# Dissecting the binding mechanisms of transcription factors to DNA using a statistical thermodynamics framework.

Patrick C.N. Martin

A thesis submitted for the degree of Doctor of Philosophy in Biological Sciences

School of Biological Sciences

University of Essex

May 2020

# ABSTRACT

At the heart of gene regulation are Transcription Factors (TFs), proteins which bind to DNA in a sequence specific manner and drive the activation or repression of genes. Statistical thermodynamics has shown to be a promising avenue to describe the binding mechanisms of TFs. Here, I present ChIPanalyser, an R/Bioconductor package that models and predicts binding of TFs to DNA using a statistical thermodynamic framework. First, I show that goodness of fit metrics are an important consideration in TF binding predictions as well as demonstrate ChIPanalyser's high performance compared to other tools and frameworks. Then, I focused on investigating the binding mechanisms of three TFs that are known architectural proteins CTCF, BEAF-32 and su(Hw) in three Drosophila cell lines (BG3, Kc167 and S2). I demonstrate that architectural proteins show varying affinities towards DNA accessibility and that protein abundance plays a lesser role in their binding. While BEAF-32 binds in open chromatin, CTCF and su(Hw) showed increased binding is less accessible DNA. Furthermore, the model was able to recover binding preferences of three Hox TFs with respect to DNA accessibility. However, DNA accessibility showed some limitations to describe the full scope of TF binding affinities. I developed a genetic algorithm to investigate the binding affinity of the aforementioned TFs with respect to chromatin states. The improved model recovered chromatin state affinities and showed a more nuanced picture of TF binding. Finally, I examined the binding mechanisms of Su(H). The model was able to recover known binding mechanisms with respect to both chromatin state affinity and TF abundance. Overall, ChIPanalyser provides accurate TF binding predictions as well as insights into the mechanisms of TF binding.

# ACKNOWLEDGMENTS

All aboard! Thomas the Thank engine is about to leave thesis station.

First stop, Dr. Nicolae Radu Zabet. I would like to thank Radu for his supervision and help throughout this PhD. I appreciate his guidance, good-will and kindness through out my PhD.

Next stop, the OG gang: Tyler, Mohab, Louis and Stuart. Tyler, thank you for the laughs, the support , the R help and for the Brisket. Mohab, I am sorry that I sold myself out for a piece of brisket but I greatly value your friendship and appreciate your constant support even after you left. Louis, Thank you for being you. Every day I got to gaze upon your face as my collection of pictures of you is disturbingly high. Stuart, thank you for all your help through out the years and in a way making many of our shenanigans possible.

D & D station. Thank you Myles for introducing me to D & D. This might not be academically related but from a personal level, I will forever be grateful. Thank you Sam and Tyler for also being part of this Journey. Along those lines, I would like to thank the D & D party and friends Jareth, Lena, Istvan, Howard, Jess and Drew for taking part in this glorious activity that is table top gaming.

I would also like to thank the genomics group for their help, support and feedback over the years. I would especially like to thank Dr Antonio Marco, my board member, for his help and useful feedback.

I would also like to thank my parents for their support through out the years. They put up with my shenanigans and helped me get to where I am now.

Finally, last but most certainly not least, I would like to thank Alma for being part of my life.

# CONTENTS

II MECHANISM OF TRANSCRIPTION FACTOR BINDING

III CONCLUDING REMARKS

# LIST OF FIGURES

# LIST OF TABLES

# Part I

# FOUNDATIONS

# 1

# THE BIOLOGY OF GENE REGULATION

> Problems worthy of attack prove their worth by fighting back.
>
> Paul Erdos

## IN THE BEGINNING, THERE WERE PEAS

Humanity has understood the concept of inheritance for thousands of years. Our ancestors used their empirical knowledge to improve crops via selective breeding. One of my favourite example is Teosinte. Teosinte is a tall grass native to Mexico. After thousands of years of selective breeding, this tall tough grass became a staple in your summer BBQs: maize. Despite the great insight our ancestors had, it was not until the mid 1800's that a formalisation of inheritance was proposed by Gregor Mendel. Mendel was an Austrian (former Austrian Empire - now Czech Republic) friar born in 1822. Between 1856 and 1863, Mendel studied trait inheritance in English peas. After 7 years of research, Mendel gave us the concepts of dominant and recessive traits. Unfortunately, Mendel's work was forgotten for many years and only uncovered years after his death when others claimed to have reached similar conclusions. At the time, a "gene" was considered as the basic unit of inheritance. Interestingly, the origin of the word *gene* and *genetics* are attributed to one of Mendel's contemporaries. None other than biology superstar Charles Darwin coined the term *pangene* from the greek *pan* ("all") and *genesis* ("birth") or *genos* ("origin"). Darwin's definition described a general character of inheritance. Later, Wilhelm Johannsen and William Bateson introduced the term *Gene* and *Genetics* as the basic entities of inheritance. And thus, the field of genetics was born.

## FROM INHERITANCE TO DNA

Over the years, the very definition of genes evolved (no pun intended). Starting from the basic unit of inheritance, genes became a *linearly organised entity*. By studying the segregation of mutations in *Drosophila*, Morgan and his students were able to establish the mechanisms of genetic inheritance. Genetic linkage and physical location on chromosomes was later established by Barbara McClintock in 1929. By the 1940s, genes had become the blue print for a protein or "one gene, one enzyme" as described by Beadle and Tantum in 1941. Interestingly, the next decade brought us the certainty that somehow genes were a physical molecules and that mutations in this physical molecules were responsible for trait variation. It is only in 1953 that DNA was discovered by Watson and Crick (and Franklin). This newly discovered molecule explained how genetic information was stored and transmitted to the next generations.

In the 1940s and 1950s, Barbara McClintock discovered the action of jumping genes in Maize. She postulated that these genes were responsible for turning other genes on and off. This was the first hint towards gene expression and genetic regulation. Unfortunately, as it was often the case that the time, her work was received with scepticisms. It was only a decade later that the concept of gene expression was accepted after Jacobs and Monods work on the lac operon in *Escherichia coli*.

## UNDERSTANDING GENE EXPRESSION

The precise regulation and expression of genes is key to most if not all cellular processes. Basic cellular functionalities such as cell cycle or response to the environment rely on on the correct expression of genes [Rowicka et al., 2007, Passegué et al., 2005, Whitfield et al., 2002, Abe et al., 2005, Lee et al., 2014b]. Embryonic development in *Drosophila* requires the precise expression of developmental genes ( HOX genes) in order to elicit the correct longitudinal patterning[Alexander et al., 2009, Mallo and Alonso, 2013]. Correctly expressed, these genes will lead to limb formation. The *Drosophila* embryo will grow from egg to fully formed organism, ready to live the exciting life of a fly. Misregulation of developmental genes often leads to some unwanted consequences. In certain case, this would lead to ectopic expression of certain tissues [Halder et al., 1995, Schneuwly et al., 1987, Phelps and Brand, 1998, Akiyama-Oda et al., 1998] Disrupted gene expression and regulation patterns is also often symptomatic of disease [Honrado et al., 2006, Hernandez et al., 2000, Liang and Pardee, 2003, Sekar et al., 2015]. Understanding when, why and how genes are expressed is potentially one of the most important questions in modern molecular biology. But what exactly drives the expression of genes?

## TRANSCRIPTION FACTORS ARE AT THE HEART OF GENE EXPRESSION

Transcription factors (TF) are proteins involved in the transcription of DNA to RNA. These proteins both initiate and regulate the expression of genes [Latchman, 2001, Lambert et al., 2018]. Generally speaking, TFs bind to DNA in a sequence specific manner [Ptashne and Gann, 1997, Spitz and Furlong, 2012]. This is possible thanks to their DNA binding domain. TFs can be categorised based on their specific molecular function. General Transcription Factors (GTF) bind to DNA in promoter regions close to the transcriptional start site. GTFs are directly implicated in the recruitment of the transcriptional pre-initiation complex (TPIC). This complex correctly places the RNA polymerase on the transcription start site and prepares DNA for transcription. Other TFs bind to regulatory sequences located either in enhancer regions or promoter regions. Their binding can either activate or repress gene transcription. Some TFs , called Pioneer Transcription Factors are responsible for "priming" transcription by binding into closed chromatin [Aguilar-Arnal and Sassone-Corsi, 2015, Voss et al., 2011]. Pioneer TFs can bind into closed chromatin and open chromatin either through the recruitment of ATP- dependant chromatin remodellers or by direct competition with nucleosomes [Soufi et al., 2015, Mayran et al., 2018, Donaghey et al., 2018]. However, the term pioneer does not only refer to their ability to bind into closed chromatin but also to their key role into priming transcriptional regulation. While chromatin remodellers may be recruited by other TFs to increase DNA accessibility, pioneer TFs are required for a specific transcriptional event to occur [Zaret and Carroll, 2011]. They are actively required in the transcriptional regulation of a given pathway. In certain instances, pioneer TFs would bind to DNA in order to prime a gene for later activation [Iwafuchi-Doi, 2019]. By reducing the number of steps prior to transcriptional activation, pioneer TFs ensure a fast response time when the final regulators come into play. This denotes the double role of pioneer TFs: priming DNA for transcription and actively participating in regulatory pathways.

Finally, it is also important to consider other DNA binding proteins that do not directly participate in transcription but rather influence and modulate DNA confirmation [Beagan et al., 2016, Hansen et al., 2019]. CCCTC-binding factor (CTCF) is a highly conserved zinc finger protein that bind to DNA and can indirectly control gene expression by enabling or inhibiting the communication between enhancers and their target promoter. CTCF also plays a role at a larger scale by stabilising larger chromatin structures [Kim et al., 2015].

Understanding when and where TFs bind to DNA is key to understanding gene expression. *In vivo* experiments such as ChIP-seq (chromatin immunoprecipitation followed by sequencing) or ChIP-chip (chromatin immunoprecipitation on tilling arrays) have become the gold standard or determining TF binding events at a genome wide scale scale [Solomon et al., 1988, Park, 2009]. Simply put, ChIP-seq relies on anti-body recognition of TF-DNA complexes. The resulting DNA is then subject to Next Generation sequencing (NGS). In the past decade, there has been an ongoing effort to map the binding of TFs to DNA in various organisms. The ENCODE and modENCODE project provide ChIP-seq ( or ChIP-chip) experiments of numerous TFs [ENCODE Project Consortium, 2012, modENCODE Consortium et al., 2010, Landt et al., 2012a, Davis et al., 2018].

Despite our ability to determine TF binding events, we still lack a complete understanding of the mechanisms driving TF binding. The following sections will describe the various factors thought to impact TF binding.

