specific TFs of interest using *in vivo* genomic extracts (Wu, 1984). In short, TF-bound DNA can be subjected to DNAse treatment and the resulting cleavage pattern is compared to a treatment control with unbound DNA (Figure 1.1 on page 8)(Vierstra and Stamatoyannopoulos, 2016).

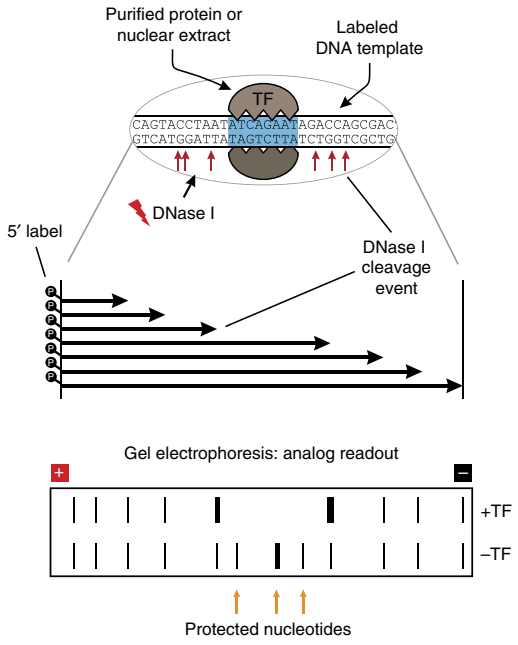


Figure 1.1: Traditional Footprinting Method: *(Figure reproduced from Vierstra and Stamatoyannopoulos, Nature Methods 2016)*

The method of footprinting has been in particularly useful for locating TF binding sites in an *in vivo* context or in testing whether certain genomic elements could be recognized when incubated with *in vitro* purified TFs. It complemented the structural studies and served as a starting point for identifying sequences suitable for crystallization trials. However, in contrast to the structural studies, the footprinting method gave little insight into the mechanistic details of how a TF recognizes a specific site. Recently, a high-throughput adaptation of this method was developed that overcomes some of these limitations; it will be discussed in a section below.

Electro-Mobility Shift Assay (EMSA)

Another commonly used and simple technique to characterize the binding of proteins to DNA relies on the reduction in electrophoretic mobility of DNA ligands in a native poly-

acrylamide gel upon protein binding. These gel retardation assays are generally referred to as Electro-Mobility Shift Assays, or EMSAs for brevity, and allow detection of the relative amounts of protein-bound and -unbound DNA by either radioactive or fluorescent labeling of the DNA and subsequent imaging or ethidiumbromide staining of the gel (Figure 1.2 on page 9). There are several applications in which EMSAs have been used (Lane et al., 1992):

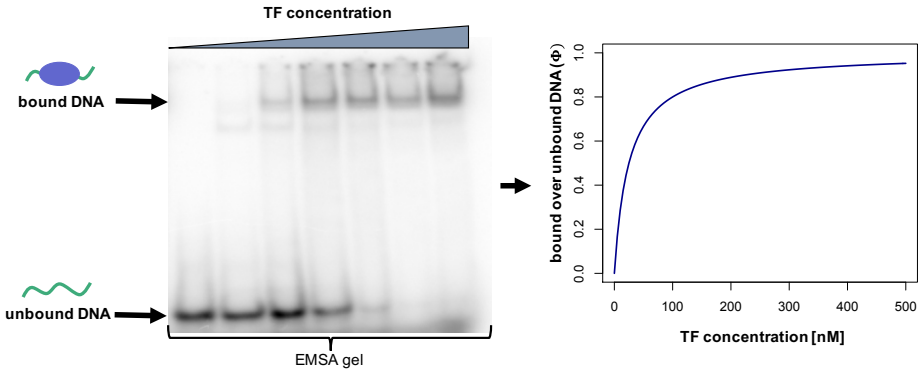


Figure 1.2: EMSA and Kd Determination: Varying concentrations of TFs are incubated with a DNA probe and the TF-DNA complex is separated from free DNA by electrophoresis. Quantifying the bound to unbound DNA fraction and plotting it agains the protein concentration in each well allows K_d estimation.

1. **Quantification of binding constants:** Both association and dissociation rates can be measured by adding either one component of the complex or competitor DNA in time intervals and loading them on a running gel. Perhaps the more standard measurement is that of the equilibrium constant K_{eq}, more commonly expressed by its inverse: $K_d = \frac{1}{K_{eq}}$. Here, the binding reaction is incubated until equilibrium is reached and reactions with identical amounts of labeled DNA but increasing amounts of TF are loaded on the gel. Calculation of a K_d is done by constructing a binding curve, where the fraction Θ of bound $(TF : DNA)$ to unbound (DNA^{free}) DNA is plotted against

the total protein concentration (TF^{total}) (Figure 1.2 on page 9). Using the following relationship,

$$\mathrm{K_{eq}} = \frac{\mathrm{TF : DNA}}{\mathrm{TF^{free} * DNA^{free}}} = \frac{1}{\mathrm{K_d}} \tag{1.1}$$

the K_d can be determined from the binding curve in the limit where the TF concentration is much larger than the DNA concentration and the free TF concentration is roughly the total TF concentration ($TF^{free} \sim TF^{total}$). Using $DNA^{total} = DNA^{free} + DNA^{bound}$, the equation then simplifies to:

$$\Theta = \frac{1}{1 + \frac{\mathrm{K_d}}{\mathrm{TF^{total}}}} \tag{1.2}$$

2. **Competitive EMSAs:** The differences in K_d across different DNA ligands can be experimentally quantified by adding increasing amounts of unlabeled "cold" DNA to constant amounts of TFs and labeled "hot" DNA. Comparing two cold DNAs can then be achieved by plotting a dose-repsone curve – the ratio of bound and unbound hot DNA versus the log-concentration of cold DNA – and determining the concentration of cold DNA (inhibitor) at which half of the hot DNA is outcompeted (inhibitor concentration at 50% binding, or IC50) (Craig, 1993). The ratio of the obtained IC50 values equals the ratio in K_d only if the concentrations used for the hot probe and the protein are kept identical between competitor experiments.

3. **Cooperative binding:** As reduction in electromobility of DNA is influenced by mass, charge and conformation (Lane et al., 1992), it is feasible to resolve different complex compositions with an EMSA. This makes it suitable to study the relative thermodynamic stability of multi-TF complexes as the different subcomplexes (e.g. monomeric, dimeric, or oligomeric) will shift DNA to a different extent. The intensity of the shifted and super-shifted bands (comparison of ratios) can then be used to compare subcomplex stability.

One of the benefits of using an EMSA over other forms of electrophoresis is the stabilizing effect that the gel matrix provides for labile complexes compared to their kinetic stability in solution. This effect is known as the "cage" effect (Lawn et al., 1981) and has primarily been contributed to the increase in local concentration (within a gel) due to the decreased accessibility of reagents and the exclusion of volume by the gel (Lawn et al., 1981). In addition, the gel matrix also decreases the effective dissociation by limiting the "escape" rate of the free TF from the gel compartment that the protein-DNA complex migrates in at any given time (Cann, 1989). In summary, EMSAs are useful for quantifying kinetic and thermodynamic constants of TF-DNA interactions and provide some insight into their conformational state (e.g. complex composition). Similar to classical footprinting, one limitation is the restricted number of measurements that can be done at a time. In recent years, however, a few high-throughput adaptations of the gel-shift assays have been developed, which will be discussed in the next section.

1.2.2 *In vitro* High-Throughput Methods

Microarrays

With the arrival of robotic systems for molecular biology and the invention of cDNA cloning in the 70's and 80's (Auffray and Rougeon, 1980), a new technology – DNA arrays

– emerged in the late 90's and early 2000's (Bumgarner, 2013). Common to all arrays is the idea that fragments of known cDNA sequences are attached to a surface in clusters, which allows hybridization of fluorescently labeled target molecules to the array and subsequent quantification via imaging. The arrays were eventually compact enough to fit on a small chip, ultimately referred to as microarrays. Three different styles exist: i) spotted arrays, where sequences are deposited on glass plates with micro droplets (Derisi et al., 1997), ii) self-assembled arrays using beads with specific DNA fragments that are randomly dispersed into wells, with location-to-sequence mapping achieved via optical encoding or using fluorescent barcodes (Ferguson et al., 2000), and iii) in-situ synthesized arrays, such as those produced by Affymetrix or Agilent. A few years after application of these arrays to expression profiling, they were also used to map TF binding sites by chromatin-immunoprecipitation of a TF of interest and subsequent hybridization of the TF-bound DNA fragments (Iyer et al., 2001; Horak and Snyder, 2002). Only a few years later, this technology was adapted in a way that allowed direct probing of TF binding to double-stranded DNA molecules – the so called Protein Binding Microarray (PBM) (Bulyk et al., 2001; Mukherjee et al., 2004). Quantification of binding affinity for specific sequences was achieved by fluorescently labeling the TF of interest and reading out the fluorescence intensity after a first incubation step and subsequent removal of excess TFs to avoid non-specific binding. TF binding preferences and relative affinities for many sequences could thus be tested simultaneously, describing the first high-throughput assay of TF-binding specificity. Since then, the array has been applied in varying ways, using two different fluorophores to study TF dimerization, synthesizing PBMs with genomic sequences to restrict the analysis to relevant *in vivo* sites (Gordân et al., 2013), and probing TF binding to epigenetic 5-methylcytosine marks added by a post-array-production enzymatic reaction (Mann et al., 2013).

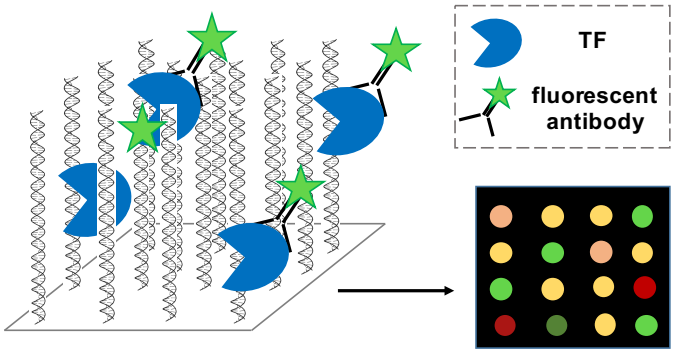


Figure 1.3: Protein-Binding-Microarray Technology: TF binding strength to dsDNA molecules, spotted on a microarray, is quantified by reading out the intensity of fluorescently labelled antibodies specific to the probed TF

One-Hybrid Systems

Using the ease by which plasmids can be transformed into bacterial or yeast cells, the one-hybrid methodology detects which sequences are recognized by a given TF by utilizing overexpression of TFs fused to an activator domain (bait). Simultaneously, a vector system containing a randomized stretch of sequences and a weak promoter – the prey – is used to allow expression of a downstream reporter gene. When a suitable binding site is present in the randomized sequence stretch, the TF binds and recruits the activator domain to the weak promoter (bait-prey interaction), inducing gene expression. Combinations of downstream genes, such as the HIS3/URA3 system, can be used for simultaneous, positive and negative selection by providing a growth advantage for successful interactions (see Figure 1.4 on page 14). Depending on the organism, the Yeast 1 Hybrid (Y1H) (Li and Herskowitz, 1993; Deplancke et al., 2004) and the Bacteria 1 Hybrid (B1H) (Meng and Wolfe, 2006; Noyes et al., 2008) are distinguished. Many adaptations have been developed to accommodate more complex interactions. The advantage of such systems is the ease of use: no prior purification of individual TFs, or microarray synthesis is needed, and the experimental setup resembles an actual "*in vivo*-type" setting . However, the method only probes TF-DNA interactions

indirectly through the downstream bait-prey expression response, thus requiring controls for background activation and potential interactions with endogenous factors. Selection for TF pairs is also more difficult, due to a more complicated cloning scheme, transformation limitations, and the need for a split activator domain that is only functional when both factors bind cooperatively.

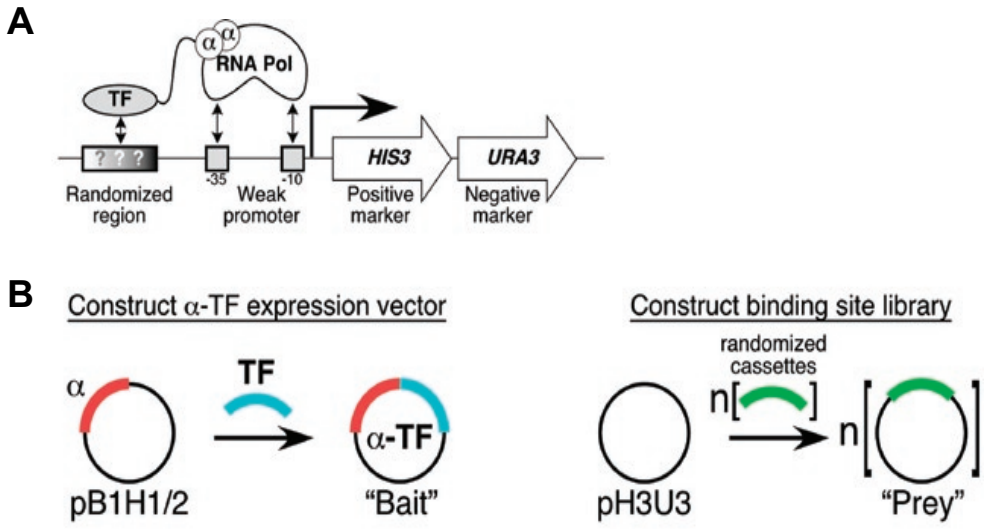


Figure 1.4: Schematic for Using the B1H Assay: *(Figure by Meng andWolfe, Nature Methods 2006)* **(A)** Design of a the B1H. A TF fused to the alpha subunit of the RNA polymerase recognizes its cognate binding site within the random pool of sequences. Binding initiates expression of the HIS3 URA3 system. **(B)** Cloning strategy for "bait" and "prey" vectors.

SELEX-based Methods

In parallel with the microarray technology, another high-throughput approach emerged that has revolutionized genome-wide molecular studies. Its prototype, sequencing-by-synthesis was first developed by Shankar Balasubramanian and David Klenerman at Cambridge University and then commercialized by the Solexa-Illumina company. Today the technique is commonly referred to as Next-Generation-Sequencing (NGS) and its cost-effectiveness,

turn-around-time, accuracy, and unprecedented sequencing depth has made it the method of choice for many applications, including expression profiling or TF binding (Git et al., 2010; Lee et al., 2013). To study TF binding *in vitro*, NGS is combined with Systematic Evolution of Ligands by Exponential Enrichment(SELEX) techniques (Tuerk and Gold, 1990), using several different approaches to enrich for the TF-DNA complex. In essence, a TF of choice is incubated with a randomized sequence pool, followed by a selection step that separates bound from unbound DNA, amplification of the enriched DNA and subjection of the enriched pool to another round of SELEX. Before and after each stage of enrichment, NGS-libraries are prepared and sequenced. The enrichment of different sequences can thus be followed over multiple rounds. From equation 1.1 and 1.2 we know that the K_d of an individual sequence S_i is inversely related to the fractional occupancy (Θ_i). When starting with N such sequences, equation 1.1 can be rewritten by using the relation $S_i/S = f_i(S_i)$, where S denotes the total sequence concentration and $f_i(S_i)$ the frequency of sequence i:

$$\frac{S * TF^{free}}{K_d + TF} = TF : S = \sum_{i=1}^{N} TF : S_i = TF^{free} * S \sum_{i=1}^{N} \frac{f_i(S_i)}{K_{di} + TF} \tag{1.3}$$

The equation above shows that the overall K_d is related to the sum of individual K_{di}'s and $f_i(S_i)$. Since relative fractions don't change under PCR conditions, we can relate the post-selection frequency of sequence i $f_i(S_i)'$ in a SELEX experiment to the pre-selection frequency by:

$$f_i(S_i)' = \frac{K_d(TF) + TF^{free}}{K_{di} + TF^{free}} * f_i(S_i) \tag{1.4}$$

In the limit where the TF^{free} is much smaller than the K_d ($TF^{free} << K_d(S_i)$) and when comparing S_i to a references sequence (S_{ref}, we obtain a simple relationship relating the ratio of frequencies of two sequences to their respective K_d's:

$$\frac{f_i(S_i)'}{f_{ref}(S_{ref})'} = \frac{K_d(S_{ref})}{K_d(S_i)} * \frac{f_i(S_i)}{f_{ref}(S_{ref})} \tag{1.5}$$

One therefore only needs to follow the frequencies across different SELEX experiments to compare relative K_d's (Levine and Nilsen-Hamilton, 2007). Using the sequencing counts pre- and post-selection as estimates, relative affinities can be calculated for any given sequence with respect to a reference sequence. Usually the most enriched sequence is chosen as a reference, such that relative affinities range between 0 and 1.

Depending on the experimental setup, enrichment step, and post-analysis chosen for a SELEX experiment, several different approaches need to be distinguished (Figure 1.5 on page 17):

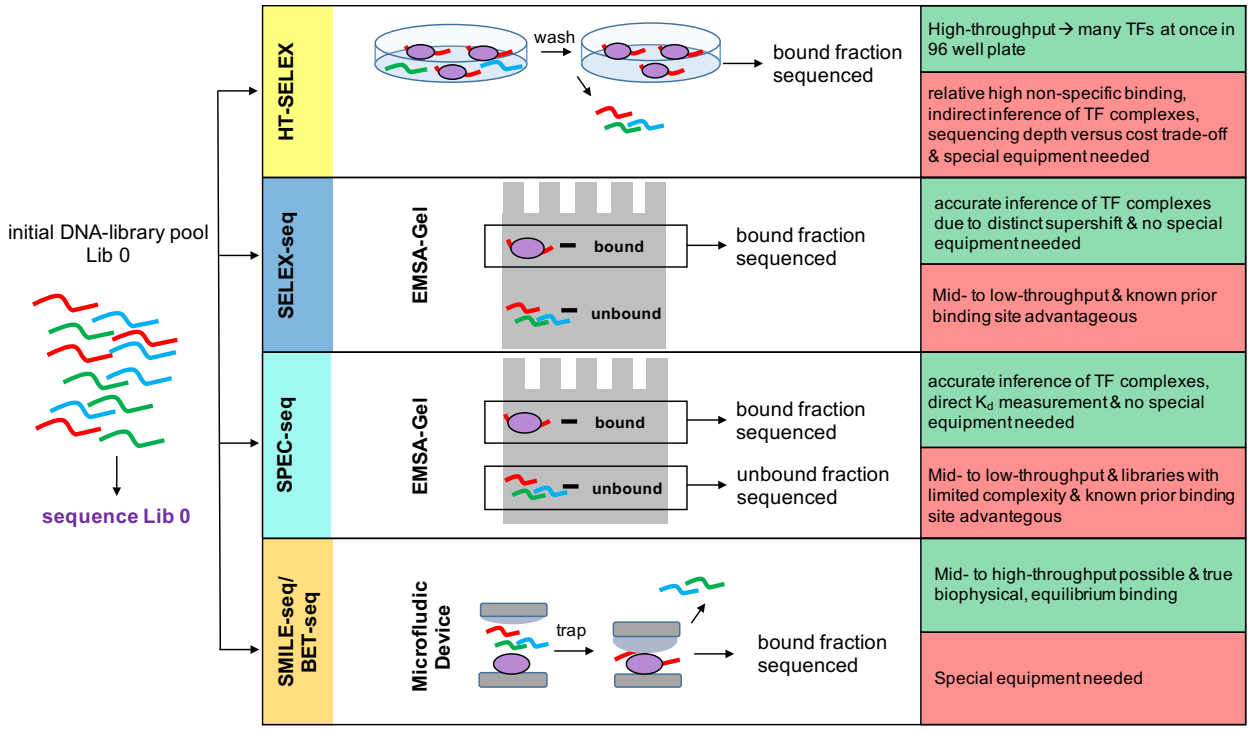


Figure 1.5: Overview of Types of SELEX Experiments: The table provides an overview of the different types of SELEX experiments that exist. Outlining the experimental strategy and advantages and drawbacks of each method.

1. SELEX-seq & SPEC-seq

   SELEX-seq and SPEC-seq probe TF-binding specificity for many DNA ligands simultaneously by using EMSAs to separate bound from unbound DNA ligands. Using EMSAs limits the throughput in terms of testing many TFs simultaneously, but has several advantages compared to the solution-based assay: i) Supershifting of a complex already serves as a quality control that the TF is functionally active (e.g. has a properly folded DBD), ii) potential remnants of bacterial proteins are removed due to a the minimal chance they would share the same electro-mobility within the gel matrix, and perhaps most importantly iii) it allows for separation of TF-DNA complexes with different composition or stoichiometry, e.g. monomer versus homo-/hetero-dimeric TF complexes. The latter application is problematic in solution-based assays, as cooper-

ativity can only be distinguished from DNA-sequence-assisted or entirely independent binding of two TFs through sophisticated data analyses (Rastogi et al., 2018). Two different flavors or EMSA-based SELEX experiments exist. To allow identification of the best binding site and to accurately quantify all sequence affinities in reference to the top binder, SELEX-seq (Slattery et al., 2011) sequences the initial library (R0) and constructs a markov model to account for sequence biases, which can occur during library synthesis, .

The other flavor – SPEC-seq – uses a library with a reduced number of sequences to guarantee accurate counting in both the bound and unbound fractions. This in return allows accurate calculation of relative affinities, as the free protein concentration drops out of the equation to determine relative K_d's (Stormo et al., 2015). Although both approaches use EMSAs to separate bound from unbound ligands, SPEC-seq has the advantage of calculating exact ratios for each individual probe in the pool, and does not rely on a prior bias model of the initial library or non-linear fits to obtain relative K_d's. However, it is limited to a subset of DNA sites and thus requires prior knowledge of the potential TF binding site and might miss subtle, yet important differences in core and flanking regions that were not considered when designing the semi-random library.

2. HT-SELEX

In High-Throughput (HT)-SELEX, TFs are immobilized on beads, incubated with DNA libraries, washed to remove unbound DNA and the bound fraction is extracted by a final elution step (Jolma et al., 2010). Using 96-well plates, many TFs (either full-length or DBDs) can be assayed simultaneously. Barcoding each library with a TF-specific short sequencing tag allows pooling of all reactions. This method is capable of testing hundreds of TFs, since only low amounts of protein are required, which

can be purified on beads from minimal bacterial cultures. The scalability makes it a great approach for identifying binding motifs for large sets of TFs. Indeed, the method has been used to create binding models for a large numbers of TFs, including fly and human proteins (Jolma et al., 2010; Nitta et al., 2015; Jolma et al., 2013). In addition it has been modified to test TF-dimer binding by using two different protein tags (Jolma et al., 2015). A downside to the large-scale approach, however, is that the sequencing depth for most TF libraries is rather low, requiring several enrichment cycles (often 4-8 cycles) to attain robust sequence enrichment. Accurate quantification of subtle differences between members of the same TF family might thus be difficult. Another factor contributing to the observed variability in selection round requirements might be the lack of a stringent quality control as a result of the automated purification protocol and the in solution incubation approach. As is known for other bead-based selection methods, non-specific binding can be pervasive, which poses a problem for TFs with K_d's in the higher nanomolar range. Nevertheless, HT-SELEX has been an invaluable resource for identifying initial TF motifs.

3. MITOMI, SMiLE-seq & BET-seq

Perhaps the only method that can achieve absolute quantification of TF-binding affinities across many sequences in intermediate-throughput is MITOMI (Maerkle and Quake, 2010), short for mechanically induced trapping of molecular interactions. MITOMI combines microarray technology with a microfluidics approach, spotting distinct sequences onto a plate and transferring them into separate compartments of a microfluidics device. The sequences are labeled with fluorophores and incubated with *in situ* synthesized histidine-tagged TFs until equilibrium is reached. TFs are deposited under constant flow, guaranteeing equal TF concentrations across microfluidic cells. To remove any void volume while preserving the equilibrium concentrations, the protein-

DNA complexes are trapped by a downward-pushing button that constrains the complexes to a precise area. To prevent any TF-DNA complex from escaping the area, each cell is additionally coated with anti-histidine antibody and bovine serum albumin. Measuring the intensities for each well allows absolute quantification of K_d's. In the past year, two high-throughput adaptations of MITOMI emerged that use the idea behind MITOMI but add the SELEX aspect of selection of DNA ligands from a random pool. In both methods, SMiLE-seq (Isakova et al., 2017) and BET-seq (Le et al., 2018), instead of spotting distinct sequences to each well, an entire library of sequences is added. The bound DNA is still captured at equilibrium with mechanical trapping, but quantification is achieved using NGS of the trapped fragments rather than measurement of fluorescent intensities. Each well can thus either have a different TF, allowing for larger TF-throughput or a different TF concentration. The main difference between the two is similar to the difference between SELEX-seq and SPEC-seq or between HT-SELEX and SPEC-sec. SMiLE-seq sequences the initial library (like SELEX-seq does) to analyze enrichment, and has limited sequencing depth as it probes many TFs, whereas BET-seq, sequences the bound (trapped) and unbound (washed away) fractions to obtain more accurate and direct quantification. Therefore, like SPEC-seq, BET-seq is limited to a library with reduced complexity and higher sequencing requirements to guarantee occurrence of all sequences in both bound and unbound fractions.

Deciding which of the above techniques is ideal depends on the specific scientific question and the equipment at hand. Both HT-SELEX and SMiLE-seq or BET-seq need specialized platforms, such as building a microfluidics device or having an automated platform to do purification, washing and library preparation in a 96-well format. SELEX-seq and SPEC-seq on the other hand only require making and running a native acrylamide gel, which is standard lab equipment. For large screens and motif discovery, HT-SELEX or SMiLE-seq

are well suited, but for a more detailed analysis of individual factors, one might prefer a method with less throughput, which allows for higher sequencing depth and more accurate binding models, such as SELEX-seq, BET-seq or SPEC-seq. If accurate quantification is key, there are two potential approaches: Either using methods such as SPEC-seq or BET-seq that provide direct quantification by taking ratios, but are limited to reduced-complexity libraries or by using accurate mathematical models that can capture both initial library bias and selction process, such as the one described in Rastogi et al., which will be discussed in the algorithm and analysis section below (Rastogi et al., 2018). In general, since MITOMI is the gold standard for quantifying TF binding in absolute terms with the least amount of bias (that could arise from a non-linear selection processes e.g. during the wash steps or non-specific binding), methods that use mechanical based trapping also seem to provide the highest-quality data (see comparison in (Rastogi et al., 2018)). SMiLE-seq ot BET-seq might thus be the method of choice for future applications. However, the lack of standardization and/or commercialization of microfluidic devices to benefit a larger group of researchers still remains a limiting factor.

1.2.3 *In vivo* High-Throughput Methods

Direct Methods – ChIP-on-Chip & ChIP-seq

One method to probe TF binding *in vivo* is ChIP-on-Chip, short for chromatin immunoprecipitation (IP) on a microarray chip – a technology that was briefly mentioned earlier in the introduction (Iyer et al., 2001; Horak and Snyder, 2002). In this approach, cross-linked chromatin is extracted from cells, fragmented using sonication or enzymatic digestion and immunoprecipitated with an antibody raised against the TF of interest. After the IP step, the TF-DNA fragments are reverse-crosslinked, DNA is isolated and subsequently hybridized to a DNA microarray (containing e.g. promoter regions of interest). The IP-coupled microarray allows identification of genes that are potentially regulated by the TF *in vivo*. Like

any other method, it also has a few specific limitations: i) resolution limited by the probes on the array, and ii) partially confounded motif discovery as a consequence of other genomic features correlating with TF binding (e.g. CG content on hyper-accessible promoter regions).

Similar to other microarray techniques, the array-based design was eventually replaced by high-throughput sequencing, with the first protocol for ChIP-seq becoming available in 2007 (Barski et al., 2007; Johnson et al., 2007; Mikkelsen et al., 2007). ChIP-Seq follows the same steps for IP as ChIP-on-Chip protocols, but rather than hybridizing the DNA to an array, NGS-sequencing libraries are prepared from the DNA fragments, followed by Illumina sequencing and sequence mapping to the respective genome. Several adaptations of ChIP-seq have been developed since then, using different fragmentation approaches, IP conditions, or crosslinking methods (Rhee and Pugh, 2011; Schmidl et al., 2015; Skene and Henikoff, 2017).

One of the methods, CUT & RUN (Skene and Henikoff, 2017) provides an alternative to classical ChIP-seq approaches, as it isolates bound TFs from intact cells *in situ*, omitting both crosslinking and sonication. After permeabilizing cells and incubating with primary TF antibody, the cells are treated with micrococcal nuclease, which is attached to the secondary antibody and cuts around the TF. The resulting short, TF-bound DNA pieces are subsequently isolated by a simple centrifugation step. CUT & RUN has been demonstrated to result in significant background reduction and has the unique advantage of capturing bound TFs in their natural state.

Indirect Probing – ATAC-Seq & DNase-Seq

An alternative, indirect way of identifying sites occupied by TFs *in vivo* is achieved by exploiting the differences in observed cut- or insertion-rates of endonucleases or transposases when encountering accessible versus inaccessible genomic DNA. One commonly used method is described in (Crawford et al., 2006) and relies on cutting by DNase I to identify DNase

hypersensitive sites genome wide. Cells are treated with DNase, the resulting fragments are extracted and sequenced, before mapping the reads back to the genome. Identification of accessible chromatin is achieved by searching for regions with increased cut frequencies, harboring more mapped reads than the surroundings. If a TF's residence time is long enough to withstand DNase treatment, it will leave a "footprint", a short stretch with decreased DNase cut-rates. In combination with TF motifs that are available from databases such as JASPAR (Khan et al., 2018) those footprints can be scored and grouped by their sequence signatures and assigned to the TF with the highest likelihood of being bound in a given cell type. Another technique, ATAC-seq, uses the Tn5 transposase instead of DNase I to probe chromatin accessibility (Buenrostro et al., 2013). Since its first description five years ago, ATAC-seq has become the method of choice due to a significant reduction in experimental steps by Tn5's ability to cut and simultaneously insert the pre-loaded sequencing adapters required for library amplification.

Although ChIP-seq or other *in vivo* experiments are often used as gold standard when evaluating methods and algorithms, there are a few things to keep in mind when interpreting *in vivo* binding. Besides the obvious enrichment biases that result from the use of different antibodies or binding resins, it is important to note that ChIP-seq relies on irreversible crosslinking of proteins to DNA and therefore measures TF occupancy rather than affinity, which requires equilibrium conditions. Binding of a TF to a particular site will thus depend on both the accessibility of the site across a potentially heterogeneous cell pool as well as the local TF concentration within the nuclear compartment a site is located in. Highly accessible regions (such as promoters of strongly expressed genes) therefore show strong binding signatures for nearly every TF tested. This phenomenon is commonly known as "hyper-ChIPability" or "colocalization hotspots" (Moorman et al., 2006; Ward et al., 2014). One extreme example is given by the apparent recruitment of repressor TFs to highly expressed

gene promoters (Teytelman et al., 2013) lacking any meaningful biological function. This might be attributed to the lower apparent K_d when a site is highly accessible and huge amounts of recruitment factors are present. Another explanation, which has gained more attention recently, is that sites can fall within transcriptional hubs, creating an environment with increased local TF concentration (Mercer and Mattick, 2013), which could likewise results in a lower apparent K_d and potentially spurious binding without biological relevance. These hubs however, can also function to selectively recruit TFs to low-affinity binding sites that are important for gene regulation (Tsai et al., 2017). Further details shall be discussed in section 1.4. As a consequence, less accessible, yet biologically functional sites might end up on the other end of the spectrum, with seemingly no enrichment above background, as their fraction only contributes marginally to the entire pool of enriched sequences. In addition, sites with decreased accessibility are less likely to be properly fragmented and thus tend to not to be included in the final sequencing library, which is usually generated after a rigorous size-selection step. Therefore, ChIP-seq data should be interpreted with caution, ideally making use of as much orthogonal information as possible, such as ATAC-seq signal, scores from *in vitro* binding models or gene expression information. The convolution of the true TF-binding signature and non-specific, yet correlated genomic features increases the need for accurate binding models to differentiate functional (with perhaps low levels of TF occupancy) from non-functional sites.

1.3 Models & Algorithms

Alongside the development of experimental *in vitro* and *in vivo* methods to identify TF binding sites, another line of research has focused on the mathematical representation of such sites and their *de novo* discovery through statistical means. In the early days, researchers already noted reoccurring sequences within the promoters of bacterial genes and discovered "by eye" the first DNA motifs used to regulate gene expression (Pribnow, 1975; Rosenberg and Court, 1979). Identifying motifs based on their sequence pattern alone however, turned out to be rather challenging as it soon became obvious that TFs tolerate DNA sites with a certain degree of degeneracy. Therefore, the question arose how to best quantify and represent the observed degree of TF sequence specificity in a systematic way that allows quantification of TF binding strength as well as providing a means to discover unknown motifs.

1.3.1 Motif Representation

Consensus Methods

The different half sites of the γ - operator or the *Escherichia coli* promoter -10 element are examples of binding site degeneracies as they display variability at several positions within the protein binding sites (Pribnow, 1975; Maniatis et al., 1975). Despite the lack of sequence identity, those few early known examples could often be summarized by a "consensus site". Consensus sites are DNA sequence representations utilizing an expanded base code indicating conserved as well as degenerate base positions within a site (Day and Mcmorris, 1992). The letter N for instance described complete degeneracy, whereas letters such as R or Y restricted mutations to purines or pyrimidines respectively. However, with rapid increases in DNA sequencing capacity many more such potential binding sites were discovered, revealing that a simple code was not sufficient to capture the variable degree of TF binding affinity and its effect on downstream gene expression.

Genomic Sequences	ATGTCGA
	TTGATGA
	CTGACGA
	GTGTCGA
	ATGATGA
	GTGTCGA
	ATGTCGA
Consensus Motif	NTGWYGA

Table 1.1: Example of a consensus motif representation for sites found *in vivo*

Weight-Matrix Representations

Perhaps, the most commonly used way to represent a TF's preference for different sequences is the Position Weight Matrix (PWM) or variants of it (Stormo and Schneider, 1982). In essence, a PWM is a matrix with four rows representing the alphabet (in case of DNA the four bases A,C,G,T) and as many columns as needed to encompass the entire TF footprint. Each entry in the matrix is a score for a particular base at a given position within the binding site and the score for any particular sequence is computed by summing up the respective entries in the PWM. The matrix has the property to assign the highest score to the consensus sequence and lower scores for sequences that deviate from the top site. There are several ways to compute the weight each base obtains at any given position (for a detailed review see (Stormo, 2000)), but the most commonly used one is simply the normalized negative logarithm of the frequency for each of the four bases in a given set of positive training data (Staden, 1984). Other methods made use of neural nets to find a matrix that best separates positive from negative examples (Stormo and Schneider, 1982) or included quantitative expression data to test model performance (Hertz and Stormo, 1996).

Base Alphabet	position 1	position 2	position 3	position 4	position 5
A	-5	-41	**10**	-37	-10
C	**5**	-4	-34	**9**	-15
G	-15	**11**	-2	-37	-4
T	-10	-25	-27	-29	**7**

Table 1.2: Example of a PWM representation. Bold sequences represent the top binding site CGACT.

An important aspect to consider, however, is to what extent a given PWM can classify a site above background sequences and how much each position contributes to classifying a true binding site. For this purpose, Schneider et al. came up with an approach taken from information theory and computed the information content for particular PWMs (Schneider et al., 1986) which is defined as:

$$\mathrm{I_i} = \sum_{\mathrm{b=A}}^{\mathrm{T}} \mathrm{f_{b,i}} \log_2(\frac{\mathrm{f_{b,i}}}{\mathrm{p_{BG}}}) \tag{1.6}$$

with $f_{b,i}$ being the frequency of a base at position i and p_{BG} the background probability for a given base. Assuming a uniform distribution for p_{bg} this would simplify to:

$$\mathrm{I_i} = \sum_{\mathrm{b=A}}^{\mathrm{T}} 2 + \mathrm{f_{b,i}} \log_2(\mathrm{f_{b,i}}) \tag{1.7}$$

As genomes are rarely uniformly distributed and often have varying AT versus CG content,

the background model should not be neglected.

While using frequencies and information content has been successful in characterizing and identifying TF binding sites, it is only a proxy for the underlying free energy contribution of each base within the binding site. In order to capture true biophysical binding specificities, a model must be based on the underlying physical selection process that gives rise to the data. Examples include biophysical models fit directly to low-throughput experimental binding data (K_d measurements) (Liu and Clarke, 2002), gene expression levels or from modeling protein-DNA interactions (Endres et al., 2004). Keeping the same matrix representation, it is possible to construct a position-specific affinity matrix (PSAM) that maps the relative change in affinity (or fold change in K_d) for each possible point mutation away from the sequence with highest TF affinity (Foat et al., 2005, 2006). For instance, one would only have to measure the K_d for $3 * k + 1$ sequences (the top sequence and its $3 * k$ point mutations), with k being the number of specified positions within the binding site, to be able to predict any possible sequence of length k. Predicted relative affinities are simply obtained by multiplying over the entries in the PSAM representative for each base position in the binding site. Matrix entries for the consensus site take the value 1, making the highest achievable relative affinity 1. The resulting PSAM can then be directly transformed into $\Delta\Delta$G values using the relation:

$$\Delta\Delta\mathrm{G}(\mathrm{S_i}) = -\mathrm{RT}\ln(\frac{\mathrm{K_d}(\mathrm{S_i})}{\mathrm{K_d}(\mathrm{S_{ref}})}) \tag{1.8}$$

It is important to notice that the above mentioned methods are all assuming independence of individual positions and do not consider dinucleotide interactions. More complex models can be used to include higher order interactions and will be discussed in a later chapter.

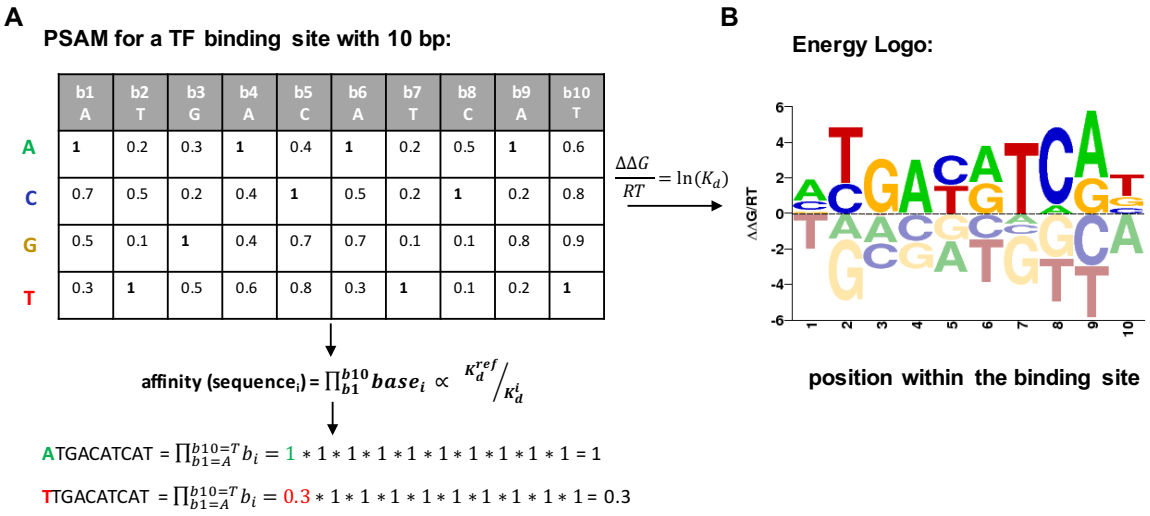


Figure 1.6: Example PSAM and Energy Logo: **(A)** A position-specific-affinity-matrix is given for a 10bp long TF binding site. Entries in each cell represent the relative change in K_d for all possible, single point mutations away from the top binding site. Example for one such mutation at position 1 is given below. **(B)** An energy logo representing a PSAM by transforming relative K_d ratios to $\Delta\Delta G/RT$ of binding.

1.3.2 Motif Discovery

In general, motif discovery methods are based on the same concepts as used for motif representations and often produce a motif as an output. In the following, the most commonly used methods are discussed.

Enrichment-based Methods

A simple, yet powerful way of identifying unknown motifs is by searching for overrepresented sequences. This can be achieved via counting how many times a kmer of a given length occurs in a region of interest (e.g. sequences obtained from an experimental selection process) and comparing it to a control set. The first examples of such algorithms were used by Galas et al. (1985), looking for patterns in *Escherichia coli* promoters, and (van Helden et al., 1998), analyzing oligonucleotide frequencies within promoters of coregulated

genes. However, the authors already noted that there is no biological significance to the scoring parameters (Galas et al., 1985). Therefore, pattern recognition methods need to be evaluated with caution, especially given the often extensive sequence biases present in many genomes. Using overrepresentation of kmers is also frequently used in high-throughput TF binding data in order to reduce the level of complexity and to improve the signal to noise of full probe count data tables. Most of the time, kmer counting is used to seed the initial model and thus find a good starting point from which the final motif model is inferred. For SELEX-data for instance, the obtained kmer counts can directly be transformed into a PSAM by taking the frequency ratios of a given sequence S_i to a reference sequence S_{ref} (see equation 1.5). Similar to genomic data, high-throughput data can display a high degree of sequence bias that is independent of the selection process and that needs to be corrected for.

Direct Inference of Motif Models

Instead of identifying the top sequences, it is possible to search directly for the PWM or PSAM that best explains the data. A commonly used model (MEME) is based on an expectation maximization algorithm (Lawrence and Reilly, 1990; Bailey and Elkan, 1994), that given a starting PWM computes the score for each possible motif start in a given sequence and thus produces a weighted alignment, which subsequently can be used to update the initial motif model. Those two steps are repeated until the algorithm converges. Using the more biophysical model represented by the PSAM, the MatrixREDUCE method also uses an iterative minimization algorithm to directly infer PSAMs from hybridization/expression data by computing the correlation of the log_2 expression and either the occurrence (Foat et al., 2005) or the exact score of a motif across a given sequence range (Foat et al., 2006). Another method that in its simplest form was already used in the early days of motif discovery (Stormo and Schneider, 1982) is a neural net. In recent years, with increasing computational

capacity, the neural net has become a widely used method for many applications, including motif discovery (Alipanahi et al., 2015; Ching et al., 2018).

Generalized Linear Modeling & NRLB

Since the kmer count tables obtained from high-throughput binding experiments, such as SELEX-seq, HT-SELEX or SMiLE-seq, exhibit a high degree of complexity and therefore, are difficult to both interpret and visualize, recent effort has been devoted to the development of algorithms that simplify the kmer tables based on the underlying equilibrium-thermodynamic selection process that generated the data. A simplified, yet generalizable and easily extendable method is described in Chapter 2 and 3 of this thesis. First, an initial seeding and alignment step is performed by scanning the most representative PSAM over the entire probe sequence. Only probes where a unique offset accounts for most of the selection process are retained; and sequences arising from non-specific binding or those where multiple, partial binding sites contributed to the selection in an non-linear manner are eliminated. In a next step, a generalized linear model is used to relate the counts for a particular binding site to a set of features:

$$\mathrm{f_1(S_i)} \propto \mathrm{f_0(S_i)} * \exp \frac{-\Delta\Delta \mathrm{G(S_i)}}{\mathrm{RT}} \tag{1.9}$$

where the frequency of a sequence i in round 1 of selection $f_1(S_i)$ is proportional to the frequency of the probe within the initial pool $f_0(S_i)$ and the relative affinity of the interaction $\Delta\Delta G(S_i)$. The $\Delta\Delta G(S_i)$ values can then be expressed as the sum of features X present at any given position ϕ within the binding site:

$$\frac{\Delta\Delta G(S_i)}{RT} = \sum_{\phi} \beta_\phi X_\phi(S_i) \tag{1.10}$$

where β_ϕ represents the weight of a given feature at position ϕ.

The approach above is an approximation of the true selection probability for a given probe, as it assumes that a single binding mode explains the observed selection and that flanking sequences do not contribute significantly. This might be valid for many applications, however, in order to model binding with high accuracy, a full model should be used, accounting for every possible way a TF can interact with a specific DNA sequence, including non-specific binding:

$$\mathrm{f_1(S_i)} \propto \mathrm{f_0(S_i)} \sum_{\mathrm{views(v)}} [\sum_{\mathrm{j}} \exp \frac{\Delta\Delta \mathrm{G_j(S_{iv})}}{\mathrm{RT}} + \exp \frac{\Delta\Delta \mathrm{G_{ns}}}{\mathrm{RT}}] \tag{1.11}$$

Each view in the above equation represents a distinct binding offset, including binding on either strand, and simplifies to equation 1.9 if a single view is considered and the non-specific binding term removed by requiring the selected view to contribute much more to the selection than any of the other views within a probe. Solving the above equation requires the use of a feature-based, log-linear multinomial model, which has recently been implemented with the No Read Left Behind (NRLB) algorithm (Rastogi et al., 2018). The features used in either the simplified or the full model can be nucleotide or dinucleotide indicators, DNA shape parameters, and even epigenetic DNA modifications. Thus, the model not only builds upon the underlying biophysical equilibrium conditions, but is also capable of incorporating a wide range of predictors. NRLB can not only predict accurate binding affinities over the entire sequence space, but it also allows inference of multiple, simultaneous binding modes (such as monomeric and dimeric binding contributions within a single experiment). Therefore, although it cannot be easily implemented and is currently only available for mono-

and dinucleotide and shape features, NRLB should be the method of choice when accurate quantification (such as the prediction of *in vivo* expression strength) is required.

Most motif discovery models used to construct PWMs rely on some form of seeding and a few sets of assumptions depending on the underlying algorithm. Which model is best suited, depends on the question at hand and perhaps, availability and the ease-of use. For a more detailed summary of different machine learning approaches used for motif discovery see (Li et al., 2015).

1.4 Secondary Mechanisms

Leaving the general concept of TF binding behind and considering the biochemical interactions taking place at the interface between TF and DNA - between residues of TF amino acids and the DNA bases and backbone - it becomes obvious that sequence features alone are only a proxy for the complex set of binding forces (electrostatic, steric and hydrophobic) at work. In Pabo's 1992 review article, the challenges to build accurate models for TF binding were already foreshadowed and can be summarized as the difficulty to characterize and quantify a three-dimensional interface in simple terms. If we want to truly capture binding specificity for a given TF, we have to find a model that fully encompasses the underlying biophysical properties of the interaction surfaces. Simplifying it to a sequence motif is a first, perhaps valid approximation, as the sequence dictates the biochemical landscape of the individual building blocks that make up DNA, however, it will not capture features of the interaction surface that go beyond mononucleotide recognition. For instance, neighboring or even longer, specific stretches of DNA base pairs can impact the interaction surface seen by a TF, via changing the structural properties of the DNA molecule and thus influencing the overall electrostatics or the accessibility of certain bases (Rohs et al., 2009a). Moreover,

any modification to individual DNA bases (or protein amino acids) will affect the binding interface and thus the affinity of a TF to a cognate site. Given the observation that structurally related TFs within the same family tend to recognize seemingly identical or highly similar motifs, there is a need to go beyond sequence identity. Rationalizing the concepts of DNA shape, DNA modifications, multi-TF complexes and their impact on TF binding are the first steps to improve our current understanding of TF binding and specificity. Ultimately, we need to incorporate all these aspects to properly address the concept of adaptive DNA binding – the various different conformations and readout strategies a given TF can deploy to interact with DNA in different contexts. Whether context is defined by variations in sequence, structure, cofactor-mediated, or epigenetic DNA modifications, TFs will adapt to allow optimal target site recognition. No "one-size-fits-all" conformation will be sufficient to characterize binding in different settings. The following will therefore give an overview about the different types of context-dependent TF binding:

1.4.1 DNA Modifications

Chemical modifications of DNA bases, most prevalently DNA methylation, is an ancient mechanism found in all three kingdoms of life. These epigenetic (i.e. beyond genetic) marks have diverse functions and mechanisms of regulation, which vary greatly among kingdoms and even species. This diversity, and the apparent lack of DNA methylation in commonly used model organisms such as *Drosophila melanogaster* or *Caenorhabditis elegans*, has made it challenging to identify a universal mechanism for this epigenetic mark. In prokaryotes, methylation (of adenines) is part of the restriction-modification system used to protect against foreign viral DNA by exclusively cutting foreign and not the methylated host DNA (Roberts et al., 2015). In plants, the dominant methylation mark is 5-methyl cytosine (5mC), which occurs in both CpG and non CpG contexts. In particular, the model organism *Arabidopsis thaliana* is an important resource for studying DNA methylation. For a detailed

review see (Huang and Ecker, 2017). As in mammals, the epigenetic pattern in plants is inherited through generations (Heard and Martienssen, 2014) and many strains exist to study both genetic and epigenetic inter-individual variability (Schmitz et al., 2013). Up to this date, several high-throughput methylation profiling techniques have been developed for this purpose (Laird, 2010) and one of them – whole-genome bisulfite sequencing (WGBS) – can map individual methylation marks at nucleotide resolution. An important result from the studies in plants was the finding that between early and late generations, the rate of spontaneous epimutations at methylated versus unmethylated CpGs was estimated to be roughly five orders of magnitude higher than that of spontaneous nucleotide mutations (10^{-4} versus 10^{-9} per site per generation) (van der Graaf et al., 2015). This result is interesting, as it argues for a rate high enough to allow uncoupling of genetic and epigenetic variation but still low enough to be subjected to selection across generations (Huang and Ecker, 2017). The dynamic nature of the methylation mark makes it both a promising candidate for playing a critical role in gene regulation, as well as a difficult subject to study due to the high degree of variability observed among different cell types or even among cells of the same type.

In animals the dominant modification is 5mC in a CpG context, with 60-80% of all CpGs being methylated in mammalian genomes, leaving high GC-content CpG island promoters aside (Edwards et al., 2010; Lister et al., 2009). On the other extreme, as mentioned above, some model organisms have lost methylation (^{5}mC context) all together, perhaps contributing to making those organisms a popular – since simple – study object.

Although the mechanism of DNA methylation in mammals and the machinery for establishing and removing ^{5}mC marks (Kohli and Zhang, 2013) is fairly well documented (Figure 1.7 on page 36), there is to date no consensus on the overall role and function of this mark.

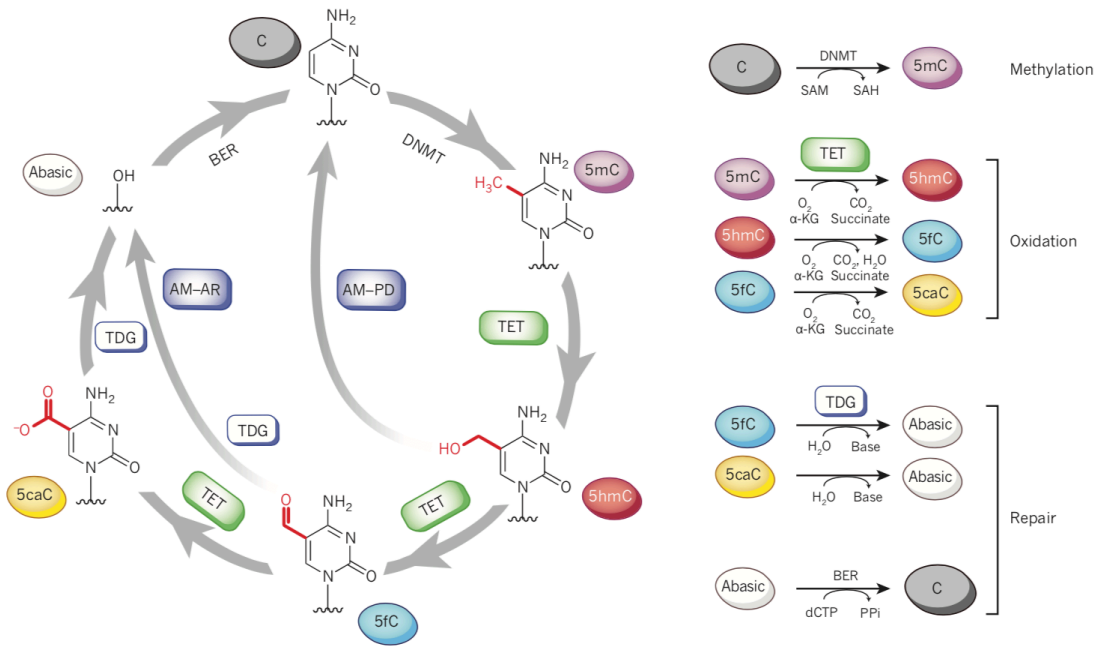


Figure 1.7: The Mammalian DNA Methylation Cycle: *(Figure by Kohli and Zhang, Nature 2013)* Cycle of establishing, erasing and reparing DNA methylation marks. Schematic illustrates the enzymes involved and the underlying chemical reaction type in each process (DNMTs provide the methylation machinery, TET enzymes the demethylation activity and TDGs (thymine DNA glycosylase) are involved in repair). AM, PD AR and BER stand for active modification, passive dilution, active restauration and base excision repair.

Among the few well supported functions are the regulation of allele-specific expression of imprinted genes (Bourc'his et al., 2001), the silencing of retrotransposable elements (Walsh et al., 1998), and control of X-inactivation via Xist RNA (Panning and Jaenisch, 1996). The roles in gene silencing, together with the observation that CpG-island promoters and in general promoters of highly transcribed genes tend to have no or low levels of methylation has led to the belief that there might be a direct role of methylation in the downstream expression of genes. However, the apparent negative correlation between methylation and transcript levels is not proof for a causal role for methylation in gene regulation. Arguments against

it are that some genes expressed in one but not another cellular context still display the same low level of promoter/enhancer methylation and that conversely, demethylation might rather be a consequence of high transcription (Bestor et al., 2015). In addition, embryonic stem cells (ESCs) that are mutant for DNA methyltransferase 1 (Dnmt1) (Li et al., 1992) or all known methyltransferases (Domcke et al., 2015) do not show activation of genes thought to be repressed by methylation, nor do they have wide-spread differences in their genome architecture as measured by DNase hypersensitive sites (Domcke et al., 2015) .

Another argument against the general role of methylation is the conservation of core regulatory components across species with and without methylation such as mice versus fruit flies. However, due to the simplicity of those few laboratory organisms, it is possible that redundant mechanisms have evolved to compensate for the lack of methylation. For instance, recent studies have identified low levels of ^{6}mA methylation in both *Drosophila melanogaster*, *Chlamydomonas* and *Caenorhabditis elegans* (Greer et al., 2015; Fu et al., 2015; Zhang et al., 2015), supporting the notion of a simplified methylation system or utilization of a different mark. In addition, the discovery of fully functional DNA methylation machinery in the invertebrate honey bee (Wang et al., 2006) has demonstrated that methylation is indeed widespread throughout animal taxa (Yi, 2017).

In another line of arguments, evidence for methylation playing a role in gene regulation at some level stems from the observations that Dnmt1-free ESCs have lost their capacity to differentiate and that mice with targeted mutations in Dnmt-1 are dying at embryonic stage (Li et al., 1992). In fact, there is no known differentiated cell type that is viable without functional methylation machinery, underlining the importance of methylation in genome stability. Interestingly, ESCs cultured in serum have much higher methylation levels than the inner cell mass (ICM) from which they are derived from (80% vs. 40%) (Ambrosi et al., 2017). Recent studies have indicated that serum-cultured ESCs might represent a mix of cells with varying degree of pluripotency and thus may be more "primed" for

differentiation than mESCs grown in *naïve* conditions (Habibi et al., 2013). The increase in methylation upon serum addition and the high degree of hypomethylation in the *naïve* state might provide an explanation for both the viability of methylation-free ESCs and the need of methylation in differentiation. Moreover, recent studies have indicated that the repressive histone mark H3K9me3 plays an important role in maintaining genome integrity by silencing retrotransposable elements in ESCs even upon removal of methylation marks (Ambrosi et al., 2017). This dual back-up implies that the pluripotent state is unique and that ESCs harbor epigenetic plasticity in preparation for differentiation, and thus need to rely on a redundant control mechanisms keeping them in the self-renewal state. This back-up mechanism might also provide an explanation why certain species have lost methylation by finding a way to bypass the differentiation requirement. The fact that DNA methylation can be both dispensable as well as indispensable, depending on the cell-type at hand, supports the idea that DNA regulation is context-dependent. It is plausible, that DNA methylation in mammals might be important for transitioning into new differentiated cell states, but takes a passive role once a new chromatin architecture is set up. This hypothesis is in line with the observation that methylation is indispensable for differentiation, with distinct methylation patterns observed in specific cell types and with the aberrant methylation profiles seen in many cancers (Jones and Baylin, 2007). Given the dependency of the methylation pattern on the underlying cell state, any input in the regulation of that state and thus the downstream gene regulatory network must show the same level of specificity, complexity, and uniqueness. Moreover, methylation marks need to somehow be readout to provide input in the setup of such a regulatory system.

When trying to identify a plausible mediator between methylation mark and chromatin state, DNA binding proteins, and specifically TFs, are a natural first guess. Not only do they bind DNA directly with methylation of a potential binding site inevitably altering the chemical properties of the interaction interface, but they also provide specificity and uniqueness to a given desired differentiation program. The above-mentioned role of DNA-

binding proteins in the methylation-mediated bacterial defense system is a good example of how methylation can impact binding of proteins to DNA. Both restriction enzymes that are indifferent to methylation and those that have blocked activity are known. Furthermore, DNA methylation has been shown to impact cleavage by DNase I by altering the shape of the recognized DNA site (Lazarovici et al., 2013).

Traditionally, DNA methylation is thought to block TF binding and thus negatively impact expression of downstream genes. However, due to the dynamic nature of methylation marks, potential cell-to-cell heterogeneity and the difficulty to assign a causal direction between transcription output and demethylation, the observed correlations are not backed up by hard evidence. One exception is again found in plants, where the existence of plant populations with well-characterized methylation profiles allowed identification of epimutations (affecting the methylation status of individual CpGs) that influence downstream gene expression. These methylation quantitative trait loci (meQTLs) were shown to overlap in 20-25% of cases with TF binding sites in cis (Huang and Ecker, 2017). More generally, regulation by TFs is temporally and spatially constrained, and specific to a given TF and cell type. Methylation patterns follow the same logic, so that by combining the two systems an even higher degree of specificity can be achieved. By the same token, the combinatorial nature increases the complexity tremendously, making it difficult to pin-point distinct gene-regulatory mechanisms that involve the recognition of specific methylation marks by TFs. In the past, most studies have analyzed global and aggregate effects of DNA methylation, without taking individual TF specificities and single-cell states into account. Moreover, rewriting chromatin states might involve previously inaccessible regions that generally have high levels of methylation but, since they are hard to study, are often omitted when searching for links between methylation marks and TF regulation of genes. It is thus important to first establish methods and models that accurately capture and quantify how methylation marks influence the interaction with specific TFs. Chapter 2 of this thesis will discuss the efforts we have made to characterize TF binding to methylated DNAs, and will outline a

method that allows accurate quantification and modeling of TF binding modulation by ^{5}mC (Kribelbauer et al., 2017; Yin et al., 2017).

To summarize, DNA methylation is a versatile mark for which there will likely be no one-size-fits-all mechanism that captures each individual aspect of its regulatory input. Studies have been mainly restricted to ESCs, which for reasons discussed above might not be an ideal model system to study methylation-dependent regulation of gene expression. Given the lethality of methylation removal in somatic cells and the many confounding correlated features, such as the given *in vivo* chromatin landscape or genomic sequence patterns, it is important to first establish *in vitro* based models of TF binding to methylated DNA. Only then, we might be able to shed light on the causal links between this mark and gene regulation. It is likely that only a few key TFs might be used to mediate between a given methylation pattern and the gene regulatory network, or that the methylation marks are influencing gene expression indirectly by remodeling the overall cell state and chromatin structure. In either case, specific proteins are likely to recognize the set up epigenetic marks, either by actively binding to them or by the failure thereof. The potential ^{5}mCpG recognition motif (RH motif) in many zinc-finger proteins (Blattler and Farnham, 2013) is yet another indicator that binding to methylated DNA might serve a specific function, whether to recruit histone remodelers, demethyltransferase or to directly influence gene transcription. This is particularly interesting, as zinc-finger TFs are a highly expanded class in humans.

1.4.2 DNA Shape

Double-stranded DNA (dsDNA) as seen by TFs *in vivo* is thought to be predominantly in B-DNA form, representing a more or less rigid structure defined by specific constraints and parameters. The initial structure of a dsDNA molecule was obtained by Rosalind Franklin using X-ray crystallography and interpreted by Watson and Crick (Watson and Crick, 1953).

While their work revolutionized the field and their finding was of "considerable biological interest" and indeed suggested a "possible copying mechanism for the genetic material", it had one limitation – it was inferred from a crystal structure, which imposes strict constraints on the conformation and symmetry of the entire set of molecules used for solving it. The first cue that DNA might, perhaps, not be as stiff and uniformly shaped as initially thought, was provided ten years later, by Kaarst Hoogsteen, who reported an alternative base pairing that differed from the one described by Watson-Crick (Hoogsteen, 1963). Hoogsteen base pairs would require helical dsDNA to adopt a shape substantially different to the one postulated by Watson and Crick and could thus impact sequence recognition by TFs. Although they were rarely observed, they were found in certain structures of protein-DNA complexes (Aishima et al., 2002) implying a potential role in TF binding. Only in recent years, with the advancements made in nuclear magnetic resonance (NMR) spectroscopy, it was possible to detect transient geometries in canonical duplex DNA that deviated from classical Watson-Crick base pairing and resembled the pairing described by Hoogsteen (Nikolova et al., 2011; Honig and Rohs, 2011). Despite the low population of those states, these new findings suggest that an equilibrium of different DNA geometries exists and a specific DNA shape could be recognized and trapped by DNA-binding proteins.

This experimental evidence served as an additional conformation for a line of research devoted to understanding the complex ways proteins interact with DNA, and how TFs make use of DNA's intrinsic shape. Similar to the belief that Hoogsteen base pairs are a rare species and thus not biologically relevant, protein interactions with the DNA minor groove have also been thought to confer no sequence-specificity in TF binding. As in contrast to the major grove, there is no hydrogen-bonding signature unique to specific base-pairs (Seeman et al., 1976). Therefore, sequence-specificity conferring interactions with proteins were thought to occur predominantly along the major groove (Pabo and Sauer, 1984). However, many crystal structures painted a different picture and showed that some protein-DNA complexes had undergone quite extensive structural changes in an effort to widen or narrow the minor

groove. One extreme example, with extensive hydrogen bonding occurring at backbone phosphates along the minor grove, but with complete absence of direct interactions with bases in the major grove that could have explained the sequence specificity, is that of the *trp* repressor-operator complex (Otwinowski et al., 1988). This "indirect" read-out was explained by *trp's* ability to detect variations in the geometry of the phosphate backbone, which itself was dependent on the underlying DNA sequence. Another mechanism by which proteins can sense the shape of the minor grove is by inserting positively charged arginines, as seen for the *D. melanogaster* Hox TF Sex combs reduced (Scr) (Joshi et al., 2007). That this property is not specific to Scr, but rather widely used in a range of proteins was then demonstrated by compiling and comparing all available structures of free DNA and Protein-DNA complexes (Rohs et al., 2009b) (Figure 1.8 on page 43). Calculating the average minor groove width for all available tertranucleotide sequences in free and bound DNA structures respectively, revealed a tendency of minor grove width narrowing in AT-rich sequences and an enrichment of arginines in exactly those locations. This sequence dependency of minor groove width, together with a few additional structural parameters were then systematically tabulated for all possible pentamers, using Monte Carlo simulations to obtain the required structural information on free DNA (Zhou et al., 2013). The tabulated tables allowed the high-throughput prediction of DNA shape features by using a sliding window approach.

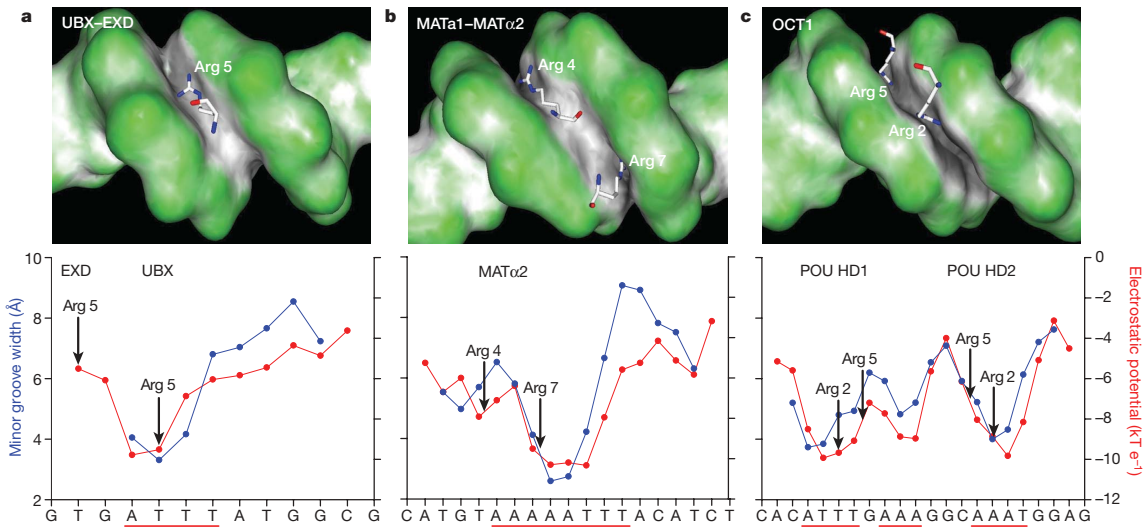


Figure 1.8: MGW Readout by Arginines: *(Figure by Rohs et al., Nature 2009)* **(a-c)** Examples for MGW readout by arginines as seen in different crystal structures. Below each structure, the minor groove width and the electrostatic potential along the DNA is plotted. Arginines shown in the structures are indicated with arrows.

With the ease of assigning a "shape" number for any given sequence, more and more studies included shape as predictors in their models for TF binding and often found that adding shape features could improve model performance (Mathelier et al., 2016). However, one caveat of using shape along with sequence features is that the two predictors are not independent of each other and thus should not be used simultaneously without properly accounting for the mathematical structure relating the two. Only recently, an in depth analysis of the latter could demonstrate that about fifty to sixty percent of variation seen in the pentamer tables can be explained by fitting a simple mononucleotide sequence model, with a dinucleotide model accounting for up to 98% (Rube et al., 2018). Therefore, it is advisable to fit a sequence model first, followed by a post-hoc analysis of the underlying shape readout, which will be further discussed in Chapter 3. To understand how TFs sense DNA shape one should not incorporate it into a model that explains which sequences a TF will bind to, but rather one should deduct the recognition mechanism from the preferably bound sequence

by using the tabulated sequence-to-shape relationships. Importantly, the ability to correlate bound sequences with their underlying shape features does not only aid in identifying binding mechanisms but can also be used to obtain information about the interface of protein-DNA complexes for which no structural information is available. And since TF-DNA crystal structures usually exist for only one or maybe a few select DNA sequences, such information is extremely valuable in obtaining a proper picture on the variety of subtly different structural conformations a TF will adopt when encountering different sequences. This will ultimately bring us one step further in inferring the complex rules governing protein-DNA interactions and in being able to specifically manipulate and design TF-DNA binding systems.

1.4.3 Cooperative Binding, Transient Interactions & Low-Affinity Sites

As a result of the absence of a unifying code linking TF amino acid arrangement and DNA sequence recognition, a wealth of structural as well as *in vitro* and *in vivo* binding data is now available for a large number of individual TF-DNA complexes. When comparing the predicted genomic binding sites for a given TF based on *in vitro* motifs to those actually bound by the TF *in vivo*, the latter ones generally fall behind the prediction (Slattery et al., 2014). This finding could be suggestive of a few things, namely, i) that there might be additional mechanisms *in vivo* that refine binding of individual TFs to a small subset that go beyond sequence recognition, ii) that not enough sites with high enough affinity to match the motif are actually available, and iii) that with the motif alone we are not capturing the entire sequence specifcity of a TF. Given the rather short sequence motifs, we can almost certainly exclude ii) since even though condensed chromatin can limit the accessibility of a binding site, there are still many accessible sites that remain unbound. Taking a look at the class of homeodomain TFs – a well characterized family of transcription factors that is important for development and conserved across species – we can find a good illustration

of this dilemma. In a high-throughput screen to identify the binding affinities of all 84 *D. melanogaster* homeodomain TFs, it was found that they can be sub-classified in 11 specificity classes that recognize specific variants of the about six base pair long sequence motif (Noyes et al., 2008). With the genome of fruit flies being about 120 million bases in size (and thus containing about the same number of hexamer motifs) and with $4^6 = 4096$ possible hexamers, you would expect about $30,000$ binding sites for each of the 11 specificity classes assuming a uniform sequence distribution and a zero-tolerance for binding site mutations. Even when restricting the genome to accessible regions, you would still expect to observe at least a few sites being occupied per open region. However, the high-confidence binding sites identified by ChIP-seq are more often in the order of a few thousand sites or less genome-wide. To complicate the picture even further, although amino acid variations within two regions of the 84 homeodomain DBDs specify different DNA sequence preferences (Noyes et al., 2008), the recognized motifs still share a high degree of overlap, in particular when comparing TFs within a specificity subclass. How the TFs achieve *in vivo* specificity and control expression of distinct sets of target genes is therefore an important question for which presumably no simple answer can be found.

In the past years, different studies have revealed a few ways how TFs fine-tune binding specificity *in vivo*. One, perhaps obvious, mechanism (that falls within the class i) explanation above) is the spatial and temporal separation of TF expression patterns, which for some cases can explain the observed binding selectivity of TF homologs but not for others, as demonstrated by the famous antennae-to-leg transformation observed when swapping homologous *D. melanogaster* Hox homeodomain factors (Lewis, 1978). An alternative explanation involves the divergent amino acid sequences outside a TF's DBD that can form interactions with other TFs. Such TF-complexes not only restrict the number of potential binding sites (two motifs instead of one) but they can also alter the monomeric binding preference of the TF upon dimerization. An example of such latent specificity is demonstrated by the altered binding specificities of the eight *D. melanogaster* Hox proteins upon complex

formation with their cofactor Extradenticle (Exd) (Slattery et al., 2011), but has also been described for many other factors (Jolma et al., 2015; Siggers et al., 2011). Cofactors could also act indirectly by recruiting TFs to specific genomic loci and, by trapping the TF there, create a specific "microenvironment" with increased local TF concentration (Reiter et al., 2017). As a consequence of boosting the local concentration, sites with lower *in vitro* affinity become available for TF binding. This concept of "transcriptional hubs" is supported when analyzing the organization of enhancers – genomic regions that regulate gene expression (Shlyueva et al., 2014; Lifanov, 2003). Although enhancers generally contain binding sites for many different TFs, only a subset of TFs will actually regulate enhancer activity (Arnone and Davidson, 1997). In order to identify those functional TFs, enhancers can be screened for the occurrence of multiple motifs for a single TF. The presence of such homotypic clusters of TF binding sites (HCTs) can serve as an indicator for functional binding and has been shown to be a common feature in enhancer architecture (Gotea et al., 2010). However, identifying an HCT can be tricky when using simple motif scoring as these clusters often consist of multiple low affinity sites that confer specificity by cumulatively acting on the enhancer (Crocker et al., 2015; Rastogi et al., 2018). For instance, Crocker et al. found that mutations in individual low affinity sites resulted in lowered gene expression, but mutations in two sites were required for complete abolishment. In addition, changing the affinity of a site from low to high results in robust but ectopic expression patterns (Farley et al., 2015), indicating that low affinity sites are necessary in order to confer specificity. Along those lines it has to be noted that a low affinity site is still several orders of magnitudes higher than non-specific binding (Rastogi et al., 2018) and therefore, the term "low-affinity" site is somewhat arbitrarily defined as a site whose affinity is lower than that of the consensus site and has enough mutations to not be readily recognized as such. The large number of TF binding sites within enhancers, the presence of "low-affinity" HCTs and the apparent trade-off between very high affinity and specificity for a specific TF all support a model where individual TFs concentrate in specific genomic loci and specifically recognize lower-affinity, yet

TF specificity-conferring sites to control and fine-tune gene expression. Additional support for this model comes from a recent study, that identified a correlation between the nuclear enhancer localization and the concentration of two TFs regulating the enhancer's activity (Tsai et al., 2017). The identification of functional low affinity sites that do not readily match a TF's consensus motif, has opened up the question whether amino acids thought to not be involved directly in DNA binding could confer additional specificity that perhaps, allows closely related TFs to distinguish their binding sites *in vivo*. Indeed, recent efforts in developing new experimental protocols and computational methods capable of accurately capturing additional binding specificity over an extended footprint, have shown that DNA bases flanking the core motif contribute significantly to TF binding specificity between TF homologs and even paralogs *in vitro* and *in vivo* (Rastogi et al., 2018; Le et al., 2018; Shen et al., 2018). The additional information in such extended TF binding models presumably stems from amino acids at the edge or outside of the DBD that bind DNA transiently or by utilizing different recognition modes and thus have not been identified by structural analysis or classical motif enrichment methods. The term "low-affinity binding sites" therefore might also need to be revisited, since with capturing the full range of specificity, seemingly low affinity sites might actually not be that low affinity after all.

Although initial data exists, proof of the broad relevance of such "transient" amino acid-DNA interactions for TF binding and TF specificity, has yet to be provided. Chapter 3 of this thesis will therefore be devoted to the in depth study of a tetrameric protein-DNA complex that serves as a model system to study the major aspects of context-dependent TF binding. We will investigate TF (and even isoform)-specific cooperativity, the importance of orientation and spacing of TF-complexes, sequence-preference within the DNA spacer that results from DNA shape recognition by amino acids at the edge of the DBD and finally, we will use the sequence-to-shape relationship to identify different conformational states or

recognition modes a TF can adopt or make use of depending on the underlying context (sequence or complex composition). The latter principle will be rigorously tested by the targeted design of protein mutations, thought to be responsible for the specific shape readout. Those mutations will allow us to differentiate different complex compositions, as well as distinct recognition modes a given TF complex can use to bind specific sequence classes.

The aspects discussed in this section suggest that TFs rely on a broad array of mechanisms and complex combinatorial logic to identify their cognate binding sites *in vitro* and *in vivo*, which go well beyond the classical concept of "direct" base readout. Only recently, we have started to develop experimental and computational tools to investigate some of these mechanisms in more detail. To what extent they influence *in vivo* TF binding specificity remains to be seen.

Chapter 2

Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes

The work described in this chapter is published:

Judith F. Kribelbauer, Oleg Laptenko, Siying Chen, Gabriella D. Martini, William A. Freed-Pastor, Carol Prives, Richard S. Mann, and Harmen J. Bussemaker

Cell Reports, Vol 19, Issue 11, p2383-2395, (2017)

Author Contributions:

Design, J.F.K., R.S.M., and H.J.B.; Experimental Methods, J.F.K. with help from O.L. and S.C.; Validation, J.F.K.; Data Analysis, J.F.K., Algorithm Development, J.F.K., G.D.M. and H.J.B; Writing – Original Draft, J.F.K.; Writing – Review & Editing, J.F.K., W.A.F-P., C.P., R.S.M and H.J.B; Visualization, J.F.K.; Supervision, R.S.M and H.J.B.; Project Administration, H.J.B.; Funding Acquisition, C.P. R.S. M. and H.J.B.

2.1 Summary

Although DNA modifications play an important role in gene regulation, the underlying mechanisms remain elusive. We developed EpiSELEX-seq to probe the sensitivity of transcription factor binding to DNA modification *in vitro* using massively parallel sequencing. Feature-based modeling quantifies the effect of cytosine methylation (^{5}mC) on binding free energy in a position-specific manner. Application to the human bZIP proteins ATF4 and C/EBPβ, and three different Pbx-Hox complexes shows that ^{5}mCpG can both increase and decrease affinity, depending on where the modification occurs within the protein-DNA interface. The TF paralogs tested vary in their methylation sensitivity, for which we provide a structural rationale. We show that ^{5}mCpG can also enhance *in vitro* p53 binding, and provide evidence for increased *in vivo* p53 occupancy at methylated binding sites, correlating with primed-enhancer histone marks. Our results establish a powerful strategy for dissecting epigenetic modulation of protein-DNA interactions and their role in gene regulation.

2.2 Introduction

High-throughput profiling of *in vitro* transcription factor (TF) binding specificities is a powerful approach for obtaining sequence motifs for a variety of TF families, and in several different organisms (Badis et al., 2009; Jolma et al., 2015; Weirauch et al., 2014). However, despite the growing number of known TF motifs, accurate prediction of *in vivo* TF binding and its effect on target gene expression has remained surprisingly difficult. One of the complications is that protein-protein interactions can modify the DNA binding specificities of transcription factors (Jolma et al., 2015; Slattery et al., 2011; Miller, 2009). Another potential complication is the existence of covalent modifications of DNA, particularly cytosine methylation (^{5}mCpG), which is widespread in vertebrates. Because of their potential to alter chromatin state (Hashimshony2003) or DNA shape (Lazarovici et al., 2013), an important and hotly debated question is to what extent DNA modifications can influence TF binding

and thereby contribute to changes in the epigenetic landscape and gene regulation. Such a regulatory mechanism is conceptually compelling, as DNA modifications could provide an additional layer of temporal and spatial control to fine-tune gene expression.

^{5}mCpG has been shown to be important in gene silencing in normal and cancer cells (Jones and Baylin, 2007; Stein et al., 1982), gene imprinting (Razin and Cedar, 1994), and X chromosome inactivation (Hellman, 2007; Tribioli et al., 1992). In spite of this progress, there is no general mechanism explaining the impact of DNA methylation on gene expression (Machado et al., 2015). Several studies have found that despite the overall association between promoter methylation and transcriptional silencing, some promoters can simultaneously be methylated and transcriptionally active (Gutierrez-Arcelus et al., 2013). In addition, systematic studies with cancer cell lines have found that aberrant methylation, such as hypermethylation of specific CpG islands, is a hallmark of cancer progression (Baylin and Jones, 2011; Paz et al., 2003). Recent studies have identified additional modifications such as ^{5}hmC and ^{6}mA in mammalian genomes, raising the possibility that these also influence gene regulation (Fu et al., 2015; Greer et al., 2015; Zhang et al., 2015). To identify the causal determinants of *in vivo* TF binding among all these correlated variables, detailed quantitative characterization of the effect of DNA modification on *in vitro* transcription factor binding is a prerequisite.

On a limited scale, the *in vitro* platform of protein binding microarrays (PBM) has been used to probe TF binding to methylated DNA probes (Hu et al., 2013; Mann et al., 2013). These studies demonstrated that ^{5}mCpGs can have both positive and negative effects on affinity. However, they were limited by the fact that the DNA arrays contained either fully methylated or fully un-methylated sequences (Mann et al., 2013), but not both in competition, or they only considered a select subset of sequences (Hu et al., 2013). In addition, the data analysis in these studies was restricted to oligomer-based methods, which makes it difficult to identify position-specific effects, especially for lower-affinity binding sites that deviate from the consensus motif. To study the effect of cytosine methylation on TF

binding at high resolution, a quantitative assay is required that allows for simultaneously probing of methylated and unmethylated DNA probes across all possible sequence contexts.

To address these issues, we developed EpiSELEX-seq, a method that uses a single round of gel electrophoresis to simultaneously assess binding to methylated and unmethylated DNA fragments, thus allowing methylation sensitivity to be analyzed for any TF or TF complex. We apply EpiSELEX-seq to human bZIP and Hox complexes, as well as tetramers of the tumor suppressor protein p53. Using a feature-based Poisson regression model, we quantify position-specific methylation effects on *in vitro* binding in the low affinity range. For p53, by jointly analyzing whole genome bisulfite sequencing and *in vivo* binding (ChIP-seq) data, we provide evidence that the increased *in vitro* affinity for specific DNA sequences due to methylation leads to enhanced occupancy *in vivo*. These sites of increased binding have a histone modification pattern associated with primed enhancers, supporting a role for p53 as a pioneer factor that can access methylated DNA sites.

2.3 Results

2.3.1 Affinity-based Selection from Mixed Pools of Methylated and Unmethylated DNA Ligands

To quantitatively assess the effects of DNA methylation on TF binding, we developed a method in which methylated (Lib-M) and unmethylated (Lib-U) libraries containing a randomized region of a desired length (16 bp or 26 bp) were first separately synthesized, each distinguished by a unique 4 bp barcode located near the variable region (Figure 2.2 A on page 55). After treatment of Lib-M with a DNA methyltransferase, both libraries were mixed in equal proportions, incubated with a TF of interest, and subjected to a single round of EMSA selection. Sequencing libraries were prepared from the library mix both before (R0) and after (R1) affinity-based selection (Figure 2.2 B on page 55 and Figure 2.1 A-B on page 54). For each sequenced DNA ligand, the barcode allows us to reconstruct the methylation status at the time of TF binding. For accurate affinity estimation, it is important that the two cytosines in each CpG base-pair step in Lib-M be fully methylated, as incomplete methylation would lead to underestimation of the impact of ^{5}mCpG on TF binding. We employed two separate tests to confirm full methylation: (i) methylation, bisulfite treatment, and sub-cloning of a test sequence containing four CpGs and (ii) high-throughput sequencing followed by dinucleotide analysis of a methylated library that was either treated or not treated with bisulfite. In the first test, we determined that optimal methylation efficiency is achieved after two successive rounds of methylation with $\geq 250\ ng$ of input DNA per reaction (Table 2.2 on page 85). Using larger amounts of DNA (e.g., the recommended 1 μg) resulted in incomplete methylation of the test probes. The short size of our probes ($\sim$ 50 bp) compared to typical genomic fragments (>1 kb) might be the source of this discrepancy because suboptimal conditions typically resulted in the methylation of either all four CpGs or none, arguing for a processive nature of the DNA methyltransferase. In the second test, bisulfite treatment of an unmethylated library of random 16-mers showed

depletion of all CpN dinucleotides, as expected (Figure 2.2 C on page 55). In contrast, under optimal methylation conditions, bisulfite treatment of a methylated library showed depletion of all CpN dinucleotides except CpG, which was recovered at levels identical to those observed in non-bisulfite treated, methylated libraries (Figure 2.2 D on page 55).

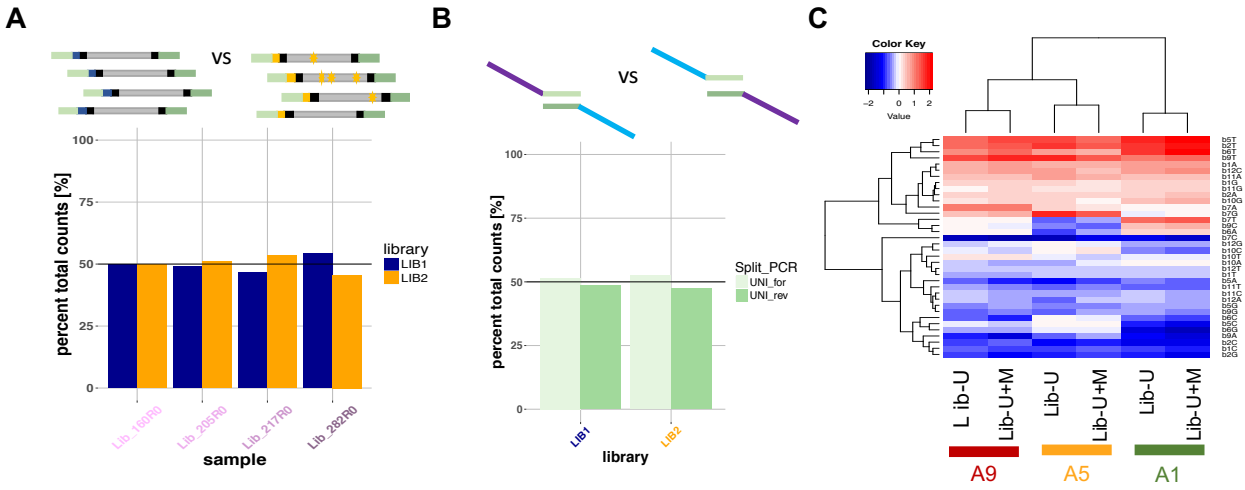


Figure 2.1: EpiSELEX-seq method and Feature-Based model performance:

Related to Figure 1 and Experimental Procedures: **(A)** Quality control for Lib-U and Lib-M proportions. Shown is the percentage of total reads that belong to Lib-U (blue) or Lib-M (orange) for four individually generated and mixed R0 pools. **(B)** Split-pool PCR. Shown is the percentage of total reads that originated from the UNI-for (light green) or UNI-rev (dark green) primer sets for each Lib-U and Lib-M. Both are roughly equally distributed, such that library diversity in the fixed flanks is maximal. **(C)** Heatmap analysis of $-\Delta\Delta G/RT$ base coefficient of either *i*) base feature (Lib-U only fit) or *ii*) the joint fit with ^{5}mCpG coefficients (Lib-U + Lib-M). Adding methylation coefficients does not change the base coefficients, arguing for a robust model.

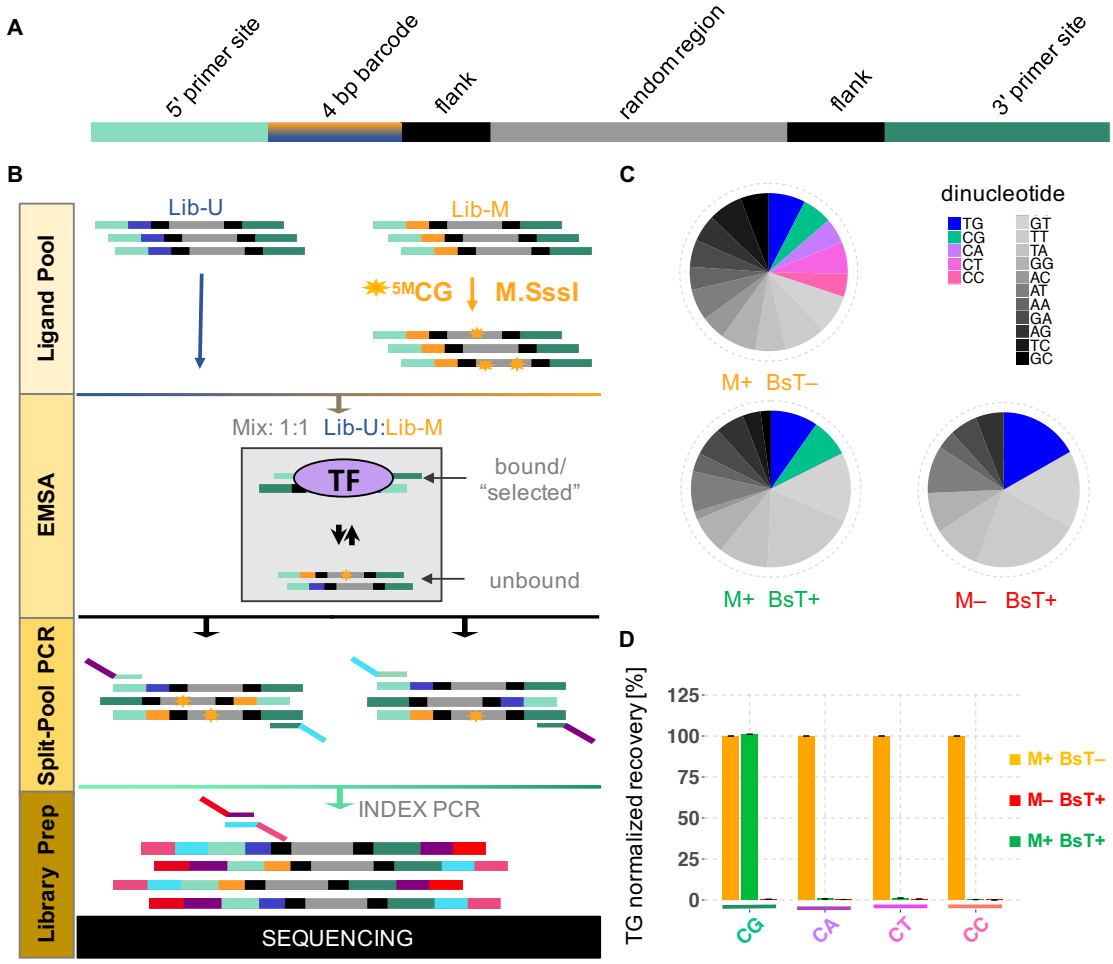


Figure 2.2: Overview and validation of the EpiSELEX-seq design: **(A)** Library design. 4 bp barcodes distinguish unmodified (Lib-U) and modified (Lib-M) DNA ligands. All libraries share a random region, reverse-complement-symmetric flanks and a pair of 5 and 3 primer sites. **(B)** EpiSELEX-seq workflow. Lib-M is methylated and mixed with Lib-U. The mixed pool is incubated with a TF of interest and the bound fraction is separated by an EMSA, purified, split, and amplified using two sets of primers. Unique Illumina barcodes are added for multiplexing. **(C)** Validation of methylation protocol. Shown are nucleotides frequencies in Lib-M after various combinations of optional methylation ($M+/M-$) and bisulfite treatment ($BsT+/BsT-$), determined by Illumina sequencing. The four CpN dinucleotides, for which the methylation status of the cytosine is unambiguous, are highlighted, as is TpG, which serves as a reference for CpN dinucleotides. **(D)** TpG-normalized recovery of the four CpN dinucletides. Only the CpGs protected by methylation are retained after bisulfite conversion.

2.3.2 EpiSELEX-seq Identifies Differences in Methylation Sensitivity Within the bZIP Family

To benchmark our method, we considered the basic leucine zipper (bZIP) transcription factors ATF4 and C/EBPβ, previously reported to be sensitive to DNA methylation (Mann et al., 2013). Many bZIP homo- or heterodimers preferentially bind to the cAMP response element (CRE) TGACGTCA and/or the C/EBP consensus TTGCGCAA (Figure 2.3 A-B on page 57). These palindromic sequences both contain a central CpG dinucleotide, creating the potential for methylation-sensitive DNA binding. For ATF4 homodimers, as expected, the relative enrichment of 10 bp sequences (encompassing the suspected TF footprint) that do not contain any CpG dinucleotides is similar between Lib-U and Lib-M (Figure 2.3 C on page 57). However, sequences that contain at least one CpG fall into distinct groups, each with a different ratio between Lib-M and Lib-U, indicative of a sensitivity to cytosine methylation that depends on the position of the CpG dinucleotide within the binding site (Figure 2.3 C on page 57) and (Figure 2.4 A on page 58). When a CpG base pair step is present at the center of the ATF4 binding site, methylation of both cytosines leads to a decrease in affinity. By contrast, sequences that contain a CpG in the flank of the motif (at positions $-3/-4$ or $+3/+4$) are bound much more strongly when methylated, leading to an alternative optimal left half-site, (^{5}mC)GAT. Interestingly, these methylation sensitivities are not observed for C/EBPβ (Figure 2.3 D on page 57), consistent with a previous observation that *in vivo* binding by this factor tolerates CpG methylation (Zhu et al., 2016). The methylation sensitivity for ATF4 is also reflected in the energy logos (Foat et al., 2006) that can be derived from the oligomer enrichment tables by considering all possible point mutations away from the optimal sequence (see Experimental Procedures). The logo derived from Lib-M, when compared to its equivalent for the unmethylated library (Lib-U), no longer has a central CpG as the most preferred sequence, and shows an increased preference for a CpG at position $-3/-4$ (Figure 2.3 E-F on page 57). Together, these findings demonstrate that sensitivity to DNA methylation can differ between paralogs from

the same structural family.

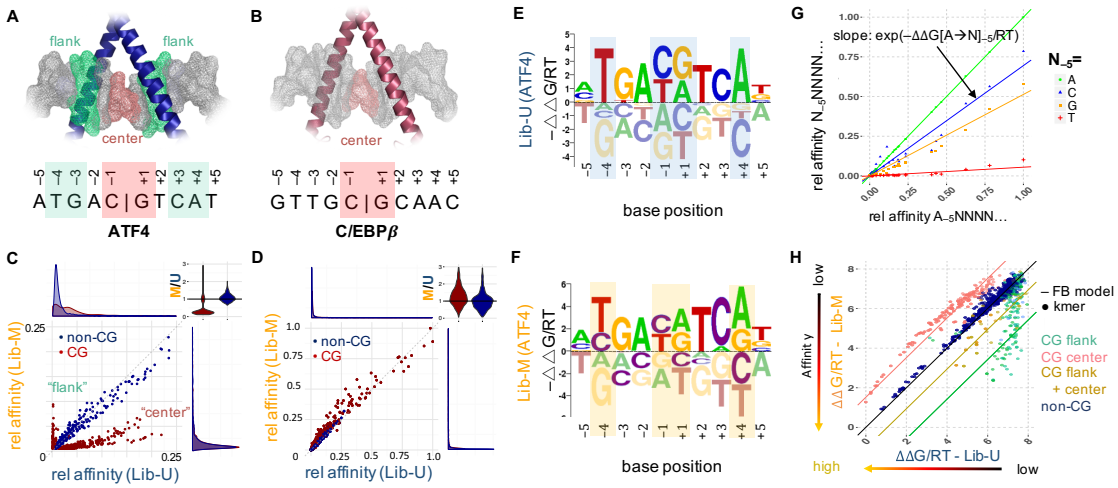


Figure 2.3: Probing methylation sensitivity for ATF4: **(A,B)** Crystal structure (PDB-ID: 1GTW) for the human bZIP homodimer C/EBPβ along with the symmetric consensus motif for ATF4 **(A)** or for C/EBPβ **(B)** and the definition of 'flank' (green) and 'center' (pink) positions in the binding sites. **(C)** Enlargement of low affinity range comparing the relative enrichment of 10 bp oligonucleotides between Lib-M versus Lib-U for ATF4. Non-CpG sequences (blue) show similar enrichment in both libraries, while distinct subsets of the CpG-containing sequences (red) are either preferred in Lib-U ("center") or in Lib-M ("flank"). **(D)** As in C but for C/EBPβ homodimers. Non-CpG and CpG-containing sequences show similar enrichments in both libraries across entire sequence range. Insets in C and D show the marginal distributions and the distribution of methylated/unmethylated ratio for all oligomers with a relative enrichment above 10^{-3}. **(E,F)** Energy-Logo for ATF4 derived from Lib-U (E) and Lib-M (F). The central CpG is no longer the top choice in the methylated library. 5mCpGs at the equivalent positions $-4/-3$ and $+3/+4$ appear as a new sequence feature in Lib-M. (G) Relative affinities (each point represents a 10 bp oligomer) containing either an A (reference base) or a point mutation (C, T, or G) at position -5. The slope of the lines represents the value of $\Delta\Delta G$ associated with each point mutation as estimated from the Lib-U read counts using a feature-based model. (H) Lib-M versus Lib-U 10-mer relative affinity plots in logarithmic scale. Lines represent the $\Delta\Delta G$ coefficients for the position-dependent methylation effects derived from the feature-based model.

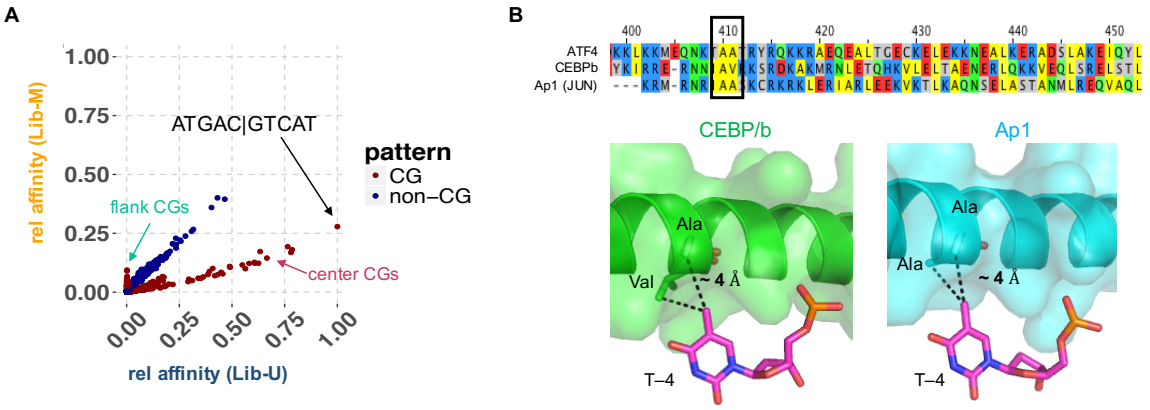


Figure 2.4: Methyl group read out by bZIP transcription factors:

Related to Figure 2.3 on page 57 **(A)** Comparison between relative affinities for Lib-M and Lib-U for ATF4 and 10-mers. Red points denote sequences containing a CpG. They fall into two groups, depending on where the CpG occurs within the protein:DNA interface. The top sequence for Lib-U is the palindromic CRE-site ATGAC|GTCAT. The affinity of the top site is decreased 4-fold upon methylation (relative affinity of 1 in Lib-U versus 0.25 in Lib-M). **(B)** The alignment shows conservation of pairs of hydrophobic amino acids (Ala-Ala or Ala-Val) across different bZIP TFs. The Ala-Val or Ala-Val pair interacts with the methyl group of a T at position −4 and is responsible for the positive effect on binding via methylation of a cytosine at position −4. Importantly, a G at position −3 is preferred by the proteins, and therefore the majority of observations for a C at position −4 will be in a CpG context with methylated cytosines.

2.3.3 Feature-Based Modeling Quantifies Position-Specific Methylation Effects

To systematically analyze the quantitative effect of cytosine methylation on binding affinity, we developed a feature-based generalized linear model to estimate the change in binding free energy associated with cytosine modification at any particular offset within the binding site. The frequency of DNA ligand S after one round (R1) of affinity-based selection, F1, is proportional to the frequency of the same probe in the initial (R0) pool, F0, as well as to the relative affinity of the interaction:

$$\mathrm{F_1(S)} \propto \mathrm{F_0(S)} * \exp^{[-\frac{\Delta\Delta \mathrm{G(S)}}{\mathrm{RT}}]} \tag{2.1}$$

We model $\Delta\Delta G(S)$, the difference in binding free energy between ligand S and the optimal ligand S_{opt} as a sum of contributions due to the specific (binary) features ϕ associated with:

$$\frac{\Delta\Delta \mathrm{G(S)}}{\mathrm{RT}} \equiv \frac{\Delta \mathrm{GS} - \Delta \mathrm{G(S_{opt})}}{\mathrm{RT}} = \sum_{\phi} \beta_{\phi} \mathrm{X}_{\phi}(\mathrm{S}) \tag{2.2}$$

Some features indicate the presence ($X = 1$) or absence ($X = 0$) of a specific base at a given position within the binding site, while others indicate the methylation status of a particular CpG dinucleotide. We estimate the corresponding coefficients β_ϕ from the data by fitting a generalized linear model based on counting statistics to the read counts in R1 while accounting for biases in R0 (see Experimental Procedures for details). To validate this modeling approach, we first inferred free energy effects for the three possible substitutions of the optimal base A_{-5} using the ATF4 homodimer data. Good agreement is observed with

the results obtained using oligomer enrichment (Figure 2.3 G on page 57). Next, we used an extended model that included features indicating methylation status. The coefficients from this fit indicate that methylation of $C_{-1}|G_{+1}$ represses binding ($\Delta\Delta G/RT = 1.5$, corresponding to $0.9\frac{kcal}{mol}$, or equivalently, a 4.5-fold reduction in affinity), consistent with the changes in oligomer enrichment between Lib-U and Lib-M (Figure 2.3 H on page 57). The coefficients for the equivalent flanking positions $C_{-4}|G_{-3}$ and $C_{+3}|G_{+4}$ are almost identical, as expected based on symmetry, and indicate a strong increase in binding due to methylation ($\Delta\Delta G/RT = -2.6$). Our model also predicts the combined effect of methylating both $C_{-1}|G_{+1}$ and $C_{+3}|G_{+4}$ (or $C_{-4}|G_{-3}$) by simply adding up the respective free energy coefficients (Figure 2.3 H on page 57).

2.3.4 Explaining the Effect of Cytosine Methylation by "Thymine Mimicry"

Although it has distinct base pairing preferences, ^{5}mC is chemically similar to thymine in that both have a methyl group at the carbon 5 position of the pyrimidine ring (Figure 2.5 A on page 61). Therefore, the total impact of a C to T transition on protein-DNA binding free energy, $\Delta\Delta G[C \rightarrow T]$, can be separated into (i) the effect of the methyl group alone, $\Delta\Delta G[\text{C} \rightarrow {}^5\text{mC}]$ and (ii) changes in charge and base pair interactions, $\Delta\Delta G[{}^5\text{mC} \rightarrow \text{T}]$ (Figure 2.5 A on page 61). Following this logic, the value of $\Delta\Delta G[\text{C} \rightarrow \text{T}]$ and $\Delta\Delta G[{}^5\text{mC} \rightarrow \text{T}]$, as estimated using the unmethylated (Lib-U) and methylated (Lib-M) library, respectively, can be subtracted from each other to obtain an estimate of the effect due to methylation $\Delta\Delta G[\text{C} \rightarrow {}^5\text{mC}]$. This approach was successful when applied to ATF4 to predict the effect of methylating the CpG dinucleotide, both at the central $(-1|+1)$ and the flanking $(-4|-3)$ positions (Figure 2.5 B-D on page 61). In agreement with these observations, many bZIP proteins contain two conserved hydrophobic amino acids that in crystal structures make van der Waals (VdW) contacts with the carbon 5 methyl group of

thymidine at position −4 in the binding site (Figure 2.4 B on page 58). ATF4, but not C/EBPβ, has a valine instead of an alanine at one of these positions, providing a possible mechanistic explanation for the increased preference of ATF4 for ^{5}mC over C, where the gain of a methyl group on the base may compensate for the loss of a methyl group in alanine compared to valine.

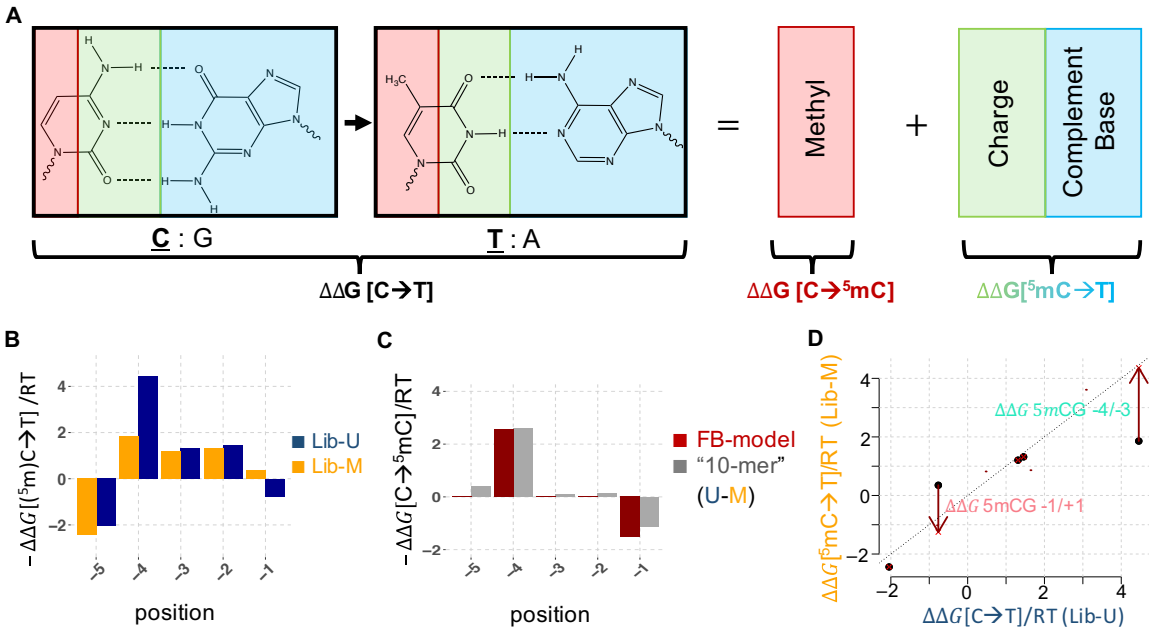


Figure 2.5: Deconvolving the methylation sensitivity for ATF4: **(A)** Decomposition of the position-specific DNA-protein binding free energy change associated with a $C \rightarrow T$ transition. The $C \rightarrow T$ change is the sum of C → ^{5}mC and ^{5}mC → T, allowing an interpretation of methylation sensitivity in terms of "thymine mimicry." **(B)** Change in binding free energy associated with C → T transition in each library as derived from an oligomer based PSAM. **(C)** Position-specific methylation effect on binding free energy, as estimated based on either the oligomer-enrichment-based approach (as in B; grey) or the feature-based-modeling approach (red). **(D)** The methylation effect as estimated using the feature-based model (red arrows) explains the differences in the C → T transition effect observed for Lib-U and Lib-M.

2.3.5 Deciphering the DNA Binding Specificity of Human Pbx-Hox Complexes

An important aspect of gene regulation is the capacity of TFs to form complexes with cofactors. A prominent example of such cooperative binding is that of Hox proteins and their three amino acid loop extension (TALE) cofactors, which play a crucial role in animal development (Merabet and Mann, 2016). As monomers, Hox family members bind to similar DNA sequences *in vitro*, but have distinct functions *in vivo*. Previously, we used SELEX-seq to capture the latent binding specificity of all eight *Drosophila* Hox proteins with their TALE cofactors Extradenticle(Exd) and the HM-isoform (HM) of Homothorax (Hth), which is required for optimal Exd-Hox interaction (Slattery et al., 2011). In mammals, where the Hox cluster has been duplicated several times in the genome, multiple cofactors from the PBC and MEIS class of TALE factors, as well as epigenetic DNA modifications, all have the potential to modulate DNA binding.

Here, we used EpiSELEX-seq to characterize the binding of human heterodimeric Pbx-Hox complexes to DNA (Figure 2.6 A on page 63). To cover the three Hox subclasses defined in Slattery et al. (Slattery et al., 2011), we performed these experiments using HoxA1, HoxA5 and HoxA9, each in complex with the cofactor PBX1, which was purified together with the HM domain of MEIS1. Comparing the pattern of 12 bp oligomer enrichment from R0 to R1 for each complex, we found similar cofactor-dependent differences in binding specificity between these Hox proteins as previously observed for their *D. melanogaster* orthologs (Slattery et al., 2011) (Figure 2.6 B on page 63 and Figure 2.7 A on page 65): the preferred central dinucleotide spacer (underlined) in the binding site consensus NTGAY<u>NN</u>AYNNN (where Y denotes C or T) is TG for anterior (Class I) factor HoxA1, TA for central (Class II) factor HoxA5, and TT for posterior (Class III) factor HoxA9 (Figure 2.6 D on page 63).

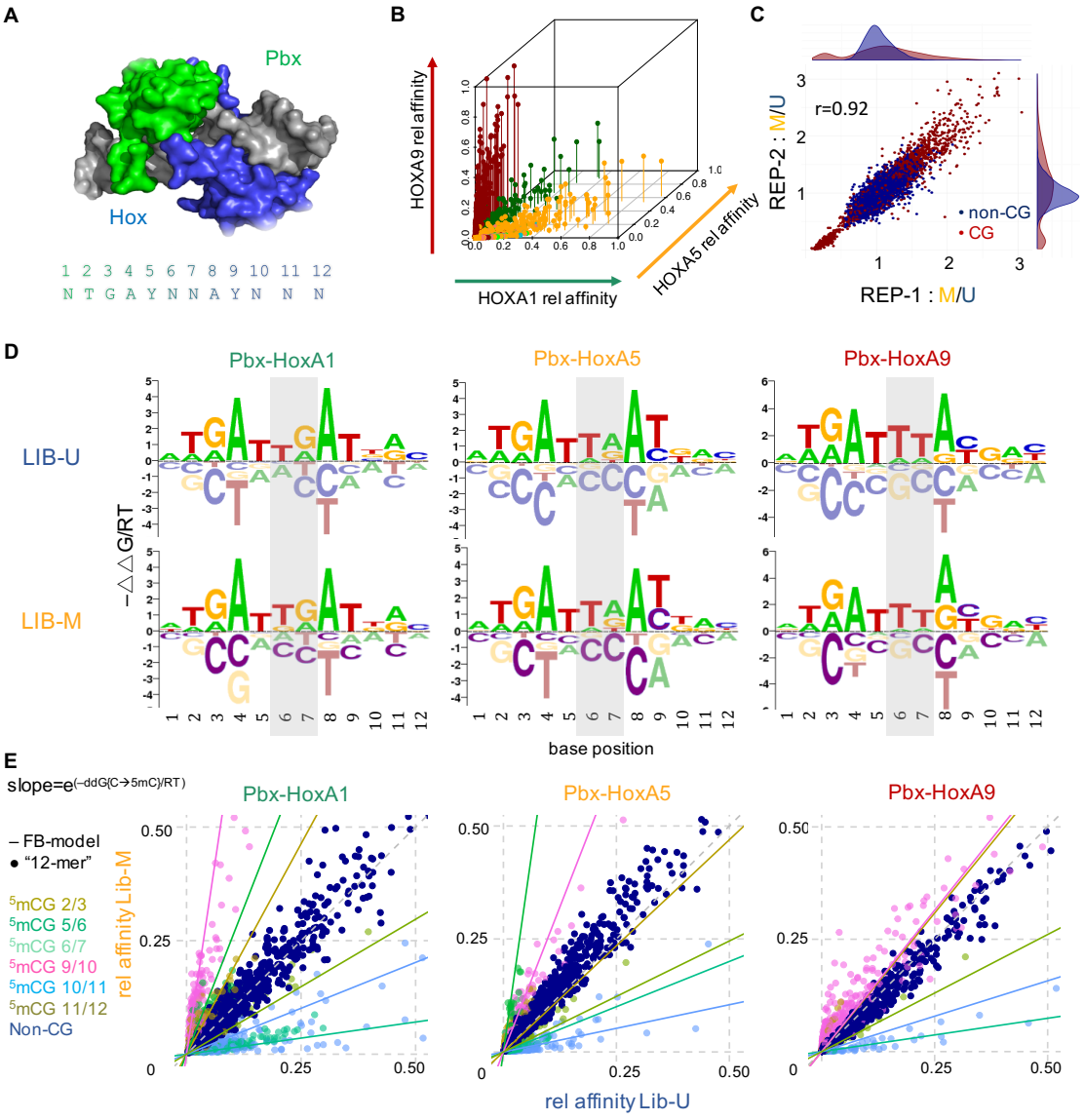


Figure 2.6: Methylation-sensitivity of human Pbx-Hox complexes:

Figure 2.6: *continued from page 63.*

(A) Crystal structure (PDB-ID: 1PUF) of human Pbx-HoxA9 with Hox shown in blue and Pbx1 in green. The consensus sequence with position labels is shown as a reference. **(B)** Relative affinity comparison of Pbx plus HoxA1, HoxA5 or HoxA9 (green, orange, red). Each Hox prefers distinct sets of 12-mers. Preferred central spacers (position 6 and 7) are TG, TA and TT for HoxA1, HoxA5 and HoxA9, respectively. **(C)** Replicate agreement for EpiSELEX-seq of Pbx1-HoxA9. Methylated/unmethylated (M/U) ratios for 12-mers are shown for one replicate versus the other. Sequences with or without CpGs are red or dark blue respectively. Pearson correlation of 0.92. Staggered density plots show a narrow distribution of non-CpG 12-mers around 1, but a much broader and bimodal distribution for CpG 12-mers. **(D)** Oligomer-based energy logos for all three Pbx-Hox complexes for Lib-U and Lib-M. No obvious differences between the methylated and unmethylated libraries are observed. Central spacer is shaded in grey. (E) Lib-M versus Lib-U relative affinity plots for all three complexes. Points are colored based on the position of the CpG dinucleotide (dark blue for non-CpG sequences). The slopes of the lines represent the exponentiated free energy coefficient for the methylation effect in the feature-based (FB) model.

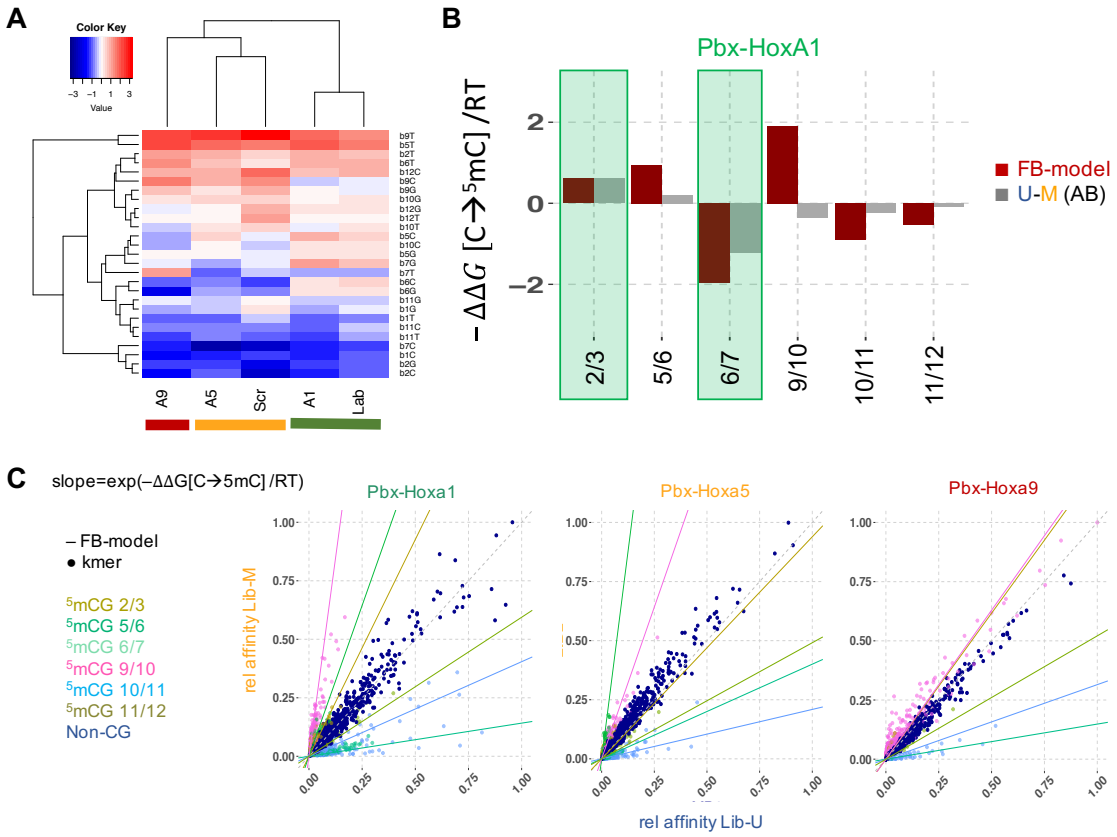


Figure 2.7: Binding preferences of Pbx-Hox complexes:

related to Figure 2.6 on page 63 **(A)** $-\Delta\Delta G/RT$ base feature coefficients of R1 data (Lib-U only) for fly and human Exd- or Pbx-Hox complexes are clustered together. Classes 1-3 are separated and orthologs not paralogs are closer to each other with regard to their sequence preference. (Abdb, the paralog of HoxA9, is not shown due to a low quality R1 dataset). **(B)** Energy differences for a C→^{5}mC transition at various CpG positions is shown for the Feature-based model (red) or as derived by taking the affinity-based difference in $\Delta\Delta G/RT$ (C → T) of Lib-U and Lib-M for Pbx-HoxA1. Only two positions (2/3) and (6/7) show agreement between the two methods. The positions both have a TG in the seed sequence used for the affinity-based analysis (ATGAYTGATTAC). All other positions do not capture the C→^{5}mC difference in the "kmer" model as they are followed by a non-G base in the seed sequence. The "U-M" difference is close to 0 as those CpNs are naturally unmethylated due to the non-CpG context. **(C)** Figure shows the full range of all 12-mer relative affinities in Lib-M over Lib-U. Compare Figure 2.6 E on page 63 for a blow-up of lower-affinity sites.

2.3.6 Human Pbx-Hox Dimers Show Position-Specific Methylation Sensitivity

The EpiSELEX-seq protocol allows us to assess the three human Pbx-Hox complexes for sensitivity to cytosine methylation. We first constructed separate energy logos for Lib-U and Lib-M by considering all possible point mutations from the most enriched 12 bp sequence (Figure 2.6 D on page 63). While paralog-dependent differences in the central spacer (shaded area) are readily apparent, the logos for the unmethylated (Lib-U) and methylated (Lib-M) human libraries are otherwise highly similar to each other and to those of their fly orthologs. However, this oligomer enrichment based approach is unable to detect methylation sensitivity for any cytosine that does not occur in a CpG context in the optimal sequence (Figure 2.9 on page 68). For ATF4, both cytosine positions at which methylation sensitivity was observed (-4 and -1) were fortuitously followed by a guanine (cf. Figure 2.3 A on page 57), but this is not the case for Pbx-Hox. Indeed, when we used our feature-based Poisson regression model to jointly analyze the Lib-U and Lib-M libraries in order to quantify the effect of ^{5}mCpG on binding, all three Hox proteins and Pbx showed significant methylation sensitivity at various positions throughout the binding interface (Figure 2.6 C,E on page 63) and Figure 2.7 B on page 65). The direction and amplitude of the methylation effect are highly position-dependent: methylation of CpG dinucleotides that start at positions 5 or 9 (underlined in the consensus sequence NTGA<u>Y</u>NNA<u>Y</u>NNN) enhance binding by several fold. In contrast, methylation of CpGs shifted by one position (positions 6 or 10, underlined in NTGAY<u>N</u>NAY<u>N</u>NN) decreases binding by up to 7-fold (Figure 2.6 E on page 63). This is reflected in both the energy coefficients (lines in Figure 2.6 E on page 63) and in the relative enrichment of 12-mers (points in Figure 2.6 E on page 63 and Figure 2.7 C on page 65). We tested these predictions using competition DNA binding experiments. Consistent with our EpiSELEX-seq analysis, using binding sites that contain a CpG at position 9/10 revealed that a higher concentration was required for unmethylated ($IC_{50} = 45.5 \pm 14.7$) than for methylated ($IC_{50} = 20.3 \pm 2.6$) binding sites to compete with a radioactively

labeled consensus probe for Pbx-HoxA1 binding (Figure 2.8 on page 67).