### *DNA sequence and shape*

TFs read genomic information in two fundamental ways. The first aspect is the DNA sequence itself [Ptashne and Gann, 1997, Spitz and Furlong, 2012, Slattery et al., 2014]. The amino acid sequence in the DNA binding domain will a create physical interaction with the nucleotides present in DNA. The nature of the interactions consists of simple hydrogen bonds, salt bridges or hydrophobic interactions. TF binding sites (TFBS) are the result of minimising binding energy between TF's and DNA sequence. The second aspect is DNA shape [Abe et al., 2015, Inukai et al., 2017]. Structural features such as DNA bending or groove width are also recognised by certain TFs.

Experimentally, specific binding motifs can be determined by using methods such as protein binding microarrays (PBM),BunDLE-seq (binding to designed library, extracting, and sequencing) or (SELEX-seq (systematic evolution of ligands by exponential enrichment followed by sequencing)(reviewed by [Lai et al., 2018]). Although high-throughput methods, these methods are *in vitro* assays . In many cases, biological context strongly impacts TFBSs. Furhtermore, these methods yield binding motifs between 8 to 20 bp in length. Given the size of genomes, there is a high likelihood of these sequences appearing at a higher rate than the TF binding events described by ChIP experiments.

### *DNA and friends*

In the context of the nucleus, DNA rarely comes alone as a naked strand of nucleotides. DNA (in 147 bp sequences) wraps itself around protein complexes known as nucleosomes each forming of an octameric complex of histones. At a higher level, DNA and nucleosomes can form tightly compacted DNA regions. Dense chromatin forms a barrier for the Transcriptional machinery. Furthermore, most TFs are unable to bind in tightly compacted DNA and prefer nucleosome depleted regions [Li et al., 2011, Chereji et al., 2016, Lamparter et al., 2017, Zhu et al., 2018, Klemm et al., 2019]. Chromatin accessibility could be defined as to which degree can nuclear proteins or macromolecules bind to naked DNA.

Large-scale cis-element studies show that the majority of TF activity is found in Nucleosome depleted regions. Interestingly, in *Homo sapiens*, nucleosomes tend to cluster around regulatory elements [Tillo et al., 2010]. DNA accessibility plays a key role in limiting the number of potential binding motifs available for TF binding. DNA accessibility can be assessed experimentally thanks to recent method such as DNase-seq or ATAC-seq. DNase-seq relies on mapping DNase I hypersensitvity sites. The DNase I enzyme selectively digests regions depleted from nucleosomes. The digested fragments are then captured and subjected to NGS [Song and Crawford, 2010]. ATAC-seq relies on the insertion of sequencing adapters in regions of accessible DNA [Buenrostro et al., 2015]. Thanks to the low DNA quantity requirement, ATAC-seq has been adapted to single cell DNA accessibility assays. However, not all TFs bind to accessible DNA. Pioneer Transcription factors have shown to bind in dense chromatin. This class of TF is thought to compete with nucleosomes by various mechanisms such as recruiting chromatin re-modellers or by altering the chemical states of histones[Soufi et al., 2015, Mayran et al., 2018, Donaghey et al., 2018, Zaret and Carroll, 2011].

### *Modified DNA and friends*

Both DNA and histones can undergo chemical modifications. The most common DNA modification is characterised by the addition of Methyl group on the 5′ side of cytosines. DNA methylation is thought to play a role in gene regulation and gene expression, more specifically gene silencing. The role of DNA methylation is thought to be two fold: (i) directly obstruct TF binding to DNA [Domcke et al., 2015] and (ii) recruit methyl binding proteins to serve as repressive complexes[Miller and Grant, 2013]. The role of DNA methylation should be taken with caution when it comes to TF binding. In *Drosophila*, methylation levels have been shown to be close to non-existent [Rae and Steele, 1979, Urieli-Shoval et al., 1982]. On the other hand, despite being present at low levels, methylated regions can still play a functional role in gene expression [Lyko et al., 2000]. This would suggest that if DNA methylation plays a role in TF binding, this mechanism might not be conserved between species.

Histone modifications on the other hand seem to be highly conserved between species [Hayes and Wolffe, 1992, Bannister and Kouzarides, 2011]. Histone modifications are post translational modifications generally occurring along N-terminal histone tails but can also occur in the histone globular domain. Modifications include methylation, phosphorylation , acetylation, ubiquitination and sumoylation. Both the nature and the location (along the polypeptide chain) are thought to determine the functional impact of a histone modification. Acetylation of lysine neutralises the positive electrostatic charge and thus decreases the binding affinity between DNA and histones. In turn, nucleosomes can be displaced leading to increased DNA accessibility and TF binding. The addition of a single aceytl group on H4K16 is sufficient to disrupt the formation of 30nm DNA fibres [Shogren-Knaak et al., 2006, Zhang and Presgraves, 2017]. The role of methylation seems to be less clear as the addition of one or more methyl group to histone tail lysine's can be associated with active or repressive marks [Bogliotti and Ross, 2012, Li, 2002, Liu et al., 2016]. H3K4me3 is associated with active transcription. It is thought that H3K4me3 can be recognised by PHD fingers and prevent the binding of nuRD, a repressive complex [Champagne and Kutateladze, 2009]. Conversely, H3K9me3 is considered as a repressive mark as it recruits Heterochromatin proteins 1 (HP1). HP1 can directly reorganise the structure of chromatin and induce chromatin compaction [Zeng et al., 2010].

The presence and/or absence of histone modifications are indicative of chromatin states. Chromatin states could be defined as the set of DNA and/or histone modifications in a given genomic region [Baker, 2011]. Repressive heterochromatin is for example devoid of acetylation and show increased levels of H3K27me3 among others. On the other side of the spectrum, regions of active transcription show increased levels of acetylation and H3K36me3.

The role of chromatin states in undeniable in TF binding. Chromatin states would indicate available regions for TF binding. If a binding motif is present within that region, TF binding events are expected. However, recent work has suggested that the relationship between histone modification and TF binding is specific to the TF family. The specific histone modification associated with a given chromatin state would allow TFs to discriminate between binding sites [Casey et al., 2018, Xin and Rohs, 2018].

### *Concentration*

Even if a binding motif is accessible for a TF to bind, this is not always sufficient to explain gene expression. In certain cases, TFs require to be in a high enough concentration in order to induce transcription. A notable example of TF concentration playing a role in gene expression is embryonic development in *Drosophila* [Moens and Selleri, 2006, Alexander et al., 2009, Petkova et al., 2019]. The longitudinal patterning of the *Drosophila* embryo is regulated by the highly conserved set of genes known as Hox genes. The product of these genes are TFs that bind to developmental enhancers and induce cellular differentiation. Spatial and temporal regulation of Hox TFs is required for correct developmental patterns to occur. It has also been suggested that TF concentration plays a key role in the correct expression of downstream developmental genes [Dubuis et al., 2013b, Dubuis et al., 2013a]. Although many studies on the subject have focused on mRNA levels as a proxy for TF concentration, recent work has confirmed the role of TF concentration by direct measure of protein abundance[Papadopoulos et al., 2019]. Furthermore, fluctuations in cell to cell signalling levels seem to be responsible for cellular differentiation in mice [Ohnishi et al., 2014]. A higher cellular signal would induce the expression of a different set of genes than if the signal occurred at a lower rate.

The role of TF concentration in TF binding has been considered for many years. Simply put, a TF can be considered as a ligand and the preferred binding motif as a receptor. In these circumstances, ligand concentration plays an essential role in its ability to bind to a specific target. Changes in concentration would not play significant roles in TF binding unless lower affinity sites are considered. Fluctuations at high affinity sites play a lesser role as demonstrated by stochastic simulation in *E. coli* [Zabet et al., 2013]. Furthermore, Lickwar proposed that TF binding turnover plays the role of a molecular clutch for transcription factor function. A longer occupancy time leads to a higher transcriptional output than a short turnover time. Optimal binding motifs could be context dependant and not only reliant on sequence [Lickwar et al., 2012]

### *Cooperativity and Competition*

The story of TF binding is further complicated by the addition of co-factors. Eukaryotes generally posses large genomes and TF binding sites alone is insufficient to explain their specific regulatory mechanisms. Cooperative binding between TFs and their co-factors serve the purpose of enhancer activation or repression [Moens and Selleri, 2006, Mann et al., 2009, Osório, 2015, Osório, 2016, Chronis et al., 2017]. Without the cooperative action of both TFs, transcriptional regulation would not occur. Different combination of TFs and co-factors induce varying transcriptional regulation mechanisms. Cooperative binding events are common and occur between different TF families.

Cooperative binding can be described by various mechanisms [Kasahara et al., 2017]. The simplest case is TF dimerisation. TF can dimerise with itself or with co-factors. Dimersation would increase the strength of the bond between TF dimer and DNA. However, cooperative binding can be mediated through DNA on its own. DNA shape features play a role in TF binding. While one TF expends energy to twist or bend DNA, this leaves other TFs to bind freely without the expense of energy. Finally, DNA binding proteins may aid the binding of other TFs by modifying chromatin states. Assisted loading has been suggested as a mechanisms of cooperative binding [Voss et al., 2011]. The binding of one TF will modify local chromatin and aid the binding of co-factors. Furthermore, chromatin opening and nucleosome displacement is thought the be one of the roles of pioneer transcription factors [Iwafuchi-Doi and Zaret, 2014].

On the other hand, TFs may also compete for target motifs. This mode of regulation would make TFs compete for a limited pool of available sites. By binding to a target site, one TF will inhibit the binding of another. This site inhibition leads to repression of gene expression and/or fine tuning. Work on stem cells differentiation have described a mechanism by which target site competition enables cells to distinguish between differentiation signal and cellular noise [Sokolik et al., 2015].

### *Chromatin structure*

Within each cell, DNA is tightly compacted and organised. The levels of organisation range from chromosomal territories to enhancer-promoter folding [Gibcus and Dekker, 2013]. HiC has become the gold standard method of determining the three dimensional structure and organisation of the genome [Lieberman-Aiden et al., 2009]. This method relies of the "capture" of DNA fragments that are in close proximity to each other. The final product of HiC experiments is a contact matrix describing the adjusted number of contacts between each genomic region in a population of cells. Contact matrices exhibit pyramid like structures than can either be described as Topologically Associating Domains (TAD) or chromatin loops [Lieberman-Aiden et al., 2009, Dekker et al., 2013, Dekker and Heard, 2015]. Similar structures are described with respect to genome anchoring to the nuclear lamina [Guelen et al., 2008, Pope et al., 2014, van Steensel and Belmont, 2017]. These structures are known as Lamina Associated Domains (LAD). The size of these structures can vary from a few thousand base (kbp) pairs to a few million base pairs (mbp). In both cases, these structures can be defined as genomic regions showing preferential contact within itself rather than other genomic regions. The main difference between these structure types is found at the tip of the pyramid. While chromatin loops show an increased number of contacts at the tip of the pyramid, TADs generally lack this high contact density cluster [Schwarzer et al., 2017, Hansen et al., 2018]. At a larger scale, TADs and loops can be contained within A and B compartments. A compartments are considered to be active while B compartments are considered inactive in terms of gene expression [Lieberman-Aiden et al., 2009, Zhan et al., 2017, Qi and Zhang, 2018].