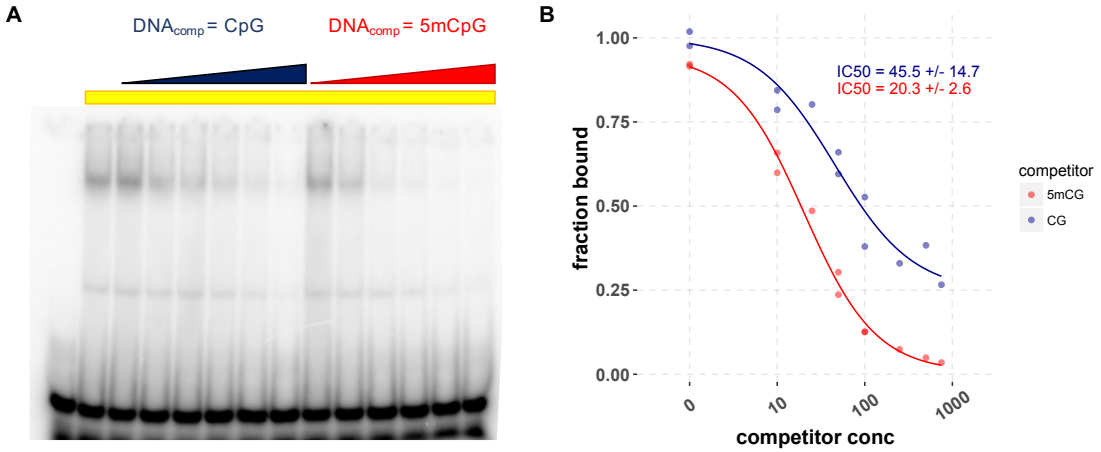


Figure 2.8: Competition assay for Pbx-HoxA1:

related to Figure 2.6 on page 63 and Figure 2.10 on page 70 **(A)** Example of a competition gel for two cold competitor probes – ATGATTGA**CG**AC (blue) and ATGATTGA^{5}**mCG**AC (red) – competing with a radiolabeled DNA ligand for Pbx-HoxA1 binding. Lane1: Hot probe only; Lane2: hot probe + complex; Lane 3-8: hot probe + complex + blue competitor DNA with increasing concentration; Lane 9-15: hot probe + complex + red competitor DNA with increasing concentration. **(B)** Dose-response curve: fraction bound, normalized by the no-competitor case, at increasing competitor concentrations. The methylated ATGATTGA^{5}**mCG**AC sequence has an IC_{50} value of $\sim$ 20 nM compared to an IC_{50} of $\sim$45 nM for the unmethylated ATGATTGA**CG**AC, indicating stronger binding ($\sim$ 2.2 fold) of Pbx-HoxA1 to the methylated probe.

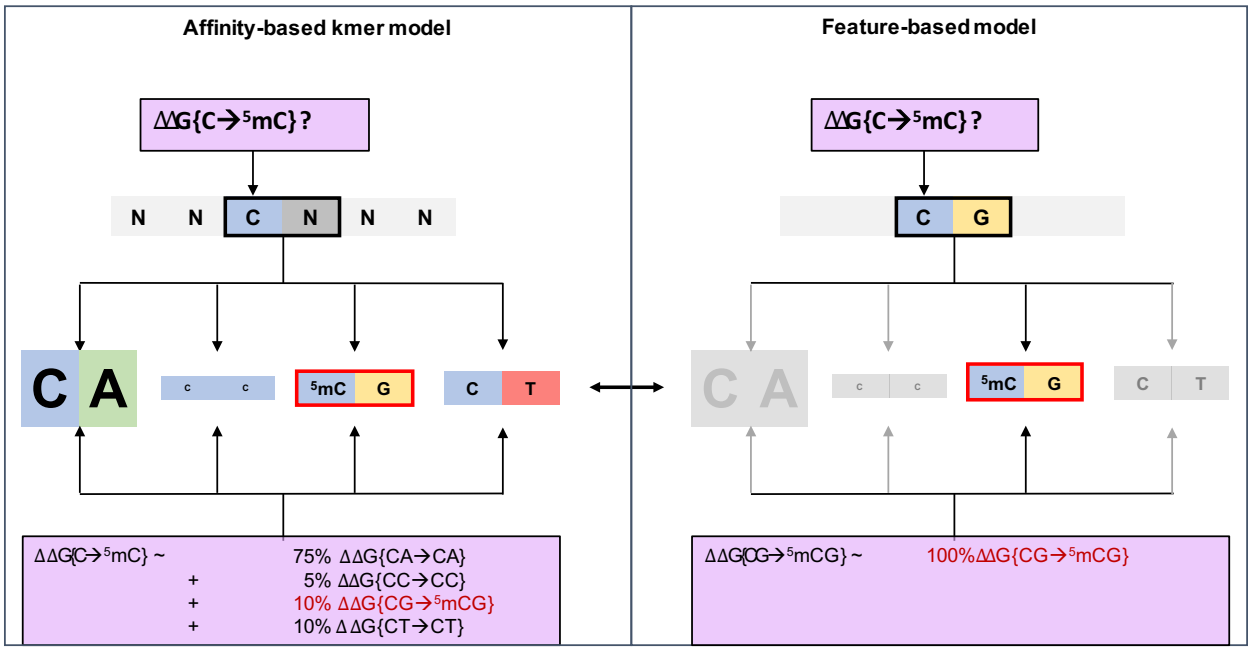
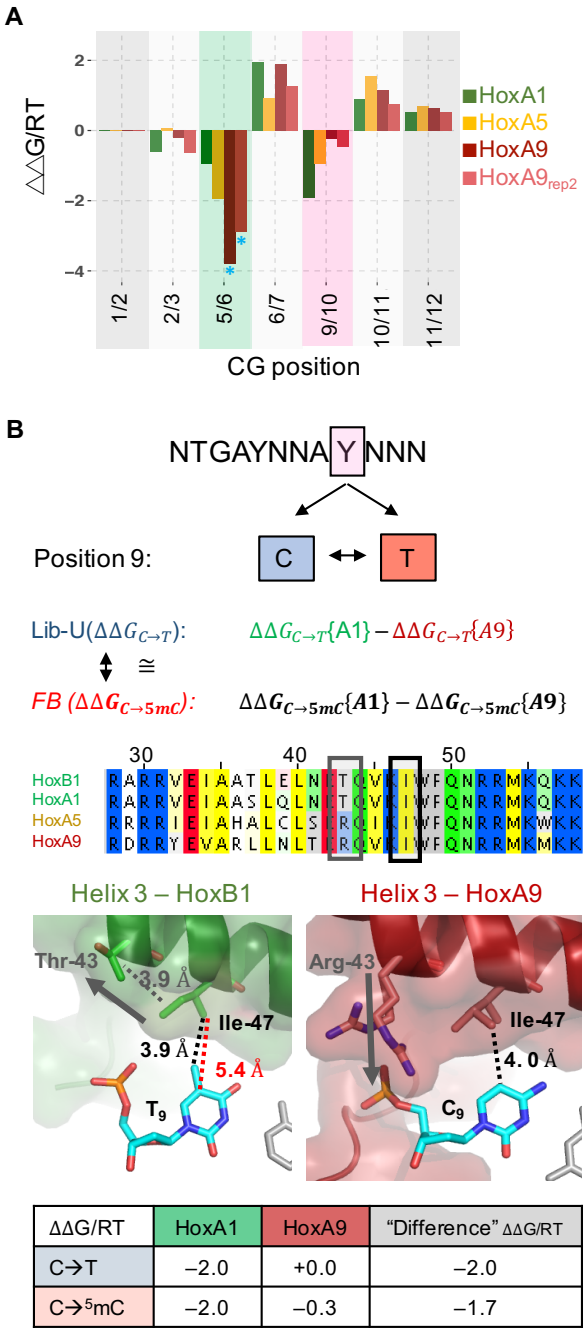


Figure 2.9: Sequence dependence of 5mCG free energy estimates in affinity-based models: *related to Figures 2.5 on page 61 and Figure 2.6 on page 63.* Estimating the effect size of a methylation mark using a "kmer"-based model (often the top kmer is used as a seed) might result in an averaging of the methylation effect, since the four CpN dinucleotides contribute differently to the observed C at the position of interest. Only in those cases where CpG (or TpG, as methylation can be seen as a "thymine mimicry") is the dominant sequence feature, the estimate will be accurate. Our feature-based model however, considers each CpG position in isolation, and estimates the methylation effect by comparing identical sequences from the unmethylated and methylated libraries, and is thus independent of sequence context.

2.3.7 Thymine Mimicry Explains Variation in Methylation Sensitivity Among Hox Paralogs

The effect of methylation on binding not only depends on the position of the CpG dinucleotide within the protein-DNA interface but also differs between Hox paralogs (Figure 2.10 A on page 70). At dinucleotide positions 5/6 and 9/10 the strength of methylation sensitivity is collinear with the Hox expression domain along the anterior-posterior axis (HoxA1-HoxA5-HoxA9), similar to other aspects of Hox function (Slattery et al., 2011). To gain more insight into the structural mechanisms underlying these differences in binding, we compared HoxA1 and HoxA9, which show distinct differences in methylation preference at position 9: Pbx-HoxA1 strongly prefers T over C ($\Delta\Delta G$[C $\rightarrow$ T] / RT), while Pbx-HoxA9 shows no such preference (Figure 2.10 B on page 70). Close examination of a Pbx-HoxB1 (a proxy for HoxA1) crystal structure reveals that isoleucine at position 47 (Ile47) within the homeodomain has a VdW interaction with the carbon 5 methyl group on base T9 of the forward DNA strand (Figure 2.10 B on page 70). In contrast, in a Pbx-HoxA9 crystal structure Ile47 is closer to and interacts with the C9 base, even without this methyl group. Accordingly, we would predict that Hox$B1/A1$ should benefit from the methylation of a C9, whereas HoxA9 should be indifferent to methylation. Indeed, $\Delta\Delta G$[C9 $\rightarrow$ T9] / RT is similar to $\Delta\Delta G$[C $\rightarrow$ ^{5}mC9] / RT for HoxA1 whereas for HoxA9 $\Delta\Delta G$[C9 $\rightarrow$ ^{5}mC9] / RT is close to zero (Figure 2.10 B on page 70). As the crystal structures show no further base-specific interactions at position 9, these differences can be fully accounted for by the relative benefit of gaining a methyl group for each paralog.



ΔΔG/RT	HoxA1	HoxA9	"Difference" ΔΔG/RT
C→T	−2.0	+0.0	−2.0
C→^{5}mC	−2.0	−0.3	−1.7

Figure 2.10: Collinearity of methylation sensitivity explained by structural differences:

Figure 2.10: *continued from page 70.*

(A) Comparison of the methylation effect for all three Pbx-Hox complexes. The two A9 replicates are shown in different shades of red and have good agreement (blue asterisks indicate that coefficients were fit at sub-optimal affinity thresholds due to low counts). Position 9/10 shows large paralog-dependent differences, with HoxA1 having high, HoxA5 medium, and HoxA9 almost no methylation sensitivity; position 5/6 shows the opposite trend. **(B)** Comparing Hox-specific C or T read-out for position 9. HoxA1 prefers a T over a C, whereas HoxA9 has equal preference. The observed difference in binding free energy associated with a $C \rightarrow T$ transition should equal the methylation sensitivity difference between HoxA1 and HoxA9. Alignment of helix3 of several Hox TFs (B1,A1,A5,A9) reveals conservation of Ile47 for the Hox family, but polymorphism at residue 43. Ile47 interacts with the pyrimidine at position 9 in both the HoxB1 and the HoxA9 structures. The distance to the aromatic carbon (C5) is 5.4Å for HoxB1, but only 3.9Å for HoxA9. Addition of a methyl group in HoxB1 reduces the distance to 4.0Å, allowing for the same VdW interaction as seen in HoxA9. Arg43 (A9) aids in bringing Ile47 closer to the DNA by interacting with the phosphate backbone at nucleotide C9, whereas Thr43 (B1/A1) does not interact with the backbone, but rather pulls Ile47 away from T9. The $C \rightarrow T$ energy difference between HoxA1 and HoxA9 is most likely driven by the methyl read-out. The table shows that the CT free energy difference is comparable to the difference in methylation sensitivity (feature-based model) between the two paralogs.

2.3.8 EpiSELEX-seq Identifies Non-Consensus P53 Binding Sequences Whose Affinity is Increased Upon Methylation

Because altered methylation patterns are observed in many cancers, we tested if binding by the human tumor suppressor protein p53 might be methylation sensitive. In vivo, p53 is thought to bind as a tetramer to two dimer sites RRRCWWGYYY (which we will refer to as CWWG) separated by a spacer of 0-13 bp (El-Deiry et al., 1992; Funk et al., 1992) (Figure 2.11 A on page 73). Consistently, the palindromic sequence GGACATGTCC site independently emerged from our data as the most enriched 10-mer in both Lib-M and Lib-U (Figure S6A). Comparing Lib-M and Lib-U directly reveals that there are three different classes of CpG-containing sequences that show altered p53 binding upon methylation (Figure 2.11 A on page 73). Methylation of a CpG occurring at the 3' end of the half site (RRRCATGY**CG**, which we will refer to as $C_{+}4|G_{+5}$, relative to the motif center) decreases binding by $\sim 20\%$, while methylation at a CpG shifted one bp to the left (RRRCATGCGY or $C_{+3}|G_{+4}$) increases binding by $\sim 50\%$. The largest effect, a $\sim 250\%$ increase in binding affinity, was observed when the CpG is in the core of the binding site (RRRCA**CG**YYY or $C_{+1}|G_{+2}$). Analysis of a p53 crystal structure (3Q06; (Petty et al., 2011)) reveals that the methyl group at carbon 5 of the T_{+1} base pyrimidine ring in the CATG core is stacked above the polar guanidinium plane of p53 amino acid R280. The latter is crucial for p53 binding as it forms a hydrogen bond with the G_{+2} base (Figure 2.11 B on page 73). The thymine methyl group might thus direct and constrain R280 towards G_{+2}, which has been proposed to serve as a methylation readout mechanism of zinc finger proteins (Liu et al., 2013). $T_{+1} \rightarrow C_{+1}$ replacement would thus eliminate the guiding methyl group, providing an explanation for the stabilizing effect of methylation at position C_{+1}.

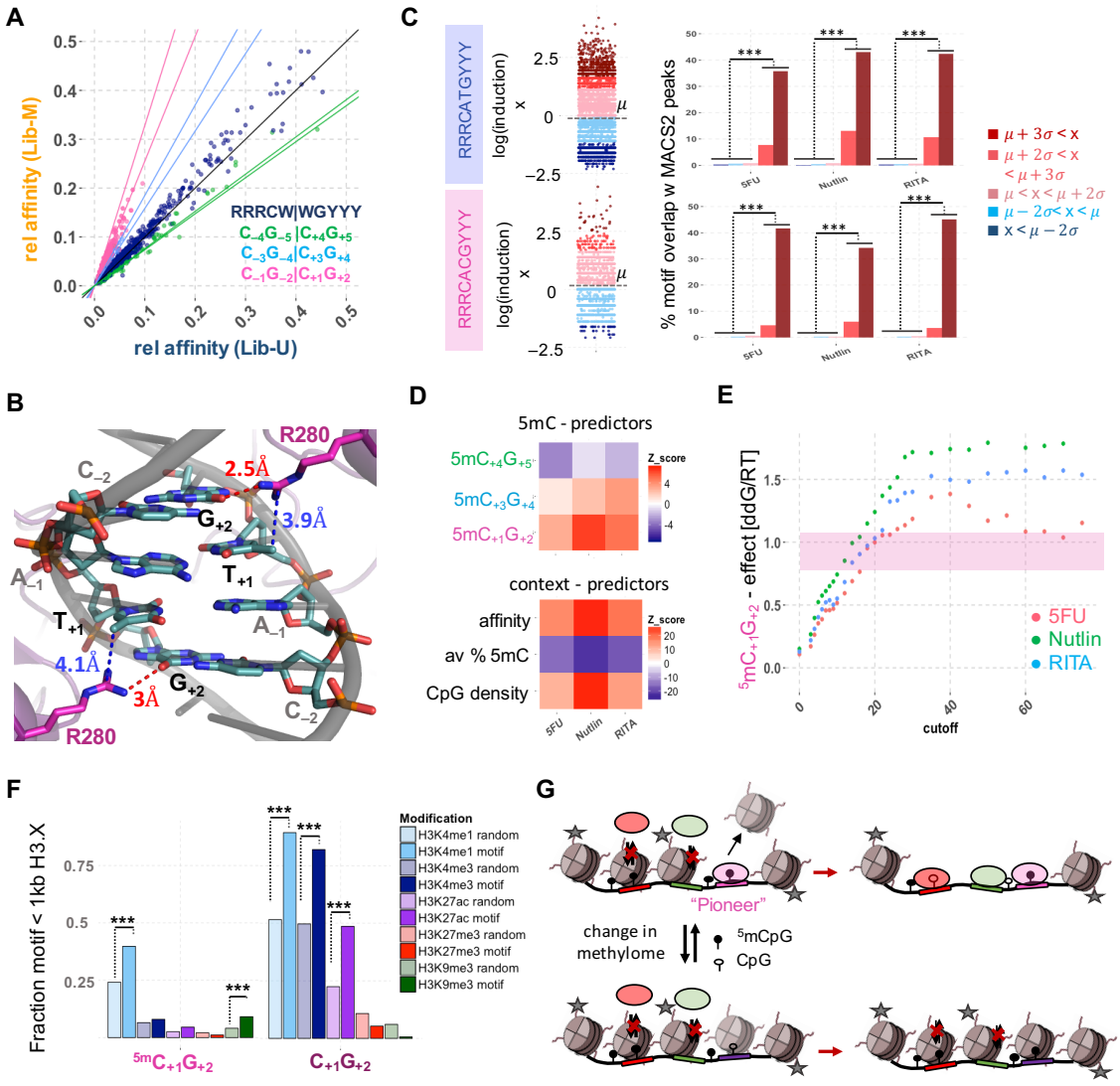


Figure 2.11: p53 differentially binds methylated motifs *in vivo* in distinct chromatin modification states: **(A)** EpiSELEX-seq 10-mer relative affinity plot showing the consensus motif (RRRCWWGYYY; blue) and 3 classes of CpG-containing motifs. CpG motifs are differentially bound upon methylation, with methylation of a) $C_{+4}|G_{+5}$ (green) halfsites reducing binding about 20%, whereas methylation of b) $C_{+3}|G_{+4}$ (cyan) and c) $C_{+1}|G_{+2}$ (pink) sites increases binding ~ 1.5 and $\sim 2-3$ fold respectively.

Figure 2.11: *continued from page 73*

Non-CpG consensus sites, as expected, show no difference between Lib-U and Lib-M. The slope of the lines represents the value of $\Delta\Delta G$ associated with methylation at each of the identified CpG positions using the feature-based model; methylation effects related by reverse-complement symmetry, estimated independently, are shown as separate lines. **(B)** p53 structure (PDB-ID, 3Q06) showing the DNA interface of a p53 dimer with the RRR**CA**|**TG**YYY core (labeled ± relative to the motif center). The two arginines (R280) form hydrogen bonds with the respective G_{+2} bases of each pentamer half sites (2.5 and 3Å; red) guided by the methyl groups of the pyrimidine carbon 5 of the T_{+1} base, which stack on top of the polar guanidinium plane (3.9 and 4Å; blue) thus constraining the possible orientations of the positive charge in favor of forming hydrogen bonds with G_{+2}. Methylation of a $T_{+1} \rightarrow C_{+1}$ substitution would therefore result in stabilization due to the regain of the position +1 methyl group. **(C)** Comparison of motif-centric analysis and MACS2 peak calling. Left panel: Distribution of log-transformed induction levels (drug-induced over uninduced) for all covered CATG or $C_{+1}|G_{+2}$ sites. Right panel: Fraction of decamer sites overlapping with MACS2 peak regions split by their log-transformed induction. For all three drugs and both the consensus CATG and the $C_{+1}|G_{+2}$ motifs there is a highly significant trend between motif-centric induction levels and MACS2 peak calling **(D)** Feature model fits of drug-induced (5FU, Nutlin, RITA), *in vivo* P53 ChIP-seq data for MCF7 using Lib-U relative affinities, average methylation levels and CpG density within a 500 bp region as context-dependent predictors and three position-specific binary methylation indicator features. Datasets were sub-sampled to 50 sites for each possible methylation-motif combination (see Experimental Procedure for details). Upper panel shows the significance of the the methylation features with red signifying positive and blue negative effects on binding. Z-scores for $C_{+1}|G_{+2}$ ranges from 3.0 (5Fu) to 6.3 (Nutlin). Lower panel shows the scores for the context dependent, confounding model predictors (highly significant across all drugs).

Figure 2.11: *continued from page 74*

(E) methylation coefficient for the most significant $C_{+1}|G_{+2}$ site was computed for increasing cut-offs on the sum of uninduced and drug-induced p53 IP coverage. Pink area shows the expected difference in binding free energy from EpiSELEX-seq results. (F) Overlap with peaks of histone modifications (< 1 kb) for methylated and unmethylated $C_{+1}|G_{+2}$ motifs (> 2 sd above mean induction, dark shade). Equally sized, methylation-matched random control sets (light shade) show the expected overlap. Primed-enhancer (H3K4me1) and heterochromatin (H3K9me3) modifications but not marks of active transcription are significantly enriched in methylated $C_{+1}|G_{+2}$ sites whereas unmethylated $C_{+1}|G_{+2}$ sites show patterns of active transcription (H3K4me1; H3K4me3; H3K27ac), perhaps reflecting increased accessibility at active promoters. **(G)** Potential mechanism how aberrant methylation patterns might contribute to altered p53 binding and thus potentially contribute to changes in chromatin landscape and gene regulation.

2.3.9 Evidence for Enhanced p53 Binding to Methylated Sites *In Vivo*

When unmethylated, sequences of type $C_{+1}|G_{+2}$ are bound by p53 at a relative affinity of $< 10\%$. However, our analysis shows that binding to these sites is strongly enhanced by cytosine methylation. To test whether this effect on in vitro binding is also observable *in vivo* we jointly analyzed whole-genome bisulfite sequencing (Consortium) and p53 genomic occupancy data – generated by ChIP-seq both before and after induction of p53 – for the cell line MCF7 (Nikulenkov et al., 2012). Using standard peak calling (Zhang et al., 2008) at a false discovery rate of 5%, we detected 40 sites that were both occupied by p53 and had an underlying DNA sequence containing a match to RRRCACGYYY, a sample too small to allow for statistical analysis of the effect of methylation status (Figure 2.12 B on page 78). Moreover, the negative effect of methylation on chromatin accessibility in vivo may obscure the positive effect on binding suggested by our SELEX analysis. To address this issue, we developed a motif-centric analysis strategy that avoids peak calling. We started by

identifying all individual matches to the most strongly bound RRRCATGYYY sites in the genome, and classifying each of these p53 half-sites in terms of the change in the number of ChIPed DNA fragments covering it before and after p53 induction. We observed a strong and statistically significant trend between motif-centric fold-induction and the probability of falling within a peak region based on MACS2 (Figure 2.11 C on page 73), indicating this approach captures the underlying p53 binding signature. In addition, this trend was robust for three different inducers of p53 activity, and was also observed for the CpG containing $C_{+1}|G_{+2}$ motif (Figure 2.11 C on page 73).

Encouraged by this observation, we used a generalized linear model that explains how the number of sequenced IP fragments covering an individual genomic match to any of the four decamer half-site motif classes (CATG, $C_{+1}|G_{+2}$, $C_{+3}|G_{+4}$, $C_{+4}|G_{+5}$) is distributed between the uninduced and induced conditions. The CATG motif, which does not match any CpG-containing decamers, serves to estimate the effect of local chromatin context, which is represented by the average methylation level and CpG content of the flanking regions as predictors in the model. To account for variation in binding affinity unrelated to methylation, we also included as a covariate the relative affinity of the 10bp half-site as derived from the DNA sequence using a scoring matrix derived from our Lib-U data (see Experimental Procedures and Supplemental Experimental Procedures for details). Finally, and most importantly, the coefficients associated with three binary indicators for the presence of a methylated CpG dinucleotide at each offset quantify the effect of cytosine methylation on the responsiveness of *in vivo* p53 binding.

When the model is fit to ChIP-seq data, the position-dependent effects of cytosine methylation within the binding site identified by our EpiSELEX-seq assay are recapitulated in MCF7 cells, with methylation of $C_{+1}|G_{+2}$ having a significant stabilizing effect (Figure 2.11 D on page 73). The coefficients for the confounding contributions in the model also behave as expected, with positive effects for CpG density and sequence-derived p53 affinity, and a negative effect for regional methylation (Figure 2.11 D on page 73). Considering that

the *in vivo* methylation effects should more closely reflect the *in vitro* effect at higher levels of ChIP enrichment, where the local chromatin context presumably is more permissive, we repeated our model fit using increasing cutoffs on the sum of induced and uninduced read counts for all consensus matches in the genome (Figure 2.11 E on page 73). The coefficient for $C_{+1}|G_{+2}$ behaves as expected, and saturates at $\Delta\Delta G/RT = +1.5$, corresponding to a ~ 4.5-fold increase in binding affinity upon full methylation of the CpG dinucleotide (Figure 2.11 E on page 73). Thus, the *in vivo* methylation effect appears to be even higher than *in vitro*, which could reflect contributions from additional methylated CpG dinucleotides within the full p53 tetramer binding site or cooperativity with other factors. For the other two motif classes ($C_{+3}|G_{+4}$ and $C_{+4}|G_{+5}$), the coverage by IP fragments is too sparse to allow quantification, consistent with the weaker *in vitro* methylation sensitivity observed for these CpG offsets with our EpiSELEX-seq assay.

It has been suggested that p53 can bind to high-nucleosome-occupancy regions and act as a pioneer factor to alter chromatin accessibility (Laptenko et al., 2011; Sammons et al., 2015). We therefore analyzed five histone modifications that in combination can be used to classify enhancers or promoters as active, closed, or primed (Calo and Wysocka, 2013). Methylated $C_{+1}|G_{+2}$ sites are significantly enriched for H3K9me3 and H3K4me1 but not H3K27ac (associated with active enhancers) or H3K4me3 (associated with active transcription), when compared to a matched control set (see Experimental Procedures for details) (Figure 2.11 F on page 73). These histone modifications have been suggested to mark either heterochromatin (H3K9me3) (Grewal and Jia, 2007) or enhancers that are primed to become active (H3K4me1) (Calo and Wysocka, 2013). We observed the same pattern for CATG sites within methylated regions (Figure 2.12 C on page 78). By contrast, unmethylated $C_{+1}|G_{+2}$ sites tend to have a strong signature of H3K4me1 and H3K4me3 or H3K27ac (Figure 2.11 F on page 73), arguing that ChIP enrichment at those loci may be due to transcriptional activity rather than specific p53 targeting. This again underscores the need to account for confounding effects when analyzing *in vivo* binding data.

Interestingly, 67 out of 90 (74%, with 44% expected, p-value=$3*10^{-10}$) of the methylated $C_{+1}|G_{+2}$ sites occur within 3 kb of a protein-coding gene (60 genes total) or a lincRNAs (20 total) (Figure 2.12 D on page 78) annotated in GENCODE (Derrien et al., 2012). The enrichment for sites occurring near lincRNAs (21/90 sites, or 23%, with 8% expected, p-value=$5 * 10^{-7}$) (Figure 2.12 D on page 78) is consistent with previous findings about p53 regulation of lincRNA expression (Léveillé et al., 2015).

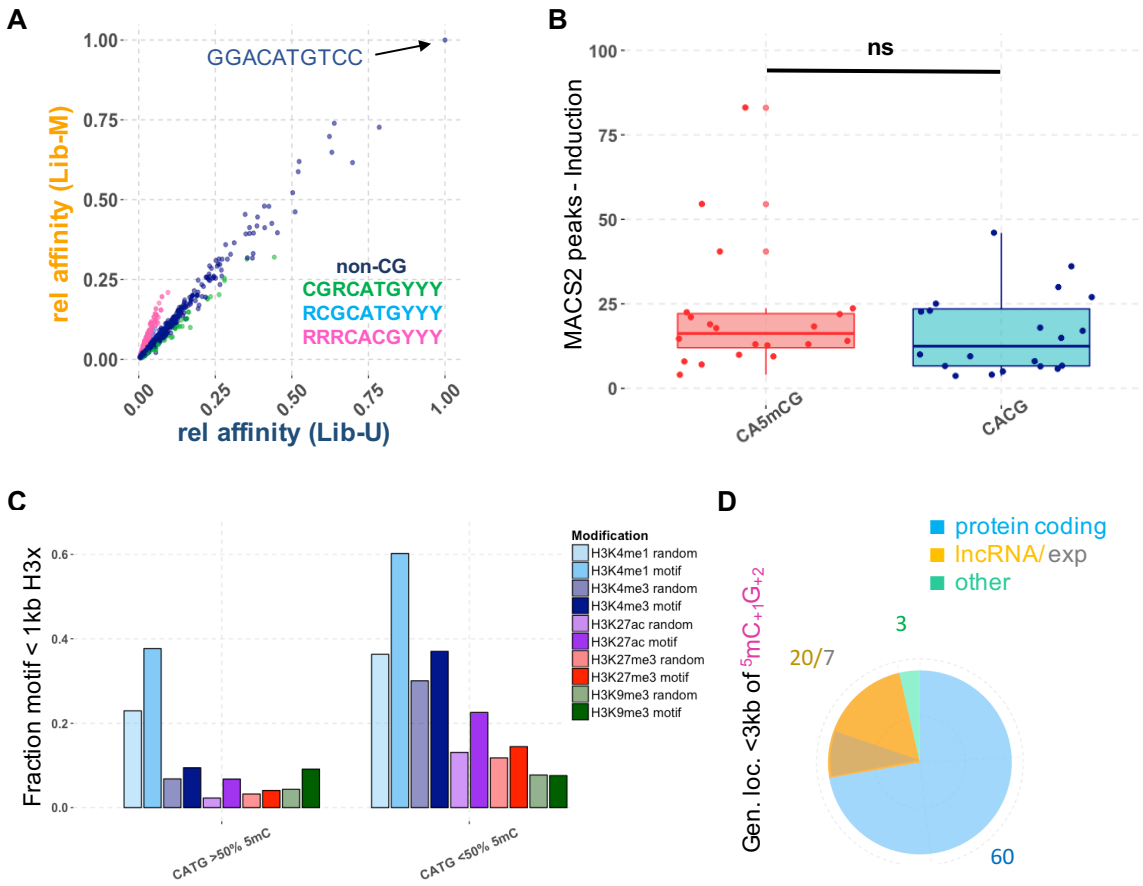


Figure 2.12: In vivo binding preferences for p53:

related to Figure 2.11 on page 73 **(A)** Comparison of normalized oligomer enrichment for 10-mers between Lib-M and Lib-U for p53.

Figure 2.12: *continued from page 78*

The top sequence (GGACATGTCC) is highlighted; it is the same for both libraries. **(B)** Motifs of type $C_{+1}G_{+2}$ split by methylation status and falling within 500bp from the ChIP-seq peak summit defined by Macs2 (q-value = 0.05). Small sample size, lack of knowledge about sequence-specific 10-mer affinities and context-dependent features are preventing a direct comparison of methylated and unmethylated motifs within peak regions. No significant difference in motif-centric induction levels (defined as the ratio of drug-induced and uninduced IP-fragment coverage) is observed (p-value = 0.3). **C)** Comparison of significantly bound CATG motifs in regions with methylation levels $> 50\%$ and $< 50\%$, respectively (500bp window), in terms of their overlap with various histone marks. The strongest enrichment is seen for H3K4me1, regardless of regional methylation levels. Enrichment for active histone marks is restricted to enhancer-specific H3K27ac modification, even within unmethylated regions lacking a strong promoter-specific H3K4me3 signature. In methylated regions the association with H3K9me3 suggests a role for p53 as a pioneer factor, in agreement with the histone modification signature observed at methylated $C_{+1}|G_{+2}$ sites. Overall enrichment levels for histone marks between are similar unmethylated and methylated regions, in contrast to the sharp difference observed between methylated or unmethylated C+1—G+2 sites (cf. main Figure 6F). (D) Characterization of the set of the 69 most highly enriched $^{5}mC_{+1}|G_{+2}$ sites in terms of GENCODE features found within 3kb. LncRNAs are enriched $\sim$ 3-fold compared to a size-matched random sample of ^{5}mCpG motifs (expected fraction indicated in darker orange).

2.4 Discussion

With EpiSELEX-seq we have developed a method that can accurately quantify the change in binding free energy associated with the presence of a methylated cytosine at any position within the protein DNA interface. A key aspect of our approach, which allows us to robustly identify methylation sensitivity, is that modified and unmodified DNA ligands are probed simultaneously in a single reaction, ensuring a direct comparison of TF occupancy. One round of selection is sufficient to accurately capture methylation effects, even for lower-affinity sites that deviate from the consensus and thus readily escape detection when binding to methylated and unmethylated ligands is assayed separately or over multiple rounds of selection. The context-sensitive nature of our analysis is essential, because opposing methylation effects can occur within a single binding site, making it difficult or impossible to detect the impact of methylation using less precise approaches such as oligomer enrichment only. This point is illustrated by our analysis of human Pbx-Hox heterodimers, whose DNA binding specificity we studied here for the first time at high resolution. The net effect of methylation on binding is close to neutral, but methylation of different CpGs in the binding site can modulate the binding affinity by up to 7-fold in either direction. This also illustrates why it may be difficult to detect methylation sensitivity by looking at motif enrichment in differentially methylated regions (DMR). Pbx-Hox sequence logos constructed separately for the unmethylated and methylated libraries were nearly indistinguishable and did not reveal significant methylation sensitivity of Pbx-Hox complexes (Figure 2.6 D on page 63). Only when we examined the consequences of methylation at specific positions were we able to identify clear effects.

Despite an ongoing debate to what extent CpG methylation is a driver of gene silencing or the consequence thereof (Ambrosi et al., 2017) , the general view is that methylation has a repressive effect on TF binding. For example, in a study that compared binding of TFs between wild-type and Dnmt1-knockout ESC cells (Domcke et al., 2015), the authors showed that removal of methylation marks at specific nuclear respiratory factor 1 (NRF1) binding sites led to increased binding and expression of nearby genes. In addition, experimentally

induced methylation reduced NRF1 binding to those sites. Here, and in agreement with recently published data (Yin et al., 2017), we demonstrate that endogenous methylated motifs containing a CpG at specific sites within the protein-DNA interface can also increase binding and that the mechanisms underlying the epigenetic control of TF binding and thus gene expression are more nuanced then previously thought.

For p53, despite a general negative effect of regional methylation on genomic occupancy, the increased binding to methylated RRRCACGYYY sites that our analysis revealed implies that methylated binding sites are functional and might direct p53 to alter previously inaccessible loci in the genome. This conclusion is supported by our finding that these occupied and methylated binding sites are associated with a histone modification pattern that indicates either compacted chromatin (Grewal and Jia, 2007) or transcriptionally poised enhancers. Additional evidence that p53 can access nucleosomal DNA *in vitro* and *in vivo*, and thus might be a pioneer factor, also supports this notion (Laptenko et al., 2011; Sammons et al., 2015). Many diseases, in particular many forms of cancer, are accompanied by aberrant methylation patterns (Kulis and Esteller, 2010) raising the question whether even subtle changes in the methylome could trigger differential TF binding and thus contribute to the onset of disease. Interestingly, H3K4me1 has also been shown to be significantly associated with loss of methylation during aging in multiple human cell types (Fernández and Bayón, 2015), providing yet additional support for the functionality and importance of such sites.

2.5 Experimental Procedures

2.5.1 Protein Expression and Purification

cDNA clones (Dharmacon) for full-length protein-coding regions for human HoxA1, HoxA5, HoxA9, C/EBPβ and ATF4 were cloned into C-terminal HIS-tagged pet expression vectors, expressed in Rosetta(DE3) cells providing additional tRNAs and purified using TALON resins (Clontech). Amino acids $8-423$ for human Pbx1 were co-purified with the HM-domain (AA $1-200$) of human Meis1 protein. p53 protein was purified as described in (Laptenko et al., 2015) containing a deletion in the C-terminal basic region to prevent non-specific DNA binding contributions outside the core DNA-binding domain.

2.5.2 Library Design

Full library sequences were as follows: 5′ -GGTAGTGGAGG-TGGG-CCTGG-**16**(**26**)x**N**-CCAGG-GAGGTGGAGTAGG- 3′ for Lib-U and 5′ -GGTAGTGGAGG-GCAC-CCTGG-**16**(**26**)x**N**-CCAGG-GAGGTGGAGTAGG- 3′ for Lib-M. The first 11 bp and last 13 bp are distinct primer landing sides for PCR amplification. Following the 11 bp at the 5′ end is a 4 bp barcode region, which can be modified as needed. Reverse complement symmetric stretches flank the random region on each side to avoid TF binding biases to a particular 5′ to 3′ orientation. Libraries were double-stranded by annealing the 3′ primer and filling in the missing nucleotides with Klenow polymerase (NEB).

2.5.3 Processing of Methylated and Unmethylated Libraries

Lib-M was methylated using the methyltransferase M.SssI (NEB) following suppliers instructions, with small adaptations to the input amount and incubation time. For optimal methylation $\sim$ 250 ng of double stranded library per 1x reaction was methylated for 2.5 hours at 37°C, followed by a second round of methylation for $>$ 1h at 37°C. Up to 400 ng of previously methylated DNA can be united in the second round of methylation. Libraries

were purified with Oligo-Clean-up columns (Zymo) and concentrations were measured by spectroscopy. Libraries were mixed in equal proportions and a small amount was set aside to serve as the R0 control to account for biases in the initial pool.

2.5.4 EpiSELEX-seq Protocol

Electromobility Shift Assays (EMSAs) and extraction of bound DNA were performed as described previously (Slattery et al., 2011). A concentration ratio of at least 10 : 1 (library versus protein) is recommended to achieve decent enrichment after a single round. Optimization might be necessary dependent on the TF used. Purified, bound DNA was PCR amplified using high-fidelity enzymes (Phusion or Q5; NEB) with overhang primers adding TruSeq Illumina adapter sites. 13 to 15 cycles are generally sufficient. The primer annealing temperature was increased sequentially from 39°C to 72°C to guarantee proper annealing to the initially short (11 and 13 base pair) primer landing sites. The Q5 enzyme (NEB) can be used at an initial annealing temperature of 47°C (8-10 cycles) followed by 3-5 cycles at the recommended enzyme extension temperature. Each amplification is set up with four 50 μl reactions total, split into two pairs of primer sets. The first set contained the Illumina universal primer landing site, followed by the Illumina adapter sequence and library $5'$ overhang site as the forward primer (UNI-for) and the library $3'$ overhang site followed by the Illumina adapter and INDEX primer landing site as the reverse primer (INDEX-rev). The second set of primers had the UNI and INDEX sites swapped (INDEX-for; UNI-rev). The split-pool PCR prevents a unique directionality, which could cause problems during sequencing due to low-diversity (identical fixed flanks). Efficient splitting was analyzed by comparing the number of reads resulting from each set of primer sequences (Figure 2.1 on page 54). Amplified PCR products contained landing sides for the universal and one of the 24 NEB TruSEQ indexing primers compatible for Illumina sequencing. Specific Illumina barcodes were added by a five cycle PCR using NEBNext Multiplex Oligos for Illumina sequencing and Phusion or Q5 polymerase. The indexed libraries were gel-purified as described previously (Slattery

et al., 2011) and the concentration measured with a Qubit spectrometer. Multiple indexed experiments were pooled and sequenced using a v2 75 cycle high-output kit on an Illumina NEXTSeq Series desktop sequencer at the Genome Center at Columbia University. For the initial R0 libraries and for each single-round enriched (R1) library, 5-35 million single-end reads were obtained. A $\sim 10\%$ PhiX spike-in was used for optimal sequencing quality.

Table 2.1: Primer-sequences used for Split-PCR: sequences matching libraries are in bold

probe		**sequence**	
SET 1			
UNI-for	5'	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**GGTAGTGGAGG**	3'
INDEX-rev	5'	GACTGGAGTTCAGACGTGTGCTCTTCCGATCT**CCTACTCCACCTCCC**	3'
SET 2			
INDEX-for	5'	GACTGGAGTTCAGACGTGTGCTCTTCCGATCT**GGTAGTGGAGG**	3'
UNI-rev	5'	ACACTCTTTCCCTACACGACGCTCTTCCGATCT**CCTACTCCACCTCCC**	3'

2.5.5 Testing for Methylation Efficiency

The probe 5′-GGTAGTGGAGGTGGG ACGGCCGTGCGCTCGA GGGAGGTGGAG TAGG-3′ was double-stranded and methylated for a positive control or unmethylated for a negative control. Methylation reactions for the positive sample were carried out with varying incubation time, DNA input amounts and repetitions. Subsequently, both positive and negative control samples were bisulfite treated, amplified and ligated into a pBlueScript vector. The ligation product was transformed into competent cells and 4-8 colonies were picked for sequencing.

Table 2.2: Testing methylation conditions with test probe: *related to Figure 2.2 on page 55*

Input per reaction	**1000 ng**	**700 ng**	**250ng**
Repetition [n]	1	2	2
Incubation time [h]	1	1	2.5
size methylated [^{5}mC/tot]	6/20	12/38	44/48
% methylation	$> 30\%$	$> 40\%$	$> 90\%$

2.5.6 Bisulfite Conversion of Lib-M

Double-stranded Lib-M was split into two parts (positive and negative control). For the positive control the DNA was methylated using optimal conditions. Subsequently, both negative and positive controls were bisulfite treated and amplified following the manufacturer's instructions (EpiMark Bisulfite Conversion Kit and EpiMark polymerase; NEB). Preparation for Illumina sequencing was carried out following the standard EpiSELEX-seq protocol and samples were sequenced to a depth of at least 1 million reads. Dinucleotide frequencies were normalized by using "TpG" as a reference, and all CpN dinucleotides were compared across the original Lib-M (methylated but untreated), positive control, and negative control in order to estimate methylation efficiency.

2.5.7 EpiSELEX-seq Data Processing

Data were collected with an Illumina NEXTSeq Series desktop sequencer and raw FASTQ files were downloaded. FASTQ files were pre-processed using the FASTX toolkit (Hannon lab). In a first step, files with a unique Illumina indexing barcode from all 4 lanes were collapsed into a single file and subsequently reverse complemented due to the split-pool approach during the library preparation. Original and reverse complemented files were merged and the $5'$ primer and $3'$ primer binding sites were trimmed such that each line starts with

the library-specific four-base barcode, followed by the 5′ flank, random region and ending with the 3′ flank. Downstream analysis was done using R software including the packages SELEX and stringi. Each data set was assigned to either Lib-U and Lib-M. A 5th-order Markov Model was generated on each round 0 (separately for Lib-U and Lib-M) to capture biases in the initial sequence pool using the R package bioconductor.org/packages/SELEX (Riley et al., 2014).

2.5.8 Analysis Based on Oligomer Enrichment Differences

Relative affinities for oligomers of length k were estimated by calculating the oligomer enrichment for R1 counts compared to the expected count as obtained from a Markov Model prediction of R0. Fold-enrichments were normalized based on the most enriched oligomer. In the case of ATF4, the most highly enriched 10bp sequence was different between Lib-U and Lib-M, due to the presence of a repressive effect of methylation at the central CpG. Therefore, libraries were normalized by the joint, most enriched oligomer. For the other libraries the most enriched sequence did not include a CpG or methylation of that CpG had no effect on binding, such that normalization could be carried out as described previously (Slattery et al., 2011). Position-specific affinity matrices (PSAMs) (Foat et al., 2006) were generated by considering all 3k point mutations away from the most enriched oligomer, and binding free energy differences between the mutated and optimal sequence were estimated as the negative logarithm of the relative fold-enrichment. To estimate the effect of methylation on binding free energy, we separately calculated $\Delta\Delta G/RT$ for a C→T transition for Lib-U and Lib-M, and that

$$\Delta\Delta\mathrm{G}[\mathrm{C} \rightarrow^{5} \mathrm{mC}] \approx \Delta\Delta\mathrm{G}[\mathrm{C} \rightarrow \mathrm{T}]_{\mathrm{Lib-U}} - \Delta\Delta\mathrm{G}[\mathrm{C} \rightarrow \mathrm{T}]_{\mathrm{Lib-M}} \tag{2.3}$$

In the case where the affinity-based model fully captured the methylation effect, the value of $\Delta\Delta G[C \rightarrow T]_{Lib-M}$ implies a methylated C. Importantly, however, the oligomer-

enrichment-based approach is incapable of capturing the methylation effect whenever the seed sequence prefers a non-G base at the position adjacent (3′) to the one being estimated. In that case, the estimation of Lib-M is dominated by unmethylated cytosines occurring in a non-CpG context (i.e., CpA, CpT, or CpC) (Figure 2.9 on page 68).

2.5.9 Feature-Based Modeling

A feature-based generalized linear model based on Poisson statistics is fit to the read counts after selection (R1). First, PSAMs are constructed from oligomer enrichment tables for each sample as described above and used to scan the random region of each probe in both orientations and for each offset relative to the constant flank. Up to 2 bp overlap with the constant flanks on each side were considered for binding. E.g. a k-mer of length 10 would have $16+2+2-(10-1) = 11$ different binding possibilities per orientation for a 16bp random region. To achieve unambiguous definition of features in the protein-DNA binding interface of each probe, we only kept probes for which a single offset/orientation contributed at least 95% to the sum over all affinities. Sequences below a certain relative affinity threshold were excluded to avoid biases due to non-specific binding; the threshold was chosen to achieve a minimum read count of 100 (corresponding to a 10% relative error). To avoid bias against unobserved reads, the R1 reads were randomly split into two equal halves, the first of which was used to define the set of oligomers that correspond to the rows in the design matrix, and the second to define the counts (including zero for oligomer-containing probes only occurring in the first half of the random split). Regression models were fit in two ways: (i) using the Lib-U R1 count for a particular motif of length k, the markov model prediction from the corresponding R0 as an offset, and 4k base indicator features for each position in the motif as independent variables; (ii) same as before, but including Lib-M and using both base and ^{5}mCpG indicator features. For the ^{5}mCpG indicator, any position not containing a CpG in Lib-M and any position in Lib-U was assigned the value zero, and all CpGs in Lib-M were assigned the value one. The overall feature set in the combined model comprised 3k non-

redundant base features and k−1 methylation-status features (one for each possible ^{5}mCpG dinucleotide position). Stability of the model estimates was assessed by comparing the base feature coefficients from i) and ii). As expected, adding CpG features and reads from Lib-M did not affect the base feature estimates (Figure 2.1 on page 54). For p53, due to the long 26 bp random region and overall low sequence selectivity a library indicator was fit.

2.5.10 Competition Assay for Pbx-HoxA1

A 25-bp probe including the top 12-mer site ATGATTGA**TT**AC was double-stranded by annealing with its reverse complement and radiolabeled using T4 PNK (NEB). Likewise, two 12-mer containing competitor probes with identical sequence – ATGATTGA**CG**AC – but different methylation status at position 9 of the Pbx-HoxA1-DNA interface were tested for their capacity to compete with the labeled probe for Pbx-HoxA1 binding in an EMSA. Pbx-HoxA1 and labeled-probe concentrations were held constant while increasing the concentrations of the cold competitor DNA over a 1000-fold range. The fraction of labeled probe bound for both methylated and unmethylated competitor was computed by quantifying the intensities of the protein-bound and unbound fractions (using ImageJ) and normalizing by the total amount bound in the absence of competitor DNA. Each competition was performed in duplicate. IC50 values were calculated by fitting a dose-response curve for each competitor DNA using the R package drc.

2.5.11 Data Processing for *In Vivo* p53 Binding

ChIP-seq fastq files for p53 binding assayed in the MCF7 cell line were downloaded from the SRA database (Sequence Read Archive Accession number: SRP007261) for the no drug control and the three drugs Nutlin, RITA and 5FU, along with the BED file containing whole-genome bisulfite sequencing data for MCF7 (GSM1328112). ChIP-seq FASTQ files were aligned to the reference hg19 genome (Bowtie2) and converted to coverage tracks (bedGraph format; extending reads by 200 nt) using the bamCoverage function from the deeptools

package. For standard peak calling, MACS2 was run with options -g hs and -q 0.05 using the uninduced p53 IP as a control. BED files of called peak regions mapped to hg19 for five MCF7 Histone modifications were downloaded from ENCODE (ENCSR000EWP, ENCSR000EWQ, ENCSR000EWR, ENCSR493NBY, ENCSR985MIB). GTF files for the current releases (v25) of human whole genome annotation and lncRNA specific annotation data were downloaded from the GENCODE database (mapped to GRCh37/hg19) and used directly in downstream analyses.

2.5.12 *In Vivo*, Motif-Centric p53 Binding Analysis

The following steps were all done in R using the following packages: Biostrings, BSgenome, rTracklayer, and ggplot2. The hg19 genome was scanned for sites mapping either to the consensus RRRCATGYYY, or the three CpG motif classes (RRRCATGYCG, RRRCATGCGY, RRRCACGYYY). Next, WGBS BED files were intersected with the CpG containing motifs and methylation status was assigned based on the percentage methylated within the WGBS sequencing data ("1" for $> 80\%$ methylated and "0" for $< 10\%$ methylated). To assure correct calling of methylation status only sites that had at least 10x coverage were considered. For all motifs, the average methylation level of the 500 bp centered around the motif and the overall CpG density was computed. In the last step, the per-motif coverage for the uninduced (control) and the drug-induced (IP) p53-IP was obtained by intersection with the ChIP-seq coverage tracks. Only motifs with at least 1x coverage in both control and IP were retained. Each individual genomic motif location represented a single row in the design matrix (X). The in vitro affinity for each 10-mer (as derived from the unmethylated EpiSELEX-seq data), the three position-specific binary ^{5}mCpG indicator, the average methylation level, and the CpG density within the 500 bp region constituted the columns of the design matrix. The glm function with family = "binomial" was used in R to fit the following model of the probability of a specific motif being bound:

$$\text{p(bound)} \equiv \frac{\text{IP}}{\text{IP} + \text{CTRL}} = \frac{1}{1 + \exp^{-\frac{\Delta\Delta\text{G}}{\text{RT}}}} \tag{2.4}$$

where

$$\frac{\Delta\Delta\text{G}}{\text{RT}} = \sum_{\phi} \beta_{\phi} \text{X}_{\phi} \tag{2.5}$$

Regression coefficients and Z-scores quantifying their statistical significance were obtained for each model fit. Positive coefficients represent increased, and negative coefficients decreased binding. To account for potential bias regarding which motifs did make it into the training data, we sub-sampled (~ 200 times) from the entire set of sequences in a way that guarantees an equal number of occurrences of methylation status and motif class (for the CATG motif, a threshold of 50% average regional methylation was used as a proxy for methylation status). The sample size was chosen such that less than one third of all instances of each possible methylation-status/motif combination were included at a time. This way, the number of motif occurrences should not influence model performance.

For models enriched for stronger p53 binding, we considered the entire data set to allow for sequential removal of rows for which the sum of drug-induced and uninduced IP fragment counts fell below a certain threshold. Statistical association between MACS2 peaks and single-motif-instance fold-enrichment values was determined by splitting motifs based on their log-transformed induction levels, computing the fraction of motifs in each group overlapping with a peak region (1kb around peak center), and performing Fisher's exact test.

2.5.13 Overlap with GENCODE Annotation and Histone Marks

To test for enrichment of a set of p53 bound motifs (defined as having a motif-centric induction level of two standard deviations above the mean) with either GENCODE annotations or Histone marks, we first computed the overlap between the motif set and either the gene annotations (within 3kb) or the histone peaks (within 1kb). Next we generated random motif sets from the WGBS data, which matched the methylation status cutoff used for the original motif set. To compute p-values, we generated > 100 random sets and calculated the probability of observing the actual overlap based on the sampling of random overlaps.

2.6 Acknowledgements

We thank Remo Rohs, Timothy Bestor, Barry Honig, as well as the members of the Bussemaker, Mann, and Rohs labs, for valuable discussions. This work was supported by an HHMI International Student Research Fellowship (J.F.K.) and NIH grants R01HG003008 to H.J.B. and R35GM118336 to R.S.M and grants P01CA087497 and R01CA196234 to C.P. Columbia University's Shared Research Computing Facility is supported by NIH grant G20RR030893 and NYSTAR contract C090171.

2.7 Addendum: Explaining the Negative Impact of ^{5}mC Methylation Within the Pbx-Hox Spacer

The work described in this addendum is published as part of the study indicated below. Only the section directly related to this thesis chapter is described in the following section. The remaining results and insights of the study are not part of this thesis, nor related to it.

Rao, S., Chiu, T. P., Kribelbauer, J. F., Mann, R. S., Bussemaker, H. J., & Rohs, R. (2018)

"Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein-DNA binding"

Epigenetics and Chromatin, Volume 11, Issue 1, 2018

Author Contributions specific to this section:

Conceptualization, J.F.K.; Data Analysis, S.R. and J.F.K., Visualization, S.R. and J.F.K.; Review & Editing, J.F.K., S.R., R.S.M., R.R. and H.J.B.

In theory, every methylation effect observed by our EpiSELEX-seq method should have an underlying biophysical explanation accounting for the differences in TF binding free energy between methylated and unmethylated DNA ligands. Failure to provide a mechanistic explanation should raise red flags and cast doubt on the experimental set-up and on whether potential biases might not have been accounted for during the analysis. However, identifying an underlying mechanism that explains the observed ^{5}mC impact on binding might sometimes not be feasible. One possible way of doing so is by analyzing existing crystal structures of the DNA-TF complex, in particular, by focusing on the local electro-chemical environment of the methylation group that is being considered. This approach was used for both Pbx-Hox and p53. The most straight forward way how proteins might favor or disfavor a methyl group is through direct interaction with protein surface charges. Either a neutral or a charged patch can be in close proximity to a methyl group (^{5}mC), in which case we expect methylation to either have no effect or be beneficial for binding (neutral interaction between ^{5}mC and the protein surface), or to be deleterious for binding (interaction with a charged patch). We inferred such a direct readout for several CpGs within the binding sites of Pbx-Hox and p53, which we coined "thymine mimicry" due to its similarity to the methyl group in thymidine. It even explained the differences we detected at position 9/10 (ATGAYNNA<u>YN</u>NN) within the binding site of Pbx in complex with either HoxA1, HoxA5 or HoxA9 (compare Figure 2.10 on page 70 and Figure 2.11 on page 73).

Despite its success in many cases, this "direct" readout could not explain the negative impact methylation had at position 6/7 (spacer region of Pbx-Hox) or 10/11 (Hox flank) (see Figure 2.10 A on page 70 and Figure 2.13 A on page 95). Since a methyl group at base 6 in the spacer (and base -7 on the reverse strand for fully methylated ^{5}mCpGs) is not close enough to the interaction surface spanned by amino acids from either Pbx or Hox (Figure 2.13 A on page 95), thymine mimicry is unlikely to explain the observed negative impact on binding when $C_6|G_7$ is methylated (Figure 2.13 B and C on page 95). In addition, the most preferred base at position 6 is a thymine and thus a methyl group at this position cannot

per se be bad for binding.

In an attempt to find an alternative explanation, we considered two previously documented cases of DNA shape recognition by proteins: i) the sensing of DNA minor groove width by N-terminal residues within the Hox homeodomain (HD) (Abe et al., 2015) and ii) the effect DNA methylation can have on DNA shape (in particular MGW narrowing) and thus on protein binding (examplified by different DNase I cleavage rates between methylated and unmethylated hexamers) (Lazarovici et al., 2013). The observed negative impact Pbx-Hox spacer methylation has on binding might therefore be mediated by a change in intrinsic DNA shape. Such a deviation in geometry could result in widening of the Pbx-Hox DNA spacer minor groove, which ultimately is readout by Hox N-terminal arm arginines. Using high-throughput DNA shape predictions (Rao et al., 2018) and comparing the average minor groove width in the context of methylated and unmethylated CpGs should thus display a selective widening of the minor groove in the methylated sequence context for the Pbx-Hox spacer. Indeed, (as shown in Figure 2.13 D on page 95) MGW for the Pbx-Hox spacer is significantly larger for the methylated sequence context (NAY^5**mCG**AY) than the unmethylated one. In contrast, MGW for the sequence context for position 9/10, where methylation is beneficial for binding (direct readout), remains unaffected, supporting the idea that shape recognition only plays a role in the methylation induced reduction in binding free energy for CpG position 6/7 (Figure 2.13 D on page 95).

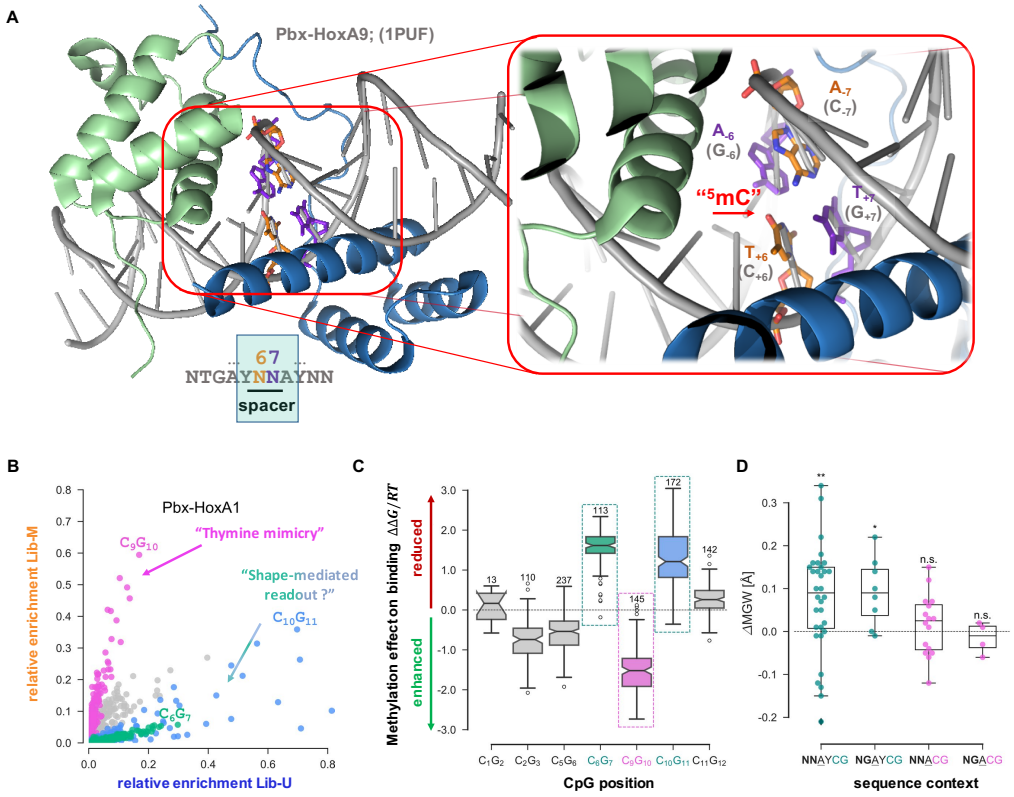


Figure 2.13: Methylation widens minor groove within the Pbx-Hox spacer:

Panel B-D of this figure have been adapted from (Rao et al.*, 2018). Original figures were generated by S. Rao based on data generated by J.F. Kribelbauer and S. Rao. The theoretical framework for the analysis of the Pbx-Hox spacer methylation sensitivity was conceived by J.F.Kribelbauer with input from S. Rao, H.J.Bussemaker, R.S.Mann and R.Rohs.*

(A) Crystal structure of Pbx in complex sith HoxA9 (representative, PDB-ID: 1PUF) highlighting the DNA spacer between Pbx and Hox. The highlighted basepair represents position 6/7 within the Pbx-Hox interface (orange and purple; labeling for hypothetical CpG positions in grey). The potential location of a methyl group at base position 6 is highlighted in the close up in red (^{5}mC) and is not within reach of amino acids of either Pbx or Hox and thus presumably not directly contacted by the proteins.