The structure of DNA within cells is believed to impact gene expression [Rao et al., 2014, Lupiáñez et al., 2015, Sexton and Cavalli, 2015, Zhou et al., 2019]. Numerous examples show how altered DNA folding will induce ectopic expression of certain genes (reviewed by [Stadhouders et al., 2019, Ghavi-Helm et al., 2019]). The simplest case to consider is enhancer-promoter interactions. DNA will be folded in order to bring an enhancer close to its designated promoter thus inducing transcription [Andersson and Sandelin, 2019]. It remains unclear what drives the folding of DNA. One hypothesis involves Transcription Factors. By binding to DNA, they would drive the folding of DNA and bring two sections of the genome closer together. The other side of the argument would be that DNA is already folded in a way that would assist enhancer-promoter interaction [van Steensel and Furlong, 2019, El Khattabi et al., 2019]). This mechanism would be aided by DNA binding proteins that may not directly be involved in transcriptional regulation but rather promote gene expression via genome structural maintenance. In *Drosophila*, there are examples of DNA binding proteins that play the role of architectural maintainers as well as gene expression regulators [Li et al., 2015, Cubeñas-Potts et al., 2017]. This helps to distinguish between permissive chromatin contacts and instructive chromatin contacts. While permissive chromatin contacts are independent of cell type and may be unrelated to gene regulation, instructive chromatin contacts refers to DNA-DNA contacts that are cell type or tissue type specific and induce transcriptional activation [De Laat and Duboule, 2013]. A subset of TAD boundaries were shown to be cell type specific suggesting a role in transcriptional regulation and cell differentiation [Chathoth and Zabet, 2019]. This would also suggest that only a subset of TADs are conserved between cell types. Conserved TAD boundaries would induce permissive chromatin contacts and non-cell specific gene expression while cell-types specific TAD would play a role in cell type specific gene regulation.

The folding of the genome could also create pockets or hubs that increase local TF concentration. DNA hubs would help TFs to bind to target genes [Boija et al., 2018].

Work on phase separation has shown to be a promising avenue to describe the influence of genome architecture on TF binding [Hnisz et al., 2017, Boeynaems et al., 2018]. However, our understanding of genome architecture is still in its infancy. It is still unclear how TFs find their way through the complex maze of chromatin in order to find their target sites.

# 2

# MODELS OF TF BINDING

## THE DAWN OF TF BINDING MODELLING

With the increase of available data, there has been a need to find convenient and elegant ways to both analyse and conceptualise biological data. In genome biology, this is especially true as large genomes are impossible to analyse by hand. After being sequenced, the human genome was printed into a set of books. The whole human genome could scarcely fit in 130 volumes of double sided size 4 font print outs!

In recent years, the increase in computational power has made the analysis of genomic data somewhat trivial (I insist on the "somewhat"). In coordination with next generation sequencing, the field of bioinformatics has exploded and provided a new understanding of TF binding. However, the mechanisms of TF binding still remain a vibrant field of research. In particular, over the past decades, there has been an ongoing effort to describe, model and predict TF binding to DNA. In the following sections, I will give an overview of the various models developed to predict TF binding. The sections are broken down into the various biological or physical features used to model TF binding.

## SEQUENCE BASED MODELS

The most fundamental aspect to consider when modelling TF binding is binding motif. In 1987, Berg and Von Hippel proposed a statistical-mechanical model based on binding energy contributions of each base pair [Berg and von Hippel, 1987a]. Simply put, the model statistically described binding motifs (using experimentally determined binding motifs) and the distribution of similar motifs in promoter regions. Motifs that shared similar base pair composition to known motifs were highly predictive of gene activity in other promoter regions (for the same TF). This model is however based on a few assumptions. Firstly, binding motifs have been selected through evolution to conform to protein binding specificity. Secondly, more than one sequence is capable of fulfilling these binding requirements. It is also important to consider that these high affinity binding motifs are equally likely to occur. By considering sequence statistics alone, the model demonstrated reasonable agreement with experimental data available at the time. However, due to the low number of binding sites used to develop this model, Berg and Von Hippel conceded that this would result in small sample statistical bias. Similarly, Robert Harr also proposed a pattern matching algorithm based on the assumption that genetic information follows a statistically predictable pattern [Harr et al., 1983]. All four base pairs have , theoretically at least , the same chance of occurring in the genome.

In 1982, Gary Stormo proposed a model incorporating information theory (originally developed for ribosomal binding on RNA) [Stormo, 1982]. This model made use of the perceptron algorithm to distinguish translation initiation sites in *E. coli*. The accumulation of binding motifs for a certain protein offered the possibility of creating a "Position Frequency Matrix"(PFM). By dividing base pair frequency by the total amount of binding sequences, one will obtain a "Position Probability Matrix"(PPM). The original model proposed by Stormo assigned a weighted probability to the matrix under the following assumptions:

- Each base pair in the binding motif independently contributes to the binding specificity of a given sequence.

- The probability of occurrence needs to be weighted by considering nucleotide frequency in the genome and that a nucleotide appearing at a higher frequency must have a higher contribution to DNA binding specificity.

Nevertheless, it is possible for some base pairs to be absent from a binding motif PFM. Zero probabilities are replaced by introducing Laplacian Operators or psuedo-counts. Laplacian Operators will ensure that null probabilities will not interfere with the probability weighting and sequence scoring. The information content (A measure of the specificity of a sequence) is given by:

$$I_{seq}(i) = \sum_{b} f_{b,i} log_2 \frac{f_{b,i}}{P_b}$$

with $I$ the information content of the sequence *seq*, $f_{b,i}$ the observed frequency of each base pair at that position and $P_b$ the background frequency of each base pair. The information content is also called relative entropy or Kullback-Liebler distance. In order to estimate the statistical significance of a binding motif, $I$ is a normalised log-likelihood ratio statistic. The probability of a site $S_\alpha$ being bound is given by:

$$P(S_\alpha Bound) = \frac{-e^{H(b,i) \cdot S_\alpha}}{Z}$$

with $H(b,i)$ being a matrix containing the binding energy of each base pair (independently contributing to binding), $S_\alpha$a particular sequence and $Z$ the sum of whole genome binding affinities. In order to justify the use of the equations above, one would need a collection of high affinity binding proteins and the complete genome of a specific organism. It would also be necessary to assume that the genome is random (considering background probability of any base pair). This assumption can be made when considering short sequences such as the length a binding site. A

Probability Weighted Matrix would then become the matrix that would maximise the binding probability of the sequence $S_\alpha$. The maximised probability matrix was formulated by [Heumann et al., 1994]. However, this model shows its limitations due to the high amount of false positive binding motifs. The high affinity binding sites proposed by this method are much more frequent than what is found experimentally. It has been suggested that a uniform correction of base pair occurrence may be misleading. Other models based on the matrix framework were also proposed by [Zhang and Marr, 1993]. Stromo revisited and improved his model over the years [Stormo, 2000, Stormo and Zhao, 2010]. Although imperfect, this model is still used today. Many modern models of protein binding to DNA still rely on PWM scores [Tompa et al., 2005, Elemento and Tavazoie, 2005]. PWM motifs have also been used to describe TF binding families, in particular motif families [Jolma et al., 2013]. TFs can be classified with respect to their binding motif. The ability of multiple TFs binding to the same motif strongly indicates other contributing factors to TF binding and the subsequent gene regulation.

### BASE PAIR CONTRIBUTION AND DNA SHAPE

As described previously, PWM scores exhibit certain limitations and important assumptions. More specifically, independent contribution of base pairs in a binding motif has been debated [Zhou et al., 2015]. Zhou investigated base pair contribution in binding motif specificity by considering base pairs as a series of k-mers. As base pair composition affects DNA structure, DNA shape features were also included into the model. These included Minor Groove Width, Propeller Twist , Roll, and Helix Twist. By using support vector regression, they demonstrated that DNA shape features strongly improve the quality of the model. Increasing k-mer length also displayed an increased ability to predict and model TF binding. This suggests base pairs do not act independently but rather as a cohesive unit. The k-mer length was limited to

3-mer as increasing the number of k-mers also increases the feature space and thus the computational power needed. Later, the same lab produced an improved model that included 13 DNA shape features. They showed that DNA shape inclusion consistently improved TF binding predictions [Li et al., 2017].

## CHROMATIN STATES

As described in the previous section, DNA accessibility and chromatin states seem to play a central role in TF binding. It comes to no surprise that these factors have been included in many TF binding prediction tools and models. In 2011, the CENTIPEDE algorithm was developed by [Pique-Regi et al., 2011]. The model was developed using Bayesian inferential statistics. The likelihood of a TF binding depended on both the binding motifs (PWM) as well as the cell specific DNA accessibility pattern. However, CENTIPEDE did not take into account DNase I fluctuations between binding sites nor did it account for variability between DNase I replicates. These caveats were later corrected with the release of msCENTIPEDE [Raj et al., 2015a].
The first major paper to use machine learning in TF binding prediction was released in 2014. PIQ (Protein Interaction Quantitation) incorporates DNase data by considering both shape and magnitude of accessibility data [Sherwood et al., 2014]. Using a method called expectation propagation, PIQ predicted TF binding sites using both accessibility and PWM motif data. At the time, this new algorithm out-performed competing tools. Interestingly, Hidden Markov Models had been used in 2011 to predict TF binding using chromatin states [Won et al., 2010]. The software ( named Chromia) included chromatin state information in order to infer TF binding sites. Machine learning algorithms have seen a huge surge in popularity throughout many fields of biology and in particular TF binding prediction [Salekin et al., 2017, Salekin et al., 2018, Alipanahi et al., 2015]. Their incredible predictive power promised a deeper understanding of biology. However, it should be noted that most machine learning al-

gorithms suffer from their lack of explainability and should be used with caution [Angermueller et al., 2016, Roscher et al., 2019]. Despite efforts to push towards explainable forms of machine learning [Tareen and Kinney, 2019], we still don't understand which factors drive predictions.

In collaboration with ENCODE, DREAM challenges released a TF binding prediction competition. The aim of the competition was two-fold. First, they aimed to identify the best performing TF binding predictions models. Second, they aimed to set guidelines in order to asses the performance of such models. ENCODE and DREAM challenges provided the competitors with ChIP-seq data for various TF in various cell lines as well chromatin accessibility data. Unsurprisingly, many tools presented were based on some form of machine learning algorithm. Some of the most notable tools were Catchitt, FactorNet and Anchor, all three scoring in the top positions [Quang and Xie, 2019, Keilwagen et al., 2019, Li et al., 2019].

The ability to predict TF binding from DNA accessibility data is undeniable. However, pioneer TF such as FOXA2 were often poorly predicted confirming that other factors are required. Furthermore, as described previously, despite strong predictions for TF such as CTCF, the binding mechanisms of TF binding remained an open question.

## BINDING KINETICS OF TF BINDING

For many years, the field of Enzymology has formalised the relationship between ligands and receptors using binding kinetics. The binding of a ligand to a receptor can also be described with respect to ligand competition and cooperativity. Grankek and Clark proposed binding kinetics model of TF binding [Granek and Clarke, 2005]. Their model also included both competition and cooperativity. They demonstrated the effect of distance on protein-protein interaction and how this would affect gene expression. However, their model of cooperative binding failed to include indirect cooperative binding mechanisms such as assisted loading [Voss et al., 2011].This mechanism also named collaborative competition describes the competition between two different transcription factors binding to the same motif. Instead of hindering the binding of its co-factor, the binding of one will increase the probability of binding of the other by increasing chromatin accessibility. Similarly, Wang proposed another kinetic model based on explicit statement of association and dissociation constants [Wang et al., 2009]. Constants were calculated by using tagged TFs. The fluorescence emission rate would change between bound and unbound states.

However, none of these models consider site specific binding but rather describe general binding kinetics between transcription factors and DNA. In 2007, a biophysical model of TF binding and computational tool (TRAP - Transcription Affinity PRediction ) was proposed by Roider [Roider et al., 2007]. At its core, the model was based on site specific binding equilibriums derived from binding kinetic equations. TRAP would predict the probability of a TF binding to a given site without requiring any threshold ; The model would describe the probability of binding at any given site along the genome. This would ensure that the whole spectrum of binding affinities was included rather than only considering high affinity binding sites. Interestingly, this model would also infer the number of bound molecules and a scaling factor by maximising correlation between predicted profiles and ChIP-chip experimental profiles.

### STATISTICAL THERMODYNAMICS

Statistical mechanics (also called statistical thermodynamics) has yielded promising insights into TF binding mechanisms. Classical thermodynamics relies on the study of a system as a whole. Any given system $S$ can be described as set of parameters such as pressure, temperature or volume (not restricted to this short list). However, thermodynamic parameters are thought to be the consequences of particle conformation within the system. In other words, a system can be described by the distribution and intrinsic proprieties of each particle within that system. The observable behaviour of a system is therefore a direct consequence of this particle distribution. Statistical thermodynamics can considered as a combinatorial model of TF binding. There have been numerous attempts to describe TF binding using combinatorial models. Analytical binding models were used to describe the competition between specific and non-specific binding of TFs to DNA [Tsodikov et al., 2001]. Combinatorial models were also used in an attempt to describe the strong but non-specific DNA binding of electrically charged macromolecules [Rouzina and Bloomfield, 1997]. More recently, Mirny described cooperative binding of non-interacting TF's through competition with nucleosomes [Mirny, 2010]. Djordjevic proposed a TF binding model using a statistical mechanics framework [Djordjevic et al., 2003]. The proposed method was based on classification of potential binding sites using sequences specific TF binding energy. Contrarily to many other methods, their model tried to brake away from information theory methods such as the PWM. This showed a significant reduction in false positive binding sites. Following the same path, He proposed a model that would consider cooperativity, competition and short range repression under a statistical mechanic framework [He et al., 2010]. However, this study was based on the assumption that proteins would only bind to high affinity sites. There was no consideration for low affinity binding. Recent developments have shown that clustering of low affinity binding sites are capable of inducing gene expression by being preferentially bound by TFs [Farley et al., 2015].

## THE ZABET AND ADRYAN MODEL

In 2015, Zabet and Adryan introduced an approximation of the statistical mechanics framework [Zabet and Adryan, 2015]. This lead to a reduction of computational complexity without loosing predictive capabilities. Furthermore, the explicit description of factors involved in TF binding may shed light on TF binding mechanisms. The model describes the probability of a transcription factor being bound to a site j by:

$$P(N, a, \lambda, \omega)_j = \frac{N \cdot a_j \cdot e^{\left(\frac{1}{\lambda} \cdot \omega_j\right)}}{N \cdot a_j \cdot e^{\left(\frac{1}{\lambda} \cdot \omega_j\right)} + L \cdot n \cdot \left[a_i \cdot e^{\left(\frac{1}{\lambda} \cdot \omega_j\right)}\right]_i}$$

with:

- $N$ , the average number of bound molecules to DNA ( considering the entire genome of an organism compacted within the nucleus)
- $a$ , the accessibility of site $j$
- $\omega$ , the binding energy required for a TF to bind to site $j$ - in the form of a PWM score.
- $\lambda$ , a scaling factor for the PWM score.
- $L$ , the length of the genome of interest
- $n$ , the ploidy level of the organism

It should be noted that Position Weight Matrices can be used to score the genome or a sequence. The score represents how strong of a binding site is a motif given the PWM for a specific TF. As described by Zabet and Adryan, both N and $\lambda$ are unknown and would need to be inferred by maximising Pearson correlation and minimising Mean Squared Error between TF binding predictions and ChIP-seq data. Their analysis was carried out on a set of five *Drosophila* TFs ( Bicoid, Caudal,Giant,Hunchback and Kruppel). The correlation between the model and experimental data was sensitive to changes in $\lambda$. Conversely, varying the number of bound molecules ($N$) prompted stronger fluctuations in Mean Squared Error (MSE). In order to accurately infer both $N$ and $\lambda$, they suggested the use of metric overlap. The best performing parameters using Pearson correlation were overlayed with best performing parameters obtained when using MSE. The resulting overlap would indicate the optimal set of parameters. One of the key innovations of their model was the inclusion of DNA accessibility. DNA accessibility would modulate the probability of TF binding in combination with binding energy (PWM), the scaling factor ($\lambda$) and the number of bound molecules. To model the probability of a site being accessible, they followed the approach proposed by Kaplan.

$$a_j = \frac{1}{1 + exp(-\beta \cdot DD_j + \alpha)}$$

$DD_j$ DNase I read density, $\alpha$ and $\beta$ are scaling parameters [Kaplan et al., 2011b]. Interestingly, Zabet and Adryan demonstrated that considering DNA as either open (accessibility value of 1) or closed ( accessibility value of 0) was sufficient to accurately predict TF binding.

The model described above is derived from combinatorial mathematics and in the context of TF binding, describes the different ways TFs can be arranged along a sequence of DNA. The model above was built upon a few assumptions. First, only one TF can bind any given site at any given time and many sites will be unbound by TFs. Second, only sites above a certain threshold are considered as potential binding sites. Third, the role of DNA accessibility is to "mask" potential binding sites under the assumption that TF binding only occurs in open chromatin. Fourth, the number of binding sites available along the genome is assumed to be much higher than the number of available TF molecules.

The model also comes with a few limitations. First, the model only describes the mechanisms by which *one* TF binds to DNA. The model does not include the full scope of mechanisms known to drive transcription such as cooperative binding. Second, using a threshold to select binding motifs will potentially filter out low affinity binding sites that may induce transcriptional activity [Farley et al., 2015]. Finally, considering

DNA accessibility as either open or closed might not capture that full scope of TF binding events.

Interestingly, the Zabet and Adryan model is very similar to the model described by Roider [Roider et al., 2007]. Both models require the use of PWMs in order to determine binding energies along DNA sequences. Both models also infer the number of bound molecules and a PWM scaling factor by maximising correlations between predicted profiles and ChIP profiles. Initially, they both considered the full spectrum of binding energies between unbound sites and highly specific binding sites. However, the analysis carried out by Zabet and Adryan only considers strong binding sites when producing predictions. Their model also differs by including DNA accessibility as a contributing factor to TF binding. The Zabet and Adryan model could also be loosely compared to the model proposed by Berg and Von Hippel (see above - [Berg and von Hippel, 1987a]) as both model are statistical-mechanical models by nature. However, the model proposed by Berg and Von Hippel considered motif sequence distribution between promoters of specific TFs and how related these sequences were to explain genetic activity in different promoters.

The following thesis directly builds upon the results obtained by Zabet and Adryan.

# 3

# AIMS

## GENERAL AIMS

The main aim of this thesis can be summarised by the following:

*Dissecting the binding mechanisms of Transcription Factors to DNA using a statistical thermodynamic framework.*

In particular, this thesis aims to dissect the contributions of various factors in TF binding, namely chromatin states, binding energy (described by PWMs), a PWM scaling factor, and number of TF molecules bound to DNA. This biophysical model aims to recover known binding mechanisms but also uncover unknown factors contributing to well studied TFs. In order to accomplish this task, I have aimed to achieve the following goals:

1. Designing a user friendly tool implementing the model described by Zabet and Adryan [Zabet and Adryan, 2015]. The creation and thorough testing of this tool will indicate how strongly this statistical thermodynamic model holds against other frameworks. Furthermore, testing will demonstrate strengths and

weaknesses within the model that will need to be taken into account for further work. This includes the incorporation of new goodness of fit metrics to evaluate model performance.

2. Dissecting the role of DNA accessibility, binding energy and specificity, and DNA binding protein abundance. This will be accomplished by a thorough inspection of a set of TFs in *Drosophila*.

3. Improve the model by including chromatin states as an influencing factor towards TF binding and implementing a genetic algorithm for parameter optimisation.

4. Tie all results together in a comprehensive analysis of a case study.

## THESIS BREAK DOWN

The following thesis will be broken down into the following chapters:

- **Chapter 1** will describe the development and testing of ChIPanalyser, an R package I have now made available on Bioconductor [Martin, 2017]. More specifically, I will describe the inner working of the package as well as demonstrate data preparation required for the use of the package. Furthermore, I will demonstrate the performance of ChIPanalyser by testing and validating its performance in a cross-validation set-up as well as comparing its performance to other TF binding predictive tools.

- **Chapter 2** will describe the insight gained in the binding of three architectural proteins in three *Drosophila* cell lines (Kc167, BG3 and S2). These architectural proteins are CTCF, BEAF-32, and su(Hw). I will illustrate that DNA accessibility plays a nuanced role in the binding of these TFs. Moreover, TF abundance plays a lesser a role in their binding, indicating a robust mechanism against protein concentration fluctuations. I will also illustrate ChIPanalyser's ability to recover known preferences towards chromatin accessibility for three Hox TFs (Ubx, Dfd, Abd-b) in *Drosophila* Kc167 cells. Ubx preferentially binds to open chromatin whereas both Dfd and Abd-b can also bind in denser chromatin.