Figure 2.13: *continued from page 95*

(B) Relative enrichment plots for methylated versus unmethylated sequences from EpiSELEX-seq experiment performed on Pbx-HoxA1, focusing on CpG-containing sequences only (compare Figure 2.6 on page 63). Pink sequences (CpG at position 9/10) are beneficial for binding and can be explained by a direct base contact. Green and blue sequences result in reduced binding and cannot be explained by a direct readout. **(C)** Summary of sequences in B (grouped by their position within the binding sites) in terms of their mean impact on binding, expressed in terms of $\Delta\Delta G/RT$. **(D)** Average difference in MGW between methylated and unmethylated sequences matching either the sequence context of the Pbx-Hox spacer or the Hox flank (NAY<u>CG</u>), or position 9/10 within the binding site (NNA<u>CG</u>). A significant difference in MGW (widening of the groove) is observed for the NAYCG (Pbx-Hox spacer) but not the NNACG (position 9/10) sequence context.

Chapter 3

Uncovering the Rules of Adaptive DNA Binding that Govern Target Specificity of a Multi-Protein Hox Complex *In Vitro* and *In Vivo*

3.1 Introduction

The Role of Different Conformational Modes in Adaptive TF-DNA Recognition

Since the discovery of double stranded DNA (dsDNA) by Watson and Crick (Watson and Crick, 1953) and the first solved structures of a protein bound to it more than two decades later (Anderson et al., 1981), it has now been established that specific interactions between transcription factors (TFs) and DNA play an essential role in gene regulation. Since those early days, much effort has been devoted to identifying the mechanisms by which TFs achieve sequence specificity and thus recognize their cognate TF binding sites (TFBS). On one hand, *in vitro* and *in vivo* binding studies using different enrichment methods, such as ChIP-seq (Barski et al., 2007; Johnson et al., 2007; Mikkelsen et al., 2007), selective evolution of ligands by exponential enrichment (SELEX) followed by deep sequencing (Slattery et al., 2011; Stormo et al., 2015; Jolma et al., 2010) or microfluidic-trapping of TF-DNA complexes (Maerkle and Quake, 2010; Isakova et al., 2017), have been used to identify the sequences recognized by TFs. The comparison of sequences by their relative affinities is then often

summarized by a scoring matrix and visualized by a motif logo (Stormo, 2000; Foat et al., 2006). On the other hand, structural studies have been used to gain insights into the binding mechanisms that govern the observed sequence specificity. Although these structural studies provide detailed insights into the biochemical features making up the interaction surface, they are limited to a single sequence. To what extent the overall conformation of the TF-DNA complex is retained when a TF is presented with another DNA ligand is largely unknown. Only a few examples exist where the same complex has been crystalized with two or more DNAs (e.g. (Joshi et al., 2007).

Naively, it was first imagined that specific TF amino acids form hydrogen bonds with cognate DNA bases lying in an exposed major groove and, by doing so, drive the differentiation between good and bad sites (Pabo and Sauer, 1984). However, an increasing number of structural studies have since then revealed the complexity of TF-DNA interaction. For example, there is a diverse set of structural folds for DNA binding domains (DBDs) that are used to recognize DNA (Garvie and Wolberger, 2001). Similarly, DNA was found to be able to deviate significantly from the canonical B-DNA shape (Rohs et al., 2009a), as highlighted by the switching between Watson-Crick and Hoogsteen base-pairs observed in a NMR structure of naked DNA (Nikolova et al., 2011; Honig and Rohs, 2011). In addition, analysis of a large number of TF-DNA complexes revealed that only one third of contacts involve direct hydrogen- or water-mediated bonds, whereas two thirds are van der Waals interactions with a dominance of DNA backbone contacts across all types (Luscombe, 2001).

Given the relatively small fraction that direct hydrogen bonding with major groove bases contributes to the total binding free energy, more and more research has focused on investigating the role of structural recognition, or "indirect readout", in TF binding (Slattery et al., 2014). Depending on binding partners or sequences, TFs might therefore have to rely on the use of subtly different structural arrangements to adjust their binding mode to a given context, referred to as "adaptive" DNA binding for brevity.

A crucial observation in this regard is that DNA sequence and shape are naturally depen-

dent on each other, such that the following questions arise: How much structural information is encoded in the underlying sequence? How many neighboring base-pairs need to be defined in order to uniquely define the geometric constraints of a given base pair? Can we use the sequence-shape relationship in predicting how TFs interact with a given sequence and thus build better TF-DNA recognition models even in the absence of structural information (as true for most sequences)?

An important step forward was achieved by the tabulation of several shape parameters for all 4^5 pentamers based on Monte Carlo simulations and using existing crystal structures of naked DNA (Zhou et al., 2013). Not only did it simplify the parameter space by assigning a single value to the overall 3D positioning of individual base-pairs in different contacts, but it could also be used to compute intrinsic DNA shape for arbitrarily long sequences, such as an entire genome. However, due to the strong dependencies between shape and mono or dinucleotide features, caution is warranted when combining the two in order to build TF recognition models (Rube et al., 2018). Rather than using shape features as parameters in TF binding models, they are a useful tool in relating the sequence recognized by a TF to the underlying binding mechanisms and thus revealing the structural recognition mode. To give an example, recognition of minor groove width (MGW) has been shown to be a mechanism by which TFs select specific DNA ligands (Joshi et al., 2007; Rohs et al., 2009b; Abe et al., 2015). Starting from a set of enriched sequences for a TF lacking structural data, the sequence-to-shape models can then be used to identify whether MGW narrowing or other shape features might contribute to the selection process. Not only can the sequence-to-shape relation provide insights into TF-DNA binding mechanisms, but it might also provide a means for the design of engineered TFs that distinctively disrupt specific TF binding modes, while keeping others unaffected.

The Role of Cooperativity in Context-Specific TF-DNA Recognition

Yet another form of "indirect readout" that complicates identification of TF binding sites (TFBS) *in vivo* is cooperative binding between TFs and their cofactors. An early example is given by the interferon-β (IFN-β) enhanceosome, a multi-protein complex which forms on segments of DNA of $\sim$100bp and whose assembly leads to synergistic transcription (Carey, 1998). Another observation that underscores the importance of cooperative binding is the fact that TFs belonging to the same structural family often share DNA recognition folds and recognize highly similar DNA sequences *in vitro* (Noyes et al., 2008). Interesting in that regard is the positive correlation between organism complexity and total amino acid length of TFs, despite the fact that the length of the DBD remains largely constant across different taxa (Charoensawan et al., 2010). Amino acids outside of the DBD are therefore important in providing a binding platform for other TFs capable of inducing cofactor-mediated latent specificity (Slattery et al., 2011). Moreover, complex formation with cofactors imposes constraints on TF binding in terms of TFBS orientation, spacing and overall complex conformation, perhaps contributing to *in vivo* TFBS selectivity. Likewise, amino acids at the edge of DBDs can contribute to DNA sequence preferences for bases flanking the core motif and the use of different TF-isoforms can add an additional layer of specificity in targeting TF-complexes to specific genomic locations.

An example of such a complex regulatory system is that of the *D. melanogaster* homeodomain proteins Homothorax (Hth), Extradenticle(Exd), and one of the eight Hox proteins. Hth exists in two major isoforms, one full-length (Hth^{FL}) and one without a DNA binding domain (DBD) (Hth^{HM} ; HM for Homothorax-MEIS domain after the human ortholog (Rieckhof et al., 1997)). Moreover, Hth forms a tight protein-protein interaction with Exd via its HM-domain and both splice variants of Hth are sufficient to recruit Exd into the nucleus (Rieckhof et al., 1997; Noro et al., 2006). The interaction is highly conserved across species (Longobardi et al., 2014; Merabet and Mann, 2016). Exd is an important cofactor for Hox proteins and can induce latent sequence specificity when forming dimeric

complexes with Hox proteins (Slattery et al., 2011). The Hth^{FL} isoform is dispensable for some of the Exd-Hox related functions, such as correct patterning of the proximal-distal axis in leg appendages, but required in other, such as distalless (Dll) regulation (Hox-dependent) (Gebelein et al., 2004) or antennal fate specification (Hox-independent) (Noro et al., 2006). The variety of different complexes that can be formed by the three types of homeodomain proteins and their splice variants is conserved in vertebrates, but with an increasing level of complexity due to evolutionary duplication events. The importance of this regulatory system is demonstrated by the severity of the knockout phenotypes (Longobardi et al., 2014), and misregulation of MEIS, PBX (the human orthologs of Hth and Exd) and human Hox proteins has also been linked to the onset of leukaemias and other types of human cancers (Kroon et al., 1998; Grubach et al., 2008). Moreover, aberrant levels of recently identified MEIS HD-less isoforms were found in colorectal cancer tissues (Crist et al., 2011), arguing for the importance of correct MEIS/Hth splicing.

The concept of a combinatorial logic governing the DNA binding site recognition of the three TFs is therefore an appealing mechanism, which could explain how in specific tissues or developmental stages fine-tuned sets of genes can be selectively turned on or off. However, the large number of different complexes that can be formed, and the difficulty to infer binding mechanisms for all of them, has largely limited our understanding of this complex regulatory system that is so crucial for healthy development.

Despite recent efforts aimed at developing and refining methods for TF binding site prediction that incorporate context-dependence of TF binding (including cooperativity or sequence-specificity resulting from amino acids outside or bordering the DBD) (Jolma et al., 2015; Le et al., 2018; Shen et al., 2018; Rastogi et al., 2018), a well-characterized example and proof of *in vivo* relevance of different conformational modes in multi-TF-DNA sequence recognition still remains to be demonstrated.

Here we use SELEX-seq to study the adaptive DNA binding behavior for all possible

complexes that can be formed by Hth, Exd and Hox. We demonstrate that i) composition, orientation, and spacing of the three homeodomains all contribute to sequence selectivity; that ii) sequence-to-shape mapping can be used to identify the mechanism by which the preferred Hth-Exd DNA spacer is selected; that iii) the identified "shape readout" of DNA bases flanking or bridging the core TFBS is not unique to the N-terminal DBD amino acids of Hox homeodomains (Abe et al., 2015) but a general feature also utilized by the N-termini of Exd and Hth; that iv) the degree of DNA shape recognition and thus of relative TF binding mode usage depends on the sequence context (high versus low affinity site), the composition of the protein complex, and the Hth isoform, and that v) engineered shape-mutant TFs can be used to selectively disrupt DNA binding in a complex-, isoform- and sequence-dependent manner.

Finally, we validate that the binding mechanisms uncovered by our *in vitro* analyses are important and fully recapitulated *in vivo* by introducing transgenic Exd, engineered to differentiate between the various binding modes described under v).

Our results show that TFs bind DNA ligands in a highly adaptive, sequence-dependent manner, utilizing different recognition modes and complex compositions. We demonstrate how the sequence-to-shape relationship can be used to identify specific TF recognition modules that, upon manipulation, are capable of distinguishing distinct binding contexts, such as the use of a specific isoform. Moreover, the mutated TFs demonstrate that the use of a specific binding mechanisms by a TF-complex can vary with DNA sequence context, e.g. high or low affinity sites. Finally, the generated "designer TFs" can also be used to distinguish different binding contexts *in vivo*.

3.2 Results

3.2.1 Complex Composition directs Conformation and Orientation Between Binding Partners

Proper development of (embryonic) body patterning is dependent on the correct spatial and temporal expression of Hox transcription factors (TFs) (Merabet and Mann, 2016). Despite the high specificity with which they regulate distinct gene sets in vivo, Hox factors exhibit similar DNA-binding specificity in vitro (Noyes et al., 2008). How they manage to discern their target genes is thus an open question in systems biology. One hypothesis is that specificity is achieved by cooperative binding with other factors. A well-studied example is the TALE-type, homeobox transcription factor Extradenticle(Exd), which can form dimers with each of the 8 *D.melanogaster* Hox factors, conferring latent specificity upon dimer binding (Slattery et al., 2011). As mentioned in the introduction, Exd is tightly regulated by the second TALE-type homeobox TF Hth with its two major isoforms Hth^{FL} (containing the c-terminal DBD) and the HD-less Hth^{HM}, both sufficient for the nuclear localization of Exd (Figure 3.1 A on page 105). Exd's nuclear import is mediated via a protein-protein interaction between the Hth HM and the Exd PBC domain (Figure 3.1 A on page 105). The Hth isoforms however, can execute distinct functions *in vivo* (Ryoo and Mann, 1999; Noro et al., 2006; Gebelein et al., 2004), and differential splicing seems of general relevance as several isoforms have also been reported in other species, including humans (Crist et al., 2011; Noro et al., 2006). As a result of the tight protein interaction between Exd and Hth, Exd-Hox target genes are believed to be controlled by ternary protein complexes with either two or three homeodomain (HD) DBDs present – Hth^{FL}-Exd-Hox or Hth^{HM}-Exd-Hox. In addition, Hth^{FL}-Exd has known Hox-independent regulatory functions, in tissues lacking Hox expression (Noro et al., 2006).

To quantify the binding specificity of different combinations of the two Hth isoforms, Exd and Hox, we carried out SELEX-seq on the trimeric Hth^{HM}-Exd-Dfd, the dimeric Hth^{FL}-

Exd and the trimeric Hth^{FL}-Exd-Dfd complex, with libraries designed to accommodate the respective complexes (variable regions of 16bp (Lib-1 and Lib-2), 21bp (Lib-3a and -3b) or 30bp (Lib-4)). To facilitate analysis of complex binding patterns when all 3 HDs are present, the two 21-bp libraries had fixed 5′ Hth sites in either orientation: **CTGTCA** (Orientation C←N; Lib3a; Hth is colored in pink) or **TGACAG** (orientation N→ C; Lib3b; Hth is colored in purple to differentiate between the binding orientations in Lib-3b and Lib-3a) (Figure 3.1 B on page 105).

We constructed position-specific affinity matrices (PSAMs) (Foat et al., 2006) for the most enriched 10- or 12-mers (Lib-2 or Lib1/3/4 respectively) for each of the tested complex compositions. This analysis indicates that in the absence of Hox, Exd-Hth forms a tail-to-head (TH) dimer resembling the Exd-Hox binding conformation (Figure 3.1 B-II on page 105). Introducing Dfd to the Hth^{FL}-Exd complex using LIB-3a or LIB-3b results in the formation of the Hox-Exd subcomplex as the dominant binding site (Figure 3.1 B-III on page 105) and Figure 3.2 on page 107). For the binding-conformation-agnostic LIB-4, Exd-Hox is still the most preferred subcomplex, but both Exd-Hth (darkblue) and sequences suggestive of Hth-dimer binding sites (deep pink) are present alongside with Exd-Hox sequences (Figure 3.1 B on page 105 and Figure 3.2 on page 107). Noticeable is also the increased specificity within the Dfd half site in the Exd-Hox energy logos compared to the logo derived from Lib-1, suggesting the presence of monomeric Hox binding. In summary, the data suggest that i) the Exd-Hox dimer is the dominant complex when adding all three HDs together, ii) trimeric Hth-Exd-Hox complex preferentially forms when a suitable Hth site is present and iii) the binding energies for trimeric Hth-Exd-Hox, monomeric Hox, dimeric Hth and Exd-Hth complexes are all similar enough to be observed (Figure 3.2 on page 107).

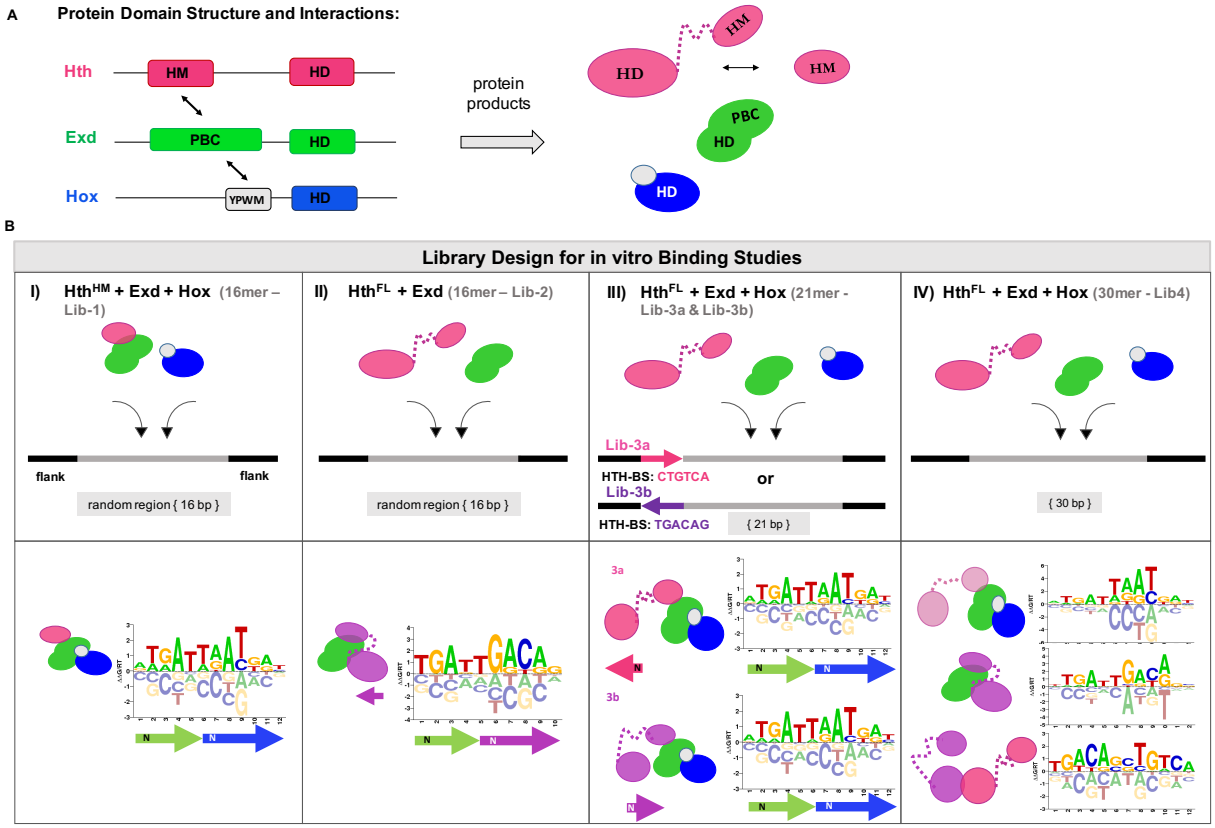


Figure 3.1: Complex composition and sequence preferences of the three homeodomain TFs Hth, Exd & Hox **(A)** Protein domain structure and splice forms for Hth, Exd and Hox. Hth contains a Homothorax MEIS domain (HM) that has been shown to interact with Exd. Two major isoforms exist, one full-length Hth including the DNA-interacting homeodomain (HD) and one HD-less isoform (HM). Exd has a HD and a PBX domain with the latter shown to interact with Hth's HM domain. Exd also interacts with the YPWM motif found in many Hox proteins. Hox contains the Exd interaction motif (YPWM) and a HD. (Schematics are not to scale). **(B)** Library Design and energy logos for sequences emerging as the top binding sites from the different SELEX experiments. i) Lib-1 has a 16 bp random region and was used for probing the Hth^{HM}-Exd-Hox sequence preference, the consensus motif is shown for Exd-Dfd (ATGATTAATGAT)

Figure 3.1: *continued from page 105;*

ii) Lib-2 also has a 16bp random region, but uses a different amplification strategy and was used to test Exd-Hth binding preference. The energy logo shows that Hth takes the same position and orientation as Hox in Lib-1 with the top site TGAT|TGACAG (first half represents Exd and second half Hth). iii) Lib-3a and Lib-3b both used a fixed flank that contained the Hth site in either orientation (CTGTCA – Lib-3a) or (TGACAG – Lib-3b) followed by a 21bp random region. In either library, the top motif is the Exd-Hox site (ATGATTAATGAT). iv) Lib-4 had a 30-mer random region to accommodate the entire Hth-Exd-Hox complex including variable spacing. Three motifs indicating different complex compositions were retained from the library – Exd-Hox, Exd-Hth and presumably a Hth dimer (TGACAG|CTGTCA). The different shades of color (pink or purple) used for Hth indicate its relative N→C orientation.

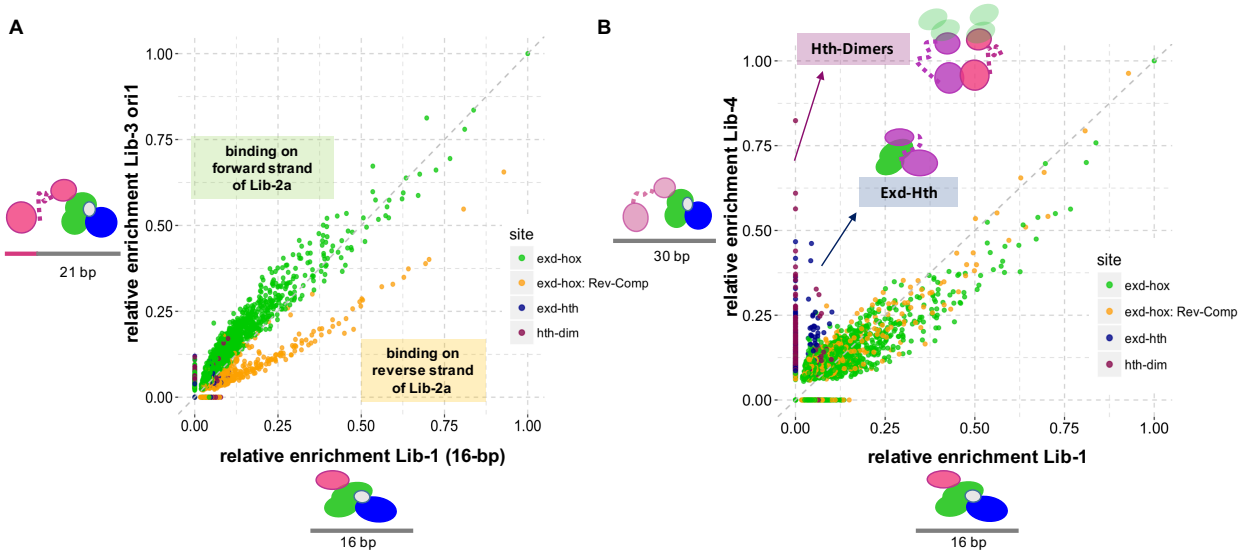


Figure 3.2: Comparing Hth^{HM} and Hth^{FL}-binding in complex with Exd-Dfd: **(A)** relative 12-mer enrichment for Lib-3a is compared to the one from Lib-1 for Exd-Dfd. Two classes of sequences are identified that can be described as either Exd-Dfd binding on the forward strand ($5' \rightarrow 3'$; green points) or Exd-Dfd binding on the reverse strand ($5' \leftarrow 3'$; orange points). The orange points were defined as sequences were the reverse-complement sequence match achieved a higher relative enrichment score, arguing for binding on the reverse strand. The split demonstrates that introducing a fixed Hth site favors binding on the forward strand (Hth-Exd-Hox) and disfavors (Hth-Hox-Exd). **(B)** relative 12-mer enrichment for Lib-4 is compared to the one from Lib-1. The split seen for Lib-3a is no longer apparent, since the symmetry-breaking fixed Hth binding site is no longer present in Lib-4. Two new groups of sequences appear, one with a signature of Hth-dimer binding (deep pink, up to 80% of the enrichment seen for Exd-Hox) and one arguing for an Exd-Hth complex composition (darkblue). The order suggests that the Exd-Hox is about twice as stable as Exd-Hth but only slightly more stable as Hth dimers.

3.2.2 Spacing and Orientation Preferences of Tetrameric Protein-DNA Complexes

To identify preferences for specific orientations of the Exd-Dfd subcomplex with respect to the fixed Hth binding site in Lib-3a and Lib-3b, we scanned the forward (F) and reverse (R) strands of each selected SELEX DNA probe using the Exd-Hox consensus match NT-GAYNNAYNNN and computed the relative enrichment on either strand for each 12-mer using the R SELEX package. Next we calculated the F/R enrichment-ratio for each 12-mer ignoring the respective distance to the Hth site. Both libraries (Lib-3a and Lib-3b) have $F/R > 1$ compared to a ratio of ~ 1 in the Hth$^{\mathrm{HM}}$-Lib-1 (Figure 3.3 A on page 110). The preference for DNA sequences harboring a Hth-Exd-Dfd over those sites where the Exd-Dfd supcomplex is bound on the R strand, thus resulting in the Hox C-terminus facing Hth (Hth-[bɬꓷ-bxꓱ]), demonstrates the presence of cooperative binding between the three-DBD-containing TFs that is independent of the orientation of the fixed Hth site. The preference for the Hth-Exd-Hox orientation is perhaps less surprising considering the direct protein-protein link that connects Hth and Exd. In reverse [bɬꓷ-bxꓱ] orientation, the Hth unstructured linker domain connecting the PBC and HM-domain would have to stretch across the Hox protein in order to reach its interaction partner Exd (compare Figure 3.3 C on page 110).

Given the presence of a flexible protein linker and the large spread of F/R enrichment ratios for Exd-Hox 12-mers in Lib-1 and Lib-2, we hypothesized that binding of the Hth$^{\mathrm{FL}}$-Exd-Hox complex not only depends on the relative orientation of the three DBDs, but also on the DNA spacing between Hth and Exd-Hox sites. Indeed, splitting the F/R 12-mer ratios by their distance from the fixed Hth site (offset in bp) revealed spacing-dependent differences in F/R ratios and thus cooperative behavior (compare Figure 3.3 B on page 110). To systematically model spacing and configuration preferences we scanned each round 2 (R2) Lib-3 21-mer probe with the Exd-Dfd PSAM obtained from the Hth$^{\mathrm{HM}}$-Exd-Dfd dataset (Lib-1) and retained only those probes where a single binding mode explained $> 95\%$ of the

probe selection (see Experimental Procedures for details; a similar strategy has also been describe here (Zhang et al., 2018)). Relative contributions for both configuration and Exd-Hox sequence affinity to the total energy of binding were inferred by fitting a generalized linear model to the filtered R2 SELEX data (Figure 3.3 C on page 110).

$$\Delta\Delta G_{\text{Complex}} = \Delta\Delta G_{\text{config}} + \Delta\Delta G_{12\text{mer}-\text{Affinity}} \tag{3.1}$$

The "configuration" model parameter was defined as Exd-Hox orientation (F or R) in combination with distance to the Hth site and values for the Exd-Hox affinity parameter were taken from Lib-1 to avoid biases resulting from Hth^{FL}-induced cooperativity. As expected, energy coefficient values most preferable for binding were obtained for combinations of DNA spacers with Exd-Hox binding in forward orientation for both Lib-3a and Lib-3b (Figure 3.3 C on page 110). However, Lib-3a and Lib-3b showed different spacing preferences, with an optimal DNA spacer length of 3-10 bp for Lib-3a and of 0-4 bp for Lib-3b (Figure 3.3 C on page 110). These preferences are in agreement with the assigned N→C terminus orientation of Hth (based on a recent crystal structure for MEIS1 (Jolma et al., 2015)). A shorter spacer is required for Lib-3b when the N-terminal arm of Hth (and with it the HM-domain) is facing towards the fixed flank of the library and thus needs to bend back over the Hth DBD to reach Exd, whereas in Lib-3a, the HM domain is facing the N-terminal end of Exd's DBD, and can thus more readily accommodate a longer DNA spacer. Rationalizing the protein orientation and spacer length with respect to the underlying protein domain configuration thus provides a means to characterize binding modes even in the absence of a crystal structure. To verify that the model coefficients represent true spacing preferences of complex binding, we performed a competition assay for the Hth-Exd-Hox complex using the Lib-3a binding orientation and three distinct DNA spacers (0, 3, or 7 bp). As predicted by our model, spacers of 3 and 7bps have similar ability to compete with a labeled DNA probe, whereas a ~ 7 times higher concentration was needed for a spacer of length 0 bp (Figure 3.3 D on page 110)

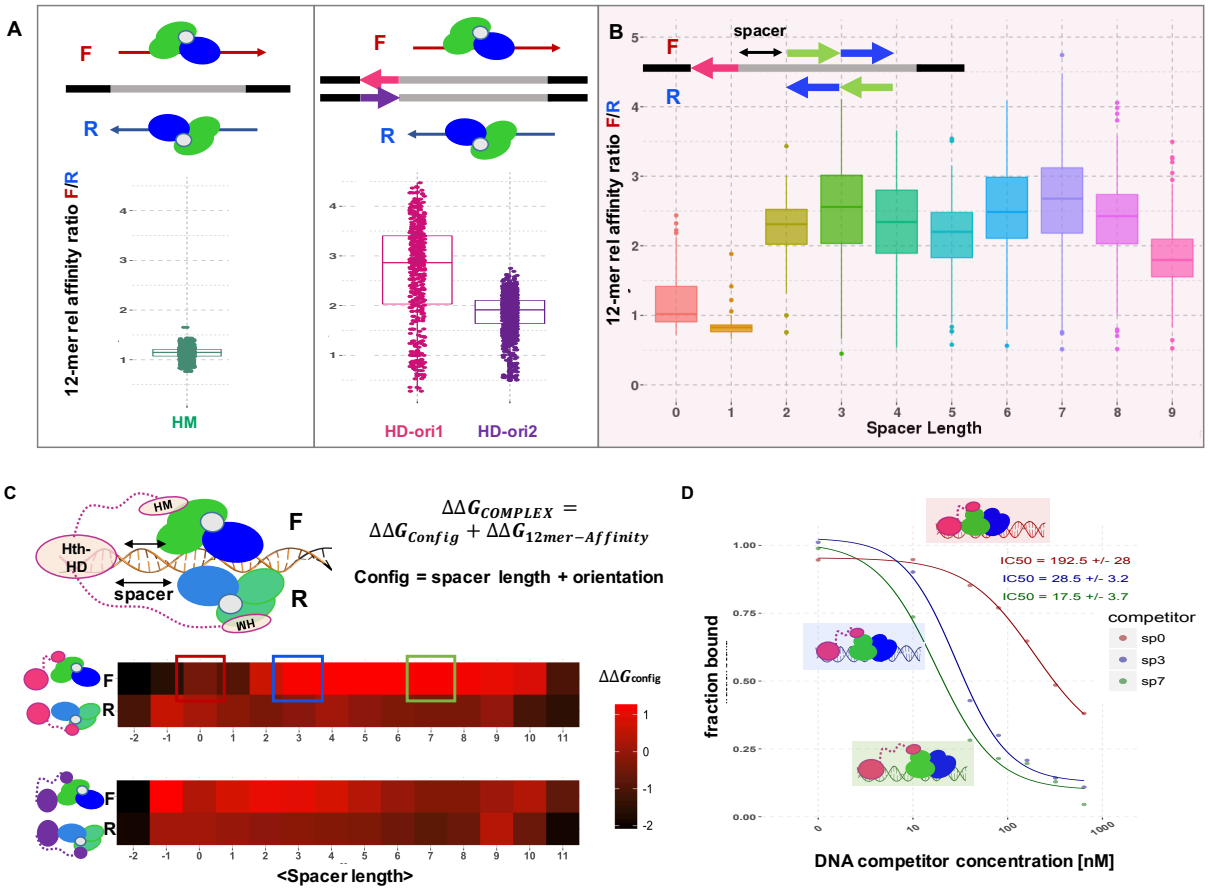


Figure 3.3: The role of complex configuration in binding site recognition – modeling of orientation & spacer length **(A)** Binding orientation preference of the Exd-Hox subcomplex with respect to the fixed Hth site. F defines binding of Exd-Hox on the forward strand (Hth-site − Exd-Dfd) and R on the reverse strand (Hth-site − [bϮD-bxƎ]). Comparing the relative enrichment ratio of F versus R for each 12-mer sequence shows a mean ratio of > 1 for both Lib-3a and Lib-3b, but a mean ratio of ~ 1 for Lib-1 where no Hth-HD is present. The F/R preference argues for a favorable Hth-Exd-Hox configuration independent of the Hth HD orientation. **(B)** Splitting the F/R ratio by offset from the fixed Hth site in Lib3-a. Starting at a DNA spacers of 2bp, an increase in the orientation selectivity is observed. **(C)** Using a generalized linear model to model the energetic contributions of different complex configurations for Hth-Exd-Hox binding for both Lib-3a and Lib-3b.

Figure 3.3: *continued from page 110;*

Configuration is defined as the combination of Exd-Hox orientation (F or R) and a specific spacer length. Lib-1 derived Exd-Dfd 12-mer affinities are used as predictors in the model, together with the configuration of the Exd-Dfd binding site. Both libraries disfavor the R orientation. Lib-3a prefers longer spacers (3-10 bp) whereas Lib-3b prefers shorter spacers ($-$1-4bp), suggestive of the Hth N-terminal arm facing Exd in Lib-3a but not Lib-3b. **(C)** Dose response curve from a competition EMSA for Hth-Exd-Dfd and a labeld DNA probe (CTGTCA-AAA-ATGATTAATGAT-flank. Three different cold probes, with spacers of 0,3 or 7bp (labeled in Panel B ; see Experimental Procedures for sequences), were used to compete out the labeled probe. The 7bp spacer was most efficient in competing with the labeled probe ($IC50 = 17.5 \pm 3.7nM$), closely followed by a 3bp spacer ($IC50 = 28.5 \pm 3.2nM$), and a $\sim$ 7-10 fold difference to the spacer of 0bp ($IC50 = 192.5 \pm 3.7nM$), agreeing with the model predictions in panel B.

3.2.3 Shape Readout of Flanking DNA Drives DNA Spacer Selection

Preferences for a specific spacer length can be attributed to geometric constraints of the multi-protein complex, such as the extent to which the Hth linker region is capable of connecting Hth's HM domain with Exd. When focusing on spacers of a given length, however, we noticed differences in enrichment of particular sequences that persisted when fixing the identity of the downstream core Exd-Dfd binding site (Figure 3.4 A on page 113). Such spacer-sequence-dependent variability in affinity is surprising as this region is presumably not

directly contacted by amino acids of either Exd or Hth. Furthermore, no obvious sequence pattern was apparent when looking at the most enriched spacer sequences for varying offsets and libraries. Therefore, we wondered whether intrinsic shape features of the DNA spacer might be responsible for the observed selection. To first model the sequence preferences of a spacer of length L, we again used our GLM framework to fit the R2 counts of $(L+12)$-kmers downstream of the fixed Hth site using the spacer sequence and the affinity of the downstream Exd-Hox 12-mer as predictors (Figure 3.4 B on page 113). Next, we ranked the 4^L possible spacers based on their predicted energetic contribution to complex binding and computed their DNA minor groove width (MGW) using the pentamer tables from (Zhou et al., 2013) (Figure 3.4 B on page 113). To test for a correlation between MGW and spacer affinity, we plotted the average (MGW) profile for sets of spacers, requiring their affinities to be above an increasingly higher threshold. The baseline MGW profile contained all spacers, whereas the highest affinity group only contained spacers of affinity > 0.95. The profile showed a strong correlation between MGW narrowing and spacer selection for particular postions (Figure 3.4 C-top on page 113). Recently, it has been shown that much of the variance observed in the pentamer tables can be explained by a simple mononucleotide model ($\sim 60\%$, see (Rube et al., 2018)). We therefore rationalized that the effects could perhaps be captured using a simple mononucleotide feature model for the entire L+12 bases in the binding site. Using a mononucleotide model has several benefits, such as greatly reducing the parameter space (instead of 4^L L-mer features, there are only $4 * L$ mononucleotide ones), and allowing the modeling of larger spacers, where observations for individual spacer sequences are sparse. Indeed, fitting a mononucleotide model to the same data and computing the spacer affinity by summing over the individual base coefficients, resulted in a similar spacer ranking (R^2 of 0.81) (Figure 3.4 B-bottom on page 113) confirming that a simpler model could be used to assess the potential shape readout. Likewise, average MGW profiles were comparable between the models (Figure 3.4 C on page 113).

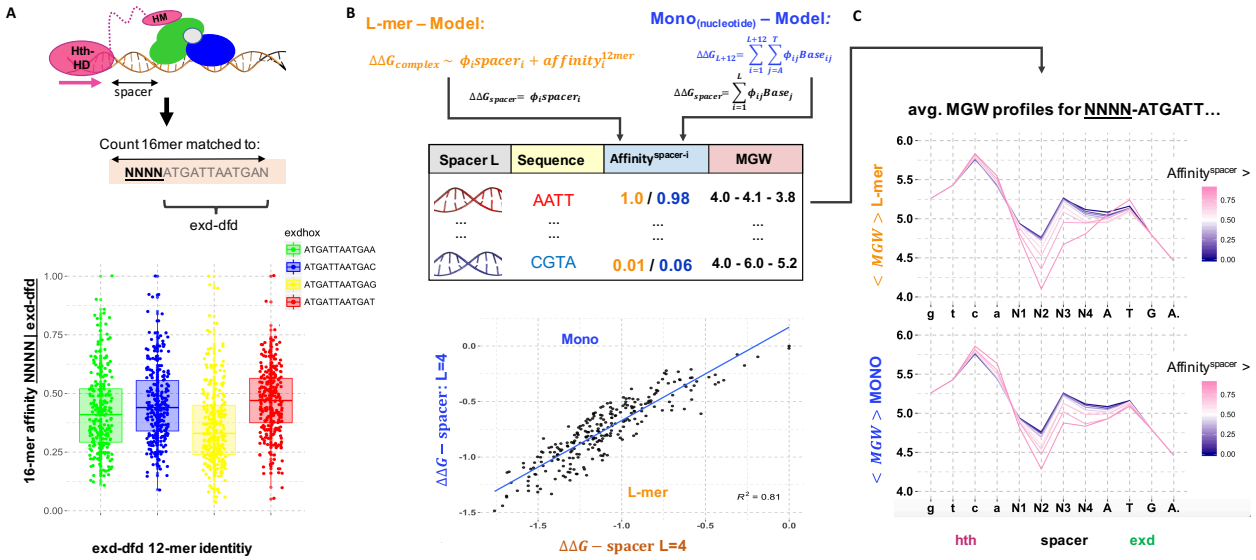


Figure 3.4: Modeling spacer sequence preferences in terms of DNA shape features **(A)** Variation in 16-mer relative enrichment for Lib-3a. Relative enrichments of 16-mer sequences directly downstream of the fixed Hth sites were computed and only those with a spacer of 4bp and matching one of four different Exd-Hox top sites from position 5-16 (ATGATTAATGA**X** with X=A,C,G or T) were retained. Boxplots show a large spread of relative 16-mer enrichments (> 20-fold differences) independent of the X base identity, which only affects the overall mean of the sequences. **(B)** Capturing the contribution of spacer sequence identity to the total complex binding energy using two different models: i) a L-mer based model using the full sequence of a spacer of length L and 12-mer Exd-Hox relative affinities from Lib-1 as predictors in the generalized linear model and ii) using base features for the entire L+12 binding site as predictors (MONO). Spacers are next ranked by either the coefficient they obtain in the L-mer model or the sum of base coefficients in the feature based model and MGW profiles for each spacer are computed. (bottom) Spacer coefficients derived from either model are compared, showing good agreement ($R^2 = 0.81$), indicating that a mononucleotide model captures much of the observed shape readout. **(C)** MGW average profiles for spacers above an increasing thresholds on spacer affinity (blue = all spacer; hot pink= only spacer > 0.95) are plotted for either L-mer or MONO model, again showing decent overlap. A strong affinity-MGW shape profile correlation indicates a potential role of shape readout in sequence selection

Further, more detailed analysis at different spacer lengths revealed two MGW minima that seemed to associate with either Hth or Exd binding (Figure 3.5 A on page 116). The base identity of the first three positions in the Exd-Dfd site was fixed to ATG to guarantee decoupling of contributions from bases within the spacer and the core of the Exd-Hox binding site. This is necessary since the shape tables are based on pentamers, thus each shape value is intrinsically conditioned on the two up-and downstream bases surrounding the central position for which shape is predicted.

To test if the apparent Exd and Hth MGW readout are independent of each other, we computed the same MGW profiles for Lib-1 using Hth^{HM} instead of Hth^{FL}. Indeed, Hth^{HM}-Exd-Dfd showed a stronger MGW selection upstream of the Exd binding site but no second MGW dip (Figure 3.5 C on page 116). Given this behavior, we further hypothesized that amino acids within the N-terminal arm of both Exd and Hth could read out the width of the minor groove, similar to what has been reported for the N-terminal arm of Hox factors (Joshi et al., 2007; Abe et al., 2015). We identified conserved, positively charged arginines in both proteins that given the existing crystal structures for MEIS and Exd-Hox are within reach of the DNA minor groove. Although they were part of the construct used for crystallization, those N-terminal regions of the homeodomains are disordered in the majority of existing crystal structures (as indicated by the B-factors), and in some cases, the N-terminal arginines are only partially modeled or even entirely missing. These observations argue for a more transient readout mechanism, such as MGW recognition, instead of base-specific hydrogen bonding, especially given that high affinity sequences were used in the crystal trials, all containing the Exd TGAY core with varying flanks (Figure 3.5 B on page 116).

To rigorously test if we indeed had identified shape-mediated spacer selectivity, we made several single and double argingine (R) to alanine (A) mutations within the N-terminal Exd DBD and also mutated all three positively charged amino acids for Hth. Binding of Hth^{FL}-Exd-Dfd was not significantly compromised for most combinations of Hth^{FL}-Exd mutations, with the most severe effect seen for the Exd- R2A & R5A double mutant (referred to as

Exd^{MUT} from here on) (Figure 3.6 B & C on page 119). At high concentrations, the Hth^{FL}- Exd^{MUT} complex started super-shifting, yet at more physiological, lower concentrations bound as well as its wild-type counterpart (Exd^{WT}). We therefore performed SELEX experiments with both Exd^{MUT} and the Hth mutant protein (K3A & K4A & R5A ; referred to as $^{MUT}Hth^{FL}$ from here on) in complex with wild-type Hth^{FL} & Dfd and wild-type Exd & Dfd respectively. Fitting mononucleotide models to the mutant data and analyzing them in terms of MGW selection confirmed that the two MGW selection minima disappeared dependent on the mutant protein used in the experiment (Figure 3.5 D on page 116). Moreover, a stronger selection for the remaining MGW minimum was observed, arguing for a model where the N-terminal arms for Exd and Hth compete for shape-mediated interaction with the DNA spacer. This result is yet another indicator that the Hth binding orientation can be inferred from *in vitro* data even in the absence of prior structural information. The dissapearance of the MGW minimum located near Hth upon mutation suggests that the N-terminal Hth arm faces Exd in Lib-3a, where the Hth site is fixed to 5′ **CTGTCA** 3′.

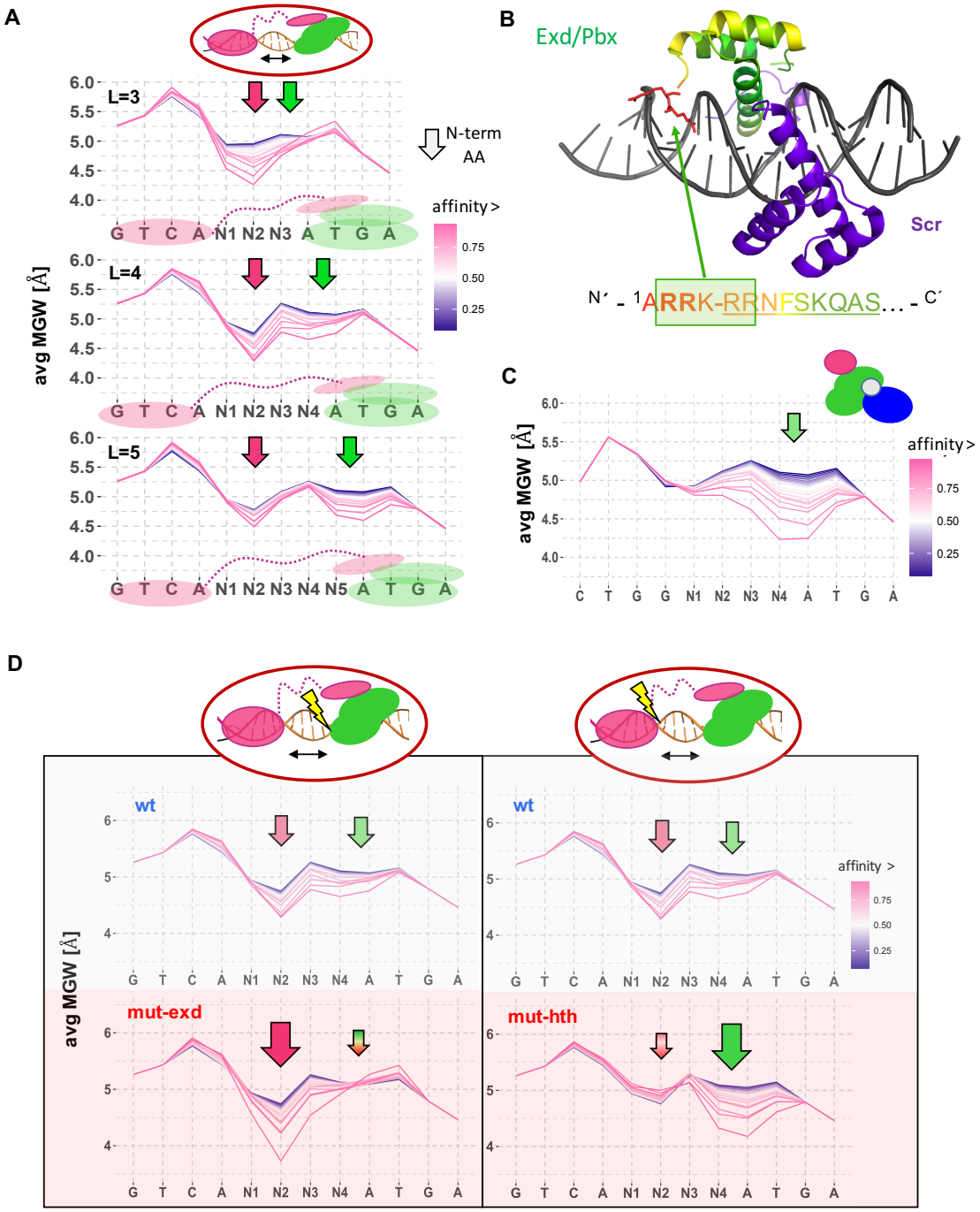


Figure 3.5: Shape readout competition between two homeodomains explains spacer sequence preferences: **(A)** Average MGW is shown for different affinity ranges for spacers of length 3,4, and 5bp (Lib-3a).

Figure 3.5: *continued from page 116*

For each spacer lengths, the average MGW of all 4^L different spacers (blue profile) were compared to the average MGW of subgroups of spacers with increasing contributions to the overall complex stability (pink gradient). Arrows mark the two observed MGW selection minima respective to the two homeodomains present – Hth (pink) and Exd (green). The Green MGW minimum appears to be moving with Exd, whereas the pink one stays localized with Hth. **(B)** Crystal Structure of Exd-Scr (PDB-ID: 2R5Y), with Exd colored by B-factors and Scr in purple. Sequence shows the first 13 amino acids of the Exd HD and conserved arginines are boxed. Underlined amino acids are resolved in the structure, with R5 and R6 shown as sticks. R5 and R6 are highly flexible (red B-factor) and located in proximity to the DNA minor groove, arguing that the extendended N-terminal end of Exd's HD (R2 to R6) are in a favorable position to read out MGW as marked with the green arrow in A. **(C)** Proposed Exd MGW readout is present in the Lib-1 Hth^{HM}-Exd-Dfd data, supporting the hypothesis that the Exd-proximal MGW minimum in A is caused by Exd. The absence of a second minimum also supports the notion of Hth amino acids being responsible for the Hth proximal minimum in A. **(D)** Comparison of affinity-MGW profile correlations (spacer = 4 bp) between i) Hth^{FL}-Exd-Dfd and Hth^{FL}-Exd^{MUT}-Dfd (mutation = R2A & R5A; left panel) and ii) Hth^{FL}-Exd-Dfd and $^{\text{MUT}}\text{Hth}^{\text{FL}}$-Exd-Dfd (mutation = K3A & K4A & R5A; right panel). Mutating two arginines within Exd selectively abolishes the Exd-proximal MGW minimum and reinforces a selection of spacers that narrow the Hth-proximal minor groove, whereas mutating arginines within Hth abolishes the Hth-proximal MGW minimum, causing the Exd-proximal to resemble the profile seen for Hth^{HM}-Exd-Dfd (compare panel C).

3.2.4 Loss of MGW Readout Impacts Complex Stability in a Hth Isoform-Dependent Manner

To investigate MGW shape readout by Exd in isolation from Hth^{FL} binding, and to test which amino acids are responsible, we mutated each of the four arginines within the N-terminal arm of Exd's HD. When performing EMSA experiments for Hth^{HM}-Exd-Dfd, we surprisingly found that two (R2 and R5) of the four single-amino-acid mutations tested abolished complex binding to the same extent as the known N51A mutation, which removes a crucial hydrogen bond between the DNA major groove and helix 3 of Exd's DBD (Figure 3.6 A on page 119). This result was unexpected, given the high degree of flexibility or complete absence of those amino acids in existing crystal structures. Adding Hth^{FL} to even double mutants containing the R5A mutation restored binding, most likely due to the stabilizing role of the third HD, reducing the entropic cost of keeping Exd close to the DNA (Figure 3.6 B on page 119).

To verify whether the most severe double mutation – Hth^{FL}-Exd^{MUT} (R2 & 5A) could still bind in the absence of Hox, we again performed EMSA experiments. Although Hth^{FL}-Exd^{MUT} shows the same supershifting behavior at increasing concentrations as Hth^{FL}-Exd^{MUT}-Dfd, it is still able to bind (Figure 3.6 C & D on page 119).

To our knowledge this is the first time, that shape readout has been demonstrated to be crucial for complex binding in a cofactor-dependent manner. As a consequence, Exd shape readout might be used to distinguish binding by Hth^{HM}-Exd from that by Hth^{FL}-Exd in an unbiased manner. This is of particular importance given the limited knowledge we have about the respective *in vivo* functions of the two isoforms, and what their downstream target genes are.

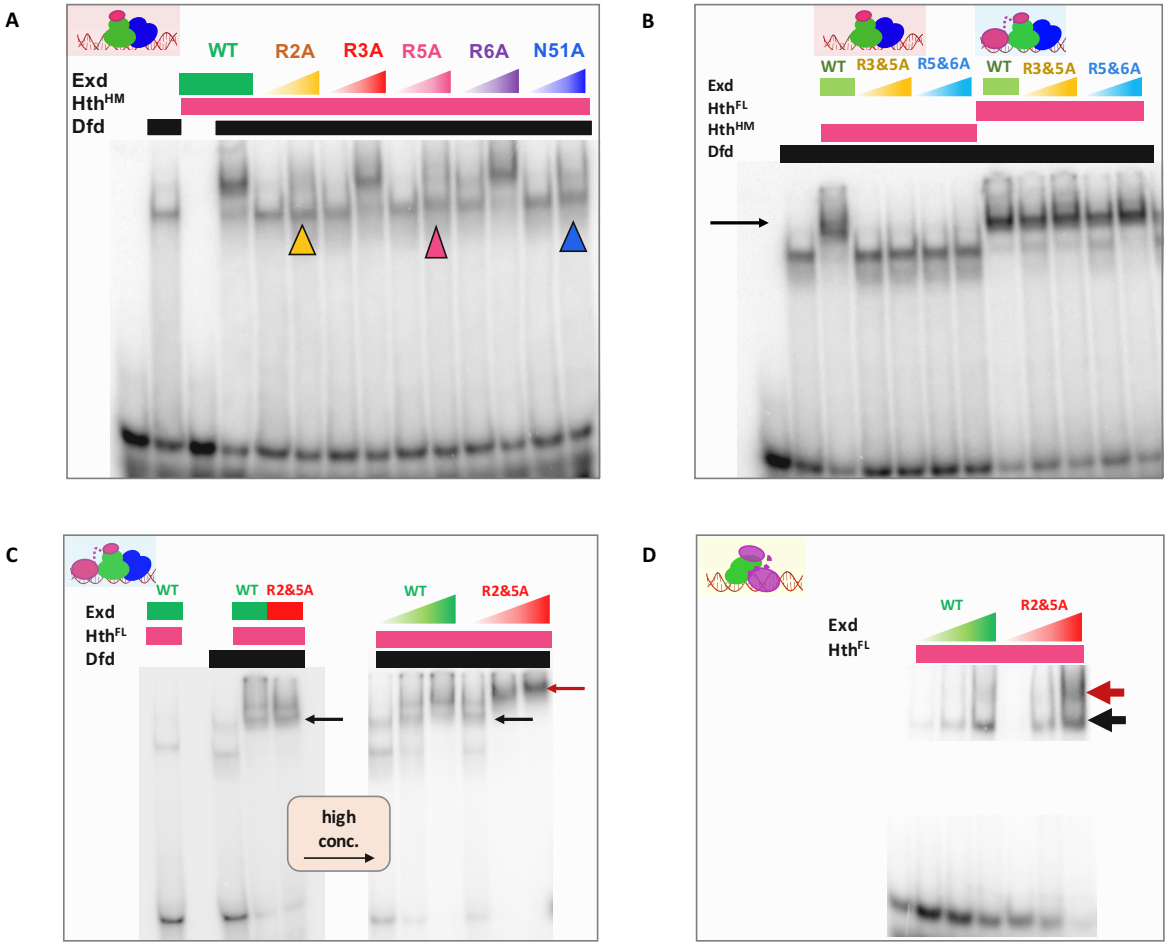


Figure 3.6: Complex destabilization due to loss of shape readout is Hth isoform specific **(A)** EMSA gel using the Hth^{HM} isoform in complex with Exd-Dfd and a labeled DNA probe matching the consensus (ATGATTAATGAT). Lane 1: probe only, lane 2: Dfd only, lane 3: Hth^{HM}-Exd, lane 4: Hth^{HM}-Exd-Dfd, lane 5-6: Hth^{HM}-ExdR2A-Dfd, lane 7-8: Hth^{HM}-ExdR3A-Dfd, lane 9-10: Hth^{HM}-ExdR5A-Dfd, lane 11-12: Hth^{HM}-ExdR6A-Dfd, lane 13-14: Hth^{HM}-ExdN51A-Dfd. Arrows indicate mutations that cause a severe binding loss, including R2A, R5A (both presumptive shape-readout mutations) and N51A (a hydrogen bond disruption in the major groove). **(B)** Hth^{FL} can rescue binding of Exd mutants containing the R5A mutations. Lane 1: probe only, lane 2: Dfd only, lane 3: Hth^{HM}-Exd-Dfd, lane 4-5: Hth^{HM}-ExdR3A& R5A-Dfd, lane 6-7: Hth^{HM}-ExdR5A& R6A-Dfd, lane 8: Hth^{FL}-Exd-Dfd, lane 9-10: Hth^{FL}-ExdR3A& R5A-Dfd, lane 11-12: Hth^{FL}-ExdR5A& R6A-Dfd.

Figure 3.6: *continued from page 119*

(C) Hth^{FL} can even rescue binding of the double Exd mutant for R2A & R5A (both capable of causing HM-loss of binding alone). Lane 1: Hth^{FL}-Exd, lane 3: Dfd, lane 4: Hth^{FL}-Exd-Dfd, lane 5: Hth^{FL}-ExdR2A & R5A-Dfd, lane 5-7 lane 4: Hth^{FL}-Exd-Dfd with increasing concentrations (50-300nM), lane 8-10: Hth^{FL}- ExdR2A & R5A -Dfd with increasing concentrations (50-300nM). mutant Exd starts supershifting at lower concentrations compared to wild-type Exd. **(D)** Testing binding for Exd-Hth in the absence of Hox (probe sequence = TGATTGACAG). Lane 1-3: Hth^{FL}-Exd with increasing concentrations (50-300nM), lane 4-6: Hth^{FL}-ExdR2A & R5A with increasing concentrations (50-300nM). Supershifting again occurs more pronounced for the Exd mutant compared to wild-type.

3.2.5 Strength of MGW-Dependent Spacer Selection Varies with Exd Binding Site Sequence

As briefly mentioned above, shape parameters, and MGW in particular, are dependent on the base identity at neighboring nucleotide positions (± 1 or ± 2). We hypothesized that fixing the first two positions (+1 and +2) within the core Exd-Hox binding site (**NT**GAYNNAYNNN) to XT, with $X = A, C, G$ or T and fitting glm spacer models separately for each XT, the dependence on neighboring nucleotides should be reflected in the coefficients at spacer position -1, if shape readout was indeed being used (Figure 3.7 A on page 123). Comparing spacer coefficients for XT = AT to either XT = TT or GT (CT was excluded due to insufficient observations), the coefficients for position -1 were indeed deviating strongest between different models, in particular when comparing AT and TT models. This was even true when including the coefficients fitted to the Hox flank (position 14-16, thought to have little influence on the overall affinity and thus are within the same order of magnitude) (Figure 3.7 A on page 123). Differences between the AT and GT model were less obvious, arguing that perhaps, sequences of type A or G-TGAYNNAYNNN might be read out differently by the protein complex compared to sequences of type T/(C)-TGAYNNAYNNN. To investigate this further, we compared the resulting average MGW profiles for each of the 3 independent models (Figure 3.7 B on page 123). Indeed, in agreement with the coefficients, the affinity-shape correlation plots were slightly different between AT and GT but deviated strongest for the TT model.

The trends observed are not sufficient to unambigiously demonstrate the presence of different recognition modes the complex might use depending on the primary sequence, since random shape features with similar complex structures could results in similarly strong correlations (see (Rube et al., 2018) for more detail). However, the Exd^{MUT} SELEX-data, where the shape readout is supposedly abolished or strongly diminished, should aid in identifying whether Hth-Exd-Hox indeed can bind DNA in subtly different conformational states, uti-

lizing shape readout with varying degree dependent on the given sequence context. In other words, whether removing the amino acids responsible for shape readout negatively impacts complex binding for specific sequences, but not others.

Direct comparison of the resulting average MGW profiles confirmed that the Exd shape readout is most strongly abolished for sequences of type AT, less so for GT, and almost not at all for TT (Figure 3.7 C on page 123). The variable impact the Exd shape mutation has on different sequences is also reflected in the energy logos obtained for Hth^{FL}-Exd^{MUT}-Dfd and Hth^{FL}-Exd-Dfd, which show a decreased weight for the Exd half-site, but more importantly, a change in the most preferred base at position +1 in the Exd-Hox core motif (see arrows in Figure 3.7 D on page 123). Plotting the 12-mer relative enrichment for sequences matching the consensus **NTGATNNATNNN** in libraries for Hth^{FL}-Exd^{MUT}-Dfd against Hth^{FL}-Exd^{WT}-Dfd (Lib-3a) illustrates the change in base preference more generally: 12-mers starting with a CT are not affected by the Exd mutation, whereas those starting with an AT are hit most severely by the shape-readout mutation (Figure 3.7 E on page 123). Together those results imply that i) we indeed observe selection of spacer sequences based on their shape (MGW) and ii) different sequences (with different shapes) are affected differently by the mutation. The latter point also implies that the complex assumes different "microstates" when encountering sequences with different shapes, and perhaps positions the Exd N-terminal arm differently along the minor groove. Considering the high flexibility of the N-terminal arm, this is indeed a plausible scenario.

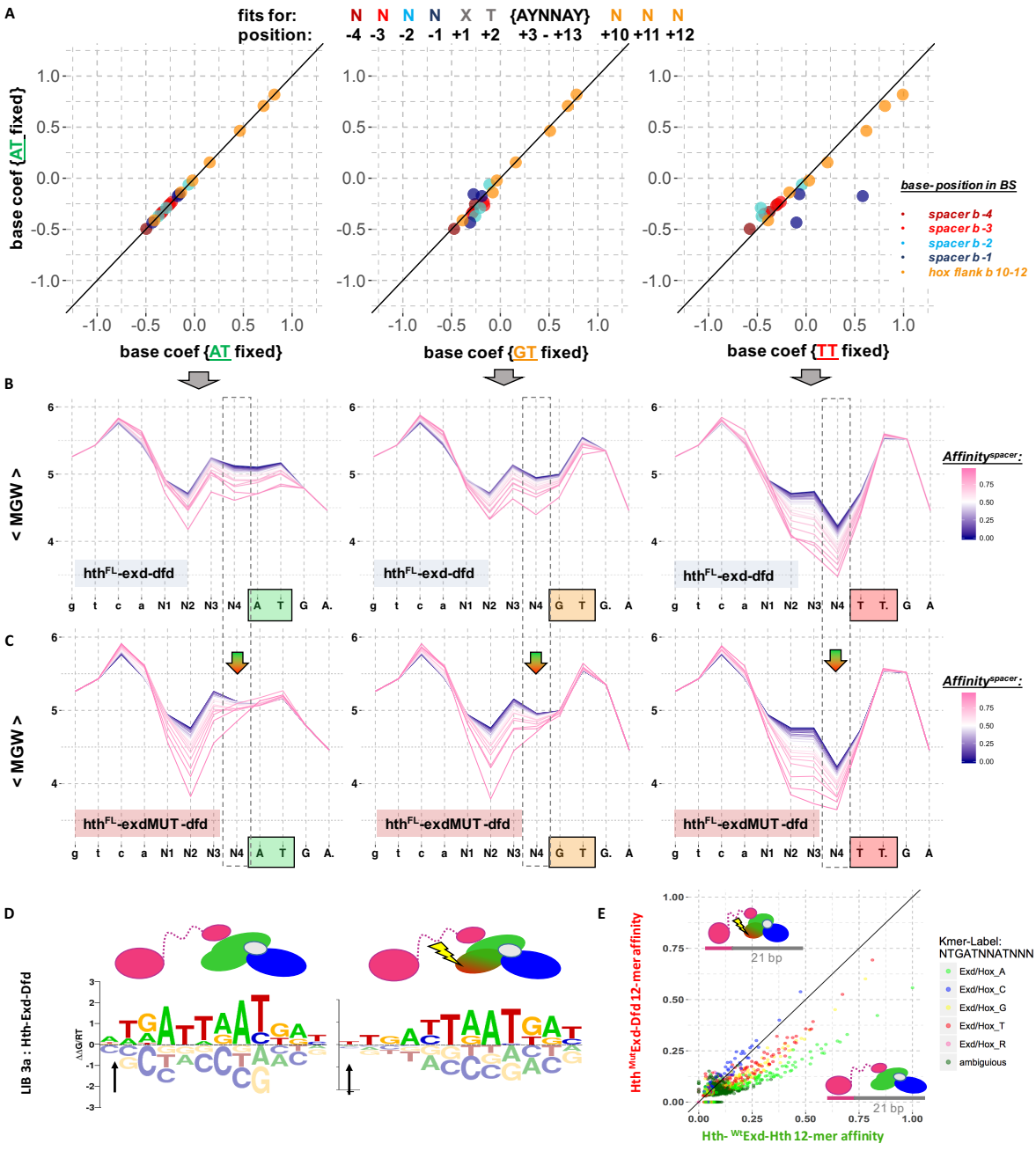


Figure 3.7: SELEX-seq analysis of Exd mutations reveals adaptive DNA binding by Exd's N-terminal Arm **(A)** base coefficient comparison from mononucleotide base-feature fits for sets of probes fixed at the N_{+1} and N_{+2} base positions ($\rightarrow XT$) of the Exd-Hox 12mer (**XT**GAYNNAYNNN).

Figure 3.7: *continued from page 123.*

The y-axis shows the $\Delta\Delta G/RT$ coefficients for bases within the spacer ($N_{-4} - N_{-1}$) and the Hox flank ($N_{+10} - N_{+12}$) for XT=AT. The x axis contains the model coefficients for either XT=AT, XT=GT or XT=TT. Due to dependency of shape on the neighboring nucleotides, fixing XT causes the strongest variation among different models for the N_{-1} base coefficients (dark blue dots), arguing that neighboring bases are not independent and shape readout might indeed cause the sequence selection. Remaining coefficients are unaffected by the conditioning as their shape is not influenced by the XT bases. N_{-1} coeffcients vary most between AT and TT model. **(B)** spacer affinity-shape correlation plots for AT, GT or TT fixed models. AT and GT show a similar baseline profile (blue lines), whereas fixing TT reveals an overall different starting shape profile. All three select for a narrow minor groove width, but use different spacer sequences to achieve this goal, as indicated by the variation in N_{-1} model coefficients. **(C)** same as B, but using the Exd "shape mutant" (R2A & R5A). Shape selection profile for sequences of type AT are most affected, followed by GT. TT-type sequences seem less affected by the mutation, suggesting that perhaps the N-terminal arm utilizes different modes to read the minor groove depending on the underlying sequence context. **(D)** Energy logos for Lib-3a 12-mers for either Exd-Dfd (left) or Exd^{MUT}-Dfd (right) show a switch in base preference for the X base (A to T), in line with the lesser degree the shape profile is affected for TT versus AT type sequences (compare C). **(E)** 12-mer scatterplot for Lib3a using either Exd^{WT} (x-axis) or Exd^{MUT} (y-axis) for sequences matching XTGAYNNAYNNN reveal the switch in the base preference observed for the X_{+1} base (A,T,G,C in the wild-type to T,G,A,C in the mutant. CT type sequences (not fit above due to a lack of wild-type observations) seem entirely unaffected by the mutation, again arguing that the N-terminal arm adopts different conformations given a specific sequence context, which are then impacted by the R2A & R5A mutations with varying degrees.

3.2.6 Shape-Readout-Mutant Reveals Adaptive DNA Binding of Exd-Hox Complexes Dependent on Sequence Context

When we take a closer look at the energy logos resulting from the Hth^{FL}-Exd^{WT} or Exd^{MUT}-Dfd SELEX-experiments in Figure 3.7 D on page 123, another change in sequence preference stood out: Intriguingly, the change occurs several bases away from the N_1 base, at the first Y (Y_5) in the consensus site **N**TGA**Y**NNAYNNN, where a cytosine is largely disfavored over thymine in the wild-type complex, whereas Hth^{FL}-Exd^{MUT}-Dfd seemingly tolerates a C (Figure 3.7 D on page 123). That mutations of flexible amino acids located at the N-terminal end of Exd (responsible for DNA shape recognition) would propagate through the complex and cause a shift in sequence preference more C-terminally was unexpected.

To investigate this more thoroughly, we first compared the Hth^{FL}-Exd wild-type and mutant SELEX data from Lib-2. We reasoned that in the absence of Hox, we could first rule out the possibility that the observed shift in sequence preference in the energy logos from the Hth^{FL}-Exd^{MUT}-Dfd library might have been caused by alternative Hth^{FL}-Exd^{MUT} complex formations. Confirming our suspicion, the Hth-Exd only data did not show any signs of enrichment of sequences that could have been classified as reminiscent of an Exd-Hox sequence. Rather, the sequences fell into two different classes: those strongly suggesting the presence of a Hth dimer (strong match to a PPSAM derived from TGACAG|CTGTCA) and those suggestive of the Exd-Hth site (strong match to a PPSAM derived from TGAT|TGACAG) (Figure 3.8 A on page 128). The preference for Hth-dimers in the Exd^{MUT}-Hth data indicates that Exd-Hth is also affected by the shape mutation, and that the super-shifting observed in the EMSA experiments is most likely caused by Hth-dimerization. Since in the absence of Hox, Hth-Exd binds DNA in a similar configuration as Exd-Hox, the destabilization is not surprising, as the Hth DBD is no longer adjacent to Exd's N-terminal arm. However, Exd^{MUT} -Hth^{FL} in contrast to Hth^{HM}-Exd^{MUT}-Hox can still bind, perhaps since the entropic cost is reduced by the direct protein-protein interaction formed between Exd and Hth. As also seen for Exd-Hox, the mutant Exd reverses the preference for the N_1

base from A-T-G-C in the wild-type to C-G/T-A in the mutant, again indicating that this sequence preference is directly linked to the positioning of Exd's N-terminal arm along the minor groove (Figure 3.8 A on page 128).

Since Exd-Hth alone could not explain the shift in sequence preference from T to C for the Y_5 position downstream of the primarily affected N_1 base, we went back to the Hth-Exd-Dfd Lib-3a,b data and compared mutant and wild-type complex composition preferences. To this end, we derived position-specific-affinity-matrices (PSAMS) from wild-type data for Hth^{HM}-Exd-Dfd (Lib-1), Hth^{FL} dimers (Lib-4) and Exd-Hth^{FL} (Lib-2) and scored each 12-mer sequence present in either Lib-3 Hth^{FL}-Exd^{MUT}-Dfd or Lib-3 Hth^{FL}-Exd^{WT}-Dfd with each of the three PSAMs. Complex composition for a given sequence was then assigned based on the highest PSAM score (see Experimental Procedures for details). Strikingly, four distinct sequence classes could be identified (Figure 3.8 B & C on page 128). Hth-dimer sequences were most preferred in the Exd^{MUT} library, as already described for the Exd-Hth libraries, followed by Exd-Hth sites and two types of Exd-Hox sequences – those matching the higher wild-type affinity consensus NTGA**T**NNAYNNN and those matching the lower wild-type affinity consensus NTGA**C**NNAYNNN. These two "affinity classes" both exhibited the same Exd-N-terminal arm shape mutant signature ("Exd^{MUT} fingerprint"), as already observed for Exd-Hth as a reversal of base preference for the N_1 base in the energy logos (Figure 3.8 B & C boxes on page 128). The same behavior was observed independent of the Hth-site orientation, which differs between Lib-3a or Lib-3b, indicating that the two classes indeed represent two distinct structural arrangements of the Exd-Hox complex.

To summarize, mutating two arginines within the Exd N-terminal arm does not only abolish the preference of a narrow MGW within the Hth-Exd spacer and a severe loss of Hth^{HM}-Exd-Hox binding, but it also reveals distinct "recognition modes" that Exd-Hox uses to accommodate binding to different DNA ligands, as summarized by the T→C switch. The fact that a mutation at a specific location within the TF-DNA interface selectively destabilizes certain types of sequences, characterized by a switch in base identity more than

three bases away from the primarily targeted base (N_1 versus Y_5), underscores that TF-complexes can adopt more than one conformational state to bind DNA ligands with different affinities. Presumably, the N-terminal arm in the "high affinity" TGAT-class is deeply buried inside the minor groove, therefore causing a severe loss of binding upon arginine removal, whereas it does not make extensive minor groove contacts in the "low affinity" TGAC-class, therefore not being impacted by the mutation as much. Moreover, the absence of such a splitting when Exd is bound to Hth alone implies that dependent on the binding partner Exd can adopt different conformations. Intriguingly, the selective destabilization of "high-affinity" Exd-Hox sites is induced by a mutation that was demonstrated to recognize structural DNA features and not base identity per se. Evolutionary changes in amino acids that do not directly contact individual base pairs can apparently cause a global shift in sequence specificity. In this case particular, it causes a shift in sequence preference upon cooperative binding with another TF, underscoring the complexity and importance of TF cooperativity. Moreover, the fingerprint observed for Exd that is independent of the binding partner (Exd-Hth vs Exd-Hox) is an example how amino acid variation at the edge of the DBD can fine-tune sequence specificity outside the core binding site.

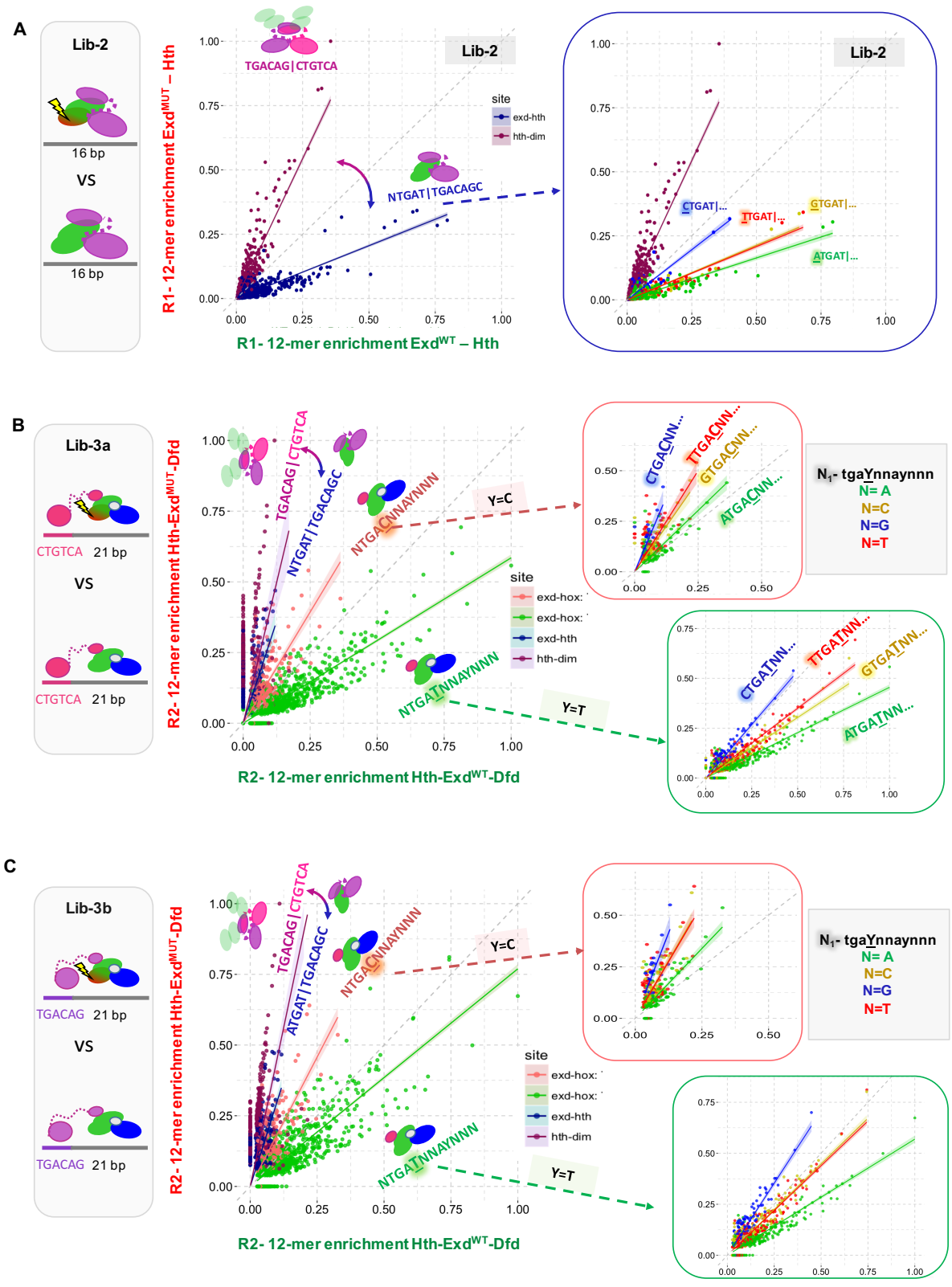


Figure 3.8: Exd^{MUT} causes switch in sequence selectivity upon dimerization with Exd-Hox:

Figure 3.8: *continued from page 128.*

(A) relative enrichment of 12-mer sequences from Lib-2 for Exd^{WT}-Hth (x-axis) is compared to that of Exd^{MUT}-Hth (y-axis). Sequences of type Exd-Hth, defined by a higher score for a PSAM derived from the top Exd-Hth site (ATGAT|TGACAGC), are all destabilized in the mutant and sites predictive of Hth-dimer binding (higher score for a PSAM derived from the sequence TGACAG|CTGTCA: dark pink) appear as the strongest bound sequences. Close up shows the same sequences, but with Exd-Hth type sequences colored by the the base identity of the first base (**N**TGAT|TGACAGC). The same change in sequence preference for the N_1 base, as previously seen for Exd-Hox type sequences is found, arguing that the mutated arginines predominantly read the MGW at this location. **(B)** relative enrichment of 12-mer sequences from Lib-3a (Hth fixed site) for Hth^{FL}-Exd^{WT}-Dfd (x-axis) is compared to that of Hth^{FL}-Exd^{MUT}-Dfd (y-axis). Sites are selectively destabilized in the mutant dependent on i) the complex composition (Hth-dimer < Exd-Hth < Exd-Hox type sites), ii) on two types of Exd-Hox sequences – those sites with a C at the Y_5 position within the Exd-Hox site (NTGA$\underline{Y}$NNAYNNN) and those with a T at position Y_5. The generally lower affinity C_5-type sequences are less destabilized compared to the higher affinity T_5-type sequences. Boxes show close up for both Exd-Hox type sequences and reveal, again, the direct impact of the Exd^{MUT} on the N_1 base preference. **(C)** like (B) but using the Lib-3b with the TGACAG Hth site orientation instead of CTGTCA.

3.2.7 Shape Readout is Important *In Vivo*

Based on our newly gained insights into the complex binding behavior of Hth-Exd-Hox, we predicted that we could use Exd^{MUT} as a tool to probe the use of different complex configurations and DNA sequence classes in a cellular context. Summarizing the different effects the Exd^{MUT} has on binding, we made the following predictions regarding *in vivo* binding:

1. Sites with strong Hth motifs should be bound similarly in both Exd^{MUT} and Exd^{WT}, given the tendency of Hth dimer formation in the Exd^{MUT} SELEX data;

2. Exd^{WT} but not Exd^{MUT} should prefer to bind as part of a trimeric Hth^{FL}-Exd-Hox or Hth^{HM}-Exd-Hox configuration and less so as part of a Exd-Hth^{FL} complex;

3. Exd^{MUT} on the contrary should prefer dimeric Exd-Hth^{FL} over trimeric Hth^{FL}-Exd-Hox binding and show the strongest loss of binding for Hth^{HM}-Exd-Hox sites;

4. When focusing on Exd-Hox sites, we should see a reversal between the "high affinity" class of TGA**T**NNAY sites and the "low affinity" class of TGA**C**NNAY sites between Exd^{MUT} and Exd^{WT}

A detailed overview is given in Figure 3.9 on page 131.

Given the varying degrees to which the shape mutation (R2A & R5A in Exd) impacts stability of different complex compositions and conformations, we reasoned that we could use the mutant protein as a "sensor" to interrogate complex composition *in vivo* and to learn something about their distinct functions. When considering *in vivo* TF binding, there are many more factors that need to be accounted for, e.g. local accessibility, molecular crowding, or the presence of cofactors or other DNA-binding proteins that could alter intrinsic DNA shape properties. Nevertheless, the observed magnitude of the identified Exd shape readout, in particular for the Hth^{HM}-isoform relative to Hth^{FL}, makes it a reasonable candidate to study to what degree adaptive binding mechanisms are utilized and relevant *in vivo*.

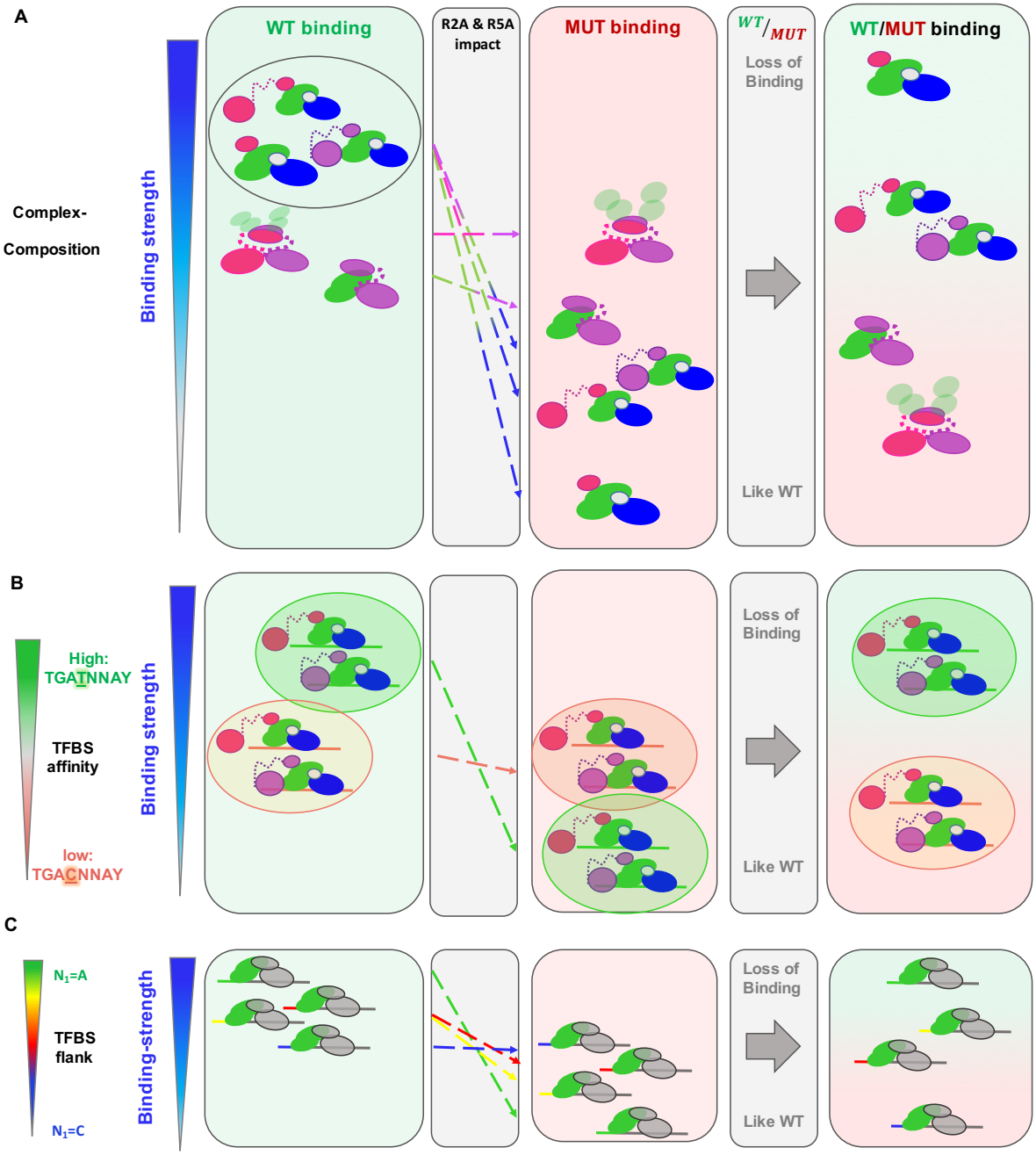


Figure 3.9: Overview: Impact of Exd^{MUT} on composition- and sequence-dependent Hth-Exd-Hox *in vitro* binding: **(A)** Summary of binding preferences for Exd^{WT} (green background) and Exd^{MUT} (red background) along with the predicted degree of loss of binding when Exd is mutated (ratio Exd^{WT}/Exd^{MUT} : green over red background).

Figure 3.9: *continued from page 131.*

Wild-type Exd prefers complex formation with Hox proteins over Hth, followed by Hth-dimers ("floppy" Exd attached) and Exd-Hth (compare Figure 3.2, Lib-4). R2A and R5A mutations reverse the order and cause Hth-dimers to be the new top binders (as binding is Exd-independent). Exd-Hth dimers only are slighty impacted by the mutation, followed by Hth-Exd-Hox trimeric DNA binding (compare Figure 3.8), where Hth^{FL} can stabilize the Exd-Hox complex such that binding is not completely lost. Finally, Hth^{HM}-Exd-Hox binding is most affected by the mutation as demonstrated by the binding loss comparable to that of the N51A mutation (compare Figure 3.6). From this follows that binding of the Hth^{HM}-Exd-Hox will be lost strongest, followed by Hth^{FL}-Exd-Hox, Exd-Hth and with no loss of Hth-dimeric sites. **(B)** Binding preference prediction as for (A) but now focusing on the DNA sequence context. Wild-type Hth^{FL}-Exd-Hox prefers sequences of type NTGA**T**NNAYNNN (green circle) over those of type NTGA**C**NNAYNNN (orange circle), whereas in the mutant the latter C_5-type sequences are getting boosted and are even surpassing many of the T_5-type seqeunces. Therefore, T_5-type sequences are lost strongest in the mutant compared to wild-type Exd. **(C)** As in B, but focusing on the N_1 base identity. N_1=A sequences are lost strongest in Exd^{MUT}, followed by N_1= G or T and N_1=C is least affected by the mutation.

We therefore generated transgenic flies by injecting either wild-type (Exd^{WT}) or mutant (Exd^{MUT}, using the R2A & R5A Exd cDNA construct) Exd tagged with a V5 epitope and under the control of a tubulin promoter into the attp40 landing site on chromosome 2L (Tub-Exd^{WT}or$^{\text{MUT}}$-V5; Figure 3.10 A on page 134). To verify functionality of the transgene, we crossed the Exd^{WT} or Exd^{MUT} flies to ones carrying an Exd^{-} allele (loss of function) and counted the number of males resulting from the cross. For the wild-type cross, we obtained several males, resulting in the successful generation of fly stocks that are ho-

mozygous null for the endogenous Exd and rescued by the presence of the Tub-Exd^{WT}-V5 transgene. The mutant cross, however, produced no males (0 out of 89; P-value=0.0008), indicating that Exd^{MUT} cannot rescue endogenous Exd expression. Since Exd function relies on Hth^{FL} or Hth^{HM}-dependent nuclear import, we first had to rule out the possibility that the observed lethality was due to failure of nuclear import of Tub-Exd^{MUT}-V5. Staining for both endogenous Exd and the V5 epitope in third instar imaginal wing discs demonstrated that both wild-type and mutant Exd transgenes overlap with endogenous Exd expression and are indeed nuclear (Figure 3.10 B on page 134). As the overall fluorescence signal is weaker for the V5 stain, we generated Exd^{-} clones in the Exd^{MUT} background to verify that nuclear import of Exd^{MUT} is efficient in the absence of endogenous Exd. Three clonal genotypes within the generated wing discs that differ in their relative ratios of endogenous to transgenic Exd need to be distinguished: i) one copy of endogenous Exd and one copy of the transgene (1:1 background); ii) no copy of endogenous Exd and one copy of the mutant transgene (0:1 background; "clone"); and iii) two copies of endogenous Exd and one copy of the transgene (2:1 background; "twin spot"). As expected, V5 intensity varies with the underlying cellular genotype, with increased levels inside the clones, but decreased levels within the twin spot (Figure 3.10 C on page 134). Staining with the Exd-specific antibody, however, is invariant to the genotype as it does not distinguish between endogenous and transgenic Exd. From this we can conclude several things: i) nuclear Exd levels are tightly regulated and constant independent of the genomic source, ii) nuclear, transgenic Exd levels depend on the ratio of endogenous versus transgenic Exd, which also implies that nuclear import might be proportional to the cytoplasmic ratio of the two and iii) the amount of Tub-Exd^{MUT}-V5 in the absence of *endogenous* Exd occurs at similar (if not identical) levels, demonstrated by invariant total Exd levels (as measured by the α -Exd antibody) across clone boundaries.

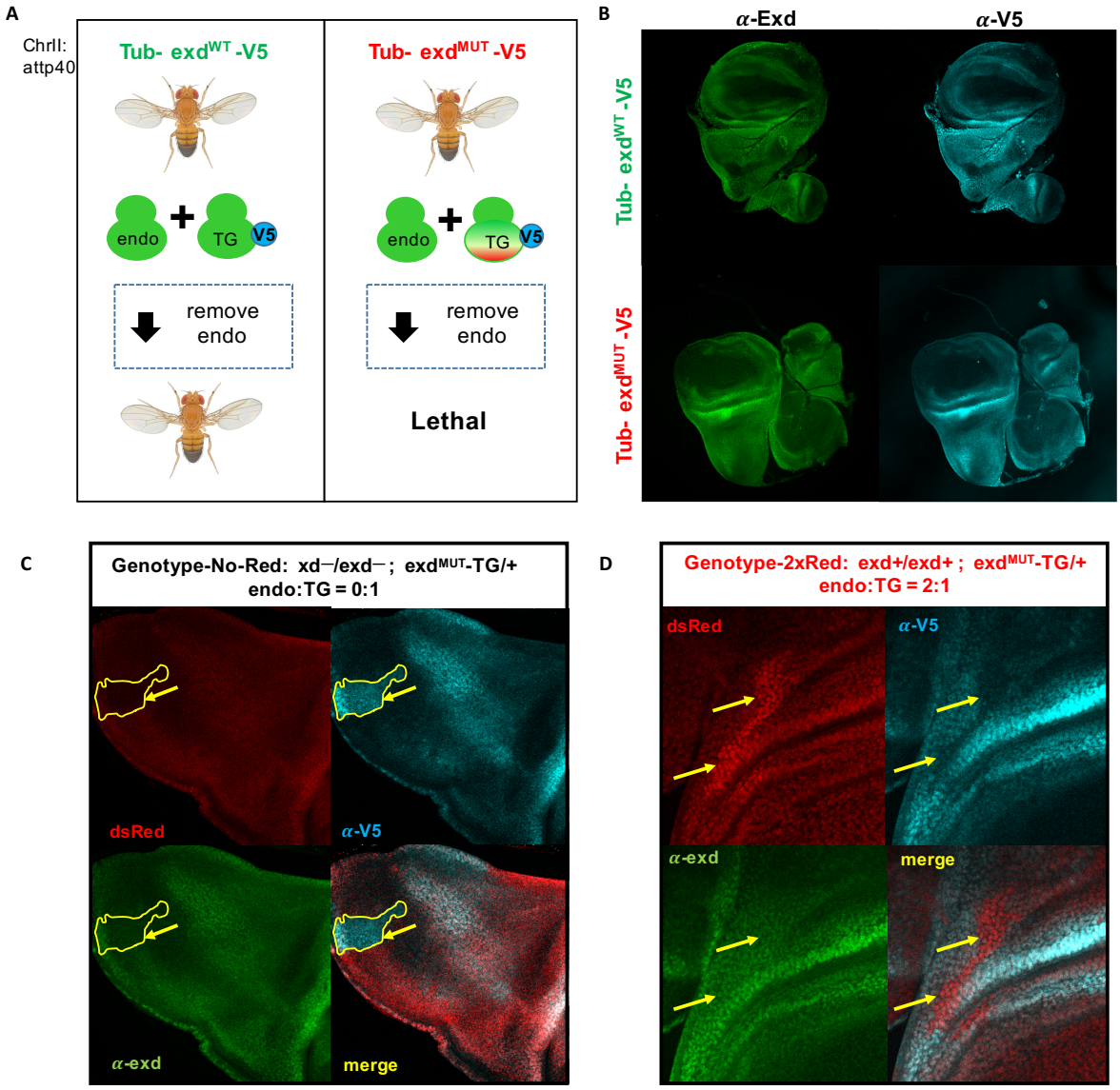


Figure 3.10: Nuclear import is unaffected in transgenic Exd^{MUT} flies: **(A)** Generation of transgenic flies and test for lethality. Tub-Exd-V5 either wild-type or mutant is injected into the attp40 landing site on chromosome II and offspring is crossed against flies carrying and Exd null mutation on the X chromosome. Transgenic Tub- Exd^{WT}-V5 rescues the Exd null mutation, whereas Tub-Exd^{MUT}-V5 does not. **(B)** Staining wing discs with antibodies against anti-Exd and anti-V5 for flies homozygous for either Tub-Exd^{WT}-V5 (top) or Tub-Exd^{MUT}-V5 (bottom) with the endogenous Exd in the background. V5 staining overlaps with the Exd signal for both genotypes.

Figure 3.10: *continued from page 134.*

(C) FLP-mediated clones removing the endogenous Exd (absence of dsRed) in flies heterozygous for Tub-Exd^{MUT}-V5. Clones carry 1 copy of V5-tagged Exd^{MUT} an no endogenous Exd, sourrounding has one copy of the endogenous and one copy of the V5-tagged Exd^{MUT}. Clones have an increase in V5 signal above background, whereas total Exd stain (green) remains constant across clone boundaries. **(D)** Same as C, but focusing on the twin spot, carrying two copies of endogenous Exd and one copy of V5-tagged Exd^{MUT} (twice dsRed). V5-signal decreases compared to background, whereas total Exd remains constant across twin spot boundaries. Levels of V5-tagged, nuclear Exd depends on the ratio of endogenous and transgenic Exd.

3.2.8 Exd^{MUT} Causes Genome-Wide Loss of Binding to Exd-Hox Sites

Given the viability of homozygous flies carrying either an Exd^{MUT} or Exd^{WT} transgene in addition to the endogenous Exd, we reasoned that despite the overall lethal phenotype, we can use the V5-tag to distinguish transgenic from endogenous Exd binding in those flies and thus assess whether the rules inferred from our *in vitro* binding study also govern binding *in vivo*. To do so, we collected ∼100 third instar larval wing discs from flies with either wild-type or mutant Exd genotype and performed ChIP-seq in replicates (Barski et al., 2007; Johnson et al., 2007; Mikkelsen et al., 2007) against the V5 tag. In addition we performed ChIP-seq in the wild-type genotype with antibodies raised either against the N-terminal or the C-terminal end of Hth, to potentially obtain information about Hth^{FL} specific sites, and we also assayed Antennapedia (Antp) binding (the Hox protein expressed in the wing) by performing ChIP-seq on third instar larval wing discs from flies homozygous for GFP-tagged Antp but otherwise wild-type.

In particular, we wanted to assess if and to what extent the following three *in vitro* phenomena were recapitulated *in vivo*:

1. the ability to use Exd^{MUT} V5-IP signal loss across different genomic sites as a "sensor" to detect Hth isoform-specific or Hox-dependent and -independent binding;

2. the presence of adaptive binding dependent on the underlying sequence context (differences between "high and low" affinity sites);

3. signs of cooperative binding at Hth-Exd-Hox trimer sites;

Using MACS2 (Zhang et al., 2008) and calling peaks with a false discovery rate of 0.01 showed good replicate agreement with 92% for Exd^{WT} & 83% for Exd^{MUT} peak overlap. As suspected, fewer peaks were called in the mutant (∼ 20% compared to wild-type). However, despite the reduction in total peaks called in the Exd^{MUT}, certain peaks were observed to

a similar degree as for Exd^{WT}, whereas other peaks appeared to be lost completely (Figure 3.11 A on page 139). In addition, when including the Antp IP signal, we could identify different categories of peaks: i) those with an Antp signal and either severe or ii) mild/no loss of Exd^{MUT}-V5 signal (perhaps indicating Hth^{HM} versus Hth^{FL} dependent sites), and iii) those without Antp signal and generally only a mild loss of Exd^{MUT}-V5 signal, perhaps indicating Hox-independent binding (Figure 3.11 A on page 139).

To systematically characterize the variability in Exd^{MUT} signal loss, we used the called peaks for the more deeply sequenced Exd^{WT} replicate and determined the raw read coverage at the peak summits (combined from both replicates) for both mutant & wild-type V5-IP (see Experimental Procedures for details). Plotting Exd^{MUT} versus Exd^{WT} V5 coverage shows an overall reduction, yet with a wide variability in signal loss for Exd^{MUT} (Figure 3.11 B on page 139). To test if the degree of signal loss might be explained by the presence or absence of an Exd-Hox or Hth site we divided the peaks into four groups based on their WT/MUT ratio (Figure 3.11 B on page 139). We next extracted the underlying peak sequences (50 bp centered around summit) and performed qualitative *de novo* motif discovery using HOMER (Heinz et al., 2010). Some of the motifs resembed the expected Exd-Antp or Hth motif, but we also observed a varying degree of significance and enrichment for Exd-Hox and Hth sites between peak classes (Figure 3.11 C on page 139). For the classes with least signal loss (labelled "highly stabilized" & "moderately stabilized"), the most significantly enriched motif – "primary motif" – strongly resembled a Hth site (TGACAG). However, in the two classes with stronger signal loss for Exd^{MUT} (labelled "moderately lost" & "highly lost") the posterior-type Exd-Hox sequence appeared as the new primary motif, which we assumed to be due to Exd- Antp binding (Figure 3.11 C on page 139). In addition, the significance of Hth site enrichment decreased with increasing mutant signal loss. Hth sites were also identified in all peak classes, suggesting that there might be an additional hidden layer contributing to signal loss. Nevertheless, the presence of Exd-Antp sites as the strongest

predictive feature of Exd^{MUT} signal loss is in line with our *in vitro* predictions. The peaks that are stabilized have a strong Hth signature and the presence of Hth sites in some of the lost sites may be indicative of differences in complex composition or DNA sequence, such as Exd-Hth, or Hth-Exd-Hox binding to either "high" and "low" affinity sites, which is expected to confound the correlation between Hth site presence and Exd^{MUT} signal stabilization.

To more generally analyze the different peak signatures, we ranked the peaks by the WT/MUT V5-IP ratio (how much they were lost) and plotted the raw IP coverage within ±1kb of the peak center for all five ChIP experiments (Exd^{WT}, Exd^{MUT}, Antp-GFP, N-terminal-Hth and C-terminal Hth ; Figure 3.11 D on page 139). Confirming our hypothesis, we see that the Antp ChIP-signal is strongest for the peaks most significantly lost in the Exd^{MUT} IP. However, within the bottom half of peaks (those with a milder signal loss) we observe a gradient of Antp IP-signal strength, perhaps indicating the presence of either trimeric Hth-Exd-Hox or lower-affinity Exd-Hox sites. In the latter case we would expect a weaker Antp IP signal and smaller differences between Exd^{WT} and Exd^{MUT} V5-IP signal (Figure 3.11 D on page 139).

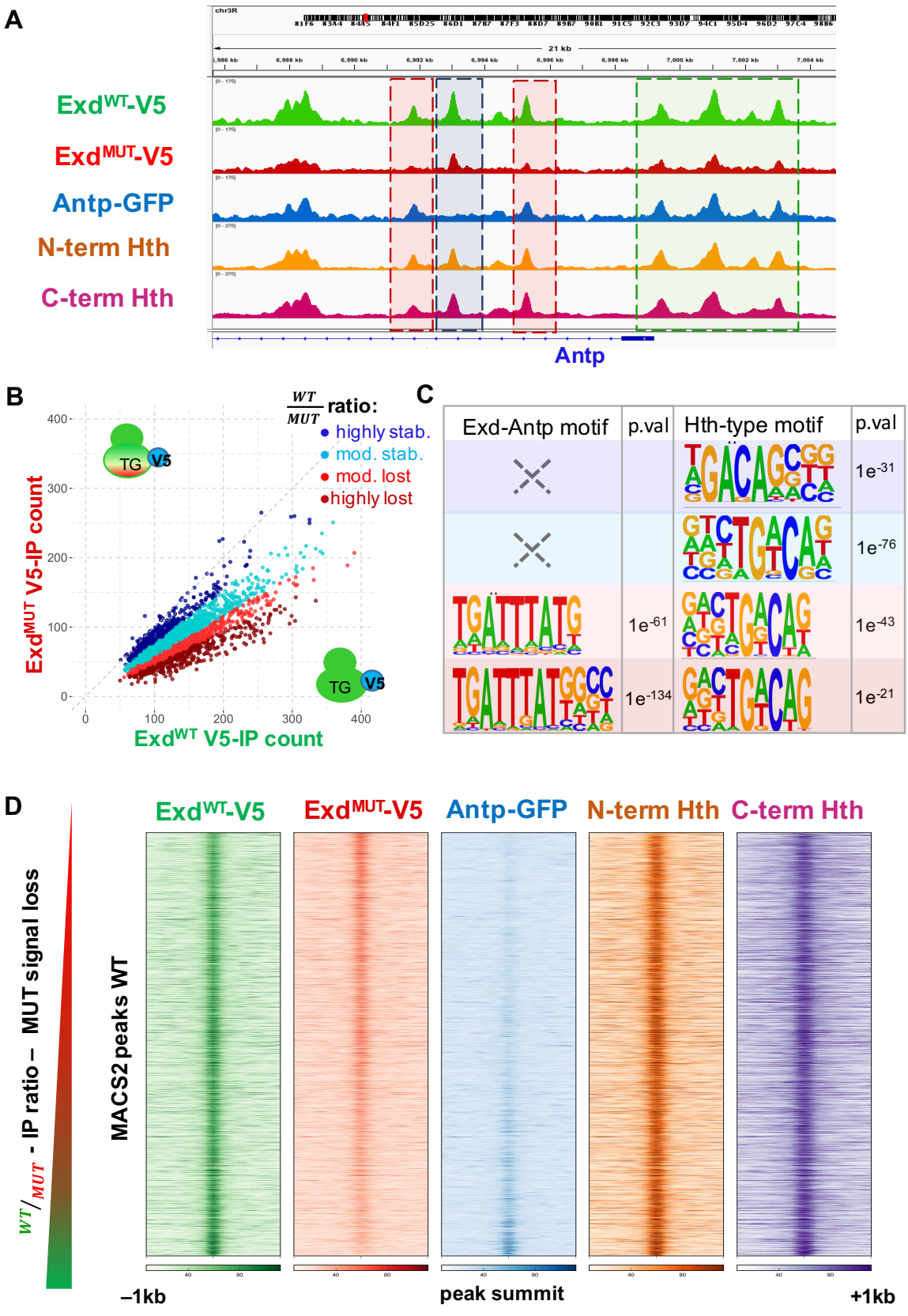


Figure 3.11: Exd^{MUT} causes site-specific loss of binding *in vivo*:

Figure 3.11: *continued from page 139.*

(A) ChIP-seq coverage at the Antp locus for Exd^{WT}-V5 (green), Exd^{MUT}-V5 (red), Antp (blue), Hth (antibody against N-terminus, orange) and Hth^{FL}-specific (antibody against C-terminus, purple). Boxes show three different types of binding sites. Red box: signal for all three homeobox TFs, but Exd^{MUT} signal is severely lost, similar to background. Blue box: no Antp signal and Exd^{MUT} signal is only mildly affected. Green box: signal from all three homeobox TFs (like red box), but Exd^{MUT} signal is still retained above background. **(B)** Raw IP coverage for Exd^{WT}-V5 (x-axis, combined replicates) and Exd^{MUT}-V5 (y-axis, combined replicates), colored by MUT/WT ratio (X, μ =mean & σ = standard deviation) : stab= $X > \mu + \sigma$, med-stab = $\mu + \sigma > X > \mu$, med-lost= $\mu > X > \mu - \sigma$, lost= $X < \mu - \sigma$. **(C)** De-novo motif discoveries on peaks in each class using 50bp centered around the peak summits. The two lost classes show an increasing Exd-Antp signature. Hth is present throughout the classes, yet to varying degrees. **(D)** Raw coverage signal for Exd, Antp, and Hth IPs for Exd^{WT} peak regions ($\pm 1kb$) ordered by their WT/MUT IP-ratio – signal loss in the mutant. Sites strongest lost in the mutant show a strong Antp IP signal.

3.2.9 Using Exd ^{MUT} to Detect a Hth^{FL} DNA Binding Signature in the Absence of Distinct Sets of Genomic Locations Bound by the two Hth Isoforms

One hypothesis regarding the function of the two Hth isoforms (Hth^{HM} and Hth^{FL}) is that they bind distinct genomic locations and therefore regulate distinct gene sets. Analyzing the IP signal obtained from two ChIP-seq experiments using antibodies raised against either the

N-terminal (N-Ab; able to detect both Hth^{HM} and Hth^{FL}) or the C-terminal (C-Ab: able to only detect Hth^{FL}) end of Hth (see Figure 3.12 A on page 143) however, showed constant signal strength for both antibodies, independent of the degree of Exd^{MUT} signal loss (compare Figure 3.11 D on page 139). If the two isoforms indeed bind separate genomic locations, we would expect to see less binding for C-Ab (Hth^{FL}-specific) with increasing Exd^{MUT} signal loss. This however is not the case. To rule out that the C-Ab might perhaps not be FL-specific, we performed western blots (WB) for both N-Ab and C-Ab using recombinant Hth^{HM}-Exd, recombinant Hth^{FL}-Exd, and third instar wing disc extracts. While N-Ab identified both isoforms of Hth, the C-Ab only showed one band, suggesting that it is indeed specific for Hth^{FL} (Figure 3.12 B on page 143). The extracts showed the presence of two isoforms, confirming that both isoforms are present in imaginal wing discs.

The finding that Hth C-Ab signal strength seems constant across all Exd binding sites is perhaps not too surprising considering that both Hth^{FL} and Hth^{HM} are attached to Exd, and can both form complexes with Hox proteins. The two resulting Hth-Exd-Hox complexes cannot necessarily know where to bind in the genome unless they are guided by an additional factor. Since motif enrichment for the identified peaks did not reveal a potential additional mediator, it is conceivable that Exd-Hox can exert its function with Hth^{FL} attached to it, even in the absence of a direct Hth-DNA interaction. As a consequence, the use of ChIP-seq signals for wild-type Exd, Hox and C-Ab and N-Ab is not sufficient to distinguish between Hth^{HM}-Exd-Hox and Hth^{FL}-Exd-Hox sites. Only in the presence of the Exd^{MUT} can we start to tease apart the two scenarios, as is demonstrated by the two distinct subsets of peaks, boxed in red or green, in Figure 3.11 A on page 139. Both types of peaks have signal from all three homeodomains and might thus be labeled Hth-Exd-Hox sites. However, in the red class the Exd^{MUT} signal is severely lost, whereas in the green class it is retained, suggesting that the red classes are Hth^{HM} sites and that the other ones are stabilized by direct binding of Hth^{FL} to DNA.

As both isoforms are present in wings discs, we reasoned that DNA Hth-Exd binding sites that harbor a Hth site, but not an Exd-Hox site, should only be bound by the full-length Hth isoform (recognized by the C-Ab), and not by the HD-less isoform, as the latter cannot bind DNA by itself in the absence of Exd-Hox complex formation. As a consequence, the ratio *C-Ab/N-Ab*, which represents the fraction of $\mathrm{Hth}^{FL}/(Hth^{FL} + Hth^{\mathrm{HM}})$ should be different at sites, where $\mathrm{Hth}^{\mathrm{HM}}$ can contribute to binding (Exd-Hox sites) and sites where it cannot (Exd-Hth or Hth-only sites ; where the fraction simplifies to $\mathrm{Hth}^{\mathrm{FL}}/\mathrm{Hth}^{\mathrm{FL}}$) (Figure 3.12 A, D on page 143).

A first indication that this might indeed be true for *in vivo* binding was obtained when plotting the ChIP IP signal of N-Ab against C-Ab at the Exd peak summits and coloring them by either the maximum Exd-Antp binding score (using an NRLB binding model derived from Lib-1 and finding the maximum score within ± 50bp around the peak summit, see Experimental Procedures for details) or the Hth binding score (model from Lib-4 using a PSAM seeded on the sequence TTGACAGC). No obvious shift was observed for the Exd-Antp binding model, whereas the Hth scores tended to accumulate towards higher C-Ab signal (Figure 3.12 C on page 143).

To quantify this more thoroughly, we grouped the peaks into three categories: i) peaks with a high scoring Exd-Antp site (not more then 10-fold less than the top score across all peaks) and lacking a decent Hth site (at least 5-fold less than the top score); ii) peaks with a high Hth (PSAM>0.9) and no high-affinity Exd-Antp site (score more than 10-fold less than the top score); and iii) peaks that were ambiguous. As expected, monitoring the ratio *C-Ab/N-Ab* showed a significant difference between the Exd-Antp only class and the Hth only class (p-value $= 1.2 * 10^{-6}$, t.test; Figure 3.12 E on page 143).

This finding demonstrated that the two Hth isoforms can both bind at Exd-Hox sites, even in the absence of a direct Hth-DNA interaction. Presumably, a "floppy" $\mathrm{Hth}^{\mathrm{FL}}$-Exd-Hox complex is still functional, which rules out the possibility that the two isoforms target distinct genomic locations. Rather, as demonstrated by the ratio of *C-Ab/N-Ab*, the HD-less

Hth^{HM} isoform competes with Hth^{FL} at Hox-dependent sites and thus might redirect Hth^{FL} binding to Hox-independent sites. Given the constant levels of nuclear Exd-Hth (compare Figure 3.10 C on page 134), this also implies that the ratio of the two isoforms might be crucial to control the level of "active" Hth^{FL}. An increase in Hth^{HM} will favor Hox-binding as it both reduces the total level of nuclear Hth^{FL} and guarantees efficient recruitment of Exd to Exd-Hox-dependent sites. In summary, our findings imply a buffering role for the two isoforms, where the ratio of the two controls how much nuclear Exd is recruited to either Exd-Hth or Exd-Hox sites.

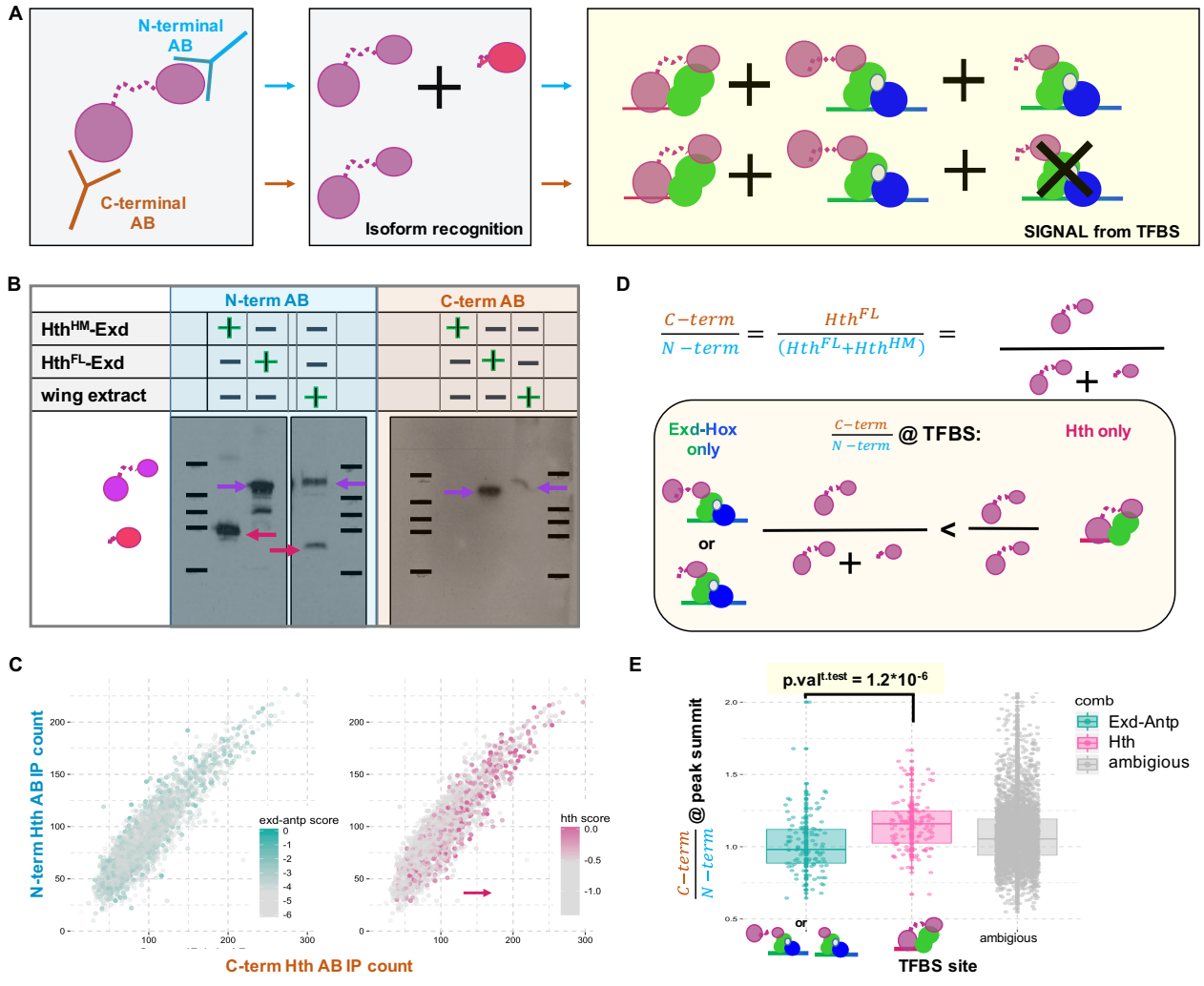


Figure 3.12: Hth isoforms bind to the same locations independent of Hth-HD-DNA interaction

Figure 3.12: *continued from page 143.*

(A) Schematic representation of the recognition mode for the two antibodies used in the IP. The C-terminal antibody (C-Ab) recognizes the homeodomain and thus only detects the Hth^{FL} isoform. The antibody raised against Hth's N-terminus (N-Ab) can recognize both isoforms (Hth^{HM} and Hth^{FL}) and therefore recognizes all complex compositions where Hth is involved. **(B)** Western Blot for both N-Ab and C-Ab. Lane 1 and lane 2 of each blot contains purified recombinant Hth^{HM} and Hth^{FL} isoforms respectively. Lane 3 (separate in the N-Ab blot due to higher exposure times needed) contains wing disc extracts. N-Ab recognizes both recombinant Hth isoforms – one higher (purple arrow) and one lower (pink arrow) molecular weight bands. And two bands are observed in the wing extracts. The C-Ab only recognizes the Hth^{FL} isoform – one high molecular weight band (purple arrow) and also only detects one protein in the wing extracts (at same molecular weight as seen for the N-Ab). **(C)** Raw Coverage plot for N-Ab (y axis) versus C-Ab (X-axis) including all called Exd^{WT} peaks. Peaks are either colored by their NRLB Exd-Antp score (green, left side) or PSAM-derived Hth-only score (pink, right side). Exd-Antp scores did not distinguish N-Ab and C-Ab signal, but Hth scores appeared scewed towards higher C-Ab signal. **(D)** Predictions for the behavior of signal ratio *C-Ab/N-Ab*: at Exd-Hox sites, the ratio should be smaller as the Hth^{HM} isoform contributes to binding and thus increases the denominator. At Exd-Hth or Hth-only sites, only Hth^{FL} should contribute to binding and not Hth^{HM}, therefore the ratio should be larger. The expected difference depends on the overall ratio of the two isoforms. **(E)** Peaks from C were split into three groups based on the NRLB or PSAM-derived binding site predictions: i) peaks with a high Exd-Antp score but no good Hth site (green), ii) peaks with a high Hth-only score but absence of a good Exd-Antp site (pink) and iii) peaks neither belonging to one of the other classes; thus ambiguous (grey). Exd-Antp only peaks (green) can have signal from both Hth^{HM} and Hth^{FL} and thus the ratio *C-Ab/N-Ab* should be lower compared to sites where Hth^{FL} is the sole contributor. Comparing the distributions of *C-Ab/N-Ab* for both peak classes (i) and ii)) show a significant lower ratio (p-value=$1.2 * 10^{-6}$, t.test) confirming that sites are indeed bound by Hth^{FL} to different extends due to the competiton with Hth^{HM} at Exd-Hox sites.

3.2.10 Disentangling *In Vivo* Exd Binding Contributions Using *In Vitro* Inferred Mechansims

Having identified a correlation between IP signal loss in the mutant and Antp-IP signal strength, yet no obvious pattern for the Hth C-Ab signal, we wondered whether we could use the binding rules identified by our *in vitro* studies to disentangle the respective contributions of Hox and Hth to *in vivo* Exd binding. We hypothesized that the observed Exd IP signal strength for each peak depends on the presence, the strength, and perhaps the configuration of the potential complexes Exd can participate in: i) Exd-Hth, ii) Exd-Hox and iii) Hth-Exd-Hox trimer, or iv) alternative Hth-Exd conformations, indicated by the presence of a Hth site. To analyze the peak locations in this "3D affinity space", we scored each peak (on both strands; ±50 bp from the summit) with an Exd-Antp model (using an NRLB model on data from Lib-1), an Exd-Hth model (NRLB model on data from Lib-2) and a Hth-only model (PSAM derived from Lib-4; see Experimental Procedures for details). We next computed the cumulative peak score for each model by summing up the scores for each binding site position (choosing the maximum of the forward ($5'\rightarrow3'$) and reverse ($5'\leftarrow3'$) score per position) (Figure 3.13 A on page 148). Since the presence or absence of Hox is a major determinant to separate Exd-Hox from Exd-Hth sites, we chose to stratify the Exd^{WT}/Input by the *Antp-GFP/Input* enrichment for each peak (Figure 3.13 B top panel on page 148). To analyze the relative contributions of each binding mode to Exd^{WT} binding, we colored each peak according to the cumulative score of either Exd-Antp (blue), Exd-Hth (green), or Hth-only (pink) and computed the correlation between Exd^{WT} or Antp-GFP binding strength and the respective model score (Figure 3.13 B top panel on page 148). As expected, the Exd-Hox model score was strongly correlated with Antp-GFP IP-signal strength ($\rho = 0.38$, p-value=$< 2.2 * 10^{-16}$), whereas Exd-Hth was not significantly correlated ($\rho = -0.02$, p-value=0.37). Perhaps surprisingly, the Hth-only score showed a negative correlation with Antp-GFP signal strength (weak correlation, ρ=-0.04; p-value=0.03). This latter result indicates that much of Hth binding occurs independent of Exd-Hox, manifesting itself as a

negative correlation. However, given the biased sample, with sites that are clearly bound by Exd, the overall slightly negative correlation does not rule out that there might be trimer sites that contribute to the Antp-IP signal. The three models are also not independent of each other, as a Exd-Hth site will also achieve a high Hth-only score and sites with lower relative affinity for Exd-Hth might obtain a similar score from the Exd-Antp model, resulting in imperfect classification. Lastly, Antp levels vary throughout the disc, with areas of complete absence, which compresses the Antp IP signal relative to Exd-Hth.