- **Chapter 3** describes the development of the model by including chromatin states as a driver towards TF binding. In order to uncover chromatin state preferences, I developed a genetic algorithm using the core functionalities offered by ChIPanalyser for parameter optimisation. I demonstrate that architectural proteins show distinct preferences towards chromatin states. The same was applied to Hox TFs to demonstrate that binding preferences could also be recovered.
- **Chapter 4** ties all the previous results and insight together by exploring the performance of the model on the well-known Notch activator Su(H). In this chapter, I examine the performance of the model on Su(H) binding predictions with respect to Accessibility, chromatin states, induction of the signalling pathway and finally, protein abundance in partial knock-downs. I demonstrate that Su(H) shows distinct preferences towards chromatin states and that as opposed to architectural proteins is sensitive towards changes in proteins abundance.
- **Chapter 5** will summarise the results of this thesis and critically discuss the findings with respect to both the model and the underlying biology.
- **Conclusion** will offer some concluding remarks with respect to the results described in this thesis as well as their context in biology.
- **Looking Forward** will describe avenues that could be explored to further improve the model and the package.

Part II

# MECHANISM OF TRANSCRIPTION FACTOR BINDING

# 1

# BUILDING CHIPANALYSER

## CHAPTER SUMMARY

The following chapter will describe the development of ChIPanalyser, an R package published and available on Bioconductor. First, I will describe the input data required by ChIPanalyser. Second, I will describe how the model was implemented and the ChIPanalyser work flow. Finally, I will describe the performance of ChIPanalyser with respect to parameter selection, a cross validation set up and a comparison to other available frameworks.

## INTRODUCTION

The formal description of the natural world has been a key aspect of science for centuries if not millennia. Probably one of the most well-known formalisation of the natural world was proposed on the 5th of July 1687. Newton proposed a formal description of the laws of motions that still hold to this day. Biology has not been exempt from this trend and in particular Transcription factor (TF) binding. As described in the introduction, TFs tend to bind in a sequence specific manner[Ptashne and Gann, 1997, Spitz and Furlong, 2012]. Most TFs (albeit not all) show a preferred binding motif along the genome. In many cases, the binding motif be adequately described by a Position Weight Matrix (PWM). The idea was developed by Stormo back in 1982 to describe the binding of ribosomes to RNA [Berg and von Hippel, 1987b, Benos et al., 2002, Stormo and Zhao, 2010]. The principle behind PWMs was that the binding preference of a given DNA binding protein can be described by a weighted matrix. Every position in the binding motif is weighted based on the frequency of occurrence of this base pair at that position in the motif but also the likely-hood of that base pair occurring within the genome[Berg and von Hippel, 1987b, Roider et al., 2007].

However, considering the binding motif alone yields a higher rate of binding sites than observed experimentally. Many of the binding sites do not seem to be bound by TFs[Zhang et al., 2005, Li et al., 2008, Farnham, 2009, Skalska et al., 2015]. This is where the crux of the TF binding prediction problem lies: How to distinguish bound binding sites from unbound ones?

There have been many approaches to describe TF binding and predict binding location of TFs along the genome. Recently, it was suggested that using a statistical thermodynamic model could help filter out bound from unbound binding sites. Statistical thermodynamics also known as statistical mechanics describes the many configuration that a system $S$ may have with respect with to each particle contained within $S$. In the context of TF binding, the system is equated to a strand of DNA and the particle conformation to the TF repartition within the system. Both the location of the TF and the number of TFs bound to DNA both play an important role in how we can describe the system. In 2015, Zabet and Adryan derived a model of TF binding from statistical thermodynamics [Zabet and Adryan, 2015]. The model was formalised as follows:

$$P(N,a,\lambda,\omega)_j = \frac{N \cdot a_j \cdot e^{(\frac{1}{\lambda} \cdot \omega_j)}}{N \cdot a_j \cdot e^{(\frac{1}{\lambda} \cdot \omega_j)} + L \cdot n \cdot [a_i \cdot e^{(\frac{1}{\lambda} \cdot \omega_j)}]_i} \tag{1}$$

describing the probability of a site $j$ being bound by a TF with:

- $N$ , the average number of bound molecules
- $a$ , the accessibility of site $j$
- $\omega$ , the binding energy required for a TF to bind to site $j$ - in the form of a Position Weight Matrix
- $\lambda$ , a scaling factor for the Position Weight Matrix.
- $L$ , the length of the genome of interest
- $n$ , the ploidy level of the organism

The model describes the different possible conformation of TF binding along the genome by considering four main factors.

- $\omega$, the binding energy of a given transcription factor. In this instance, the score associated to a PWM is equivalent to binding energy of that site.
- $\lambda$ represents a scaling factor applied to the binding energy. The scaling factor describes how well a TF discriminates between low and high affinity sites. If two sites $a$ and $b$ are similar but not equal, how much more does a given TF prefer site $a$ over site $b$. In the context of this equation, the lower the value assigned to lambda, the higher the affinity of a TF to high affinity sites, the higher its ability to discriminate between low affinity and high affinity sites.
- $N$ details the average number of molecules bound to the genome. The number of bound molecules could be seen as a proxy for TF concentration.
- $a$ represent the accessibility of a given site. This model makes the assumption that TF binding occurs in accessible DNA. If a stretch of DNA is tightly compacted (considered inaccessible for TF binding), we would not expect any binding despite potentially containing a binding site. Accessibility serve the purpose of reducing the number of possible binding sites.

ChIPanalyser is an implementation of the model described above available on Bionconductor [Gentleman et al., 2004, Martin, 2017] .The four main factors believed to influence TF binding are not always given by experimental methods. Namely, the number of bound molecules $N$ and the scaling factor $\lambda$ are not given experimentally. To extract the values of these parameters, both $N$ and $\lambda$ are inferred within ChIPanalyser by finding the combination of parameters that maximises (or minimises) a goodness of fit metric. I recognise that recent advance in experimental techniques such as FRAP, single-molecule tracking or in-gel quantitative fluorescence have given insights into the number of molecules of a given protein within a nucleus. Interestingly, if the number of molecules are known by experimental means, the estimated value of number of molecules can be used as is and the model would not required any parameter optimisation step. However, this type of data is not widely available. Furthermore, for the purpose of describing the inner working of ChIPanlayser and the strengths of the model, I will consider that both $N$ and $\lambda$ are unknown.

ChIPanalyser as a tool serves multiple purposes. The package can infer the number of bound molecules and a scaling factor from ChIP-data within biologically acceptable ranges. The final product of the package is a ChIP-like profile describing the binding of a TF to DNA. In order to both infer the optimal set of parameters ($N$ and $\lambda$) and produce ChIP-like profiles, ChIPanalyser requires input data. In the following section, I will describe the data required by the package and the pre-processing pipeline applied to said data.

## DATA INPUT

### *DNA Sequence*

In order to locate binding motifs along the genome of an organism, the package requires a reference genome. Thankfully, many reference genomes are available within the BSgenome R packages [Pages, 2018]. For the purpose of the work carried out in this thesis, I used the latest available version of the genome. For *Drosophila melanogaster*, the dm6 version of the genome was used. All data sets that required alignment were aligned to this version of the genome as well. When datasets where downloaded pre-aligned to a prior version of the genome, they were lifted over to dm6 using UCSC liftover chains. The same principle was applied when the analysis was carried out in different organisms. When required (i.e comparing frameworks), the hg38 version of the human genome was used.

### *Binding Motifs*

Binding motifs for TFs used through out this thesis were either downloaded from the JASPAR [Mathelier et al., 2014] database or extracted from the MotifDb R package [Shannon and Richards, 2018] which collects and compiles PFMs and Position Probability Matrices (PPM) from various online repositories. Each Position Frequency Matrix was also selected based on the organism used for the analysis. For the purpose of method comparison (msCENTIPEDE - binding motif requirements), TF binding sites were also extracted using FIMO from the MEME-suit tool kit [Grant et al., 2011]. PFM and PPM are converted to PWMs by ChIPanalyser using the method described by Stormo[Stormo and Zhao, 2010] . Figure 1 describes the binding motifs (also known as Motif Logo) used through out most of the work carried in Chapter 1, 2 , and 3. More specifically, three *Drosophila* architectural DNA binding proteins were selected (CTCF, BEAF-32, and su(Hw) ) as well as three Hox TF (Dfd, Abd-b, and Ubx). When required, CTCF motif for *Homo sapiens* was also selected.

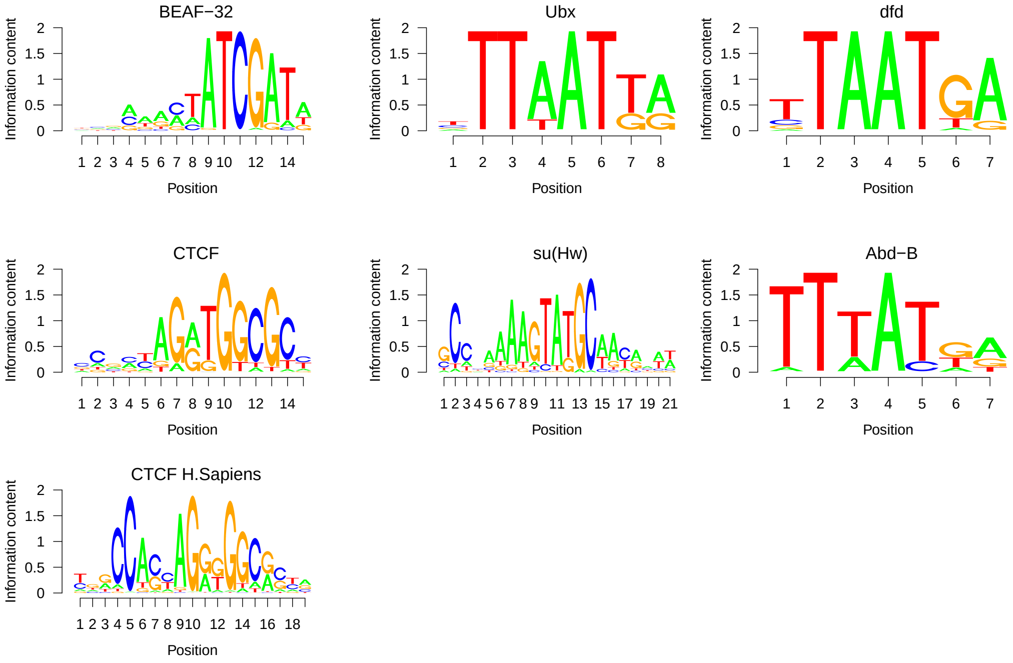


Figure 1: **Motif Logos.** Motif Logos of Architectural Proteins and Hox TFs

### *Genome binding profiles*

Genome binding profiles and peaks were downloaded from modEncode in three *Drosophila* cell lines: Kc167, S2, BG3 [ENCODE Project Consortium, 2012, Davis et al., 2018, modENCODE Consortium et al., 2010]. As ChIPanalyser mostly requires ChIP score pile-up, I considered both ChIP-seq and ChIP-chip to be similar enough for the purpose of this thesis. As all modEncode data sets were aligned to the dm3 version of the *Drosophila* genome, all data sets were lifted over to the dm6 version of the genome using UCSC liftover chains. When required, supplementary data sets were downloaded from Gene Expression Omnibus (GEO) database. GEO datasets were aligned to the dm6 genome using bowtie-2 (–non deterministic) [Langmead and Salzberg, 2012]. *SAM* files were converted to *BAM* files using samtools [Li et al., 2009]. Peaks and pile-up signal were called using macs2 with a 0.01 FDR [Zhang et al., 2008]. A summary table of published data sets used in this thesis can be found in Appendix A. Human data was downloaded from GEO and did not require further processing. It should be noted that genome binding profiles are required for model validation and inferring the optimal set of parameters by optimising a goodness of fit metric between the predicted ChIP-like profile and experimental ChIP profiles. As described above, if these parameters are known by other means, ChIP profiles are not required for the package to produce a prediction.