Looking at the contributions each model has for Exd^{WT} binding, we obtain significant positive correlations for each model (ρ=0.13, p-value=$5.2*10^{-14}$; ρ=0.07, p-value=$3.0*10^{-5}$; ρ=0.26, p.val< $2.2*10^{-16}$), with Hth contributing strongest to the observed binding, followed by Exd-Hox and lastly Exd-Hth. That the correlation with Hth is the strongest is not surprising, given that i) Exd by itself can not bind DNA, however is directly connected to Hth by a tight protein-protein interaction and ii) the patchy expression of Antp in wing discs, reducing the fraction of cells with Antp expression that contributes to the overall Exd-binding. The dominance of the Exd-Antp score driving the Exd^{WT} IP signal (compared to Exd-Hth) is consistent with our *in vitro* observations, where Exd-Hox was also the dominant configuration when all three homeodomains were present (compare Figure 3.2 B on page 107).

Having established the contribution of each possible complex to Exd^{WT} and Antp *in vivo* binding, we next wanted to verify that the "shape-mutant" – Exd^{MUT} is able to distinguish the different complex compositions *in vivo* in a similar fashion as *in vitro*. To do so, we correlated the peak score for each model with either the Exd^{MUT}/Input enrichment or the Exd^{WT}/Exd^{MUT} IP-ratio (Figure 3.13 B middle panel on page 148). As predicted, the correlation with Exd binding strength and Exd-Antp score is lost in the mutant, due to the complete loss of Hth^{HM}-Exd-Hox binding as observed *in vitro* (ρ=−0.04, p-value=0.01). There is no significant correlation (ρ=0.02, p-value=0.33) with the Exd-Hth score in the mutant, indicating that Exd-Hth binding is also affected to a certain degree by the mutation

and agreeing with *in vitro* findings (compare Figure 3.8 A & B on page 128). The only strong correlation with Exd^{MUT} binding is found for the Hth-only score (ρ=0.28, p-value$< 2.2 * 10^{-16}$), recapitulating what we found *in vitro*, i.e. the stabilizing influence of Hth-DNA binding on the Exd-Hox complex and the emergence of Hth-dimer sites as the optimal DNA sequence in the Exd^{MUT}-SELEX experiments (compare Figure 3.8 A, B on page 128).

When comparing Exd^{WT} and Exd^{MUT}, we predict that the signal for loss of binding of the mutant should be driven most strongly by the presence of an Exd-Hox site, that the correlation should be weaker for Exd-Hth sites, and that a significant negative correlation should be observed for a Hth-only site. Plotting the WT/MUT-IP ratio against the $Antp\text{-}GFP/INP$ enrichment, and again coloring the peaks by their respective model scores, confirms the prediction ($\rho_{\text{Exd-Antp}}$=0.37, p-value$< 2.2 * 10^{-16}$; $\rho_{\text{Exd-Hth}}$=0.11, p-value=$5.1 * 10^{-11}$; ρ_{Hth}=-0.04, p-value=0.03; Figure 3.13 B bottom panel on page 148). The lack of a strong negative correlation for the Hth-only score and loss of Exd^{MUT} binding can be explained by i) the presumably small fraction of sites that are actually bound by trimeric Hth^{FL}-Exd-Hox and ii) the confounding relationship between Exd-Hth and Hth model scores. Since Hth does not actively drive the signal loss in the mutant, the stabilizing effect of Hth^{FL} on Exd-Hox binding can only be established when separating Hox-dependent from Hox-independt binding, which will be demonstrated in the section below.

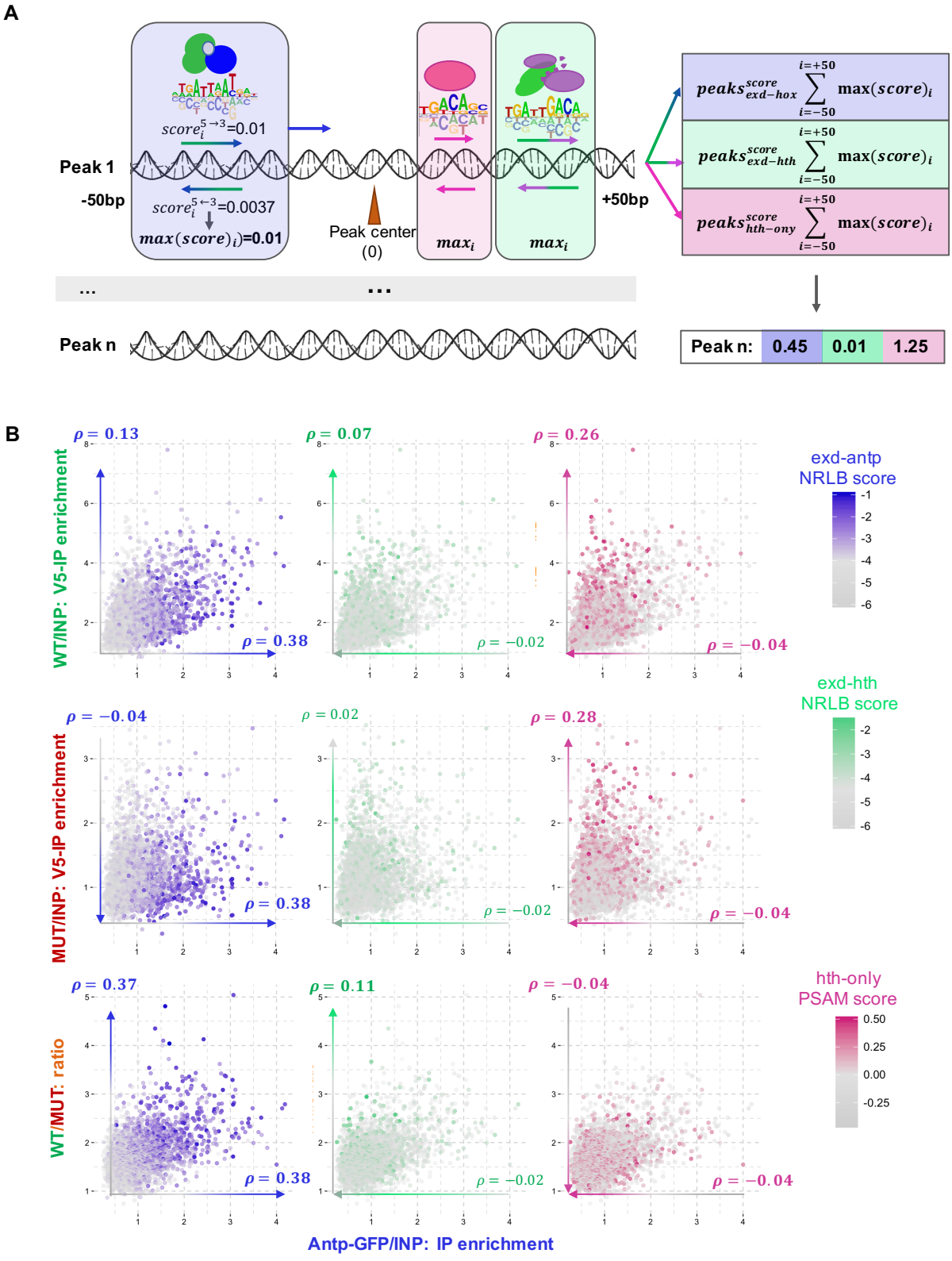


Figure 3.13: Relating the affinity signature for different complexes to ChIP-seq enrichment:

Figure 3.13: *continued from page 148.*

(A) Procedure to score each called Exd peak 50 bp up- and downstream of the peak summit. Each peak sequence is scored with the three *in vitro* derived affinity models for i) Exd-Antp (NRLB form Lib1, blue), ii) Exd-Hth (NRLB from Lib-2, green) and Hth-only (PSAM from Lib-4, pink). For each binding-site start position, two scores are computed per model, forward ($5'\rightarrow 3'$) and the reverse complement of the sequence ($5'\leftarrow 3'$) and only the maximum of the two is considered. The cumulative peak score for each model is computed by summing up the contributions across all possible binding sites. **(B)** IP-enrichment - model score correlation plots. Antp-IP enrichment is plotted on the x-axis for each panel to stratify peaks signals by the presence of Hox binding. Colors used in the plots represent the three models used in (A) (Antp-Exd = blue, Exd-Hth= pale green and Hth-only = pink). The Y-axis is different for the top, middle and bottom row and represents either Exd^{WT} (green), Exd^{MUT} (red), or the $Exd^{\text{WT}}/\text{Exd}^{\text{MUT}}$-IP ratio (orange). Correlations between motif score and IP-enrichments are shown by arrows along the axis of the respective IP-enrichment tested. Bold font indicates significant correlations. Antp is strongly correlated with a high Exd-Antp motif score, but not with the other two models. Exd^{WT} is positively correlated with all three models, since Exd is always part of the complex considered. Exd^{MUT} looses Exd-Antp correlation, Exd-Hth is dampened and Hth-only scores remain the sole correlator for Exd^{MUT} signal (stronger as seen in the wild-type). The change between Exd^{WT} and Exd^{MUT} correlation is reflected by the mutant signall loss (as measured by the $\text{Exd}^{\text{WT}}/\text{Exd}^{\text{MUT}}$-IP ratio), with a the strongest correlation with Exd-Antp peak scores, followed by Exd-Hth. Hth-only scores are slightly negatively correlated with signal loss.

3.2.11 Exd^{MUT} Serves as a Sensor for Hth/Exd/Hox Composition and Conformation *In Vivo*

Since IP signals from antibodies targeting either the N- or C-terminus of Hth overlap almost perfectly, they are not useful for distinguishing between true trimeric Hth-Exd-Hox sites, with a direct Hth-HD-DNA interaction, and sites with a "floppy" Hth, and can therefore also not unambiguously establish the presence of cooperative binding between Hth and Exd-Hox. Using Exd^{MUT}, however, should allow us to establish such a relation after all, based on the *in vitro* observation that DNA-binding by Hth^{FL} can prevent the severe binding loss observed for Hth^{HM}-Exd^{MUT}-Hox (Figure 3.6 A & B on page 119). As mentioned above, correlating the Hth-only binding score for each peak with the overall signal loss for Exd^{MUT} only established a slightly, yet significant, negative correlation, which can be attributed to the presence of different complex compositions that are all impacted differently by the mutation and thus confound the model-to-signal relationship.

To demonstrate that Hth^{FL} indeed stabilizes Exd^{MUT}-Antp binding *in vivo*, we focused on the sites that had a clear signature of Exd-Hox binding, namely a site matching the Exd-Hox consensus **TGAYNNAY**. In total, 720 of the 3546 Exd peaks ($\sim 20\%$) and 142 of the 279 Antp peaks ($\sim 51\%$, called at q-value= 0.05) contained a TGAYNNAY site. Focusing on the subset of Exd-Antp peaks with a reasonable Exd-Hox site, and ranking them based on their WT/MUT-IP ratio (equivalent to how much binding is lost in the Exd^{MUT}) indeed revealed the stabilizing effect of Hth^{FL} DNA-binding. Not only did the analysis show an increase in correlation between Exd-Antp binding strength and signal loss when focusing on the "high-confidence" Exd-Hox peaks (from ρ=0.37 for all peaks to ρ=0.39 for Exd-Hox peaks only), but it also showed a strong and significant negative correlation between signal loss and presence of a strong Hth binding site (ρ=-0.19, p-value=$3.2 * 10^{-7}$) (Figure 3.14 A on page 152). As a control, there was no correlation between mutant signal loss and predicted Exd-Hth binding strength in the Exd-Hox subset ($\rho = 0.05, p.val = 0.13$).

This result demonstrates the presence of several trimeric Hth-Exd-Hox binding sites, as

well as the selective loss of Exd^{MUT} binding, dependent both on the Hth-isoform and whether a suitable binding site for Hth^{FL} is present (see binding model in Figure 3.14 A on page 152).

Our *in vitro* studies had identified two categories of Exd-Hox binding sites affected differently by the Exd shape-readout mutations and that were suggestive of two distinct DNA recognition modes. We therefore expected that the Exd^{MUT} protein should also reveal the presence of these two modes *in vivo*. To test this hypothesis, we again focused on the subsets of peaks harboring a **TGAYNNAY** site and sorted the peaks by the Antp-GFP, the Exd^{WT}, the Exd^{MUT} or the Exd^{WT}/Exd^{MUT} binding strength. We next colored the peaks by the affinity class they belong to; orange for the "low-affinity" TGA**C**NNAY and green for the "high-affinity" TGA**T**NNAY type sequences. As expected, the high-affinity TGAT-type sites on average had a significantly higher IP-signal for both Antp and Exd^{WT} binding compared to their low-affinity TGAC-type counterparts (p-value$< 2.2 * 10^{-16}$ for Antp, p-value=0.01 for Exd^{WT}, t-test) (Figure 3.14 B on page 152). Exd^{MUT} however showed a reversal of site preference, with "low-affinity" sites now stronger bound compared to high-affinity ones (p-value=0.001, t-test). The switch in site preference was strongest when considering the signal loss of Exd^{MUT} compared to Exd^{WT}, with high-affinity sites being lost to a much greater extent than low-affinity ones (p-value$< 2.2 * 10^{-16}$, t-test) (Figure 3.14 B on page 152).

In summary, by using the Exd^{MUT} as a tool, we could establish: i) the presence and use of trimeric Hth^{FL}-Exd-Hox binding sites, and ii) the use of different Exd-Hox complex recognition modes when binding to low- and high-affinity binding sites. This result shows that mutations of amino acids within the N-terminal arm of homeodomain proteins can cause a switch in binding-site selectivity which in the case of Exd is revealed upon dimerization with Hox proteins.

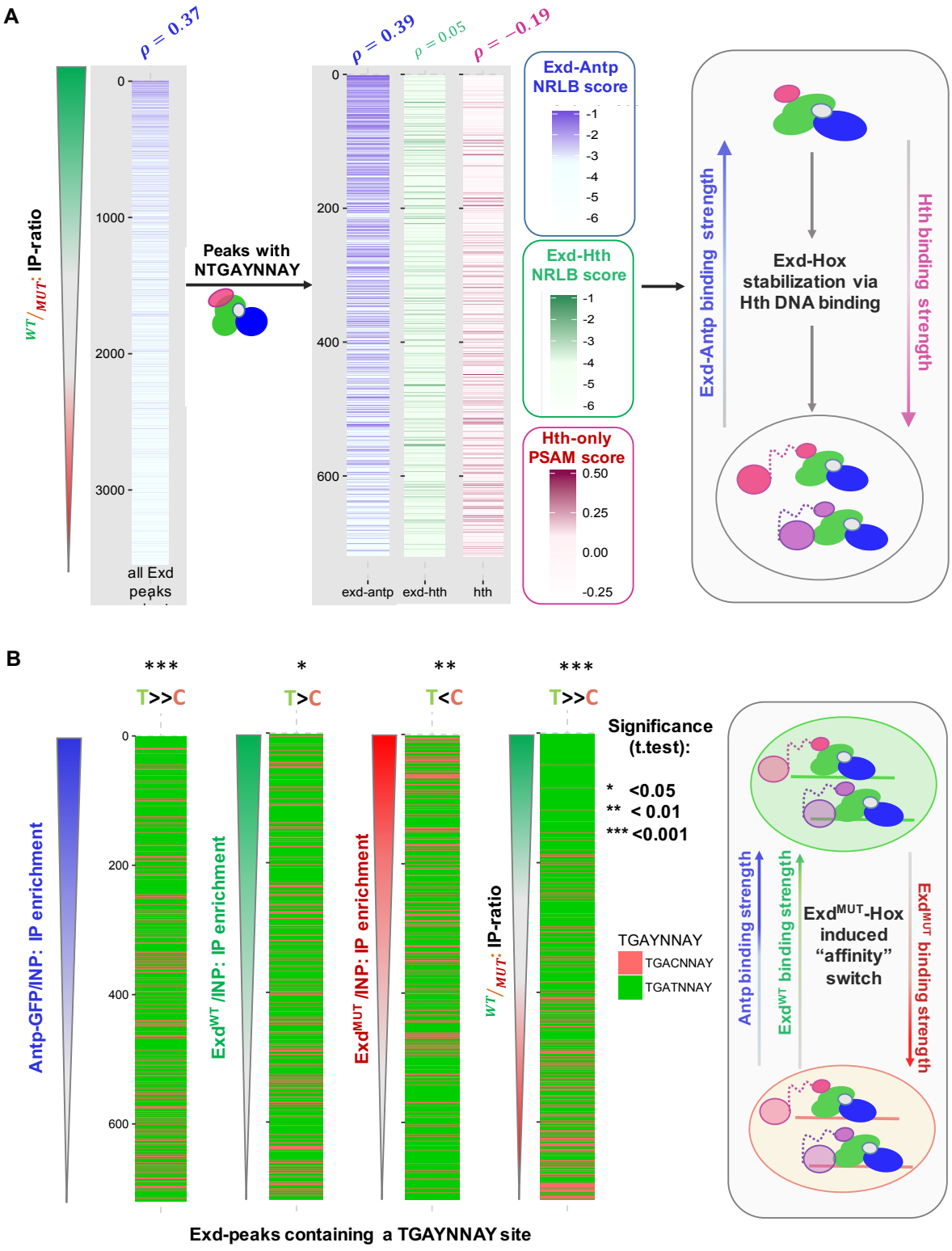


Figure 3.14: Exd^{MUT} distinguishes different complex compositions and affinity classes:

Figure 3.14: *continued from page 152.*

(A) Exd peaks with a clear Exd-Antp binding site (match for NTGAYNNAY) are isolated from all peaks and ranked by the degree of Exd^{MUT} signal loss (Exd^{WT}/Exd^{MUT}-IP-ratio). The peak affinity score for each of the three models – Exd-Antp (NRLB, blue), Exd-Hth (NRLB, pale green) and Hth-only (PSAM, pink) is indicated by the color gradient. Correlation between peak affinity score and Exd^{MUT} signal loss are shown above each strip-plot and bold face indicates a significant correlation. Limiting the peak set to peaks with clear Exd-Antp signature, reveals a negative correlation between Exd^{MUT} signal loss and Hth binding site strength, demonstrating that Hth^{FL} binding stabilizes the Exd^{MUT}-Hox complex not only *in vitro*, but also *in vivo*. Scheme on the right summarizes the effect of Hth^{FL} binding on Exd^{MUT}-Antp complex stability. Peaks at the bottom of each strip chart have a stronger Hth affinity signature compared to those at the top. **(B)** Comparing Exd-Antp sequence preferences in peaks with a clear Exd-Hox sequence signature (presence of NTGAY_5NNAY within $\pm 50bp$ from peak summit). Sites with Y_5=T (green) are more strongly enriched for both Antp-IP enrichment (blue gradient, peaks ordered by $Antp/Input$) and Exd^{WT}-IP enrichment (green gradient, peaks ordered by Exd^{WT}/Input), but preference is changed in the Exd^{MUT}. C_5-type sequences have a higher Exd^{MUT}-IP enrichment (red gradient, peaks ordered by Exd^{MUT}/Input) compared to T_5-type sequences. Ordering peaks by their mutant signal loss reveals that T_5-type sequences drive the signal loss strongest (p-value=$< 2.2 * 10^{-16}$, t-test). Panel on the right summarizes observed "affinity-selectivity" switch observed in the Exd^{MUT}: Antp and Exd^{WT} *in vivo* binding strength (from ChIP-seq signal) is higher for sequences of type NTGA**T**NNAY whereas Exd^{MUT} *in vivo* binding strength is higher for NTGA**C**NNAY sequences. T_5 identity strongly drives signal loss in the Exd^{MUT} suggestive that two different binding modes of Exd-Hox complexes are also present *in vivo*.

3.3 Discussion

In this study, we have presented an in-depth analysis of the adaptive DNA binding behavior of a multi-protein TF complex formed between three different homeodomain TFs – Hth, Exd and Hox (Dfd or Antp). The presence of more than two TFs increases the binding complexity, as different complex compositions can coexist, with a range of different orientations and spacings between the three binding partners occurring, all within a reasonable *in vivo* affinity range. The resulting combinatorial logic, which goes beyond simple dimer-formation, is plausibly a commonly utilized theme in TF-DNA recognition, especially in species where TF gene duplication events occurred and TF amino acid sequence length, not including the DBD, has increased (Charoensawan et al., 2010). The example focused on in this study – the cooperative binding between Hth, Exd and Hox proteins – is of particular importance, as these proteins and their biological functions are highly conserved across animal taxa, and an expanded set of orthologs for all three TF classes are present in mammals. To this date, only little is known about the distinct gene regulatory networks controlled by specific combinations of these three TFs. However, misregulation of any component is associated with severe developmental defects or cancer (Lewis, 1978; Crist et al., 2011). Several splice forms of Hth and its paralogs have been identified, falling into two major classes: those containing the homeodoamin DBD and those who do not (Noro et al., 2006). The impact of alternative splicing of Hth (MEIS) on complex formation with its coregulators Exd (Pbx) and Hox, and downstream gene regulation has remained largely elusive.

The DNA sequence requirements for complexes of specific subsets of three homeodomain factors to form are increasing with every partial binding site and configurational constraint added, and therefore provide a mechanism by which binding sites can be tuned and differentiated *in vivo*. Furthermore, cooperative binding between TFs can greatly facilitate binding

to suboptimal sites, in a scenario where the binding site for one TF allows recruitment and tethering of cofactors and thereby an increase in their local concentration, allowing low affinity sites to be bound to a significant degree (Crocker et al., 2015; Reiter et al., 2017; Tsai et al., 2017).

To systematically dissect these phenomena, it is important to first characterize the DNA sequence requirements for each possible subcomplex that can form in the presence of (in this case) three TFs. Besides quantifying sequence specificity, it is also important to establish the relative contribution of each complex to DNA ligand selection. This in turn should increase our ability to make predictions *in vivo*, where all three proteins are present and different conformations compete with each other for binding.

Using different SELEX-seq strategies, we established that Exd can form heterodimeric complexes with both Hth and Hox, in a manner where the C-terminus of Exd faces the N-terminus of Hth or Hox. However, Hth is readily displaced in the presence of Hox proteins and then preferentially binds upstream of the Exd site using a range of possible DNA spacer lengths and either orientation of Hth relative to Exd. When the N-terminal end of Hth faces Exd, longer spacers (3-9bp) are preferred, whereas small spacers (0-4bp) are optimal when Hth's C-terminus faces Exd. This behavior is in line with the presumptive three-dimensional structure of such a complex, where Hth's N-terminal HM domain is attached to Exd's N-terminal PBC domain and thus naturally allows for more flexibility when both N-termini are facing each other. Using SELEX-seq data to identify spacing and orientation preferences can aid in identifying the overall structural arrangement and positioning of these protein domains, which may be too flexible to be captured by crystallography. It also provides a means to assign TF orientation within a given multi-domain architecture.

In addition to the overall configuration, we also identified that sequences previously not believed to be directly contacted, in this case the spacer between Exd and Hth, are contributing to the overall sequence selectivity to a high degree. We were able to attribute the preference for specific spacer sequences to their intrinsic DNA shape, which in turn

determines the electrostatic potential sensed along the minor groove by positively charged arginines present in both the N-terminal arm of Exd and Hth. Such MGW interactions have previously been observed in crystal structures for Hox in complex with Exd, and also occurr for other factors (Joshi et al., 2007; Rohs et al., 2009b). Moreover, the spacer sequence preference seen in Exd-Hox was recently linked to the recognition of intrinsic variation in naked-DNA minor groove width (Abe et al., 2015). As shown in this study, this appears to be a mechanism used by homeodomains more generally and even applies to larger flanking sequences, such as the DNA spacer between Hth and Exd. Correlating the observed sequence selectivity in a SELEX experiment with the predicted intrinsic DNA shape can therefore help identifying specific TF readout mechanisms. However, caution is warranted when using sequence to shape mapping in such a way, as a correlation does not neccessarily imply an underlying shape-mediated readout. DNA shape naturally depends on sequence and can therefore result in correlation that does not reflect a true binding mechanism, but rather emerges from a direct amino acid base interaction, such as a crucial hydrogen bond. Combining structural data with the observed affinity-shape correlation, i.e. analyzing an existing crystal structure for the presence of arginines near a MGW selection minimum, can serve as an initial screen for "true" shape readout (if no structure is available for the specific TF, sequence alignments with other related TFs for which a structure exists can be used). However, as shown in this study, structural perturbations provide perhaps the best platform to overcome this ambiguity.

Although it has long been known that shape recognition is a crucial part in TF-DNA recognition (Luscombe, 2001), sequence selectivity is thought to predominantly result from specific hydrogen bonding between amino acids and DNA bases (mostly in the major groove). Crystal structures have contributed to this belief in the sense that they only capture rigid complex conformations and cannot resolve more flexible protein domains, which might adopt a few distinct, yet energetically similar sub-conformations. Structure-guided TF engineering therefore often focuses on amino acids directly interacting with bases in the major groove,

and the resulting mutations often lead to a global loss of binding. One example is the known homeodomain asparagine 51 mutation that disrupts a crucial hydrogen bond with a major groove adenine and was therefore used in this study to establish Exd-Hox binding loss.

Here, we not only demonstrate that amino acids disordered in the majority of existing crystal structures can be crucial for binding, but we also present a new strategy for designed TF engineering, moving the focus from direct "base recognition" to "shape recognition". Mutating amino acids at the N-terminal end of the homeodomain that are involved in MGW shape readout, we have built a powerful tool that goes beyond simply abolishing binding. Our shape-readout-perturbing mutation to Exd selectively destabilizes different complex compositions of Hth, Exd and Hox. The severity of DNA binding loss not only depends on complex compositions, but in the case of Exd-Hox also on the underlying sequence context. The shape mutant therefore reveals at least two recognition modes utilized by Exd-Hox to bind high and low affinity sites respectively. The latter finding is surprising, as the differential sequence selectivity occurs 4bp away from the base dominantly impacted by the Exd mutation. It demonstrates in a powerful way that the overall conformation of TF-DNA complexes is not as cast in stone as a crystal structure might suggest, but rather that several conformational modes are utilized by TFs to adapt their binding to different sequence contexts. Moreover, it shows that mutations that affect shape recognition within the DNA flank can have a much bigger impact on overall TF binding and sequence selectivity than previously thought.

Creating such designed TFs is useful, as they provide a means to selectively perturb distinct aspects of TF-binding *in vivo*, while circumventing the complete loss of binding often seen in hydrogen-bond-breaking mutations. Here, we showcased the differential impact such a shape mutation has on distinct Exd-containing TF complexes, both *in vitro* and *in vivo*. For instance, the mutant was used to detect the presence of trimeric complexes *in vivo*: Exd^{MUT}ChIP-seq-signal loss is counteracted by the presence of a strong Hth DNA binding site, suggesting that trimeric Hth-Exd-Hox binding is functional and actively used

for a subset of Exd-Hox sites *in vivo*. The presence of Hth^{FL} DNA binding together with Exd-Hox could not be verified by simply using an antibody that detects Hth^{FL} but not the HD-less Hth^{HM} isoform: Both isoforms co-localize to Exd-Hox sites even in the absence of a Hth binding site. This finding indicates that Hth^{FL} can remain passively attached to Exd when the latter cooperatively binds to DNA together with Hox proteins.

This brings us to another important aspect of regulation by homeodomain proteins: the outstanding question about Hth isoform function. Using Exd^{MUT} as a tool, we obtained several new insights into the regulatory logic behind the Hth-Exd interaction: Using V5-tagged Exd, and generating FLP-mediated clones null for endogenous Exd, we found that nuclear levels of Exd, and presumably also Hth due to the tightness of the interaction, are constant, and independent of the genomic location or level of Exd produced. Moreover, the ratio of endogenous to tagged Exd varied depending on the ratio of their expression levels as was demonstrated when we compared the three different genotypes present in clone-induced wing discs. Together with the finding that all Exd-Hox sites have a binding signal for the Hth^{FL} isoform, yet vary in the degree to which Exd^{MUT}-ChIP-seq-signal is lost (depending on the presence of a Hth DNA binding site), we can distill a few important rules about Hth isoform usage:

1. Given the constant nuclear level of Exd and the mutual dependency of Hth and Exd for nuclear localization, we can assume that nuclear levels of Hth are constant as well. Therefore, the amount of Hth^{FL} presumably depends on the ratio of expression levels of the *full-length* and *HD-less* isoforms. This is similar to what we observe for Exd when Exd protein is provided from two different sources.

2. There are three major Exd binding classes, one Hth^{FL}-dependent and one Hox-dependent, the latter one defined by the Hth^{HM} isoform being sufficient for binding. A third class requires DNA binding by all three homeodomains as part of the ternary Hth^{FL}-Exd-Hox complex.

3. Hth^{FL} and Hth^{HM} do not bind to distinct genomic locations. Moreover, Hth^{HM}, while sufficient, is not required for Exd-Hox binding in the absence of a Hth^{FL} binding site.

Together these findings suggest that the role of the HM-isoform might not be to recognize different sets of target genes, but rather to serve as a buffer of nuclear Hth^{FL} levels. Since Hth^{HM} cannot bind to Hox-independent Hth^{FL}-Exd sites, its amount naturally dictates how much of the fixed nuclear Exd is available for either Exd-Hth or Exd-Hox binding. As Exd levels remain constant, a large amount of Hth^{HM} isoform will favor Exd-Hox sites (see bottom panel of Figure 3.15 on page 159), whereas low levels will free up Exd to locate to Hth-Exd sites (see top panel of Figure 3.15 on page 159). By changing the expression level of the short HM-isoform, cells can therefore achieve a fast switch between a "Hox-favored" or a "full-capacity-Hth" state.

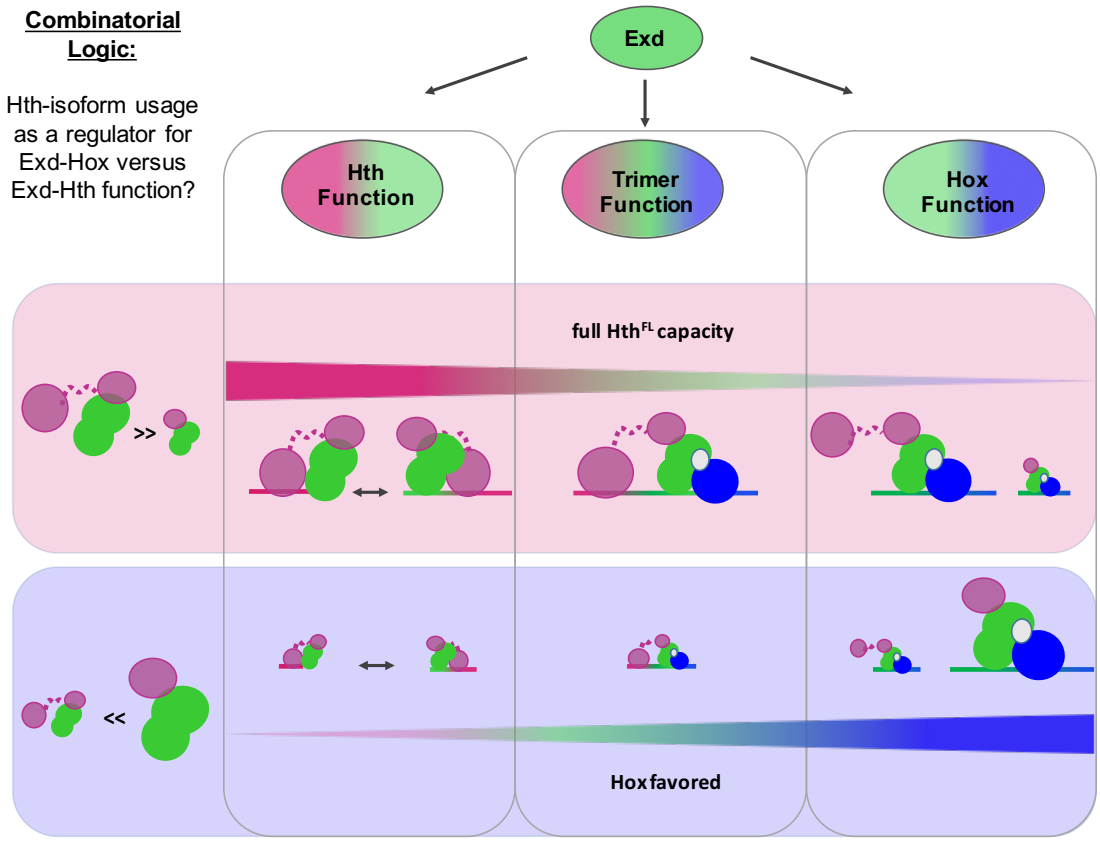


Figure 3.15: Proposed mechanism for Hth isoform usage and function

Lastly, we demonstrated that the overall ChIP-seq signal seen for a TF is seldom the result of a single binding conformation and thus a single binding site preference, but rather the combination of several different complex compositions and conformations all contributing to varying degrees to the overall IP-signal. To give an example, scoring all Exd^{WT} peaks with in-vitro derived binding models for Hth-only, Exd-Hth and Exd-Antp binding revealed that all three models were predictive for IP-enrichment to a varying degree. Using the Exd "shape mutant" we were able to validate the *in vivo* binding contributions from different complexes, by observing that individual model-to-signal correlations were selectively removed only for those complex compositions that were demonstrated to be lost in the mutant *in vitro*.

In summary, we developed a general in-depth analysis strategy for high-throughput, *in vitro* binding data of higher-order TF complexes that can aid in the targeted design of TFs with altered shape recognition properties. Our "shape mutants" revealed the use of different TF-DNA binding conformations by differentially destabilizing distinct states, thus providing an avenue to obtain insights into alternative DNA recognition modes, even in the absence of crystal structures for every possible DNA context. Importantly, the shape mutant can be used to test whether the same binding mechanisms that rule *in vitro* binding also govern *in vivo*, and to interrogate complex composition at different genomic loci. Only recently have we started to more carefully characterize the complex binding behavior of TF pairs (Slattery et al., 2011; Jolma et al., 2015). Yet, we can already begin to fathom how many other factors might use a similar set of mechanisms to regulate their target genes similar to the one identified for Hth-Exd-Hox. As demonstrated in this study, the failure of a monomeric TF binding model to explain the observed ChIP-seq signal might perhaps be more reflective of our limited understanding of the different TF configurations and complexes contributing to binding, than mysterious, higher-order chromatin effects. Understanding such complex regulation will require more detailed, in-depth analysis of *in vitro* and *in vivo* binding data and can be greatly facilitated by the use of accurate, *in vitro* binding models (Rastogi et al.,

2018) and mutant proteins designed to abolish specific aspects of TF-DNA recognition *in vivo*. Although every TF-DNA complex has a distinct interaction surface, structural readout mechanisms such as MGW recognition are often reused in TFs with similar domain structures (as shown for the MGW readout of Hth, Exd and Hox & (Rohs et al., 2009b)) and thus might provide a way to perform future studies in a higher-throughput format in which many different TF complexes are simultaneously perturbed. This study is therefore intended as a manual for the many different aspects of adaptive TF-DNA binding that can be exploited to investigate *in vivo* binding and downstream function of TF complexes or even alternative TF binding modes.

3.4 Experimental Procedures

3.4.1 Protein Purification and Mutagenesis

Fly proteins were obtained and purified as described in (Slattery et al., 2011). Briefly, PET-expression vectors containing coding regions for full-length Hth (Uniprot-ID O46339), Exd (Uniprot-ID P40427), Dfd (Uniprot-ID P07548) and Hth HM-domain (amino acids 1-242; (Uniprot-ID O46339) with hexa-histidine tags (except for Exd, which was always co-purified with full-length or HM-domain Hth) were transformed into Bl21 cells. Cells were grown for 5-7 hours, lysed and proteins extracted with affinity purification using Cobalt-Talon beads (Clontech). Site-directed mutagenesis for Exd and Hth was performed via amplification of the original plasmid with primers harboring single amino acid replacements (arginine to alanine) using Taq-polymerase (NEB). Double mutations were generated consecutively. Table 3.1 contains a summary of the mutations made and Figure 3.16 on page 162 an overview of the exact location and context within the proteins HDs.

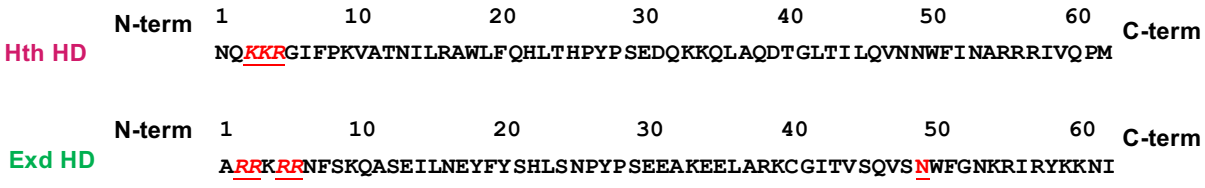


Figure 3.16: Homedomain sequences for Hth and Exd: Mutations are indicated in red

Protein	Mut1	Mut2	Mut3	Mut4	Mut5	Mut6	Mut7	Mut8
Exd	R2A	R3A	R5A	R6A	N51A	R2A&R5A	R3A&R5A	R5A&R6A
Hth	K3A&K4A &R5A	.	.	.	.	.	.	.

Table 3.1: Mutagenesis – Hth and Exd

3.4.2 Binding and Competition Assays

Electro-Mobility shift assays (EMSAs) were performed using 2nM radiolabeled DNA and protein concentration between $10 - 300nM$ in 1x Tris-running buffer and 5% TBE gels. Proteins were incubated for at least 30 min prior to loading in binding buffer (final concentration: 2% Glycerol, $30\frac{\mu g}{\mu l}$ polydIdC, 40 mM NaCl, 40 nM Tris pH=8.0, 0.4 nM MgCl2, 1mM DTT, 0.5 nM EDTA). For competition assays, a labeled probe was competed with increasing concentrations of unlabeled competitor DNA while keeping protein concentrations constant. Dose-response curves were fitted using the R package *drc* and IC50 values were obtained from the fits. Spacers with zero, three and seven bases between the Hth and Exd-Dfd sites were tested as well as spacer of length three with the top and worst sequences as obtained from SELEX-enrichment. Sequences used are summarized in Table 3.2.

probe		**sequence**	
Exd-Dfd top	5'	ATCTGG CTGTCA AAA ATGATTAATGAT CCCGGG	3'
Exd-Dfd worst	5'	ATCTGG CTGTCA CCC ATGATTAATGAT CCCGGG	3'
spacer 0	5'	ATCTGG CTGTCA ATGATTAATGAT CCCGGG	3'
spacer 3	5'	ATCTGG CTGTCA AAA ATGATTAATGAT CCCGGG	3'
spacer 7	5'	ATCTGG CTGTCA AAAAAAA ATGATTAATGAT CCCGGG	3'

Table 3.2: DNA probes used in EMSA and competition assays

3.4.3 SELEX Library Design

Library-1 contained a 16-mer random flank with no fixed sites and data for Exd-Dfd were taken from (Slattery et al., 2011). Library-2 followed the library design described in Chapter 2 of this thesis (unmethylated Library) and did not have fixed binding sites. Library-3a and-3b contained a fixed Hth site (CTGTCA or TGACAG) followed by a 21bp random region. Library-4 had a 30 bp random region and again no fixed site. Full library sequences, as seen by the protein, are listed in 3.3.

probe		sequence	
Library-1	5'	GTTCAGAGTTCTACAGTCCGACGATCTGG (16xN) CCAGCTGTCGTATGCCGTCTTCTGCTTG	3'
Library-2	5'	GGTAGTGGAGG TGGG CCTGG (16xN) CCAGG GAGGTGGAGTAGG	3'
Library-3a	5'	GTTCAGAGTTCTACAGTCCGACGATC CTGTCA (21xN) CCCGGGTCGTATGCCGTCTTCTGCTTG	3'
Library-3b	5'	GTTCAGAGTTCTACAGTCCGACGATC TGACAG (21xN) CCCGGGTCGTATGCCGTCTTCTGCTTG	3'
Library-4	5'	GTTCAGAGTTCTACAGTCCGACGATCTGG (30xN) CCCGGGTCGTATGCCGTCTTCTGCTTG	3'

Table 3.3: Library sequences

3.4.4 SELEX-Experiments

For Lib-3 and Lib-4 using wild-type or mutant homeodomain proteins, SELEX experiments were carried out following the experimental procedures described in (Slattery et al., 2011; Riley et al., 2014) and two rounds of enrichment were performed for each set of experiments. For Lib-2, a single round of selection was performed using the methodology and library design described in Chapter 2 of this thesis. Lib-1 data was obtained from a previous study (Slattery et al., 2011). In brief, for each experiment, proteins to a final concentration of ~ 50 nM were assembled and incubated with excess DNA ($\sim 10-20$ fold) for 30 minutes. After each round of selection the DNA was extracted from the gel at the respective shift heights for the complex assayed and amplified by either using Ilumina's small RNA primer sets or the set of primers described in the Experimental Procedures section of chapter two. Sequencing barcodes were added in a five cycle PCR step and the final library was gel-purified for quality control, separating the fully indexed sequences from the unindexed ones.

3.4.5 SELEX-library Sequencing and Data Processing

Libraries for Lib-2 were sequenced using a v2 75 cycle high-output kit on an Illumina NEXTSeq Series desktop sequencer at the Genome Center at Columbia University. Libraries Lib-3a and Lib-3b with either Hth or Exd mutant in complex with the respective other wild-type protein and Dfd, as well as the Lib-4 Hth-Exd-Dfd experiment were all sequenced at the New York Genome Center using separate lanes on a Illumina HiSeq 2000 sequencing System. Lib-IIIa and Lib-3b with wild-type proteins were also sequenced on a HiSEQ instrument at a different facility. Libraries were trimmed, removing Illumina and library internal adapter sequences and loaded into the R environment using the R package bioconductor.org/packages/SELEX (Riley et al., 2014).

3.4.6 Data Analysis of Complex Composition and Orientation

Simple relative enrichment tables for all libraries were generated using the R package bioconductor.org/packages/SELEX (Riley et al., 2014). To color the individual kmers based on the complex composition most likely explaining their enrichment, position-specific-affinity matrices were generated for Hth^{HM}-Exd-Dfd (using the most enriched 12-mer from Lib-1 matching the consensus NTGAYNNAYNNN), for Hth^{FL}-Exd from Lib-2 and using TGATTGACAG as the seed) and for dimeric Hth^{FL} (from Lib-4 using TGACAGCTGTCA as a seed). Each sequence was scored with each PSAM and complex composition assigned based on the highest achieved PSAM score. To remove shifted binding sites that do not encompass the full TF footprint, only sequences with a score > 0.01 for one of the three PSAMs were retained.

To test for preferences in Exd-Hox complex orientation with respect to the fixed Hth site in Lib-3a and Lib-3b, overall 12-mer relative enrichment tables were generated as described above and forward or reverse-complement orientation was assigned by comparing the relative enrichment of each 12-mer to that of its reverse-complement. Sequences with a higher

score for the forward strand (as obtained from the sequencing run) were assigned a "F" and sequences with a higher score for their reverse complement a "R". Average ratios of $\frac{F}{R}$ for Lib-1 (Hox used = Dfd), Lib-3a and Lib-3b were shown as boxplots. To account for different offsets of the Exd-Hox complex, 12-mer enrichment tables were generated for each offset respectively (using the SELEX function selex.affinities(kmer=12 , offset=x) ; with x=0 to 9) and F and R assigned as previously.

3.4.7 Feature-Based Modeling Using GLM

To model the orientation and offset preferences for the Exd-Hox subcomplex quantitatively in a unified model, each 21-mer probe (including ±2 bp into the library flank) was first scored on both strands with a PSAM obtained from the Hth^{HM}-Exd-Dfd data set from Lib-1. Only probes where a unique binding site could account for the probe selection with $> 95\%$ confidence were retained. Probes with identical 12-mer Exd-Hox sequences, spacer length and strandedness were collapsed to one entry in the design matrix. The collapsed R2 counts were used as the dependent variable in the generalized linear model, log-transformed respective R1 counts were used as an offset and both log-transformed Lib-1 derived relative enrichments for the Exd-Hox subcomplex and the overall configuration (as defined by the combination of spacer length and the DNA strand the binding site was located at) were used as predictors/features in the model. We used the glm() function in R with family="poisson" with the following model, where S_i represents the sequence of the Exd-Hox 12-mer with a specific configuration:

$$\frac{\Delta\Delta\text{G}(\text{S}_\text{i})}{\text{RT}} = \phi_i\text{config}(\text{S}_\text{i}) + \phi_\text{i}\text{affinity}(\text{S}_\text{i}) \tag{3.2}$$

Kmer based models for sequence preferences within the spacer were obtained using the same modeling framework. The full set of confidence-filtered probes was first subsetted by offset and orientation. Choosing a specific offset L (e.g. spacer of length L=4) and Hth-Exd-Hox orientation, sequences identical over L+12 bases where first collapsed and the total R2 occurrence was used as the response variable in the model. The log-transformed Markov model predictions for the R0 initial bias of each $(L+12)$-mer was used as an offset and the spacer sequence and the relative enrichment value for each 12-mer, were used as predictors, resulting in $4^L + 1$ model predictors.

$$\frac{\Delta\Delta\mathrm{G}(\mathrm{S_i})}{\mathrm{RT}} = \phi_i \mathrm{spacer}(\mathrm{S_i}) + \phi_\mathrm{i}\mathrm{affinity}(\mathrm{S_i}) \tag{3.3}$$

For the mononucleotide model, the kmers were represented by 4*(L+12) base identity indicators, reducing the parameter space. Model comparison was done by computing the R^2 (based on a linear model) between the spacer coefficients from the kmer model and the sum of the base coefficient making up the respective spacer sequence in the mononucleotide model.

$$\frac{\Delta\Delta\mathrm{G}(\mathrm{S_i})}{\mathrm{RT}} = \sum_{j=1}^{L+12}\sum_{b=A}^{T} \phi_{j,b}\mathrm{Base}_{\mathrm{j,b}} \tag{3.4}$$

Models with fixed $N_1|N_2$ base identity were obtained by further subsetting the probes, such that the Exd-Hox binding site would start with either AT, GT, TT or CT (CT was sub-

sequently excluded due to insufficient instances for most of the base predictors within the spacer). Mononucleotide models were fitted for each subset as described above, while excluding the first two base positions within the Exd-Hox site from the feature set.

3.4.8 Affinity-to-Shape Correlation

To identify, whether shape might be responsible for the observed spacer selection, we first computed the theoretical model score ($\Delta\Delta$G/RT) for each possible spacer, by summing up the respective base coefficients:

$$\text{SpacerScore} = \frac{\Delta\Delta\text{G}}{\text{RT}} = \sum_{\text{j=1}}^{\text{L}} \sum_{\text{b}} \phi_{\text{j,b}} \text{X}_{\text{j,b}} \tag{3.5}$$

With a score for each spacer in hand, we next used the pentamer shape table (Zhou et al., 2013) to compute the predicted minor groove width for each spacer. Since the score for each base in the pentamer table is dependent on the ± 2 bases, we extended the spacer $5'$ with the Hth fixed binding site, present in the library, and $3'$ by the base identity of the fixed $N_1|N_2$ used in the model. The resulting MGW profiles for each spacer were ranked by their computed $\Delta\Delta G/RT$ and average MGW profiles were obtained by taking the position-wise average. To test for a role of MGW in selection, we first computed the average MGW profile including all spacers, setting a reference point of random selection. We then subsequently increased the threshold for spacers included in the analysis based on their $\Delta\Delta G/RT$ ranking and recomputed the average MGW profile. With sequentially removing "bad" spacers from the pool, any apparent selection for a specific MGW profile should become obvious, as it mimics the underlying, biophysical selection process.

3.4.9 Generation of Fly Lines

The full-length cDNA sequence for either wild-type of R2A & R5A mutant Exd (obtained by PCR from the protein-expression vectors), followed 3′ by the sequence coding for the small V5 peptide, was ligated into the multiple cloning site (MSC) of a vector with attb sites for ϕC31-mediated integration. The vector contained a tubulin (Tub) promoter and a poly-adenylation signal surrounding the MSC. Purified vectors were sent for injection into the attp40 site on chromosome 2L, additionally marked with w+. The resulting flies were crossed with respective balancer males or females (sp/CyO; MKRS/TM2) and progeny with successful integration of the transgene (marked by w+) were crossed once more to obtain balanced stocks. Sucessful balancer removal for both wild-type and mutant was achieved by selecting progeny not carrying the CyO balancer. For flies carrying the wild-type Tub-Exd-V5 transgenes, strains lacking the endogenous copy of Exd were obtained by crossing males y/+; Tub-$^{\mathrm{WT}}$Exd-V5/Tub-$^{\mathrm{WT}}$Exd-V5 against females Exd^{-}/FM7-GFP; sp/CyO and selecting for males lacking the FM7-balancer chromosome, yet marked with w+. The resulting males were crossed with females Exd^{-}/FM7-GFP; sp/CyO to obtain homozygous Exd^{-} females. In a final step Exd^{-}/Exd^{-} ; Tub-$^{\mathrm{WT}}$Exd-V5/CyO females were backcrossed with the first generation males y/Exd^{-}; Tub-$^{\mathrm{WT}}$Exd-V5/CyO, to obtain a stable line of (y or Exd^{-})/Exd^{-} ; Tub-$^{\mathrm{WT}}$Exd-V5/Tub-$^{\mathrm{WT}}$Exd-V5 flies.

3.4.10 Immunochemistry and Genetic Manipulations

The following antibodies for Immunohistochemistry were used: rabbit anti-Exd (Abu-Shaar et al., 1999) and mouse anti-V5 (Invitrogen, R960-25). Imaginal wing discs were collected from third instar larva, fixed in 4% formaldehyde for 25 minutes and stained with the antibody overnight in a 1:500 dilution. Discs were imaged at 20x magnification using con-

focal microscopy and processed using ImageJ software. Mutant clones for endogenous Exd where generated by FLP mediated recombination. Males carrying FRT19a-ubi-Red on the X chromosome were crossed with females <FRT19a Exd^{-}/FM7-GFP ; Tub-^{MUT}Exd-V5/+ with Exd^{-} clones marked by the absence of ubi-Red signal. The progeny was heat-shocked for 40 min at 37°C 48 h after egg laying (AEL) and imaginal wing discs were dissected 72 hours later from third instar, wandering larva. Imaginal discs were stained with both rabbit anti-Exd and mouse anti-V5 following the procedure described above. For the Western Blots on recombinant Hth^{HM}, recombinant Hth^{FL} and wing disc extracts, guinea-pig anti-Hth (raised against the N-terminus of Hth; GP52) (Ryoo and Mann, 1999) and goat anti-Hth (C-terminal specific; Santa Cruz, dg-20; no longer available) were used.

3.4.11 ChIP-seq

The following antibodies were used in ChIP-seq experiments: mouse anti-V5 (Invitrogen, R960-25), rabbit anti-GFP (Invitrogen A-11122) for Antp-GFP, guinea-pig anti-Hth (raised against the N-terminus of Hth; GP52) (Ryoo and Mann, 1999), goat anti-Hth (C-terminal specific; Santa Cruz, dg-20; no longer available). About $\sim$ 100 third instar larval wing discs were used for each ChIP-seq sample. All buffers contained protease inhibitor (cOmplete, Roche). Inverted larvae were cross-linked at room temperature (RT) for 10 min in 10 ml 1% formaldehyde solution buffered with 50mM HEPES (ph=8.0), immediately quenched with 1 ml 2.5M Glycine and washed for 5 minutes in quench-solution (125 mM glycine, in 1X PBS and 0.01% Triton X-100). Inverted and cross-linked larvae were washed twice with Buffer A (10mM HEPES, pH=8.0; 10mM EDTA, pH=8.0, 0.5mM EGTA, pH=8.0; 0.025 % Triton-X) and twice with Buffer B (10mM HEPES, pH=8.0; 200mM NaCl, 1mM EDTA, pH=8.0; 0.5mM EGTA, 0.01 % Triton X-100). Wing discs were detached on ice in Buffer B and transferred into a final volume of 1 ml Buffer C (10mM HEPES, pH=8.0 ;1mM EDTA, pH=8.0; 0.5mM EGTA, pH=8.0). Chromatin was sheared into fragments by using

a probe sonicator at 15% amplitude (total time: 12 min with 15 seconds on and 40 second off intervals) and flash-frozen in liquid nitrogen for storage at -80°C until further processing (no more than 1 week). Sheared chromatin was diluted in 5X RIPA dilution buffer (1x RIPA: 140mM NaCl; 10mM HEPES, pH=8.0; 1mM EDTA, pH=8.0; 1% Glycerol; 1% Triton X-100; 0.1% DOC) and blocked with 10μg of the respective IgG-coated magnetic beads (Dynabeads, ThermoFisher) for 1h at 4°C. Beads were removed with a magnetic stand and supernatant was transferred into a new, low-binding tube. At this point, 10% of the sample was set aside to serve as an input control. Specifc antibody (10 μg for mouse anti-V5, 8 μg for rabbit anti-GFP and 3-4 μg for either Hth antibody) and 1 % of Bovine Serum Albumine (BSA) was added to the remaining chromatin and incubate over night (o/n) at 4°C. The next day, $\sim 30\mu g$ of IgG-coated and pre-blocked (with 1 % BSA) Dynabeads were added to each chromatin antibody solution and incubated for another 2 hours. Antibody-bound TF-chromatin complexes were isolated by magnetic separation (5min on a magnetic stand) and beads were washed twice with 1x RIPA, once with high salt RIPA (500mM NaCl), once with LiCl-Buffer and once with TE (10 mM Tris-Base, pH=8.0; 1mM EDTA, pH=8.0). Beads with chromatin and the input sample were redissolved in 0.5 ml Elution-Buffer (TE with 0.5% Sodium Dodecyl Sulfate (SDS) and 50mM NaCl) and incubated for 30 min at 37°C with RNase, followed by 2 hours at 55°C with proteinase K (ThermoFisher). Remaining DNA-protein complexes were decrosslinked by incubating for 16 hours at 65°C. DNA was separated from the Dynabeads by magnetic separation and purified by phenol:chloroform extraction and DNA precipitation using 1x volume of isopropanol in 100mM ammonium acetate and $1\mu l$ glycogen. Precipitated DNA was redissolved in $30\mu l$ TE.

3.4.12 ChIP-seq Library Preparation and Sequencing

ChIP-seq libraries were constructed using the NEBNext Ultra DNA Library Prep Kit for Illumina with NEBNext Mulitplex Oligos (one separate index per sample) and following

standard instructions. For the PCR amplification, 14-15 cycles were used depending on the amount of starting material, which was generally between 3-10 ng of precipitated DNA. For the input samples no more than 10 ng of DNA was used to match them as closely as possible to their respective IPs. For the final size selection, AMPure xp beads (Agencourt) were used and larger ($> 550bp$) and smaller ($< 150bp$) fragments were removed by a double-sided size selection with first 0.6x volume of beads to DNA and retaining the supernatant, followed by a final concentration of 0.9x beads to DNA and retaining the DNA-bound to the beads. Quality control was done by checking the DNA size distribution with a Bioanalyzer. ChIP-seq libraries were diluted to 2 nM, using a Qubit to verify the final concentration, pooled and sequenced with a v2 75 cycle high-output kit on an Illumina NEXTSeq Series desktop sequencer at the Genome Center at Columbia University.

3.4.13 ChIP-seq Data Processing

The four separate, raw fastq-files (from the four lanes of the sequencing run) were first collapsed into one file and subsequently aligned to the *D. melanogaster* genome version dm6 (2014, GenBank accession: GCA_000001215.4). Alignment rates were overall high and varied between 92-97% . Aligned sam files were next converted into bam files, sorted and cleared from duplicate reads using the samtools functions view, sort and rmdup (Li et al., 2009; Li, 2011). The sorted, unique bam files were indexed and converted into bigwig files using the bamCoverage function in the Deeptools suite with parameters -bs 1 -e 125 (Ramírez et al., 2016). Peaks were called using the MACS2 (Zhang et al., 2008) function callpeak using the input samples as control files with parameters -g dm -q 0.01 or 0.05 –nomodel –extsize 125 or 175 (for dg20 due to a slightly broader fragment size distribution). For further downstream analysis, peak summits from the higher sequenced Exd^{WT}-V5 ChIP replicate with a q-value threshold of 0.01 were used.

3.4.14 De-Novo Motif Discovery Using Homer

For the de-novo motif discoveries, the 50bp sequences surrounding each peaks summit were extracted and split into four groups based on the WT/MUT-V5-IP coverage ratios. The raw, combined counts from both replicates at the called peak summit were used and peaks were split at i) ratio (X) > mean (μ) + one standard deviation (σ), ii) $\mu+\sigma > X > \mu$ iii) $\mu > X > \mu - \sigma$ and iv) $X < \mu - \sigma$. Homer (Heinz et al., 2010) was run using the findMotifsGenome.pl function and the following parameters: -size 50 -len 6,9,12.

3.4.15 Coverage Plots and Downstream Peak Analysis

Heatmaps for the raw IP coverage of the five ChIP-seq samples (Exd^{MUT}, Exd^{WT}, Antp-GFP, Hth-C-Ab, Hth-N-Ab) were generated on the Exd peak set sorted by the WT/MUT IP-ratio using the Deeptools functions computeMatrix and plotHeatmap (parameters: –sortRegions "no" –refPointLabel –missingDataColor 1). Raw read coverage was extracted at the Exd peak summits (q-value=0.01) from the bigwig files for all five ChIP samples. Pairwise-coverage plots for Exd^{MUT} and Exd^{WT} were based on the combined coverage of both replicates. For each Exd peak, sequences surrounding the peak summit (±50bp) were extracted. Each peak sequence was then scanned with i) an Exd-Antp binding model (obtained by fitting a No Read Left Behind (NRLB) model (Rastogi et al., 2018) to the Lib-1 data set for Hth^{HM}-Exd-Hox (Slattery et al., 2011)), ii) an Exd-Hth model (obtained by fitting a NRLB model to the Lib-2 data for Hth^{FL}-Exd), and iii) a Hth-only model (PSAM model derived from Lib-4, using TTGACAGC as a seed). For each model view (in total there are [100−(number of positions specified by the model)+1] possible binding sites in each 100bp peak sequence), the score was computed for the "+" and "-" strand and only the maximum of the two was considered for each view. The cumulative peak score for each model was computed by summing up the scores across all views:

$$PeakScore = \sum_{\text{views(v)}} \exp \frac{\Delta\Delta\text{G}(\text{Sequence}_\text{v})}{\text{RT}} \tag{3.6}$$

Peak scores were then correlated with the IP-signal strength (IP-signal / Input) for either Exd^{WT}, Exd^{MUT}, or Antp-GFP.

For the comparison of C-Ab/N-Ab ratios across different complex compositions, the maximum score for the Exd-Antp and the Hth-only model was inferred for each peak. Peaks were then classified into groups based on the following thresholds on the model score: i) Exd-Antp only peaks with a strong Exd-Antp site (peak maximum >1/10 of maximum across all peaks) and no good Hth site (maximum peak score <0.4), ii) Hth only peaks with a strong Hth site (maximum peak score >0.9) and no good Exd-Antp site (peak maximum <1/10 of maximum across all peaks), and iii) ambiguous peaks belonging to neither i) nor ii). The t-distribution was used to test for a significant difference in the C-Ab/N-Ab IP-ratio between the two peak groups described under i) and ii).

For the comparison between "high affinity" and "low affinity" sites, peaks were scanned for motif matches for TGA**Y**NNAY and subdivided based on the identity of the first **Y** (Y=T or Y=C). The t-distribution was used to test for significant differences in the IP-enrichments (Antp-GFP, Exd^{WT}, Exd^{MUT}, and WT/MUT IP-ratio) between the two affinity classes.

3.5 Author Contributions

Concept and Design, J.F.K.; H.J.B. and R.S.M.; Experimental Methods, J.F.K. (most SELEX-seq experiments, mutagenesis, binding assays, ChIP-seq, fly work); N.A. (intital SELEX-seq experiments for Hth^{FL}-Exd^{WT}-Dfd for Lib-3a and Lib-3b); Validation, J.F.K.; Data Analysis, J.F.K., C.R. provided the NRLB models for Exd-Hth and Exd-Antp; Visualization, J.F.K.; Supervision, H.J.B. and R.S.M.; Project Administration, H.J.B. and R.S.M.; Funding Acquisition, H.J.B. and R.S.M

Chapter 4

General Discussion: How (Epigenetic) Context Impacts TF-DNA Recognition

The questions how and with what affinity TFs identify and bind different DNA sequences is integral to predicting regulatory sites and potential gene targets *in vivo*. In the past 30-40 years, a large number of different methods have been developed to address this key aspect of regulatory genomics. These can be broadly split into two major classes: i) those focusing on the exact mechanism of TF binding given a specific sequence and ii) those aimed at identifying which sequences are recognized and to what extent they are bound by a given TF, rather than how exactly.

The first type of studies are producing detailed, atomic-resolution snapshots of TFs (mostly limited to the DBD) bound to DNA, allowing us to obtain insights into the TF's domain structure, the positioning and distances of individual amino acids towards base pairs in the major groove or the phosphate backbone and the computation of charge distributions across the TF-DNA interaction surface. Moreover, they allow us to compare different structures, with the goal of identifying general principles of DNA sequence recognition, which ideally can be simplified to a few sets of rules broadly applicable for all other TFs. Most of these structures have been obtained by X-ray crystallography, but other types of spectroscopy and high-resolution imaging have also contributed to the existing TF-DNA structural repertoire, such a NMR-spectroscopy and more recently cryo-electron microscopy. Although it

has been established that TF-DNA interactions cannot be expressed in simple terms (Pabo and Sauer, 1992; Slattery et al., 2014), such detailed atomic maps are still a valuable resource for a number of reasons. They enable us to make predictions about the impact on binding of individual amino acid mutations (e.g. those linked to disease), and aid in the design of loss-of-function mutations that can be used for downstream *in vivo* perturbation experiments. Structures allow us to truly "see" and therefore understand the mechanisms supporting a TF-DNA interface. Once we understand, we can manipulate and redesign the system for another purpose, such as investigating cellular mechanisms or disease states.

The other line of experiments addresses the problem of TF-DNA recognition from a different angle and has been driven by the above-described realization that a single structure cannot explain a TF's entire sequence repertoire. In these approaches, TFs are regarded as fixed entities with well-defined geometric constraints (presumably exactly those identified in a previous structure) and instead of visualizing the configuration a TF adopts when binding to different sequences, is focused on the outcome: complete quantitative characterization of sequence selectivity. It is important to distinguish between *in vitro* methods, which treat the TF and DNA ligands in isolation and thus measure the biophysical TF binding affinity in the absence of confounding factors, and *in vivo* methods, which measure the average (across cell population) occupancy by a TF at a given genomic site. The latter is the quantity we often want to predict and rationalize, since it is indicative of downstream gene regulatory function. However, it is obviously confounded by numerous molecular players present in the nucleus of a cell, from nucleosomes to binding partners to protein and DNA modifications. Regardless of whether the experiment is performed *in vitro* or *in vivo* and of the specific experimental setup, the readout is a quantifiable, sequence-associated signal that represents a TF's binding preference (in terms of sequencing read counts for ChIP-seq and SELEX-based assays or of fluorescence intensity for array-based technology). Those methods allow us to rank and compare sequences and are therefore complementary to structural approaches. Their ultimate purpose is to summarize the complex binding mechanisms used by TFs, without

the need to obtain a detailed structure for each possible context. A typical experiment results in a highly-dimensional and complex count table, which can be summarized by a motif representation (e.g. PSAM or PWM). Often, a few additional assumptions are made, such as a single recognition mode for all sequence contexts, and independence between nucleotide positions. It is also typically assumed that the DBD (typically used in the experiment) fully captures a TF's sequence preference, and that affinity can be treated as a "binary" variable, following the motto – a sequence is either bound sufficiently well or it stems from non-specific binding and therefore should not be considered. Despite some overlap between *in vivo* and *in vitro* derived motif matrices, *in vitro* models too often fall short when used to predict the entire set of genomic binding sites, with only a small fraction being explained by the presence of a reasonable motif match (Wang et al., 2012).

Instead of primarily ascribing this short coming to complex *in vivo* binding behavior, we might want to reconsider whether, in our attempt to simplify, we might have missed many crucial aspects of TF binding. Given that a site is accessible in the genome, there is no reason to believe that a TF interacts substantially differently with DNA in a living cell, than in a test tube. A single crystal structure exits in a 3D subspace, that defines every atomic position of a TF-DNA complex and the relevant interactions occurring at the TF-DNA interface. Probing of TF-DNA interaction using high-throughput sequencing has opened up an additional dimension, providing a quantitative measure for a TF's relative sequence selectivity over a wide range of binding contexts, not just the sequence chosen in a structural study (Figure 4.1 on page 180). The sum of all energetic contributions between a TF and a specific DNA ligand determines the ranking of that ligand in a binding experiment, but clearly a lot of detailed structural information is lost during this mapping from a 3D space to a 1D scale. Specifically, we knowingly or not, make an important simplifying assumption: that both TF and DNA adopt a rather fixed conformation, and that readout is consequently driven by a few key interaction sites. This belief has been further solidified by the seeming

rigidity of DNA molecules and the obvious lack of far-ranging sequence dependencies when analyzing sequencing data. In contrast to this stands the observation that the majority of contacts in typical crystal structures are made with the DNA backbone and not with bases, arguing for the relevance of shape recognition in TF binding. Moreover, DNA itself has been shown to be rather flexible. Sampling of different DNA micro-states could therefore be crucial for TF recognition (Nikolova et al., 2011). That the same is true for TFs is needless to say, given the well-established mechanisms of conformational switching observed for other proteins such as catalytic enzymes, membrane-bound channels or ribosomal proteins. By using crystal structures (that naturally cannot reveal flexible domains) and TFs trimmed down to the DBD, we might have therefore biased our perception of TF-DNA recognition. Perhaps, TF binding is highly adaptive, relying much more on different recognition modes than currently perceived. By considering only sites that are highly enriched in sequencing experiments to guide us, we may have severely limited our understanding of adaptive binding, especially since there is no established relationship between base identity and overall protein conformation (Pabo and Sauer, 1992).

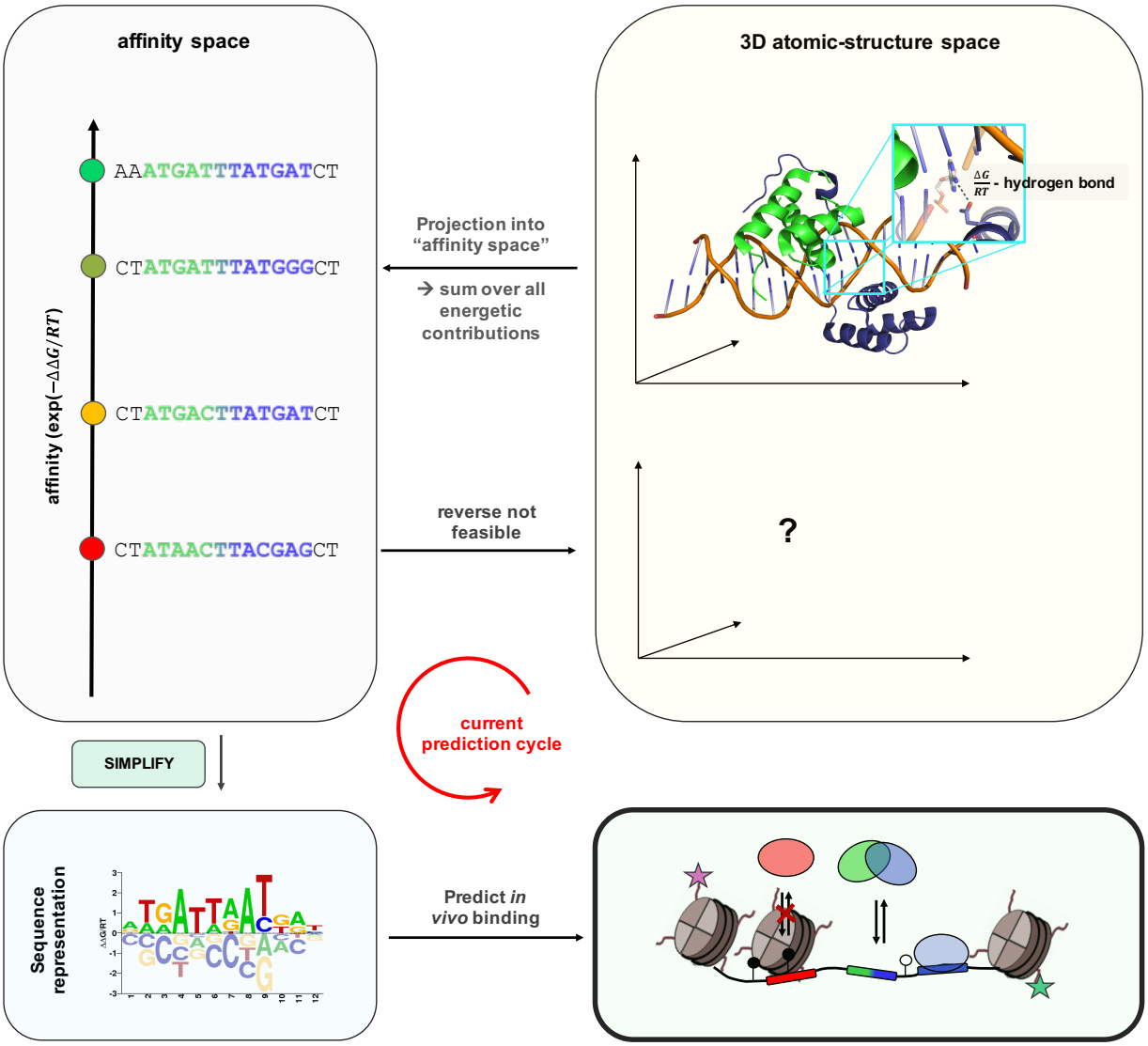


Figure 4.1: Current approach to capturing TF binding specificity: The detailed interaction surface as illustrated by a crystal structure is summarized by a single quantity (affinity or total $\Delta\Delta G/RT$ of binding) and compared to those of all other sequences. After an alignment step that uses the most enriched sequence in an sequence enrichment assay (e.g. SELEX-seq) a motif representation is constructed from the relative enrichment scores. The obtained motif matrix is then used to predict *in vivo* binding.

The work described in this thesis advocates for us to shift gears and focus on expanding our analysis tool set to truly capture the rich mechanisms through which TFs recognize their cognate binding sites. Using a simple representation for a system that in reality is highly complex certainly provides a first draft, but will ultimately limit us in our quest to truly understand *in vivo* binding behavior. Supporting this notion, more and more evidence for additional layers in TF-DNA recognition has emerged in recent years. It includes: i) the identification that cofactors can alter the binding specificity of TFs upon dimerization (Slattery et al., 2011); ii) the identification of low affinity sites that are functional *in vivo* and in fact required to confer specificity (Crocker et al., 2015; Farley et al., 2015); iii) the insight that flanking DNA sequences play a perhaps subtle, yet important role in specifying homologous and paralogous TF sequence preferences (Le et al., 2018; Shen et al., 2018); and iv) the finding that DNA shape intriniscally varies with its underlying sequence (Zhou et al., 2013).

Taking together these findings, we can formulate the challenge ahead in the following way: How can we define more holistic models that capture TF binding beyond a singular sequence representation? How do we account for the presence of cooperative binding and adaptive behavior when a TF encounters its ligand (DNA or protein) in different epigenetic contexts?

To address these questions, we first need to capture each possible epigenetic context and not just the obvious ones. Chapter 2 of this thesis therefore described a method that allows us to include epigenetic DNA modifications, which will inevitably alter the biochemical environment seen by a TF and with it the total free energy of binding (see Figure 4.2 on page 187). We have laid out potential pitfalls when using simple PWMs derived from and conditioned on a single top binding site, and provided an alternative, feature-based modeling framework that considers CpGs in their dinucleotide setting and simply adds them to the model as binary features. One advantage of our method, compared to others (Yin et al., 2017), is the simultaneous probing for TF binding to methylated and unmethylated sequences

in competition, therefore allowing a direct comparison, which is capable of detecting methylation effects on binding even in sequence contexts with relatively low affinity. Moreover, we could establish, for the well-known tumor suppressor p53, that the same binding preferences prevail *in vivo*. Therefore accounting for epigenetic effects such as DNA methylation may help resolve the discrepancy seen between *in vitro* binding models and *in vivo* TF occupancy.

A next step in improving our current methodology, is to build more accurate, biophysical models truly capturing the selection process of the entire probe, as seen by a TF in a high-throughput experiment. In addition, a model should be flexible to allow for multiple binding modes to contribute to the overall selection, such as for instance different populations of sub-complexes that can form (e.g. monomeric versus dimeric or homo- versus heterodimeric, Figure 4.2 on page 187). This task was recently achieved by (Rastogi et al., 2018) by building a model that explains the overall probe selection by considering every possible binding frame, allowing for multiple binding modes and incorporation of dinucleotide features. The latter aspect is of particular importance as dinucleotide interactions capture most of the intrinsic DNA shape properties (Rube et al., 2018). Recognition of specific, intrinsic DNA shape properties has been established to be an integral part of TF-sequence recognition (Rohs et al., 2009b; Slattery et al., 2014; Abe et al., 2015) and therefore should be captured in a comprehensive TF-specificity model.

Of particular importance is the connection that a TF's sensitivity to intrinsic DNA shape features can provide between the quantification of affinity (sum of all energetic contributions) on the one hand, and the detailed characterization of a TF-DNA conformation in 3D on the other. Knowing both the intrinsic DNA shape parameters at a given position across different sequences and the relative ranking of those sequences in terms of affinity allows us to connect the structural properties of the naked DNA ligand to a TF's selectivity and thereby infer mechanistic details of adaptive TF binding (Figure 4.2 on page 187). If sequences that share specific shape properties are selectively enriched, they perhaps minimize the free energy gap between the unbound and bound state by already resembling the preferred bound

state geometry. Preferred sequences with a distinct pattern of intrinsic DNA shape might thus foreshadow exactly this, the bound-state geometry. If this were true, we could obtain information about certain structural aspects of TF-DNA complexes in different sequence contexts, even in the absence of crystal structures, by simply correlating affinity and DNA shape parameters. However, this does not mean that existing structures are obsolete. On the contrary, they are important in ruling out the possibility that an affinity-shape correlation might be the result of a hydrogen bond that selects for a specific base pair and causes a pseudo correlation as a result of the dependencies between shape and sequence (see Rube et al. for a detailed analysis (Rube et al., 2018)). A plausible indication of true shape recognition, via the DNA minor groove, arises when the structure suggests the presence of an arginine near a stretch of bases, but seemingly not directly contacted by the protein (as seen for the DNA spacer between Hth and Exd, see Figure 3.5 B on page 116). In order to distinctively demonstrate that shape readout is utilized to fine tune TF selectivity, amino acids hypothesized to recognize such structural features must be mutated in a way that they no longer can fulfill this function. A first, definitive demonstration thereof was given by (Abe et al., 2015), showing that the mutation of specific amino acids within the Hox N-terminal arm resulted in the selection of DNA ligands that no longer exhibited the typical minor groove width profile seen in high affinity sites for the wild-type protein. Although, the changes in sequence selectivity were not as severe, it clearly demonstrated that TFs make use of shape readout with varying degrees depending on the sequences they encounter.

In this work, we have now demonstrated how the identification of such a structural recognition mechanism can be used to provide an alternative way to interrogate TF binding preferences both *in vitro* and *in vivo*. Specifically, by manipulating the TF in such a way that only the shape readout is abolished while other, core interactions can be left intact. By using this seemingly "more subtle" perturbation, and analyzing its effect on the overall binding preferences, we have created a tool that by direct comparison between wild-type and mutant, allows us to tease apart the flexibility of TF-DNA interactions: It allowed us to differentiate

between complex compositions, complex configurations, and binding modes that are distinct in their usage of either direct "base readout" or indirect "shape readout". In contrast to a direct hydrogen-bond disruption, which often globally affects binding, shape readout is much more dependent on the overall conformation of the TF of interest and its binding partners and thus provides a better means to selectively destabilize particular binding modes. By doing so for Exd and Hth, we have demonstrated that sequences outside the core binding site contribute to the overall sequence selection by tuning intrinsic DNA shape, which in turn is recognized by the TF. Furthermore, we could show that the selection of spacer sequences depends on the first base of the core binding site, in line with the dependency of shape on neighboring nucleotides. Mutating the amino acids responsible for the shape readout differentially impacted spacer sequence selection when we conditioned on the base identity of that first Exd-Hox base. This implied that different conformations of the Exd N-terminal arm exit and that they depend on the sequence context (base identity of the first base). In any other scenario, we would not expect to observe a selective destabilization for each of the four base identities.

Perhaps even more surprisingly, we discovered two populations of sequences where base identity more than 3bp away from the base primarily impacted by the shape mutant modulated the response to the mutation. The two identified sequence contexts reflect different affinity ranges and differ in the sequence identity of a base-pair that, given the currently available crystal structures with "high-affinity" type sequences, is presumably not directly read-out by a specific amino acid (at least not in every sequence context). The finding that the shape mutant destabilizes the Exd-Hox complex to a different degree in the two sequence contexts strongly suggests the existence of two subtly different, yet distinct binding modes present in the wild-type complex: one conformational mode for the "high affinity" sequence context, that heavily utilizes shape readout to optimize TF-sequence recognition, and another mode for the "low affinity" sequence context, which presumably uses a subtly different structural arrangement that perhaps is less tight, but does not rely as heavily on the shape

readout. Simply considering the PWM would never have provided this level of structural insight. The shape mutation, however, clearly separated these two recogntion modes, suggesting a crucial role of even minor structural rearrangements for TF sequence selectivity. Importantly, this preference change, seen upon removal of the shape readout, even persisted when tested *in vivo*.

Besides distinguishing individual binding modes, the shape-mutant protein also affected different TF complex compositions with varying degrees, therefore equipping us with a tool to selectively perturb the binding of different complex compositions *in vivo*. We successfully used this tool to identify the presence of trimeric Hth-Exd-Hox binding sites, but also to obtain new insight into the regulation of the two Hth splice variants. Moreover, we could demonstrate that more than one complex contributes to the overall ChIP-seq IP signal of Exd: The correlation between predicted binding affinity and IP-signal was selectively removed in our shape-mutant only for those complex compositions whose binding was strongly affect by the mutation.

The above example has demonstrated in a powerful way that TF binding to DNA is much more versatile than our current models and structural insights might suggest, and that TFs indeed exhibit adaptive binding behavior when encountering different sequence contexts. Addressing those aspects of TF binding explicitly, and ultimately including them into our binding models, will be essential for the challenging task of fully recapitulating *in vivo* binding patterns.

The role of shape recognition in TF binding does not stop at the primary sequence context, but once again, is influenced by DNA modifications. CpG methylation for instance can alter DNA shape without changing base identity. An example of this behavior is provided with the selective destabilization of a CpG Pbx-Hox spacer upon methylation. A CpG is already an unfavored spacer sequence, as it has a relatively wide minor groove, whereas Pbx-Hox prefers a narrow one (Abe et al., 2015). However, the groove is additionally widened

when the CpG becomes methylated, thus providing an explanation for the selective destabilization of methylated CpG spacers compared to unmethylated ones (Rao et al., 2018).

To summarize, the work described in this thesis paves the way to the use of novel strategies for interrogating adaptive TF binding. The description of a method that allows incorporation of epigenetic marks into TF binding models is provided, as well as an explanation for two major mechanisms by which these marks can impact TF binding: i) either directly by interacting with neutral or charged TF surface patches, resulting in increased or decreased binding respectively – termed "thymine mimicry", or ii) indirectly, by changing the intrinsic shape of the DNA molecule and thereby increasing or decreasing the energy gap between unbound and bound state. In addition to expanding the DNA context space, we also pioneered a field aimed at not only identifying shape readout mechanisms by combining the knowledge of intrinsic DNA shape and TF affinity, but also at utilizing the identified shape recognition to introduce targeted mutations. Using this strategy, we have provided a means to combine genetic manipulation and biophysical insights to probe the flexible nature of TF-DNA interactions. The resulting "shape mutants" can in turn be used to selectively perturb TF-DNA binding both *in vitro* and *in vivo*. This provided insights into different TF-DNA binding modes (adopted for different sequence contexts) and differentiated between different complex compositions (Figure 4.2 on page 187). Such targeted TF engineering represents a novel strategy for interrogating TF binding and function *in vivo*.

Instead of assuming a one-size-fits-all binding mechanism for TF-DNA complexes independent of sequence context, we should use the knowledge about a TF's sequence selectivity and the intrinsic DNA shape of those selected sequences to make structural predictions about the final bound state. Based on those initial insights, we can design mutants that selectively destabilize some (for instance MGW shape readout), but not all binding modes (intact base readout), and gain insights that go far beyond the classical motif representation. The biological relevance of such seemingly "subtle" or "minor" binding difference is demonstrated

by numerous recent studies that all illustrate how TF binding depends much more on the overall "epigenetic" context than a single crystal structure or sequence representation could possibly reveal (Le et al., 2018; Crocker et al., 2015; Shen et al., 2018; Jolma et al., 2015; Rastogi et al., 2018; Kribelbauer et al., 2017; Yin et al., 2017).

Rather than oversimplifying, we should seek to preserve the complexity of this system in our models, to reveal the hidden information in our genomes and to answer outstanding biological questions.

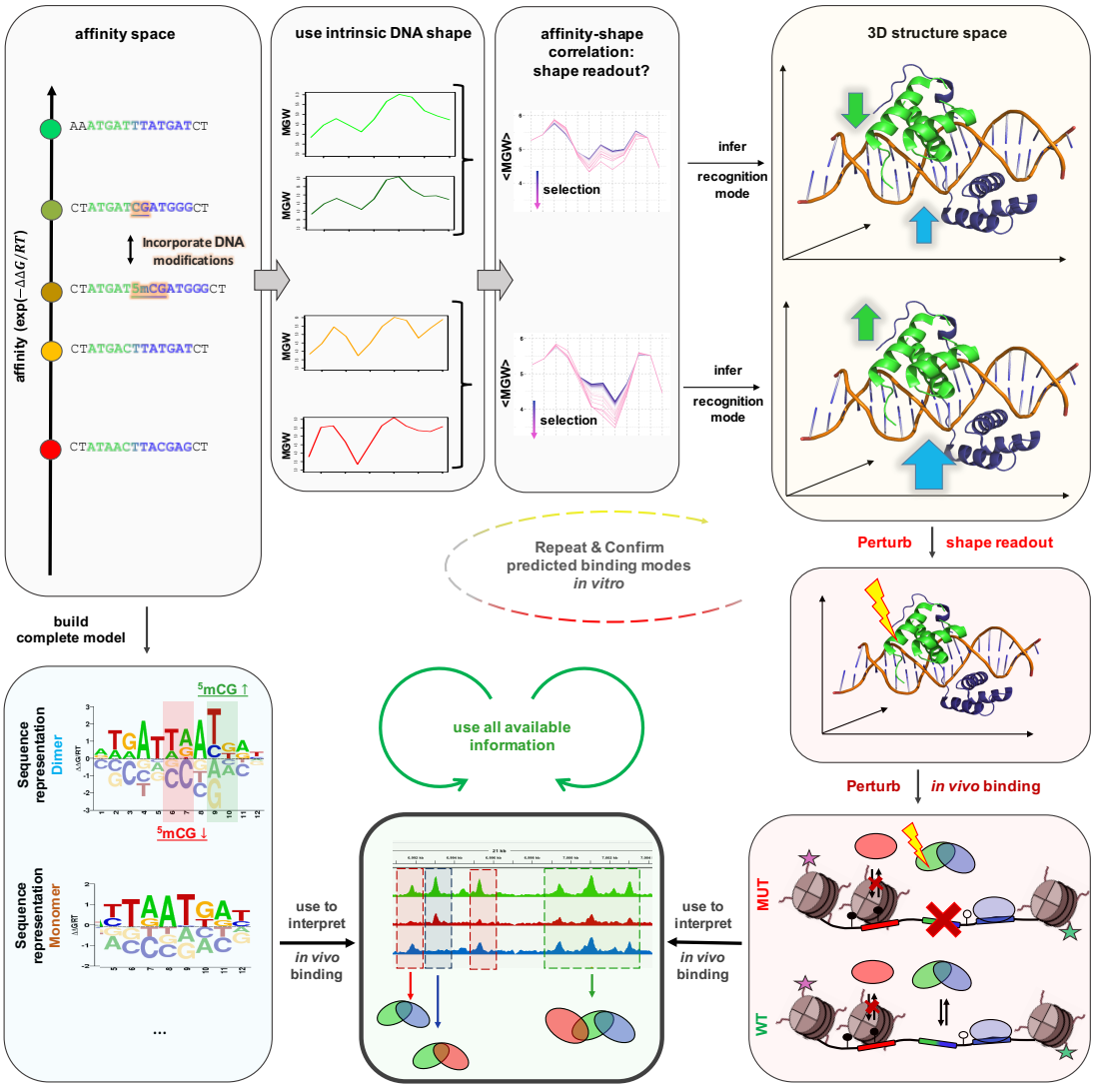


Figure 4.2: Updated approach to dissecting complex TF binding specificity:

Figure 4.2: *continued from page 187.*

We first need to probe binding for every epigenetic context (DNA modifications, orange highlight) to obtain a complete sequence model (capturing different binding modes and allowing for dinucleotide interactions, blue). We can next incorporate knowledge about intrinsic DNA shape and existing crystal structures by computing affinity-shape correlations and comparing them to the amino acid seuqence in close proximity (box 2 and 3, top). From the correlation we can infer structural recognition modes that can aid in the design of shape-readout defficient TFs (yellow and red box). Those engineered TFs can be tested *in vitro* and be introduced *in vivo* to perturb binding preferences (bottom red box). Comparing wild-type and mutant binding, together with using accurate models can reveal hidden information about complex composition or conformational changes for different genomic binding sites (green box).

4.1 Outlook

Ideally, the approach described in this work will bring us one step further to closing the gap between *in vitro* predicted and *in vivo* observed TF binding preferences. Identifying distinct structural recognition modes that are subsequently removed by targeted genetic manipulations, and comparing the binding behavior of such "shape-readout-deficient" TFs might provide a general avenue for interrogating TF binding *in vivo*.

Given the redundancy of structural motifs within TF families, the coverage of many families (with at least one member) by existing crystal structures, and the availability of a large number of high-throughput data sets capturing TF binding specificity *in vitro* (Jolma et al., 2010, 2013, 2015; Isakova et al., 2017), it might be feasible to infer a set of commonly used binding mechanisms that predominantly rely on shape recognition. Using these as

a starting point, a large number of shape-mutant TFs can be screened *in vitro* and their affinity-to-shape maps can be compared to their wild-type counterparts. Mutations, for which shape readout is confirmed, can then be used in downstream *in vivo* screening by comparing the binding patterns of wild-type and mutant ChIP-seq signals. Given the strong differentiating power observed for Exd (by just mutating 2 amino acids within the overall flexible N-terminal end of the DBD), we can suspect that similar effects will be observed for other TFs as well. With the advent of methods that allow genomic manipulation in higher throughput, such as CRISPR (Jinek et al., 2012), it might even be feasible to create *in vivo* TF-mutant libraries. The resulting wild-type and mutant ChIP-seq datasets might provide an invaluable resource in our quest to ultimately identify the complex mechanisms governing *in vivo* TF binding by perhaps differentiating peaks in terms of i) the presence of different binding partners and ii) the dependency of a site on specific binding mechanisms.

Bibliography

Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H. J., Rohs, R. and Mann, R. S. (2015). Deconvolving the recognition of DNA shape from sequence. Cell *161*, 307–318.

Abu-Shaar, M., Ryoo, H. D. and Mann, R. S. (1999). Control of the nuclear localization of Extradenticle by competing nuclear import and export signals. Genes and Development *13*, 935–945.

Aishima, J., Gitti, R. K., Noah, J. E., Gan, H. H., Schlick, T. and Wolberger, C. (2002). A Hoogsteen base pair embedded in undistorted B-DNA. Nucleic Acids Research *30*, 5244–5252.

Alipanahi, B., Delong, A., Weirauch, M. T. and Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology *33*, 831–838.

Ambrosi, C., Manzo, M. and Baubec, T. (2017). Dynamics and Context-Dependent Roles of DNA Methylation. Journal of Molecular Biology *429*, 1459–1475.

Anderson, W. F., Ohlendorf, D. H., Takeda, Y. and Matthews, B. W. (1981). Structure of the cro repressor from bacteriophage λ and its interaction with DNA. Nature *290*, 754–758.

Arnone, M. I. and Davidson, E. H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. Development *124*, 1851–1864.

Auffray, C. &. and Rougeon, F. (1980). Nucleotide sequence of a cloned cDNA corresponding to secreted u chain of mouse immunoglobin. Gene *12*, 77–86.

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C.-F., Coburn, D., Newburger, D. E., Morris, Q., Hughes, T. R. and Bulyk, M. L. (2009). Diversity and complexity in DNA recognition by transcription factors. Science (New York, N.Y.) *324*, 1720–3.

Bailey, T. L. and Elkan, C. (1994). Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology *August*, 28–36.

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I. and Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. Cell *129*, 823–837.

Bartlett, A., O'Malley, R. C., Huang, S. S. C., Galli, M., Nery, J. R., Gallavotti, A. and Ecker, J. R. (2017). Mapping genome-wide transcription-factor binding sites using DAP-seq. Nature Protocols *12*, 1659–1672.

Baylin, S. B. and Jones, P. A. (2011). A decade of exploring the cancer epigenome - biological and translational implications. Nature reviews. Cancer *11*, 726–34.

Bestor, T. H., Edwards, J. R. and Boulard, M. (2015). Notes on the role of dynamic DNA methylation in mammalian development. Proceedings of the National Academy of Sciences *112*, 6796–6799.

Blattler, A. and Farnham, P. J. (2013). Cross-talk between site-specific transcription factors and DNA methylation states. Journal of Biological Chemistry *288*, 34287–34294.

Bourc'his, D., Xu, G. L., Lin, C. S., Bollman, B. and Bestor, T. H. (2001). Dnmt3L and the establishment of maternal genomic imprints. Science *294*, 2536–2539.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. and Greenleaf, W. J. (2013). Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. Nature Methods *10*, 347–355.

Bulyk, M. L., Huang, X., Choo, Y. and Church, G. M. (2001). Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. Proceedings of the National Academy of Sciences *98*, 7158–7163.

Bumgarner, R. (2013). DNA microarrays: Types, Applications and their future. Curr Protoc Mol Biol. *6137*, 1–17.

Calo, E. and Wysocka, J. (2013). Modification of Enhancer Chromatin: What, How, and Why? Molecular Cell *49*, 825–837.

Cann, J. R. (1989). Phenomenological theory of gel electrophoresis of protein-nucleic acid complexes. J Biol Chem *264*, 17032–17040.

Carey, M. (1998). The enhanceosome and transcriptional synergy. Cell *92*, 5–8.

Charoensawan, V., Wilson, D. and Teichmann, S. A. (2010). Genomic repertoires of DNA-binding transcription factors across the tree of life. Nucleic Acids Research *38*, 7364–7377.

Ching, T., Himmelstein, D. S., Beaulieu-jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-m., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., Lengerich, B. J., Israeli, J., Lanchantin, J., Woloszynek, S., Carpenter, A. E., Shrikumar, A., Xu, J., Cofer, E. M., Lavender, C. A., Turaga, S. C., Alexandari, A. M., Lu, Z., Harris, D. J., Decaprio, D., Qi, Y., Kundaje, A., Peng, Y., Wiley, L. K., Huang, A., Gitter, A. and Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface *15*.

Craig, D. A. (1993). The Cheng-Prusoff relationship: something lost in the translation. Trends in Pharmacological Sciences *14*, 89–91.

Crawford, G. E., Holt, I. E., Whittle, J., Webb, B. D., Tai, D., Davis, S., Margulies, E. H., Chen, Y. D., Bernat, J. A., Ginsburg, D., Zhou, D., Luo, S., Vasicek, T. J., Daly, M. J., Wolfsberg, T. G. and Collins, F. S. (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Research *16*, 123–131.

Crist, R. C., Roth, J. J., Waldman, S. A. and Buchberg, A. M. (2011). A conserved tissue-specific homeodomain-less isoform of MEIS1 is downregulated in colorectal cancer. PLoS ONE *6*, 1–7.

Crocker, J., Abe, N., Rinaldi, L., McGregor, A. P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., Mann, R. S. and Stern, D. L. (2015). Low affinity binding site clusters confer HOX specificity and regulatory robustness. Cell *160*, 191–203.

Davidson, E. H., Jacobs, H. T. and Britten, R. J. (1983). Eukaryotic gene expression: Very short repeats and coordinate induction of genes. Nature *301*, 468–470.

Day, W. H. and Mcmorris, F. R. (1992). Critical comparison of consensus methods for molecular sequences. Nucleic Acids Research *20*, 1093–1099.

Deplancke, B., Dupuy, D., Vidal, M. and Walhout, A. J. M. (2004). A Gateway-Compatible Yeast One-Hybrid System. Genome Research *14*, 2093–2101.

Derisi, J. L., Iyer, V. R. and Brown, P. O. (1997). Exploring the Metabolic and Genetic Control of Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. Science *680*.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., Thomas, M., Davis, C. A., Shiekhattar, R., Gingeras, T. R., Hubbard, T. J., Notredame, C., Harrow, J. and Guigo, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs : Analysis of their gene structure , evolution , and expression. Genome Biology and Evolution *22*, 1775–1789.

Domcke, S., Bardet, A. F., Adrian Ginno, P., Hartl, D., Burger, L. and Schübeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. Nature *528*, 575–579.

Edwards, J. R., O'Donnell, A. H., Rollins, R. A., Peckham, H. E., Lee, C., Milekic, M. H., Chanrion, B., Fu, Y., Su, T., Hibshoosh, H., Gingrich, J. A., Haghighi, F., Nutter, R. and Bestor, T. H. (2010). Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. Genome Research *20*, 972–980.

El-Deiry, W. S., Kern, S. E., Pietenpol, J. A., Kinzler, K. W. and Vogelstein, B. (1992). Definition of a consensus binding site for p53. Nature Genetics *1*, 45–49.

Endres, R. G., Schulthess, T. C. and Wingreen, N. S. (2004). Toward an atomistic model for predicting transcription-factor binding sites. Proteins: Structure, Function and Genetics *57*, 262–268.

Farley, E. K., Olson, K. M., Zhang, W., Brandt, A. J., Rokhsar, D. S. and Levine, M. S. (2015). Suboptimization of developmental enhancers. Science *350*, 325–328.

Ferguson, J. A., Steemers, F. J. and Walt, D. R. (2000). High-density fiber-optic DNA random microsphere array. Analytical Chemistry *72*, 5618–5624.

Fernández, A. and Bayón, G. (2015). H3K4me1 marks DNA regions hypomethylated during aging in stem and differentiated cells. Genome Research *1*, 27–40.

Foat, B. C., Houshmandi, S. S., Olivas, W. M. and Bussemaker, H. J. (2005). Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. Proceedings of the National Academy of Sciences of the United States of America *102*, 17675–17680.

Foat, B. C., Morozov, A. V. and Bussemaker, H. J. (2006). Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics *22*, 141–149.

Frank, J. (2017). Time-resolved cryo-electron microscopy: Recent progress. Journal of Structural Biology *200*, 303–306.

Fu, Y., Luo, G. Z., Chen, K., Deng, X., Yu, M., Han, D., Hao, Z., Liu, J., Lu, X., Doré, L. C., Weng, X., Ji, Q., Mets, L. and He, C. (2015). N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. Cell *161*, 879–892.

Funk, W. D., Pak, D. T., Karas, R. H., Wright, W. E. and Shay, J. W. (1992). A transcriptionally active DNA-binding site for human p53 protein complexes. Molecular and cellular biology *12*, 2866–2871.

Galas, D. J., Eggert, M. and Watermant, M. S. (1985). Rigorous Pattern-recognition Methods for DNA Sequences Analysis of Promoter Sequences from Escherichia coli. Methods *186*, 117–128.

Galas, D. J. and Schmitz, A. (1978). DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic acids research *5*, 3157–70.

Garvie, C. W. and Wolberger, C. (2001). Recognition of specific DNA sequences. Molecular Cell *8*, 937–946.

Gebelein, B., McKay, D. J. and Mann, R. S. (2004). Direct integration of Hox and segmentation gene inputs during Drosophila development. Nature *431*, 653–659.

Gene, D. H. H.-s. and Pelham, H. R. B. (1982). A Regulatory Upstream Promoter Element in the Drosophila Hsp 70 Heat-Shock Gene. Cell *30*, 517–528.

Git, A., Dvinge, H., Salmon-Divon, M., Osborne, M., Kutter, C., Hadfield, J., Bertone, P. and Caldas, C. (2010). Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. Rna *16*, 991–1006.

Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R. and Bulyk, M. L. (2013). Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. Cell Reports *3*, 1093–1104.

Gotea, V., Visel, A., Westlund, J. M., Gotea, V., Visel, A., Westlund, J. M., Nobrega, M. A., Pennacchio, L. A. and Ovcharenko, I. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Research *20*, 565–577.

Greer, E. L., Blanco, M. A., Gu, L., Sendinc, E., Liu, J., Aristizábal-Corrales, D., Hsu, C. H., Aravind, L., He, C. and Shi, Y. (2015). DNA methylation on N6-adenine in C. elegans. Cell *161*, 868–878.

Grewal, S. I. S. and Jia, S. (2007). Heterochromatin revisited. Nature reviews. Genetics *8*, 35–46.

Grubach, L., Juhl-Christensen, C., Rethmeier, A., Olesen, L. H., Aggerholm, A., Hokland, P. and Østergaard, M. (2008). Gene expression profiling of Polycomb, Hox and Meis genes in patients with acute myeloid leukaemia. European Journal of Haematology *81*, 112–122.

Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S. B., Buil, A., Ongen, H., Yurovsky, A., Bryois, J., Giger, T., Romano, L., Planchon, A., Falconnet, E., Bielser, D., Gagnebin, M., Padioleau, I., Borel, C., Letourneau, A., Makrythanasis, P., Guipponi, M., Gehrig, C., Antonarakis, S. E. and Dermitzakis, E. T. (2013). Passive and active DNA methylation and the interplay with genetic variation in gene regulation. eLife *2013*, 1–18.

Habibi, E., Brinkman, A. B., Arand, J., Kroeze, L. I., Kerstens, H. H. D., Matarese, F., Lepikhov, K., Gut, M., Brun-Heath, I., Hubner, N. C., Benedetti, R., Altucci, L., Jansen, J. H., Walter, J., Gut, I. G., Marks, H. and Stunnenberg, H. G. (2013). Whole-genome bisulfite sequencing of two distinct interconvertible DNA methylomes of mouse embryonic stem cells. Cell Stem Cell *13*, 360–369.

Heard, E. and Martienssen, R. A. (2014). Transgenerational epigenetic inheritance: Myths and mechanisms. Cell *157*, 95–109.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. and Glass, C. K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Molecular Cell *38*, 576–589.

Hellman, A. (2007). Gene Body Specific Methylation. Science *315*, 1141–1143.

Hertz, G. Z. and Stormo, G. D. (1996). Escherichia coli Promoter Sequences: Analysis and Prediction. Methods in Enzymology *273*, 30–42.

Honig, B. and Rohs, R. (2011). Old droughts in. Nature *470*, 472–473.

Hoogsteen, K. (1963). The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. Acta Crystallographica *16*, 907–916.

Horak, C. E. and Snyder, M. (2002). ChIP-chip: A genomic approach for identifying transcription factor binding sites. Methods in Enzymology *350*, 469–483.

Hu, S., Wan, J., Su, Y., Song, Q., Zeng, Y., Nguyen, H. N., Shin, J., Cox, E., Rho, H. S., Woodard, C., Xia, S., Liu, S., Lyu, H., Ming, G. L., Wade, H., Song, H., Qian, J. and Zhu, H. (2013). DNA methylation presents distinct binding sites for human transcription factors. eLife *2013*, 1–16.

Huang, S.-s. C. and Ecker, J. (2017). Piecing together cis-regulatory networks: Insights from epigenomics studies in plants. Wiley Interdisciplinary Reviews: Systems Biology and Medicine *10*, 1–20.

Isakova, A., Groux, R., Imbeault, M., Rainer, P., Alpern, D., Dainese, R., Ambrosini, G., Trono, D., Bucher, P. and Deplancke, B. (2017). SMiLE-seq identifies binding motifs of single and dimeric transcription factors. Nature Methods *14*, 316–322.

Ito, S., Dalessio, A. C., Taranova, O. V., Hong, K., Sowers, L. C. and Zhang, Y. (2010). Role of tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature *466*, 1129–1133.

Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M. and Brown, P. O. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature *409*, 533–538.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. and Charpentier, E. (2012). A Programmable Dual-RNA Guided. Science *337*, 816–822.

Johnson, D. S., Mortazavi, A. and Myers, R. M. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science *316*, 1497–1503.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpa, M. J., Bonke, M., Palin, K., Talukder, S., Hughes, T. R., Luscombe, N. M., Ukkonen, E. and Taipale, J. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Research *20*, 861–873.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., Palin, K., Vaquerizas, J. M., Vincentelli, R., Luscombe, N. M., Hughes, T. R., Lemaire, P., Ukkonen, E., Kivioja, T. and Taipale, J. (2013). DNA-binding specificities of human transcription factors. Cell *152*, 327–339.

Jolma, A., Yin, Y., Nitta, K. R., Dave, K., Popov, A., Taipale, M., Enge, M., Kivioja, T., Morgunova, E. and Taipale, J. (2015). DNA-dependent formation of transcription factor pairs alters their binding specificity. Nature *527*, 384–388.

Jones, P. A. and Baylin, S. B. (2007). The Epigenomics of Cancer. Cell *128*, 683–692.

Joshi, R., Passner, J. M., Rohs, R., Jain, R., Crickmore, M. a., Jacob, V., Aggarwal, A. K. and Mann, R. S. (2007). Functional specificity of a Hox protein mediated by the recognition of minor groove structure Rohit. Cell *131*, 530–543.

Khan, A., Fornes, O., Stigliani, A., Gheorghe, M., Castro-Mondragon, J. A., Van Der Lee, R., Bessy, A., Chèneby, J., Kulkarni, S. R., Tan, G., Baranasic, D., Arenillas, D. J., Sandelin, A., Vandepoele, K., Lenhard, B., Ballester, B., Wasserman, W. W., Parcy, F. and Mathelier, A. (2018). JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Research *46*, D260–D266.

Kohli, R. M. and Zhang, Y. (2013). TET enzymes, TDG and the dynamics of DNA demethylation. Nature *502*, 472–479.

Kriaucionis, S. and Heintz, N. (2009). The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. Science *324*, 929–930.

Kribelbauer, J. F., Laptenko, O., Chen, S., Martini, G. D., Freed-Pastor, W. A., Prives, C., Mann, R. S. and Bussemaker, H. J. (2017). Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes. Cell Reports *19*, 2383–2395.

Kroon, E., Krosl, J., Thorsteinsdottir, U., Baban, S., Buchberg, A. M. and Sauvageau, G. (1998). Hoxa9 transforms primary bone marrow cells through specific collaboration with Meis1a but not Pbx1b. EMBO Journal *17*, 3714–3725.

Kulis, M. and Esteller, M. (2010). 2 DNA Methylation and Cancer. Advances in Genetics *70*, 27–56.

Laird, P. W. (2010). Principles and challenges of genome-wide DNA methylation analysis. Nature Reviews Genetics *11*, 191–203.

Lane, D., Prentki, P. and Chandler, M. (1992). Use of gel retardation to analyze protein-nucleic acid interactions. Microbiological reviews *56*, 509–528.

Laptenko, O., Beckerman, R., Freulich, E. and Prives, C. (2011). P53 Binding To Nucleosomes Within the P21 Promoter in Vivo Leads To Nucleosome Loss and Transcriptional Activation. Proceedings of the National Academy of Sciences *108*, 10385–10390.

Laptenko, O., Shiff, I., Freed-Pastor, W., Zupnick, A., Mattia, M., Freulich, E., Shamir, I., Kadouri, N., Kahan, T., Manfredi, J., Simon, I. and Prives, C. (2015). The p53 C Terminus Controls Site-Specific DNA Binding and Promotes Structural Changes within the Central DNA Binding Domain. Molecular Cell *57*, 1034–1046.

Lawn, R. M., Adelman, J., Franke, A. E., Houck, C. M., Gross, M. and Goeddel, D. V. (1981). Acids Research Nucleic. Nucleic acids research *9*, 1045–1052.

Lawrence, C. E. and Reilly, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins: Structure, Function, and Bioinformatics *7*, 41–51.

Lazarovici, A., Zhou, T., Shafer, A., Dantas Machado, A. C., Riley, T. R., Sandstrom, R., Sabo, P. J., Lu, Y., Rohs, R., Stamatoyannopoulos, J. A. and Bussemaker, H. J. (2013). Probing DNA shape and methylation state on a genomic scale with DNase I. Proceedings of the National Academy of Sciences *110*, 6376–6381.

Le, D. D., Shimko, T. C., Aditham, A. K., Keys, A. M., Longwell, S. A., Orenstein, Y. and Fordyce, P. M. (2018). Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. Proceedings of the National Academy of Sciences *115*, E3702–E3711.

Lee, C.-Y., Chiu, Y.-C., Wang, L.-B., Kuo, Y.-L., Chuang, E. Y., Lai, L.-C. and Tsai, M.-H. (2013). Common applications of next-generation sequencing technologies in genomic research. Translational Cancer Research *2*, 33–45.

Léveillé, N., Melo, C. A., Rooijers, K., Díaz-Lagares, A., Melo, S. A., Korkmaz, G., Lopes, R., Akbari Moqadam, F., Maia, A. R., Wijchers, P. J., Geeven, G., den Boer, M. L., Kalluri, R., de Laat, W., Esteller, M. and Agami, R. (2015). Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA. Nature communications *6*, 6520.

Levine, H. A. and Nilsen-Hamilton, M. (2007). A mathematical analysis of SELEX. Computational Biology and Chemistry *31*, 11–35.

Lewis, E. B. (1978). A gene complex controlling segmentation in Drosophila. Nature *276*, 565–570.

Li, E., Bestor, T. H. and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. Cell *69*, 915–926.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, J. J. and Herskowitz, I. (1993). Isolation of ORC6, a component of the yeast origin recognition complex by a one-hybrid system. Science *262*, 1870–1874.

Li, Y., yu Chen, C., Kaye, A. M. and Wasserman, W. W. (2015). The identification of cis-regulatory elements: A review from a machine learning perspective. BioSystems *138*, 6–17.

Lifanov, A. P. (2003). Homotypic Regulatory Clusters in Drosophila. Genome Research *13*, 579–588.

Lister, R., Mukamel, E. A., Nery, J. R., Urich, M., Puddifoot, C. A., Johnson, N. D., Lucero, J., Huang, Y., Dwork, A. J., Schultz, M. D., Yu, M., Tonti-Filippini, J., Heyn, H., Hu, S., Wu, J. C., Rao, A., Esteller, M., He, C., Haghighi, F. G., Sejnowski, T. J., Behrens, M. M. and Ecker, J. R. (2013). Global epigenomic reconfiguration during mammalian brain development. Science *341*.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q. M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B. and Ecker, J. R. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Liu, J., Zhu, Y., Luo, G. Z., Wang, X., Yue, Y., Wang, X., Zong, X., Chen, K., Yin, H., Fu, Y., Han, D., Wang, Y., Chen, D. and He, C. (2016). Abundant DNA 6mA methylation during early embryogenesis of zebrafish and pig. Nature Communications *7*, 1–7.

Liu, X. and Clarke, N. D. (2002). Rationalization of gene regulation by a eukaryotic transcription factor: Calculation of regulatory region occupancy from predicted binding affinities. Journal of Molecular Biology *323*, 1–8.

Liu, Y., Zhang, X., Blumenthal, R. M. and Cheng, X. (2013). A Common Mode of Recognition for Methylated CpG. Trends Biochem Sci. *38*, 177–183.

Longobardi, E., Penkov, D., Mateos, D., De Florian, G., Torres, M. and Blasi, F. (2014). Biochemistry of the tale transcription factors PREP, MEIS, and PBX in vertebrates. Developmental Dynamics *243*, 59–75.

Luscombe, N. M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. Nucleic Acids Research *29*, 2860–2874.

Machado, A. C. D., Zhou, T., Rao, S., Goel, P., Rastogi, C., Lazarovici, A., Bussemaker, H. J. and Rohs, R. (2015). Evolving insights on how cytosine methylation affects protein-DNA binding. Briefings in Functional Genomics *14*, 61–73.

Maerkle, S. J. and Quake, S. R. (2010). A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. Science *328*, 1576–1581.

Maniatis, T., Ptashne, M., Backman, K., Kleid, D., Flashman, S., Jeffrey, A., Maurer and Russell; (1975). No Title. Cell *5*, 109–113.

Mann, I. K., Chatterjee, R., Zhao, J., He, X., Weirauch, M. T., Hughes, T. R. and Vinson, C. (2013). CG methylated microarrays identify a novel methylated sequence bound by the CEBPB — ATF4 heterodimer that is active in vivo CG methylated microarrays identify a novel methylated sequence bound by the CEBPB — ATF4 heterodimer that is active in vivo. Genome Research *23*, 988–997.

Mathelier, A., Xin, B., Chiu, T. P., Yang, L., Rohs, R. and Wasserman, W. W. (2016). DNA Shape Features Improve Transcription Factor Binding Site Predictions InVivo. Cell Systems *3*, 278–286.e4.

Meng, X. and Wolfe, S. A. (2006). Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. Nature Protocols *1*, 30–45.

Merabet, S. and Mann, R. S. (2016). To Be Specific or Not: The Critical Relationship Between Hox And TALE Proteins. Trends in Genetics *xx*, 1–14.

Mercer, T. R. and Mattick, J. S. (2013). Understanding the regulatory and transcriptional complexity of the genome through structure. Genome Research *23*, 1081–1088.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T. K., Koche, R. P., Lee, W., Mendenhall, E., O'Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S. and Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560.

Miller, M. (2009). The Importance of Being Flexible: The Case of Basic Region Leucine Zipper Transcriptional Regulators. Curr Protein Pept Sci. *292*, 342–351.

Moorman, C., Sun, L. V., Wang, J., de Wit, E., Talhout, W., Ward, L. D., Greil, F., Lu, X.-J., White, K. P., Bussemaker, H. J. and van Steensel, B. (2006). Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. Proceedings of the National Academy of Sciences *103*, 12027–12032.

Morata, G., Botas, J., Kerridge, S. and Struhl, G. (1983). Homeotic transformations of the abdominal segments of Drosophila caused by breaking or deleting a central portion of the bithorax complex. Journal of embryology and experimental morphology *78*, 319–341.

Mukherjee, S., Berger, M. F., Jona, G., Wang, X. S., Muzzey, D., Snyder, M., Young, R. A. and Bulyk, M. L. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nature Genetics *36*, 1331–1339.

Nikolova, E. N., Kim, E., Wise, A. A., O'Brien, P. J., Andricioaei, I. and Al-Hashimi, H. M. (2011). Transient Hoogsteen base pairs in canonical duplex DNA. Nature *470*, 498–504.

Nikulenkov, F., Spinnler, C., Li, H., Tonelli, C., Shi, Y., Turunen, M., Kivioja, T., Ignatiev, I., Kel, A., Taipale, J. and Selivanova, G. (2012). Insights into p53 transcriptional function via genome- wide chromatin occupancy and gene expression analysis. Cell death and differentiation *19*, 1992–2002.

Nitta, K. R., Jolma, A., Yin, Y., Morgunova, E., Kivioja, T., Akhtar, J., Hens, K., Toivonen, J., Deplancke, B., Furlong, E. E. and Taipale, J. (2015). Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. eLife *2015*, 1–20.

Noro, B., Culi, J., McKay, D. J., Zhang, W. and Mann, R. S. (2006). Distinct functions of homeodomain-containing and homeodomain-less isoforms encoded by homothorax. Genes and Development *20*, 1636–1650.

Noyes, M. B., Christensen, R. G., Wakabayashi, A., Stormo, G. D., Brodsky, M. H. and Wolfe, S. A. (2008). Analysis of Homeodomain Specificities Allows the Family-wide Prediction of Preferred Recognition Sites. Cell *133*, 1277–1289.

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

Pabo, C. and Sauer, R. T. (1984). Protein-Dna Recognition. Ann. Rev. Biochem *53*, 293–321.

Pabo, C. O. and Sauer, R. T. (1992). Transcription Factors: Structural Families and Principles of DNA Recognition. Annual Review of Biochemistry *61*, 1053–1095.

Panning, B. and Jaenisch, R. (1996). DNA hypomethylation can activate Xist expression and silence X-linked genes. Genes and Development *10*, 1991–2002.

Paz, M. F., Fraga, M. F., Avila, S., Guo, M., Pollan, M., Herman, J. G. and Esteller, M. (2003). A systematic profile of DNA methylation in human cancer cell lines. Cancer Research *63*, 1114–1121.

Petty, T. J., Emamzadah, S., Costantino, L., Petkova, I., Stavridi, E. S., Saven, J. G., Vauthey, E. and Halazonetis, T. D. (2011). An induced fit mechanism regulates p53 DNA binding kinetics to confer sequence specificity. The EMBO journal *30*, 2167–76.

Pribnow, D. (1975). Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. Proceedings of the National Academy of Sciences *72*, 784–8.

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F. and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic acids research *44*, W160–W165.

Rao, S., Chiu, T. P., Kribelbauer, J. F., Mann, R. S., Bussemaker, H. J. and Rohs, R. (2018). Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein-DNA binding. Epigenetics and Chromatin *11*, 1–11.

Rastogi, C., Rube, H. T., Kribelbauer, J. F., Crocker, J., Loker, R. E., Martini, G. D., Laptenko, O., Freed-Pastor, W. A., Prives, C., Stern, D. L., Mann, R. S. and Bussemaker, H. J. (2018). Accurate and sensitive quantification of protein-DNA binding affinity. Proceedings of the National Academy of Sciences *115*, E3692–E3701.

Razin, A. and Cedar, H. (1994). DNA methylation and genomic imprinting. Cell *77*, 473–476.

Reiter, F., Wienerroither, S. and Stark, A. (2017). Combinatorial function of transcription factors and cofactors. Current Opinion in Genetics and Development *43*, 73–81.

Rhee, H. S. and Pugh, B. F. (2011). Resource Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. Cell *147*, 1408–1419.

Rieckhof, G. E., Casares, F., Ryoo, H. D., Abu-Shaar, M. and Mann, R. S. (1997). Nuclear translocation of extradenticle requires homothorax, which encodes an extradenticle-related homeodomain protein. Cell *91*, 171–183.

Riley, T. R., Slattery, M., Abe, N., Rastogi, C., Mann, R. S. and Bussemaker, H. J. (2014). SELEX-seq, a method for characterizing the complete repertoire of binding site preferences for transcription factor complexes. Methods Mol Biol. *1196*, 255–278.

Roberts, R. J., Vincze, T., Posfai, J. and Macelis, D. (2015). REBASE-a database for DNA restriction and modification: Enzymes, genes and genomes. Nucleic Acids Research *43*, D298–D299.

Rohs, R., West, S. M., Liu, P. and Honig, B. (2009a). Nuance in the double-helix and its role in protein-DNA recognition. Current Opinion in Structural Biology *19*, 171–177.

Rohs, R., West, S. M., Sosinsky, A., Liu, P., Mann, R. S. and Honig, B. (2009b). The role of DNA shape in protein-DNA recognition. Nature *461*, 1248–1253.

Rosenberg, M. and Court, D. (1979). Regulatory sequences involved in the promotion and termination of RNA transcription. Annual review of genetics *13*, 319–353.

Rube, H. T., Rastogi, C., Kribelbauer, J. F. and Bussemaker, H. J. (2018). A unified approach for quantifying and interpreting DNA shape readout by transcription factors. Molecular Systems Biology *14*, e7902.

Ryoo, H. D. and Mann, R. S. (1999). The control of trunk Hox specificity and activity by extradenticle. Genes and Development *13*, 1704–1716.

Sammons, M. A., Zhu, J., Drake, A. M. and Berger, S. L. (2015). TP53 engagement with the genome occurs in distinct local chromatin environments via pioneer factor activity. Genome Research *25*, 179–188.

Sauer, R. T., Yocum, R. R., Doolittle, R. F., Lewis, M. and Pabo, C. O. (1982). Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. Nature *298*, 447–451.

Schmidl, C., Rendeiro, A. F., Sheffield, N. C. and Bock, C. (2015). ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors. Nature Methods *12*, 963–5.

Schmitz, R. J., Schultz, M. D., Urich, M. A., Nery, J. R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R. B., Chen, H., Schork, N. J. and Ecker, J. R. (2013). Patterns of population epigenomic diversity. Nature *495*, 193–198.

Schneider, T. D., Stormo, G. D., Gold, L. and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. Journal of molecular biology *188*, 415–431.

Seeman, N. C., Rosenberg, J. M. and Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. Proceedings of the National Academy of Sciences *73*, 804–808.

Shen, N., Zhao, J., Schipper, J. L., Zhang, Y., Bepler, T., Leehr, D., Bradley, J., Horton, J., Lapp, H. and Gordan, R. (2018). Divergence in DNA Specificity among Paralogous Transcription Factors Contributes to Their Differential In Vivo Binding. Cell Systems *6*, 470–483.e8.

Shi, D. Q., Ali, I., Tang, J. and Yang, W. C. (2017). New insights into 5hmC DNA modification: Generation, distribution and function. Frontiers in Genetics *8*, 1–11.

Shlyueva, D., Stampfel, G. and Stark, A. (2014). Transcriptional enhancers: From properties to genome-wide predictions. Nature Reviews Genetics *15*, 272–286.

Siggers, T., Duyzend, M. H., Reddy, J., Khan, S. and Bulyk, M. L. (2011). Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. Molecular Systems Biology *7*, 1–14.

Skene, P. J. and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. eLife *6*, 1–35.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H. J. and Mann, R. S. (2011). Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins. Cell *147*, 1270–1282.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A. C., Gordân, R. and Rohs, R. (2014). Absence of a simple code: How transcription factors read the genome. Trends in Biochemical Sciences *39*, 381–399.

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Research *12*, 505–519.

Stein, R., Razin, A. and Cedar, H. (1982). In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse L cells. Proceedings of the National Academy of Sciences of the United States of America *79*, 3418–22.