### *Chromatin Accessibility*

DNA accessibility serves the purpose of limiting the number of binding motifs available. DNA accessibility data is produced by two main techniques: DNase I hypersensitivity followed by sequencing and ATAC-seq. DNase I hypersensitivity data was generated by modEncode for three *Drosophila* cell lines: Kc167, BG3 and S2 [Kharchenko et al., 2010]. I aligned fastq files to the dm6 genome using bowtie-2 (–non-deterministic) [Langmead and Salzberg, 2012]. *SAM* files were converted to *BAM* files using samtools [Li et al., 2009]. Peaks and read pile-up were called using macs2 (-broad-call -cutoff 0.05 -q 0.05) [Zhang et al., 2008]. When required, ATAC-seq data for Kc167 cells was downloaded from GEO. ATAC-seq processing was described by [Porcelli et al., 2019]. For the following chapter, I used only DNase hypersensitivity sites (DHS). The comparison between DHS and continuous DNA accessibility scores will be described in Chapter 2 (ChIPanalyser: Insights into Biology). DNA accessibility data is not required for ChIPanalyser to produce ChIP-like profiles. DNA accessibility may improve the quality of the predicted ChIP-like profiles.

## CHIPANALYSER WORK FLOW

In the following section, I will describe the general work flow of ChIPanalyser. The general work flow is described in Figure-3 and requires data as described above. A complete description of the ChIPanalyser work flow including working examples can be found in the ChIPanalyser Bioconductor vignette[Martin, 2017].

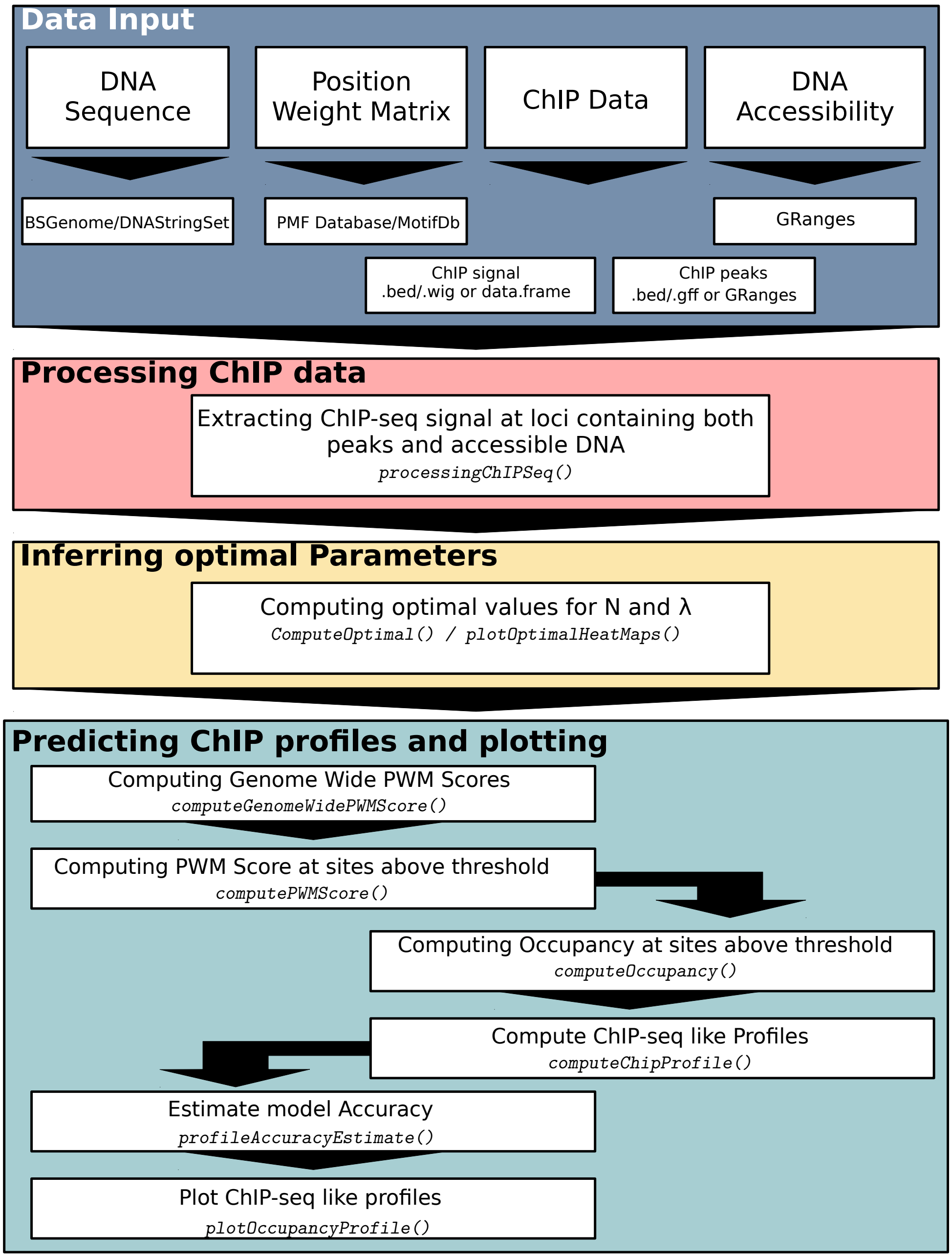

Data Input
DNA Sequence
Position Weight Matrix
ChIP Data
DNA Accessibility
BSGenome/DNAStringSet
PMF Database/MotifDb
GRanges
ChIP signal .bed/.wig or data.frame
ChIP peaks .bed/.gff or GRanges
Processing ChIP data
Extracting ChIP-seq signal at loci containing both peaks and accessible DNA
processingChIPSeq()
Inferring optimal Parameters
Computing optimal values for N and λ
ComputeOptimal() / plotOptimalHeatMaps()
Predicting ChIP profiles and plotting
Computing Genome Wide PWM Scores
computeGenomeWidePWMScore()
Computing PWM Score at sites above threshold
computePWMScore()
Computing Occupancy at sites above threshold
computeOccupancy()
Compute ChIP-seq like Profiles
computeChipProfile()
Estimate model Accuracy
profileAccuracyEstimate()
Plot ChIP-seq like profiles
plotOccupancyProfile()

Figure 2: **ChIPanalyser workflow.** ChIPanalyser follows the following work flow. **Data Input:** Data may come in various formats (e.g. bed, wig, gff etc.). **Processing ChIP-seq data:** If ChIP data is used to infer the optimal set of parameters (and/or validate model goodness of fit), ChIP data will be normalised and only regions of interest will be extracted for further analysis. **Inferring optimal parameters** Inferring optimal parameters will be achieved by maximising a goodness of fit metric. **Predicting ChIP profiles and plotting:** Armed with values for number of bound molecules and the PWM scaling factor, ChIPanalyser will produce ChIP like profiles. Both optimal parameter heat-maps and ChIP profiles can be plotted using the packages plotting functions.

ChIPanalyser will automatically convert PFMs to PWMs. It is also possible to directly provide a PWM if one is already available. If using genome binding profiles for parameter inference or model validation, the first step of the analysis is processing and extracting ChIP data. The *processingChIP* function will load ChIP data and extract a normalised ChIP score at a base pair level for the loci of interest. The loci of interest can be provided by the user. If no loci are provided by the user, the *processingChIP* function will extract loci based on the following criteria (and by extension arguments to the function):

- **Loci Width**: If no loci are provided, ChIP data will be converted into normalised ChIP scores at a base pair level and binned into regions of 20kbp (default split value). Produced bins will serve as loci of interest for further analysis.

- **Peaks**: If a set of ChIP peaks are provided, the function will select the loci of interest that contain at least one peak.

- **Chromatin State**: If DNA accessibility data is provided, the function will select the loci of interest that contain at least 100bp of accessible DNA.

- **Reduce**: The reduce argument will sort the loci of interest based on highest ChIP score contained in any given bin. The *processingChIP* function returns a ordered set of loci and scores. This argument also allows the extraction of top regions based on ChIP enrichment values. As an example, if the reduce argument is set to 50, only the top 50 regions with respect to enrichment will be returned.

Peaks and DNA accessibility data may be used in combination in order to ensure that the selected loci contain at least one ChIP peak and accessible DNA. However, for this analysis, regions were selected before hand. The reference genome was split into bins of 20kbp and all bins containing at least one peak of any TF from any data set as well as at least 100 bp of accessible DNA were used as input loci. Furthermore, black listed regions as described by USCS were also removed. This resulted in 3293 bins of 20kbp. When referring to top regions, I will be referring to the top regions selected from this set. During this step of the pipeline, ChIPanalyser provides four methods of reducing ChIP background noise: *zero, mean, median* and *sigmoid*. The current model does not consider ChIP depletion scores therefore negative ChIP scores are removed by assigning a score of zero to those positions (*zero* filtering). *Mean* and *median* assign a score of zero for any score below the mean or median score respectively. Mean and Median scores are computed after removing ChIP depletion scores. Finally, Sigmoid applies logistic weighting to every score. The logistic mid point is set at the 95th quantile of ChIP scores. The lower bound is set to zero and the upper bound is set to 2. Consequently, each score will be multiplied by a weight: if the score is above the 95th quantile the score will be weighted by values between 1 and 2. If the score is below the 95th quantile, score will be weighted by a factor ranging from 0 to 1. The logistic function (generalised sigmoid) can be described as the following:

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}} \tag{2}$$

with $L$, the curves maximum values; $k$, the steepness of the curve; $x_0$ , the sigmoid midpoint.

Once the loci of interest have been selected, the next step infers the optimal set of parameters ($N$ and $\lambda$) by maximising (or minimising in the case of Mean Squared Error) a goodness of fit metric. The goodness of fit metrics available in the package are described in Table-1.

| Metric | Description | Type |
|---|---|---|
| MSE | Mean Squared Error | Dissimilarity |
| K-S distance | Kolmogorov-Smirnov Goodness-of-Fit Test | Dissimilarity |
| Geometric ratio | $\frac{\int_a^b \lvert f(x)-g(x)\rvert dx}{\int_a^b min(f(x),g(x))dx}$ | Dissimilarity |
| Recall | $\frac{TP}{TP+FN}$ | Dissimilarity |
| Pearson | Pearson correlation coefficient between predicted and ChIP profiles | Similarity |
| Spearman | Spearman correlation coefficient between predicted and ChIP profiles | Similarity |
| Kendall | Kendall correlation coefficient between predicted and ChIP profiles | Similarity |
| Precision | $\frac{TP}{TP+FP}$ | Similarity |
| F-score | $\frac{2TP}{2TP+FP+FN}$ | Similarity |
| Accuracy | $\frac{TP+TN}{TP+TN+FP+FN}$ | Similarity |
| MCC | $\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$ | Similarity |
| AUC | Area Under the ROC (Receiver Operator Characteristic ) Curve | Similarity |

Table 1: **Goodness of Fit metrics.** ChIPanalyser offers 12 goodness of fit metrics grouped into two classes *Dissimilarity* and *Similarity*. Symbolically each metric is either a measure of how different two datasets are (Dissimilarity) or a measure of how similar two datasets are (Similarity). TP - true positives, TN - true negatives, FP - false positives and FN - false negatives. MCC represents Matthews Correlation Coefficient.

The *computeOptimal* function will infer the optimal combination of parameters by following the algorithm described in Figure-3.

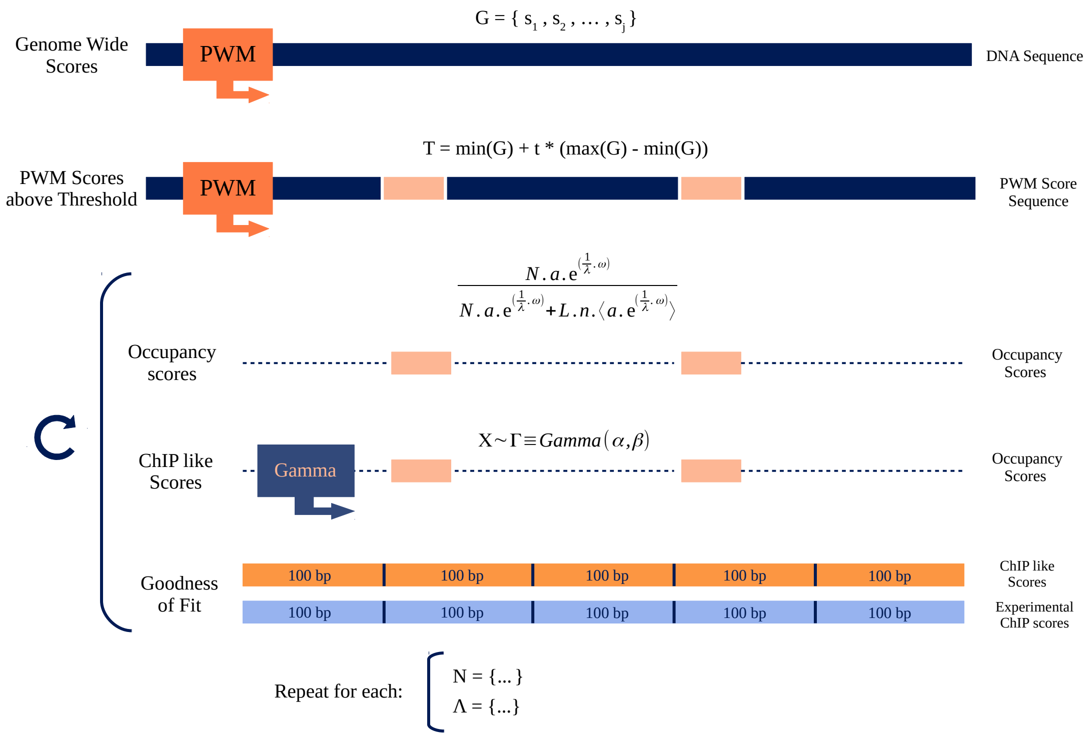


Figure 3: **ChIPanalyser Internal work flow.** In order to infer the optimal set of parameters, the package passes through the following steps. From top to bottom: (i) Score the entire genome to extract minimum and maximum PWM scores, (ii) Select scores that are above an arbitrary threshold, (iii) convert PWM scores to Occupancy probabilities using the statistical thermodynamic model and a combination of values for $N$ and $\lambda$, (iv) smooth Occupancy sites using a gamma distribution in order to produce a ChIP-like profile, (v) compute goodness of fit score for a combination of parameters by comparing the predicted profile and ChIP data in 100bp bins. Steps (iii),(iv), and (v) are repeated for each parameter combination.

The internal algorithm can be broken down into five distinct steps.

1. **Compute Genome Wide scores**: Genome wide scores are computed in order to extract minimum, maximum and average exponential scores required for the next steps. Genome wide scores are extracted from the reference genome. If DNA accessibility data is provided, Genome wide scores will be limited to only accessible DNA. It should be noted that the average exponential score is dependant on the $\lambda$ value. Minimum and maximum scores are required in order

to select only most likely binding sites. The average exponential score is required when computing the occupancy probability (see step 2 and 3).

2. **Compute Scores Above a threshold**: In order to select the most likely binding motifs, only PWM scores that are above an arbitrary threshold will be selected. The default value for the PWM threshold is set at 0.7 . This means that only the top 30% of PWM scores will be considered as potential binding sites.

3. **Compute Occupancy Probabilities**: PWM scores above threshold are converted into an Occupancy Probability by using the model described in equation 1. Different values of $N$ and $\lambda$ will yield different occupancy probabilities.

4. **Generating ChIP like profiles**: Occupancy sites are converted to ChIP-like profiles by smoothing the scores using a gamma distribution. This ensures that the prediction transitions from punctual binding site location to a continuous ChIP-like score at a base pair level. This also ensures that the predicted profile will also mimic experimental local enrichment.

5. **Computing goodness of fit**: The final step of the algorithm compares the predicted ChIP-like profile to experimental ChIP data. The comparison is done by comparing scores within bins of 100bp (user customisable - the resolution can be as high as 1 base pair). As scores are computed at base pair resolution, average occupancy score was used at lower resolutions ( i.e. average occupancy over 100 bp bins). When confusion matrices are required for scoring ( AUC, recall, F1, MCC, Accuracy,precision ), ChIPanalyser selects 20 occupancy thresholds for every regions. These thresholds represent squared occupancy scores between the minimum occupancy score and the maximum occupancy score. As occupancy scores range between 0 and 1, squaring the threshold will ensure a bottom heavy distribution. The minimum score describes the smallest occupancy score between predicted profiles and experimental profiles. The maximum score describes the largest occupancy signal between predicted profiles and experimental profiles.

Confusion matrices were constructed at each threshold. This approach ensures that ChIPanalyser would consider signal enrichment when assessing goodness of fit.

The algorithm repeats step 3 to 5 for each combination of $N$ and $\lambda$. Values for $N$ and $\lambda$ are predetermined but may also be customised by user (as long as there are at least two values for each parameter). Once all combinations have been computed, the combination of parameters that yield the best goodness of fit score will be selected and returned by the *coumputeOptimal* function. The steps described above may also be run individually if the parameters have already been selected (either by ChIPanalyser or approximated by other means). The result of parameter inference can be visualised using the *plotOptimalHeatMaps* function.

The final step of the analysis pipeline will produce occupancy profiles. Occupancy profiles show the predicted profile produced by ChIPanalyser. The plots may be enriched by the addition of ChIP data and accessibility data to serve as a point of comparison. Example plots are shown in Figure-4.