Steitz, T. a., Ohlendorf, D. H., McKay, D. B., Anderson, W. F. and Matthews, B. W. (1982). Structural similarity in the DNA-binding domains of catabolite gene activator and cro repressor proteins. Proceedings of the National Academy of Sciences of the United States of America *79*, 3097–100.

Stormo, D. and Schneider, T. D. (1982). Use of the 'Perceptron' algorithm to distinguish transational initiation sites in E. coli. Nucleic Acids Research *10*, 2997–3011.

Stormo, G. D. (2000). DNA binding sites: Representation and discovery. Bioinformatics *16*, 16–23.

Stormo, G. D., Zuo, Z. and Chang, Y. K. (2015). Spec-seq: Determining protein-DNA-binding specificity by sequencing. Briefings in Functional Genomics *14*, 30–38.

Suzuki, M., Brenner, S. E., Mark;, G. and Yagi, N. (1995). DNA recognition code of transcription factors. Protein Engineering *8*, 319–328.

Syed, K. S., He, X., Tillo, D., Wang, J., Durell, S. R. and Vinson, C. (2016). 5-Methylcytosine (5mC) and 5-Hydroxymethylcytosine (5hmC) Enhance the DNA Binding of CREB1 to the C/EBP Half-Site Tetranucleotide GCAA. Biochemistry *55*, 6940–6948.

Teytelman, L., Thurtle, D. M., Rine, J. and van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proceedings of the National Academy of Sciences *110*, 18602–18607.

Tribioli, C., Tamanini, F., Patrosso, C., Milanesi, L., Villa, A., Pergolizzi, R., Maestrini, E., Rivella, S., Bione, S. and Mancini, M. (1992). Methylation and sequence analysis around EagI sites: identification of 28 new CpG islands in XQ24-XQ28. Nucleic acids research *20*, 727–33.

Tsai, A., Muthusamy, A. K., Alves, M. R., Lavis, L. D., Singer, R. H., Stern, D. L. and Crocker, J. (2017). Nuclear microenvironments modulate transcription from low-affinity enhancers. eLife *6*, 1–18.

Tuerk, C. and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science (New York, N.Y.) *249*, 505–510.

van der Graaf, A., Wardenaar, R., Neumann, D. A., Taudt, A., Shaw, R. G., Jansen, R. C., Schmitz, R. J., Colomé-Tatché, M. and Johannes, F. (2015). Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. Proceedings of the National Academy of Sciences *112*, 6676–6681.

van Helden, J., André, B. and Collado-Vides, J. (1998). ScienceDirect - Journal of Molecular Biology : Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies1. 4 September 1998 *281*, 827–842.

Vierstra, J. and Stamatoyannopoulos, J. A. (2016). Genomic footprinting. Nature Methods *13*, 213–221.

Walsh, C. P., Chaillet, J. R. and Bestor, T. H. (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation [4]. Nature Genetics *20*, 116–117.

Wang, J., Zhuang, J., Iyer, S., Jie Wang, A., Lin, X., Whitfield, T. W., Greven, M. C., Pierce, B. G., Dong, X., Kundaje, A., Cheng, Y., Rando, O. J., Birney, E., Myers, R. M., Noble, W. S., Snyder, M. and Weng Comments, Z. (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors Repository Citation Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Research *9*, 1798 – 1812.

Wang, Y., Jorda, M., Jones, P. L., Maleszka, R. and Ling, Xu; Robertson, Hugh M.; Mizzen, Craig A.; Peinado, Miguel A., Robinson, G. E. (2006). References and Notes 1. Science *314*, 645–648.

Ward, L. D., Wang, J. and Bussemaker, H. J. (2014). Characterizing a collective and dynamic component of chromatin immunoprecipitation enrichment profiles in yeast. BMC Genomics *15*, 1–16.

Watson, J. and Crick, F. (1953). Molecular structure of nucleic acids. Nature. *171*, 737–8.

Weirauch, M. T., Yang, A., Albu, M., Cote, A. G., Montenegro-Montero, A., Drewe, P., Najafabadi, H. S., Lambert, S. A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M. G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J. S., Govindarajan, S., Shaulsky, G., Walhout, A. J., Bouget, F.-Y., Ratsch, G., Larrondo, L. F., Ecker, J. R. and Hughes, T. R. (2014). Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell *158*, 1431–1443.

Wu, C. (1984). © 1984 Nature Publishing Group. Nature *310*.

Yi, S. V. (2017). Insights into epigenome evolution from animal and plant methylomes. Genome Biology and Evolution *9*, 3189–3201.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P. K., Kivioja, T., Dave, K., Zhong, F., Nitta, K. R., Taipale, M., Popov, A., Ginno, P. A., Domcke, S., Yan, J., Schübeler, D., Vinson, C. and Taipale, J. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*, 1–15.

Zhang, G., Huang, H., Liu, D., Cheng, Y., Liu, X., Zhang, W., Yin, R., Zhang, D., Zhang, P., Liu, J., Li, C., Liu, B., Luo, Y., Zhu, Y., Zhang, N., He, S., He, C., Wang, H. and Chen, D. (2015). N⁶-methyladenine DNA modification in Drosophila. Cell *161*, 893–906.

Zhang, L., Martini, G. D., Tomas Rube, H., Kribelbauer, J. F., Rastogi, C., FitzPatrick, V. D., Houtman, J. C., Bussemaker, H. J. and Pufall, M. A. (2018). SelexGLM differentiates androgen and glucocorticoid receptor DNA-binding preference over an extended binding site. Genome Research *28*, 111–121.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nussbaum, C., Myers, R. M., Brown, M., Li, W. and Shirley, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biology *9*.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A. C., Ghane, T., Di Felice, R. and Rohs, R. (2013). DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. Nucleic acids research *41*, 56–62.

Zhu, H., Wang, G., Qian, J., Sciences, M., Miller, E., Kimmel, S., Cancer, C. and Building, T. S. (2016). Transcription factors as readers and effectors of DNA methylation. Nat Rev Genet. *17*, 551–565.

Appendix A

Constructing Short & Uniform gDNA SELEX-seq Libraries

The majority of *in vitro* methods that probe TF binding to DNA in high-throughput rely on randomly synthesized DNA libraries. Although libraries with a uniform base composition are a good starting point for probing TF-sequence selectivity in an unbiased manner, they generally do not capture genomic sequence biases a TF might encounter *in vivo*. In addition, for libraries with longer randomized regions the proportion of sequences that do not have a genomic match might be large, adding unnecessary complexity to the resulting binding data. One currently used method that utilizes genomic DNA (gDNA) in an *in vitro* TF-binding assay is DAP-seq (DNA Affinity Purification and Sequencing) (Bartlett et al., 2017). The method recovers sites bound in a ChIP-seq assay, as well as sites that could potentially be bound *in vitro* but might be inaccessible in an *in vivo* setting. One drawback of the method is the DNA fragmentation step that is part of the library construction and results in fragment size heterogeneity. Similarly to ChIP-seq, it is therefore limited in the achievable resolution and identification of TF binding preference: Deduction of the true TF-DNA footprint is challenging in the presence of confounding genomic features (such as CpG bias). Another method – gcPBM (genomic context Protein Binding Microarray) (Gordân et al., 2013) – uses custom-designed, genomic sequence microarrays and is therefore limited in the number of sequences tested and requires prior design of the array.

A remedy to both fragment size heterogeneity and restricted sequence space, is the use of type IIS restriction enzymes that cut DNA non-specifically several bases downstream of their recognition site. Combining gDNA fragmentation with subsequent type IIS recognition site ligation and DNA cutting can therefore result in genomic DNA libraries with a uniform size distribution. Such a short and uniform library might not only simplify the identification of true TF-DNA footprints, but it also reflects the true distribution of genomic binding site frequencies across a genome of choice.

Here, in a proof of principle, we demonstrate the feasibility of generating such uniform SELEX-libraries from genomic DNA. Moreover, we use the generated libraries to probe the binding of three Hth^{HM}-Exd-Hox complexes (Hox = Ubx4a, Dfd, Scr) in a gSELEX-seq experiment.

A.1 Generation of Short, Uniform, Genomic SELEX-seq Libraries using Type IIS Restriction Enzymes

To generate short genomic libraries with a uniform size distribution, two different type IIS enzymes were used – MmeI and NmeAIII. By including two enzymes into the library generation, potential cleavage biases can be remedied. Both enzymes recognize a 6bp site and reportedly cut 20bp downstream of the recognition site, creating a 2bp $3'$ overhang. Due to the $3'$ resection that occurs when repairing cut or fragmented DNA, the final blunt-ended fragments are expected to be 18bp long. In brief, genomic DNA extracted from *D. melanogaster* was fragmented (using sonication) to $\sim$ 100-400 bp long fragments, fragmented DNA was end-repaired and dA-tailed to facilitate ligation of DNA adapters. Adapters containing either MmeI or NmeAIII restriction sites were added to the genomic fragments via dA-tail mediated ligation. After the ligation step, DNA fragments were treated with either

MmeI or NmeAIII, followed by another round of end-repair and dA-tailing. Cut DNA fragments were separated from both uncut DNA and the secondary cleavage product (longer DNA fragments downstream of cut site) by gel separation and isolated by gel extraction and column purification. Illumina TruSeq adapters where added to both sides of the purified, adapter-ligated gDNA fragments and the resulting library was amplified using primers matched to the two Illumina adapter sequences. A summary of the individual steps in the protocol is described in Figure A.1 on page 209.

A.2 Analyzing the Library Properties of *D. melanogaster* gDNA Libraries

To increase library diversity, the adapter-ligated gDNA fragments derived from both type IIS enzymes were mixed in equal proportions, creating the final gDNA library. To characterize library composition and quality, the initial library was amplified with the Illumina TruSeq universal and index primers and sequenced using a single-end 75 cycle kit. The resulting fragments were analyzed for i) the correct DNA composition and insert length (type IIS adapter followed by 18-19bp of genomic DNA) and ii) biases in base composition as a function of distance to the cut site.

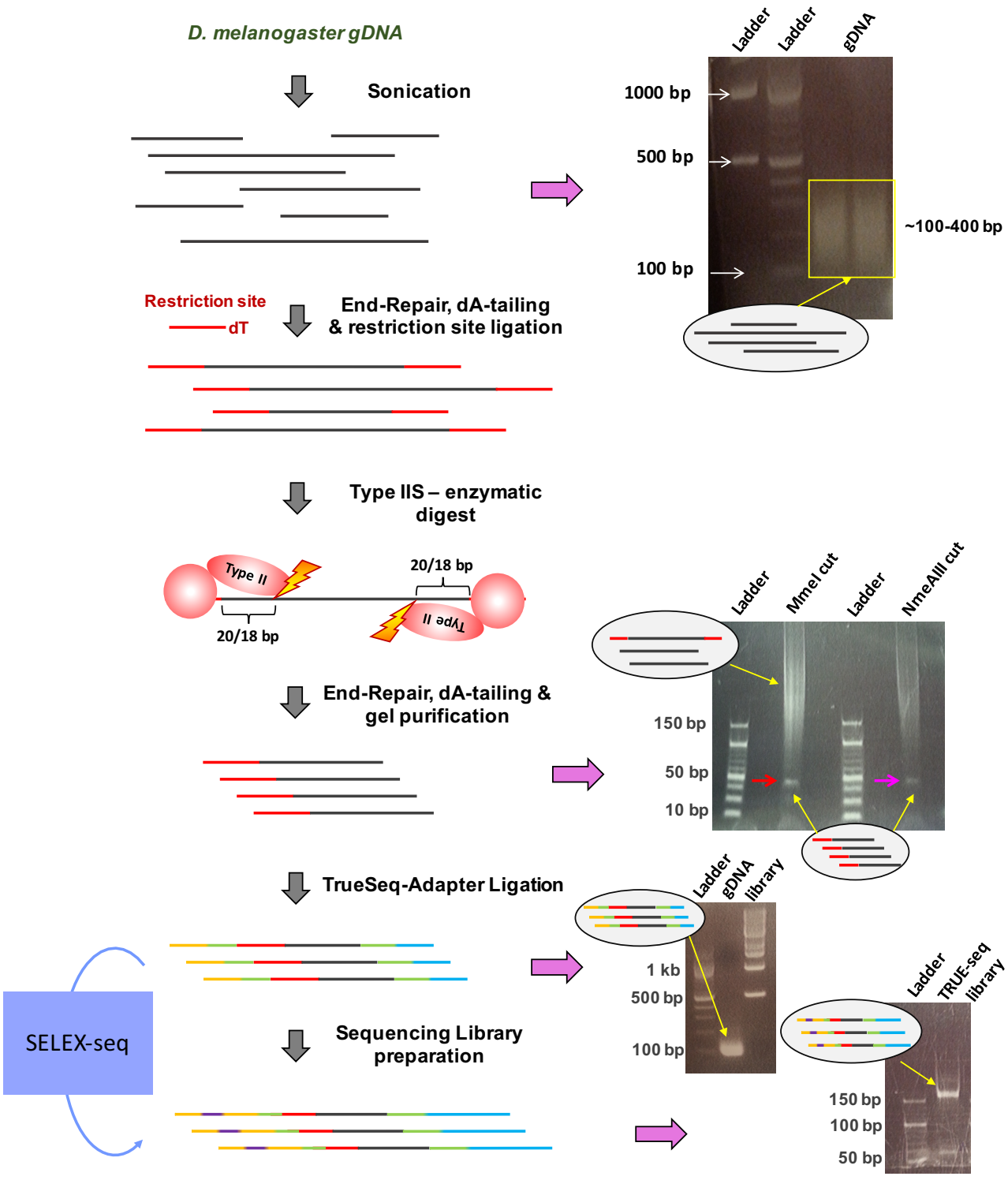


Figure A.1: Overview: Generation of Short, Uniform gDNA Libraries:

The first thing we noticed was a seeming selection for a double-ligation event between two type IIS adapters for both enzymes. Given the uniform length of the sequenced fragments, the double-ligation presumably occurred during the first ligation step and underwent selection during the enzymatic digest. Two adapter sequences were joint in reverse-complement orientation at the DNA end opposing the type IIS site, resulting in the following final library insert: GTCGGACCTAGG-CCTAGGTCCGACT - gDNA; the dash indicates the blunt-end ligation site, blue the dT overhang, red the MmeI recognition site and black the flank used in order to increase the adapter length from 6bp to 12bp to allow efficient cutting. Since the secondary ligation occurred without dA-dT mediation, we hypothesized that this rather infrequent event might have been positively selected in the restriction digest step. Evidence supporting this notion stems from the observation that type IIS enzymes cleave more efficiently when transient homodimerization occurs (see NEB documentation). We suspect that homodimer formation was greatly facilitated by the addition of a second recognition site, thereby resulting in the apparent positive selection for double-ligated adapter sequences. Although it caused initial confusion, the slightly longer library (+12bp) does not impact downstream experiments, making further speculation unnecessary. However, for further library designs it might be advisable to add a second recognition site.

First, we wanted to check the accuracy (in terms of cutting distance from the recognition site) by which the two enzymes cut their DNA substrates. To do so, we split the reads into two groups, those with a 17bp genomic match (resulting from a 20/18bp staggered cleavage event, minus the dT overhang) or those with a 18bp match (resulting from a 21/19bp staggered cleavage, minus the dT overhang) and compared their fraction in the total sequence pool. For MmeI we found that cleavage in $\sim$ 2/3 of sequences occurred at 20/18bp and in $\sim$ 1/3 at 21/19 bp. For NmeAIII the proportions were more equal ($\sim$ 52% to 48% respectively (Figure A.2) A on page 211).

Next we wanted to confirm that cleavage occurred uniformly without any severe sequence biases. We therefore analyzed the position-wise base composition in the genomic flank for

the first 17bp after the dT overhang across all sequences (both enzymes and independent of the cut distance). Except for base position 2 (which disfavored a C base), we found a rather uniform distribution with an overall preference for A/T bases across positions (Figure A.2) B on page 211). The A/T biases is expected as *D. melanogaster* genomes have a A/T skew of roughly 60% .

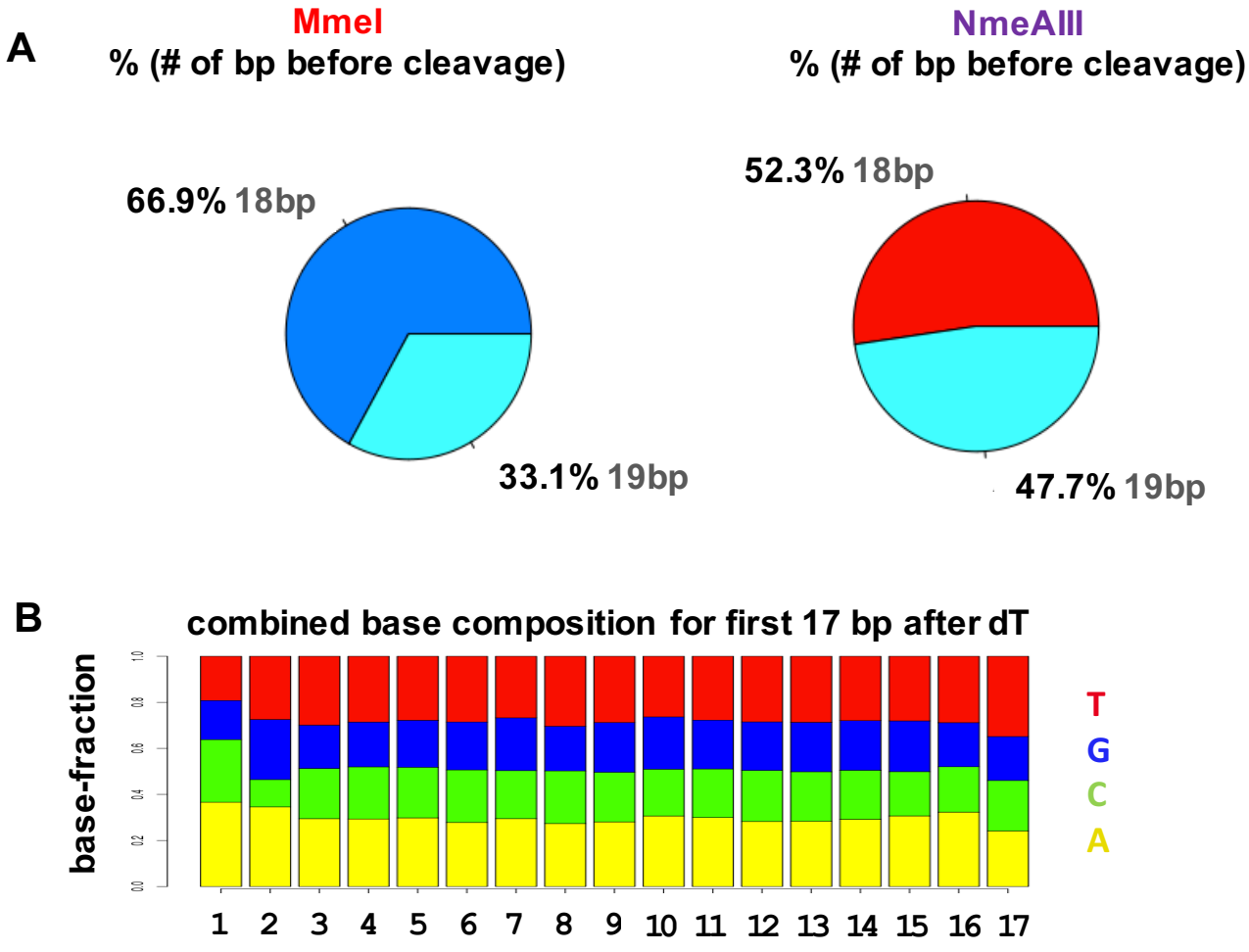


Figure A.2: Library size and sequence composition:

Testing gSELEX-seq on Exd-Hox Complexes

For a proof of principles, SELEX-seq using the genomic library (gSELEX-seq) was performed for three different *D. melanogaster* Hox proteins – ultrabithorax, Ubx4a; deformed, Dfd; sex combs reduced, Scr – in complex with Hth^{HM}-Exd. In total three rounds of selection were performed.

A.3 Experimental Procedures

Primer and Adapter Design and Usage

Restriction site adapters containing the MmeI or NmeAIII site with an upstream AvrII site and a dT overhang were ordered (Fisher Oligo) as reverse complements and annealed in equal proportions. The adapter sequences used for library generation (ligated upstream of the restriction site adapter and downstream of the cleaved genomic fragment) were identical to those used in Illumina TruSeq technology. The Illumina adapters can be purchased as part of NEBs NEBNext Multiplex Oligos for Illumina set, together with the universal and indexing primers required for generation of final sequencing libraries. The amplification primers used to amplify the initial genomic library, were sequence-matched to the Illumina adapter sequences and are describe in table A.1.

Table A.1: Adapter & Primer-sequences used for gSELEX-seq library preparation: All adapter sequences separately ordered are described here. The MmeI and NmeAIII sequences are underlined. f and r stand for forward and reverse strand.

Adapter/Primer		**sequence**	
MmeI Adapter-f	5′	CCTAGGTCCGAC∗T	3′
MmeI Adapter-r	5′	GTCGGACCTAGG	3′
NmeAIII Adapter-f	5′	CCTAGGGCCGAG∗T	3′
NmeAIII Adapter-r	5′	CTCGGCCCTAGG	3′
Illumina Adap.-Uni	5′	ACACTCTTTCCCTACACGAC GCTCTTCCGATC∗T	3′
Illumina Adap.-Index	5′	CTAGCCTTCTCG TGTGCAGACTTGAGGTCAGT	3′
Ampl. Primer-Uni	5′	ACACTCTTTCCCTACACGAC GCTCTTCCGATC	3′
Ampl. Primer-Index	5′	TGACTGGAGTTCAGACGTGT GCTCTTCCGATC T	3′

Enzymes and Reactions

MmeI and NmeAIII were purchased from NEB and their activity was tested with test probes containing the recognition site and a 38bp downstream flank (see table A.2). Restriction digests were performed following standard protocol. Likewise, end-repair and dA tailing were done using NEB's standard modules (for instance a complete end-repair and dA-tailing master mix is included in NEB's NEBNext Ultra DNA Library Prep Kit for Illumina). dA-tailing can also be done by using Klenow Fragment ($3' \rightarrow 5'$ exo-). Standard protocols were followed for each step.

Table A.2: Probes used to test type IIS enzymes: .

Test Probe		**sequence**	
MmeI probe	5'	CCTAGGTCCGAC TAGTGTGCCGTAGCGACGCGATTGCAGACTATGGACCG	3'
NmeAIII robe	5'	CCTAGGGCCGAG TAGTGTGCCGTAGCGACGCGATTGCAGACTATGGACCG	3'

Library preparation, Sequencing and Data Analysis

Libraries were prepared following standard SELEX-seq protocol (see (Kribelbauer et al., 2017; Slattery et al., 2011). Sequencing was performed at the Genome Center at Columbia University using a v2 75 cycle high-output kit on an Illumina NEXTSeq Series desktop sequencer. Data were analyzed using the R package bioconductor.org/packages/SELEX (Riley et al., 2014).

Appendix B

Expanding EpiSELEX-seq: Adding Additional DNA Modifications and Probing More TFs

In recent years, evidence has accumulated that ^{5}mC might not be the only important DNA mark present in eukaryotic genomes. Methylation of cytosines in a non-CpG contexts, their oxidized derivatives, and also methylation of adenines (the dominant mark in prokaryotes) can all be found with varying degrees in a number of different species or tissues (Kriaucionis and Heintz, 2009; Shi et al., 2017; Fu et al., 2015; Greer et al., 2015; Zhang et al., 2015). Although those studies and a few others all point to a functional role for these additional DNA modifications (Liu et al., 2016; Ito et al., 2010), little is known about the mechanisms by which they mediate their potential biological function. One hypothesis is that, similarly to what has been proposed for ^{5}mCs, they might impact binding by transcription factors (TF) and thus fine tune downstream gene expression. To this date, only a few studies have addressed the question how DNA modifications other than ^{5}mC might interact with TFs. In one recent study, using a hemi-hydroxymethylated protein binding microarray, binding of CREB1 to the C/EBP half site was found to be enhanced when hydroxy-methylation was present (Syed et al., 2016). A difficulty when attempting to probe TF binding to those modifications in high throughput is the lack of commercially available enzymes which are capable of modifying large quantities of randomized DNA ligands, and which do not require a specific sequence contexts. In addition, ^{5}mC and in particular ^{5}hmC

can occur in a non-CpG context (Lister et al., 2013), such that using the CpG methylatransferase M.SssI, followed by either chemical oxidation or enzymatic conversion to ^{5}hmC, is not ideal.

Despite those general limitations, it is desirable to design a high-throughput assay that allows probing of TF binding to additional epigenetic marks. We therefore expanded our EpiSELEX-seq method to incorporate two additional modifications – N6-methyldeoxyadenosine (^{6}mA) and 5-hydroxymethylcytosine (^{5}hmC). We tested the expanded EpiSELEX-seq libraries (that now contained up to 3 different DNA modifications, together with unmodified DNA, see Figure B.1 on page 216) using *human* MECP2, different combinations of *human* bZIP, and *D. melanogaster* Exd-Hox proteins.

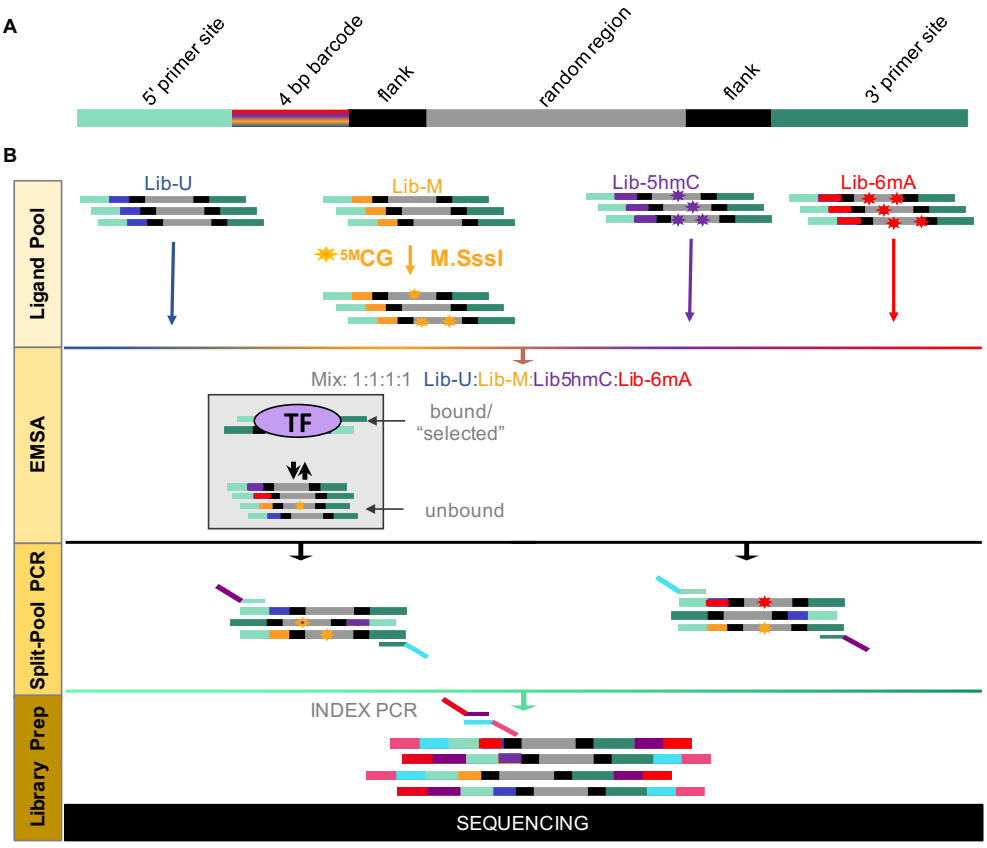


Figure B.1: Schematic expanded EpiSELEX-seq methodology:

B.1 Generation of ^{5}hmC and ^{6}mA Libraries

To generate randomized ^{5}hmC or ^{6}mA libraries that were either completely modified (epigenetic marks on both strands) or hemi-modified (epigenetic mark restricted to one strand), we ordered single-stranded randomized libraries, generated by substituting i) deoxycytidine tripohosphate (dCTP) with deoxy-(^{5}hm)-cytodine triphosphate (d^{5}hmCTP) or ii) deoxyadenosine triphosphate (dATP) with deoxy-(^{6}m)-adenosine triphosphate (d^{6}ATP) in the synthesis step (TriLink Biotechnologies). Completely modified libraries were next generated by double-stranding the single-stranded library template using a mix of deoxynucleotides, where the respective nucleotide harboring the modification in the template was again substituted by its modified counterpart. Together with the unmethylated and CpG methylated libraries described in Chapter 2, we now had a total of six differently modified libraries: unmethylated (Lib-U), ^{5}mCpG methylated (Lib-M), fully hydroxymethylated (Lib-fH), hemi-hydroxymethylated (Lib-hH), fully ^{6}mA-methylated (Lib-fA), and hemi-^{6}mA-methylated (Lib-hA). We mixed these (in equal proportions) the following way: i) UMfH-Library, containing Lib-U, Lib-M and Lib-fH; ii) UfA-Library, containing Lib-U and Lib-fA; and iii) UMhHhA-Library, containing all 4 modifications using hemi-methylated libraries for ^{5}hmC an ^{6}mA (Lib-hH and Lib-hA). A summary of the different combinations of libraries and different TF-complexes is given in Table B.1

Table B.1: Overview expanded EpiSELEX-seq experiments: .

Protein Complex	**Library**	**species**
MECP2 (Methyl-Binding-Domain-MBD)	UMfH-Library	*human*
C/EBPβ homodimer	UMfH-Library	*human*
Pbx-HoxA5	UMfH-Library	*human*
Exd-Dfd	UfA-Library	*D. melanogaster*
Exd-Ubx4a	UfA-Library	*D. melanogaster*
Pbx-HoxA5	UMhHhA-Library	*human*
C/EBPγ homodimer	UMfH-Library	*human*
ATF5 homodimer	UMfH-Library	*human*
ATF4/C/EBPγ heterodimer	UMfH-Library	*human*
ATF5/C/EBPγ heterodimer	UMfH-Library	*human*
ATF5/C/EBPβ heterodimer	UMfH-Library	*human*
MECP2 (full-length)	UM-Library	*human*

B.2 Probing *D. melanogaster* Exd-Hox for ^{6}mA Sensitivity

Since ^{6}mA has recently been detected in *D. melanogaster* (Zhang et al., 2015), we first wanted to test whether ^{6}mA marks have an impact on TF binding. We therefore performed EpiSELEX-seq using the UfA-Library on Exd-Dfd or Exd-Ubx4a. To test the global impact on binding, we compared the fraction of reads mapping to either Lib-fA or Lib-U in R0 (not selected by TF, but otherwise treated like selected libraries) and R1 (after one round of selection) for Exd-Ubx4A (Figure B.2 A on page 219). While the proportion of reads from each library are equal in R0, they were not for R1, with reads from Lib-fA reduced

to ∼13% . Only one sequence had a count >100, suggesting that the ^{6}mA mark abolished all binding by Exd-Ubx4a (similiar for Exd-Dfd, data not shown). This finding is perhaps less surprising, given that every A/T base pair was modified and Exd-Hox TF complexes preferably bind to AT-rich sequences.

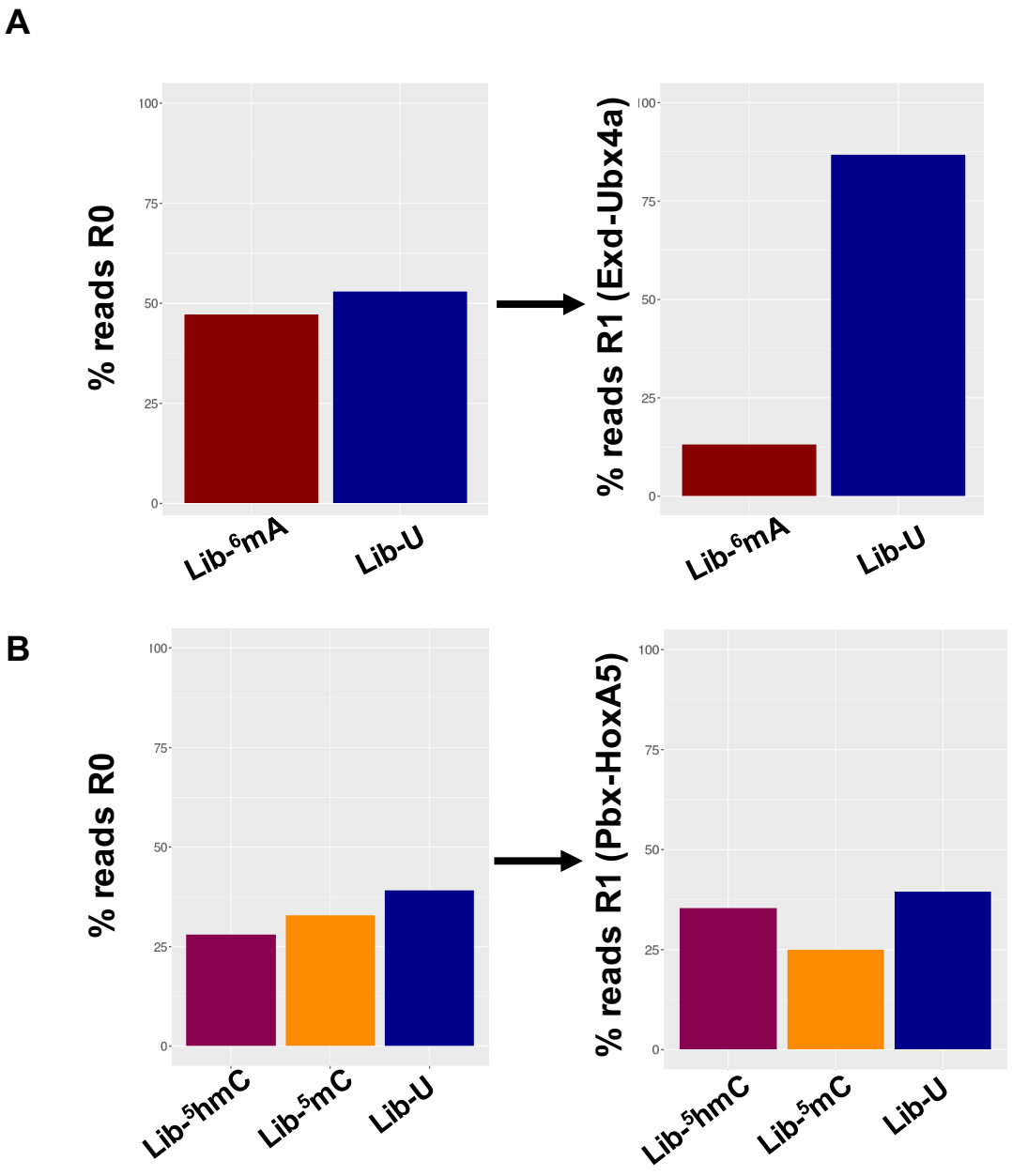


Figure B.2: Read distributions for ^{6}mA and ^{5}hmC libraries:

B.3 Probing human MECP2 and bZIPs for 5hmC Sensitivity

Like for Lib-fA, we also wanted to assess the impact of ^{5}hmC on TF binding. Analysis was carried out for the following *human* TFs: i) the methyl-binding-domain (MBD) of MECP2, ii) Pbx-HoxA5 heterodimers, and iii) C/EBPβ homodimers. To first make sure that the R0 covers all three libraries used in the UMfH-Library equally well, we first computed the proportion of reads from each library. Next we chose Pbx-HoxA5 (a motif with only a couple C/G base preferences) and compared the proportions of reads after one round of selection (Figure B.2 B on page 219). In contrast to what we observed for ^{6}mA, ^{5}hmC seemed to be benefical for binding, as demonstrated by the increase in the proportion of reads coming from Lib-fH. However, the comparison with ^{6}mA presumably does not represent a general trend, as the severe loss of binding by Exd-Hox for DNA ligands containing ^{6}mA might be the result of the large number of modified bases in the AT-rich motif.

Next we wanted to see whether the ^{5}hmC mark results in preferred binding to a distinct subset of sequences. To this end, we computed hexamer (MECP2 MBD) and decamer (C/EBPβ) enrichment tables for all three libraries present in the UMfH-Library and generated pairwise enrichment plots comparing Lib-U to either Lib-M or Lib-fH (see Figure B.3 A,B on page 221). For the MECP2 MBD, we found that hexamer sequences containing a CpG where preferentially bound when methylated. However, the preference for ^{5}hmC appeared to be following a different set of rules, with only a subset of CpG containing sequences being preferred. In addition, a distinct set of sequences in a non-CpG context stood out in terms of preferential binding upon ^{5}hmC modification among the entire sequence pool. For C/EBPβ homodimers, we observed no differential binding between Lib-U and Lib-M. However, a specific subset of sequences exhibited either increased or decreased binding upon ^{5}hmC methylation. Although preliminary, these findings suggest that TF binding preferences for other epigenetic marks are highly sequence specific and do not necessarily correlate

with the effect of methylation in a CpG context.

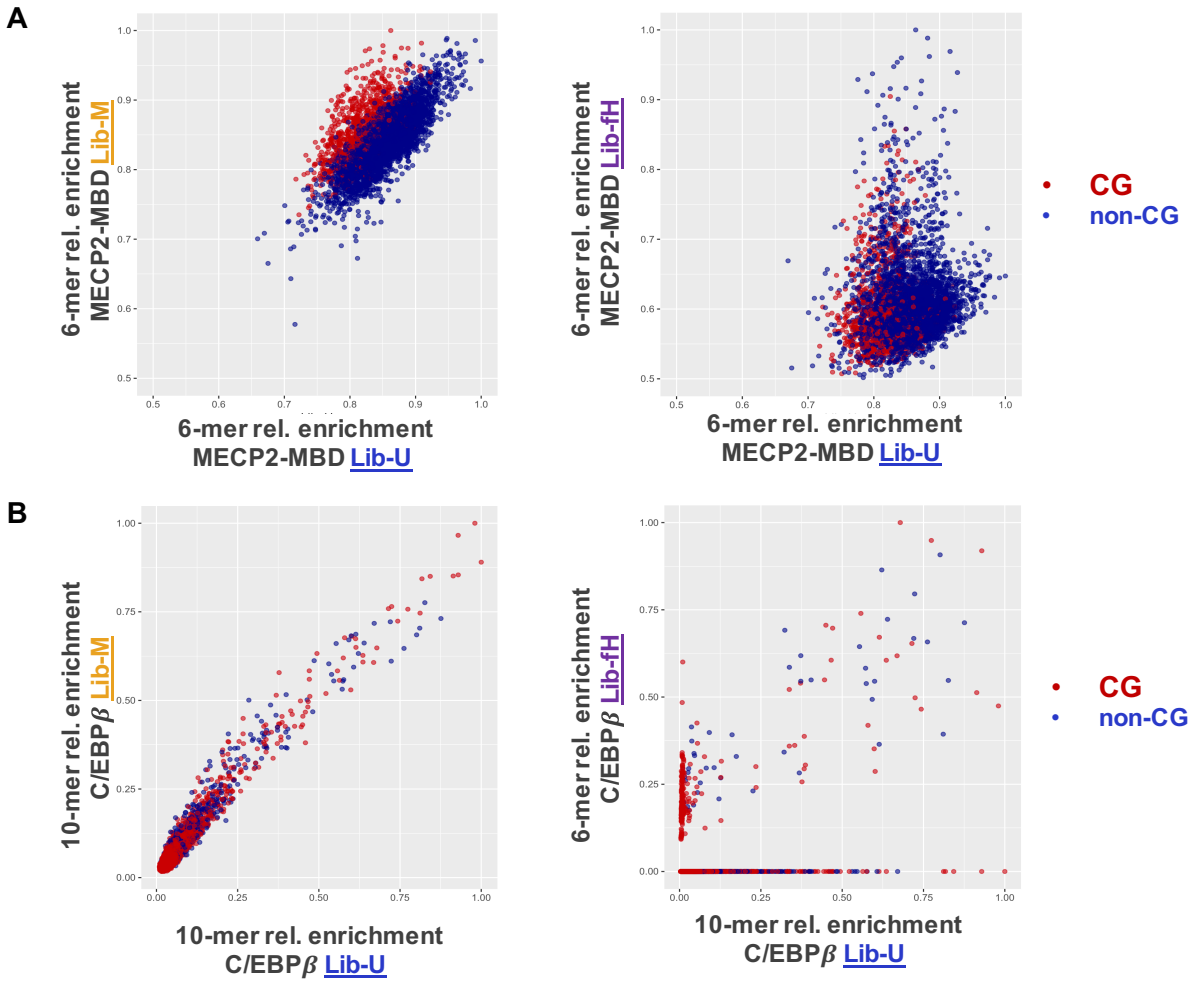


Figure B.3: Binding preference for MECP2 and C/EBPβ for ^{5}mC and ^{5}hmC:

B.4 Probing bZIP Homo- and Heterodimeric Complexes for ^{5}mCpG, ^{5}hmC, and ^{6}mA Sensitivity

Since the fully modified Lib-5hmC and Lib-6mA libraries are not reflective of an *in vivo* setting, where such modifications occur at much lower frequencies, we restricted the next set of experiments to the hemi-methylated libraries. To probe the entire modification space, we

used the UMhHhA-Library (including unmethylated and three modified libraries) and carried out SELEX-seq for different combinations of the bZIP proteins ATF4, ATF5, C/EBPβ, and C/EBPγ. We first assessed whether, and to what extent, the individual modifications might impact binding of the different homo- and heterodimeric complexes of ATF5 and C/EBPγ by comparing 10-mer enrichments between the different libraries (Figure B.4 on page 223). To our surprise, each modification impacted TF binding in a dimerization- and sequence-dependent manner. Moreover, sequences that contained a CpG did not necessarily engage in the same type of binding behavior when the CpG was either methylated or hydroxymethylated (compare first two columns of Figure B.4 on page 223, 10-mers are colored according to whether they contain a CpG or not).

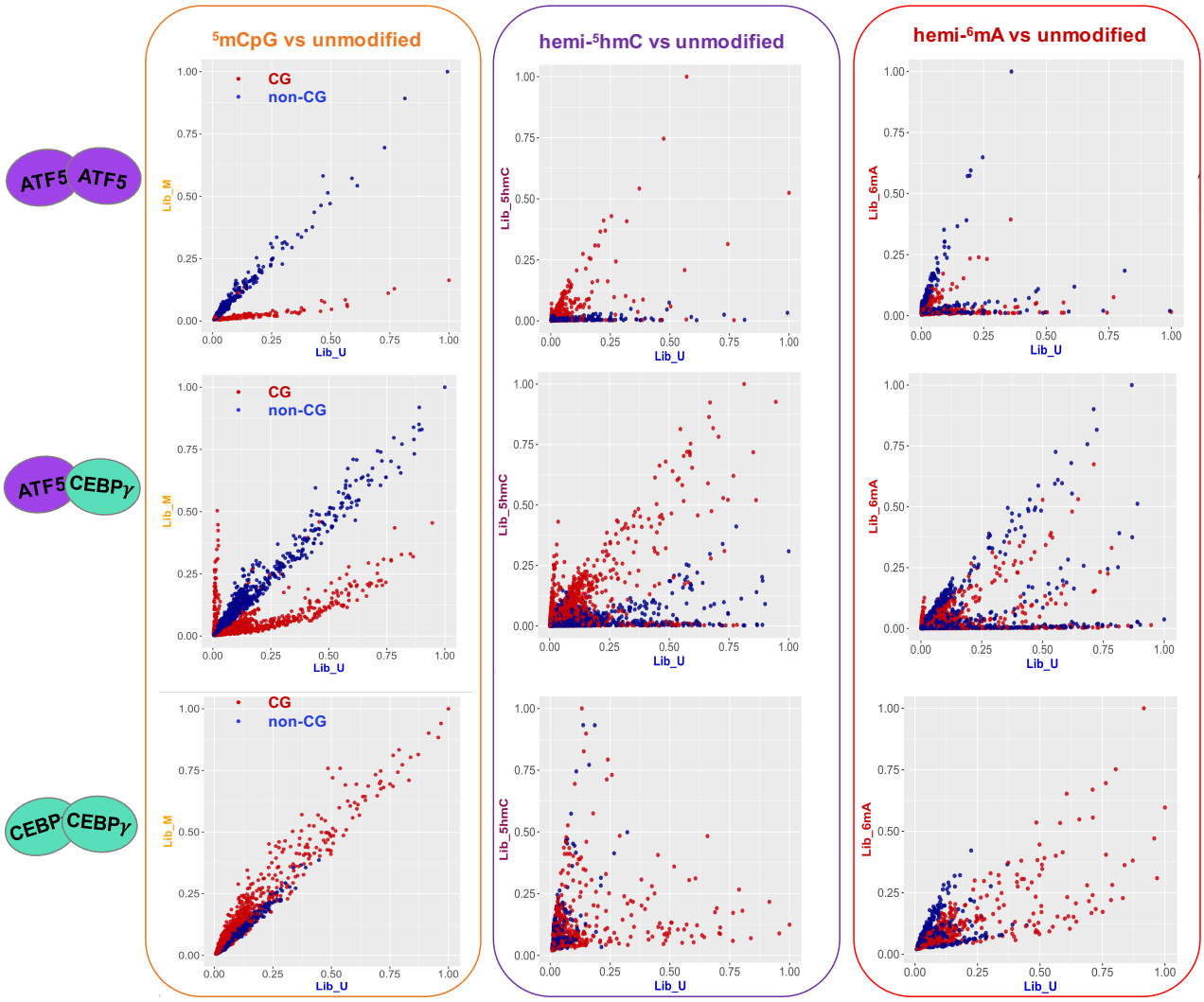


Figure B.4: Binding preference for bZIP ATF5 and C/EBPγ complexes for ^{5}mCpG, ^{5}hmC and ^{6}mA: X-axis represents the relative 10-mer sequence enrichment in Lib-U (unmethyalted). Y-axes represent the 10-mer sequence enrichments in the three different modified libraries (Lib-M, first column; Lib-hH, second column; Lib-hA, third column). Complex compositions are indicated for each row by protein schematics and 10-mers are colored according to whether they contain a CpG or not, blue = non-CpG, red = CpG-containing.

B.5 Heterodimerization Between C/EBPγ and ATF(4/5) Induces Latent Methylation Sensitivity

To our surprise, the heterodimeric complex formed by ATF5 and C/EBPγ displayed a strong binding preference upon CpG methylation for a subset of sequences that was not present for either the ATF5 or the C/EBPγ homodimer (see first column of Figure B.4 on page 223). To more thoroughly analyze this pattern, we compared the effect of CpG methylation for the four homodimeric complexes of ATF4, ATF5, C/EBPβ, and C/EBPγ (Figure B.5 on page 225) to that for the heterodimeric combinations resulting from those four TFs (Figure B.6 on page 226). The latent methylation sensitivity appeared to be induced specifically by C/EBPγ, but not C/EBPβ (right half of Figure B.6 on page 226), and was independent of the ATF-type TF used (both ATF4 and ATF5 showed the same increase in binding upon CpG methylation).

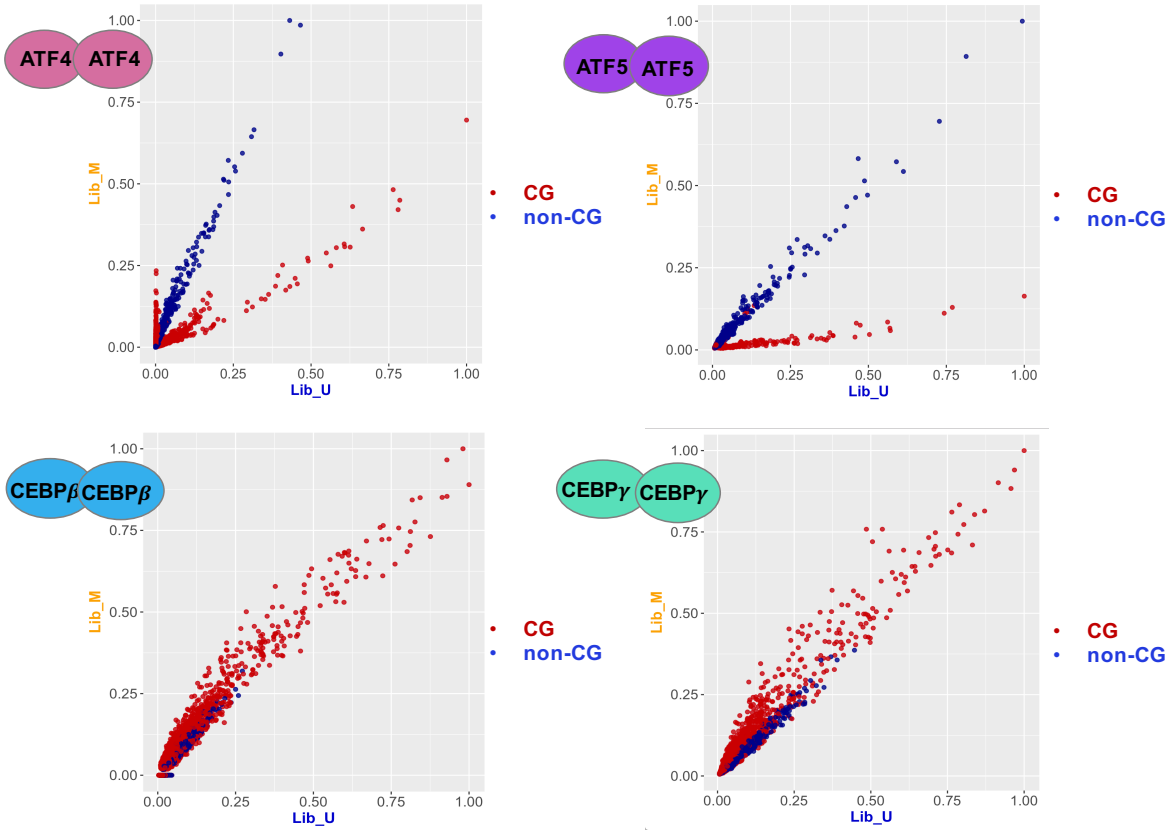


Figure B.5: Binding preferences of ATF- and C/EBP- homodimeric complexes to ^{5}mCpG: X-axis represents the relative 10-mer sequence enrichment in Lib-U (unmethyalted) and y-axis in Lib-M (^{5}mCpG).Complex composition is indicated for each cell by protein schematic.

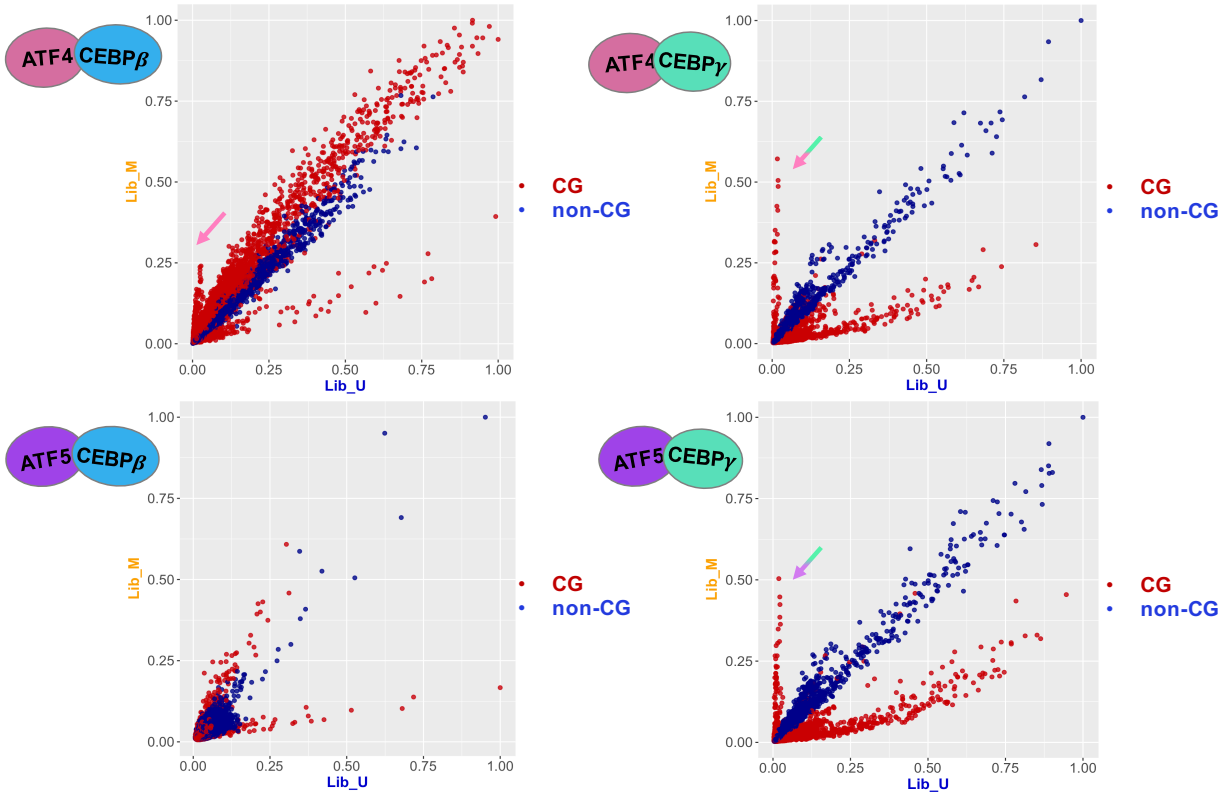


Figure B.6: Binding preferences of ATF- and C/EBP- heterodimeric complexes to ^{5}mCpG: X-axis represents the relative 10-mer sequence enrichment in Lib-U (unmethyalted) and y-axis in Lib-M (^{5}mCpG).Complex composition is indicated for each cell by protein schematic.

To rule out that the observed methylation specificity might be an artifact of homodimeric preferences (both homo- and heterodimers can form in a binding assay and their gel migration pattern is nearly identical), we next computed the methylation effect (the ratio ^{5}mCpG/ CpG) for each 10-mer sequence for ATF4 or ATF5 homodimers and ATF4/C/EBPγ or ATF5/C/EBPγ heterodimers. By plotting the homodimeric ^{5}mCpG/CpG ratio against heterodimeric ^{5}mCpG/ CpG we can identify the specifc subset of sequences impacted by either complex (Figure B.7 on page 227). We find that upon heterodimer formation of either ATF4 or ATF5 with C/EBPγ a subset of sequences is bound up to 50-fold more strongly when methylated. In addition, ATF4 homodimers prefer a non-overlapping set of methylated sequences, which was already reported in Chapter 2. ATF5, by contrast, does

not have a strong methylation preference when bound as a homodimer.

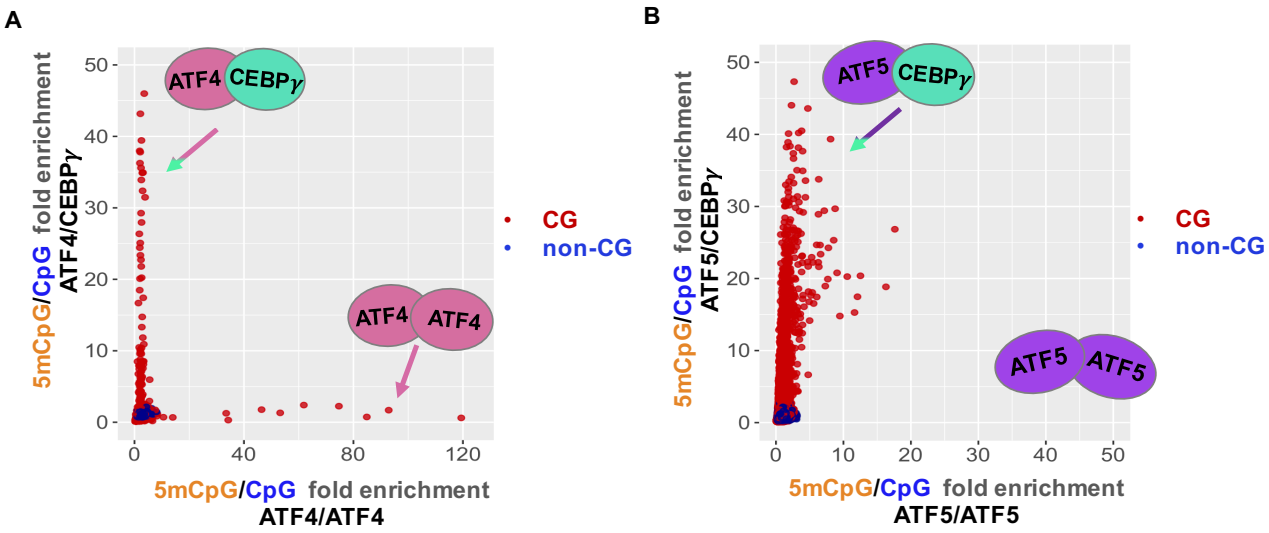


Figure B.7: Teasing apart ^{5}mCpG preferences of ATF-homodimers versus ATFx/CEBPγ heterodimers:

In summary, our findings suggest that epigenetic modifications are capable of modulating TF binding in a highly specific manner that not only depends on the identity of the modification mark itself, but also the sequence context. Moreover, we have identified "latent" methylation sensitivity between ATF-type TFs when bound cooperatively with C/EBPγ. The fact that C/EBPβ does not engage with ATFs in a similar manner underscores the importance of cooperative binding in specifying which sites are bound in a given cellular context. In this particular case, heterodimer formation might target ^{5}mCpG sites that would neither be recognized by the individual homodimeric complexes nor when the site was left unmethylated.

B.6 Experimental Procedures

Library Design and EpiSELEX-seq Experiments

The design of the additional libraries followed the same design as the experimental procedures described in Chapter 2 (Experimental Procedures). The table below contains the complete library sequences. The libraries were ordered using N_{16}= dATP, dGTP, dTTP, d^5hmCTP for Lib-^{5}hmC and N_{16}= dCTP, dGTP, dTTP, d^6mATP for Lib-6mA.

Table B.2: Overview of expanded EpiSELEX-seq experiments: Sequences in red indicate the library-specific barcode.

theadLibrary Name		**Sequence**	
Lib-^{5}hmC	5′	GGTAGTGGAGGCAGTCCTGG-N_{16} -CCAGGGAGGTGGAGTAGG	3′
Lib-6mA	5′	GGTAGTGGAGGAGTGCCTGG-N_{16} -CCAGGGAGGTGGAGTAGG	3′

Double-stranding of Single-Stranded Modified Oligonucleotides

Doublestranding of single-stranded Libraries was done using Klenow Polymerase (Thermo Fisher) at 37°C for 30 min and using either a standard deoxynucleotide mix (hemi-methylated libraries) or a mix with dATP, dGTP, dTTP, d^5hmCTP for Lib-^{5}hmC, or a CTP, dGTP, dTTP, d^6mATP for Lib-6mA. Efficient incorporation of modified deoxynucleotides was assessed using native polyacrylamide electrophoresis and comparison of the single-stranded template to the double-stranded product.

B.7 Acknowledgement

The work on ATF5 and C/EBPγ was the result of a collaboration with the Lomvardas lab, who provided the purified, recombinant proteins. We would like to thank Jerome Kahiapo and Dr. Stavros Lomvardas for their advice and sharing their insights on the potential methylation sensitivity of ATF5 and C/EBPγ, which initiated this collaboration.