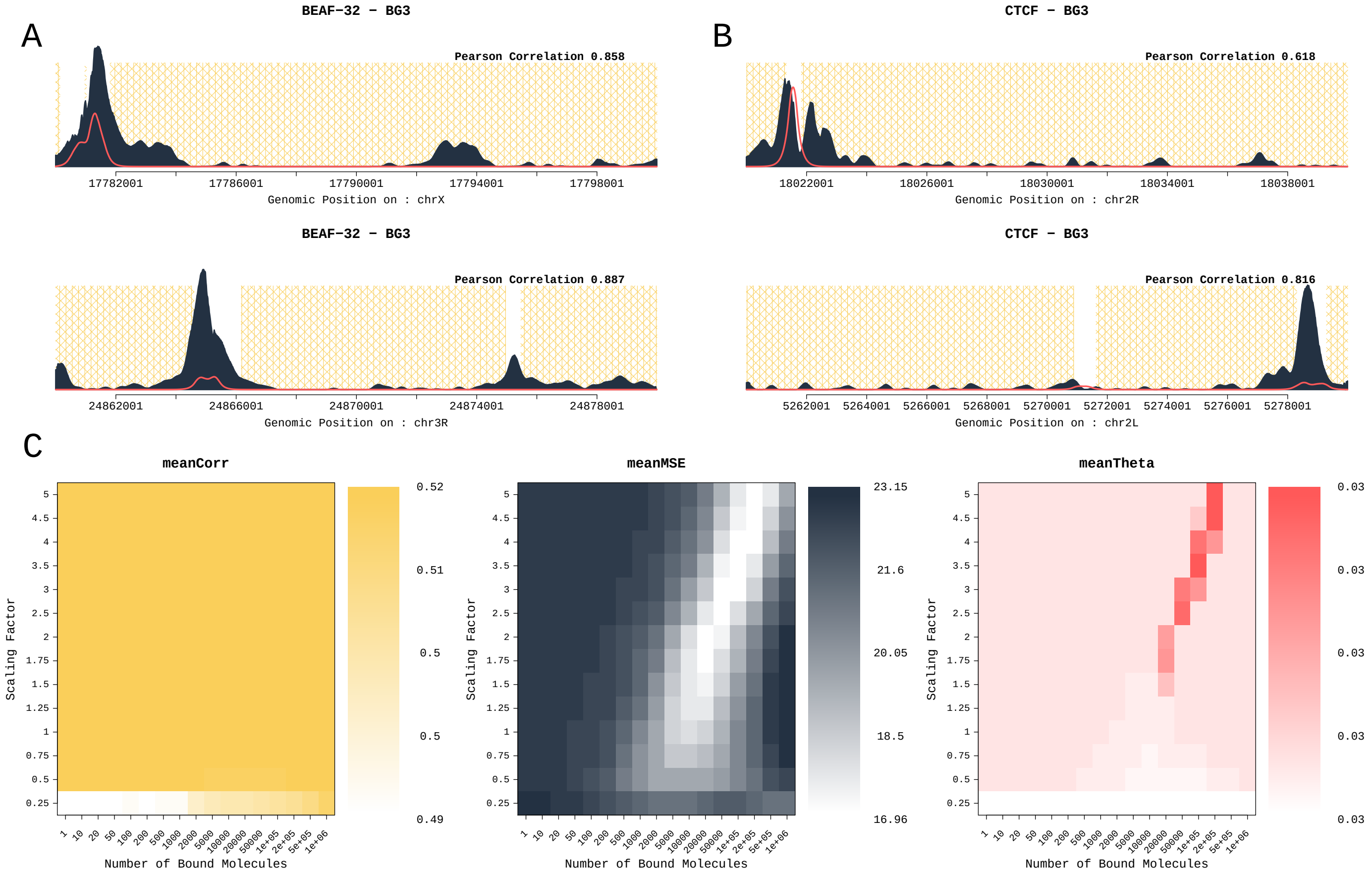


Figure 4: **Metric Unreliability.** The original model called for the maximisation of Pearson correlation coefficient and the minimisation of Mean Squared Error. When extending the analysis from the five original TF, I noticed that there were certain oddities arising from the use of Pearson correlation coefficient. As seen in (**A**) and (**B**) Pearson correlation coefficient does not seem to accurately measure the goodness of fit of the model. I observed an overestimation or underestimation of the models ability to predict TF binding. The second difficulty arising from using Pearson correlation coefficient was that it showed very little difference in correlation for some data sets. Previously, Zabet and Adryan described the optimal set of parameters by using the overlap of correlation and MSE as shown in (**C**). In many cases, MSE was the driving force behind the optimal parameters. Correlation played little to no role. For these reasons, I decided to investigate other goodness of fit methods. In (**A**) and (**B**), yellow shaded areas are regions of inaccessible DNA. Dark grey is experimental ChIP signal and finally the red line represents predicted profile. Profiles in (**A**) originate from BEAF-32 in BG3 cells. Profiles in (**B**) originate from CTCF in BG3 cells. All heat maps represent Abd-B in Kc167 cells.

The initial version of ChIPanalyser only used two goodness of fit metrics: Pearson Correlation and Mean Squared Error (MSE)[Zabet and Adryan, 2015]. Optimal set of parameters were selected by not only maximising correlation but by also minimising MSE. The goal was to find the sweet spot that would include both a high correlation and low MSE. Unfortunately, there were some limitations with this approach.
The first limitation came from the use of Pearson correlation as a goodness of fit metric. Figure-4 **A** and **B** show predicted profiles (red lines) with respect to DNA accessibility (yellow boxes) and compared to ChIP data (dark grey area) for BEAF-32 and CTCF in BG3 cells. Pearson correlation seemed to either overestimate or underestimate how well the model actually performed. The Pearson correlation associated to each profile is included on the top right hand corner of each profile. Intuitively, the two bottom profiles show an overestimation of the performance of the model.
The second limitation arose from using using the overlap of Pearson correlation and MSE as a way of selecting the optimal parameters. Figure-4 **C** shows the heat maps produced by ChIPanalyser for Dfd in Kc167 cells. From left to right: mean correlation over all regions for each parameters combination; mean MSE over all regions for each parameter combination; the overlap of Pearson Correlation and MSE for each parameters combination. In this example, Pearson correlation plays little to no role in selecting the optimal set of parameters and the selection is driven by MSE.

Based on these preliminary results, the initial version of ChIPanalyser needed to be improved. The two hypothesised culprits for this behaviour were ChIP background noise and the goodness of fit metrics used. The latest version of the package includes noise filtering methods and different goodness of fit metrics to choose from. But how well do these additions perform? Do they actually improve the package and the performance of the model?

### FILTERING NOISE

As described above, ChIPanalyser provides four noise filtering methods: *zero, mean, median* and *sigmoid*. In order to test the performance of these noise filters on ChIP data, I selected three CTCF data sets: a noisy data set (modEncode 3674), a clean data set (modEncode 2639) and finally a combination of all CTCF data sets in *Drosophila* S2 cells by adding enrichment score together at base pair level. Data sets were normalised prior to combination in order to ensure equal contribution of each data set. Noise filters are applied to ChIP data during the extraction and normalisation of ChIP signal (*processingChIP*). For the purpose of this analysis, only the top 10 regions of each dataset were selected. The optimal parameters were then inferred based on filtered ChIP profiles.

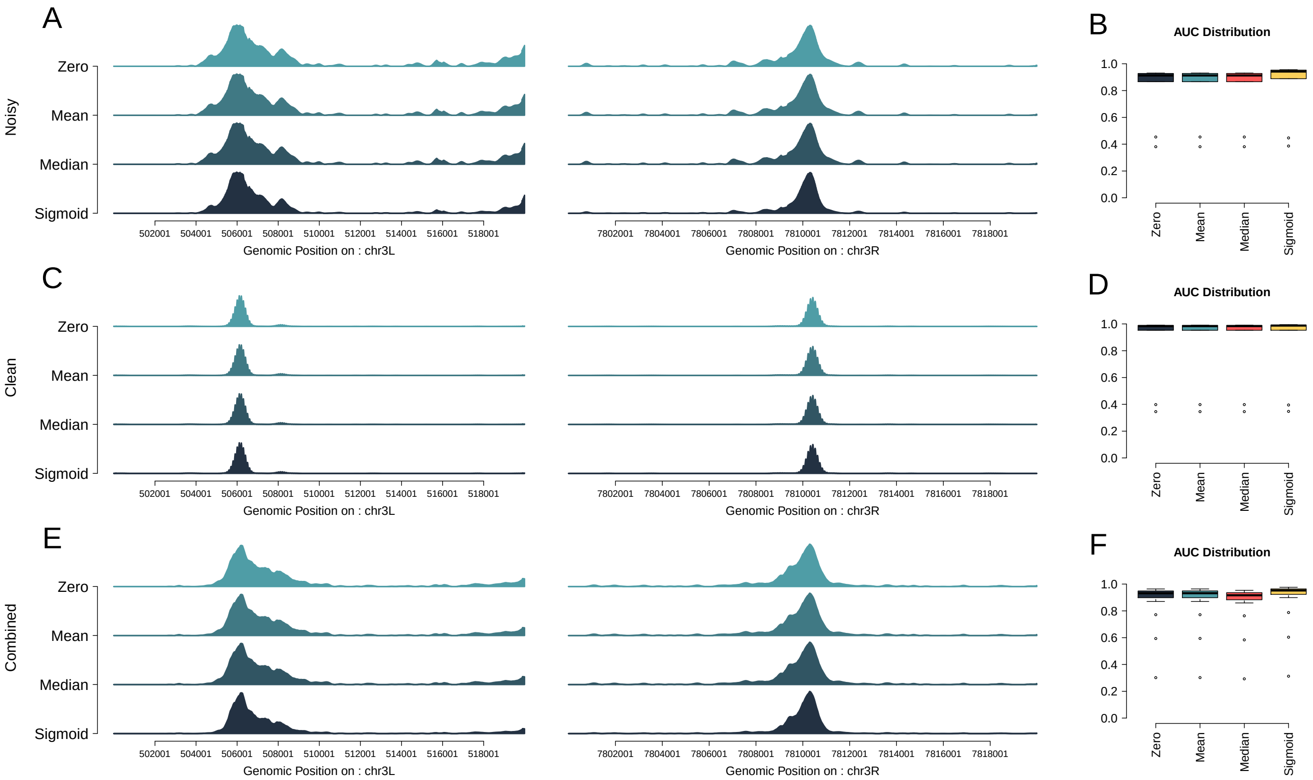


Figure 5: **Noise Filtering methods have little effect on experimental ChIP signal.** In order to improve our predictions, I sought to test four noise filtering methods: *Zero*, *Mean*, *Median* and *Sigmoid*. I tested these methods on three CTCF datasets: (**A-B**) a ChIP-seq dataset with high background noise (modEncode 3674), (**C-D**) a ChIP-chip dataset with very little background noise (modEncode 2639), and (**E-F**) a combination of all ChIP-seq datasets in S2 cells (by adding enrichment signals together at a base pair level) In (**A**),(**C**), and (**E**), I noticed that noise filtering methods have a limited effect on reducing noise. The subsequent effect on predictions was limited but showed a slight improvement when using the sigmoid method as described in (**B**), (**D**), and (**F**).

Figure-5 **A,C**, and **E** shows the effect of each noise filtering method on ChIP profiles. Figure-5 **B,D**, and **F** describes the effect of each noise filter method on the performance of predicted ChIP-like profiles. Generally, the four noise filtering methods used in ChIPanalyser have little effect on ChIP data and only sigmoid filtering slightly improves the predictions (Figure-5 **A/B**). This is especially the case when validation ChIP data sets are noisy. Unsurprisingly , there is little to no effect on the performance of the model in clean data sets as seen in Figure-5 **C/D**.

## INFERRING PARAMETERS: TRIALS AND TRIBULATIONS

The next step was to asses how different goodness of fit metrics would affect the selection of optimal parameters and the performance of the model. In particular, I compared correlations (Pearson, Spearman and Kendall), MSE, Kolmogorov-Smirnov Distance, precision, recall, accuracy, F-score, Matthews correlation coefficient (MCC) and AUC ROC (see Table-1). In addition, I also developed a novel method that describes the ratio of geometric shared area between curves and difference in area between curves. The same three CTCF datasets as described above were used (clean, noisy and combined). The resulting heat maps saw the emergence of two classes within these metrics: (i) similarity metrics that describe how similar the two curves are (correlation coefficients, precision, MCC, Accuracy, F-score and AUC ROC) and (ii) dissimilarity metrics that measure of how different two curves are (MSE, geometric ratio, recall and Kolmogorov-Smirnov distance). The results showed that depending on the metric used, the optimal set of parameters would vary significantly, but each of the two classes (similarity and dissimilarity metrics) displayed similar yet not identical values for the optimal parameters (see Figure-6 **A-F**).

Goodness of fit metrics influence the way the model selects the optimal parameters, but how does this translate to the individual predicted ChIP profile level? I further investigated this behaviour at the individual loci using the same three CTCF datasets (clean, noisy and combined). All associated scores were computed based on the comparison between the predictions and ChIP data in 100 bp windows. Figure-6 **G-I** shows that similarity metrics (dark blue shades) tend to be less prone to false positive peaks but miss the actual ChIP signal enrichment within the peak (the height of the peak). On the other hand, dissimilarity metrics (light blue shades) generate far more false positives but accurately recover the height of the peaks.

Overall, the best performing metrics were AUC ROC, MSE and geometric ratio. AUC ROC occasionally missed peak height completely but seemed to recover peak location fairly accurately, while geometric ratio and MSE rarely missed peak height but also tended to predict a higher number of false positive peaks. For much of the following analysis, I used AUC ROC and MSE, since they are more widely used estimators and performed best. More specifically, MSE was used as the training metric to select optimal set of parameters and AUC for validation.

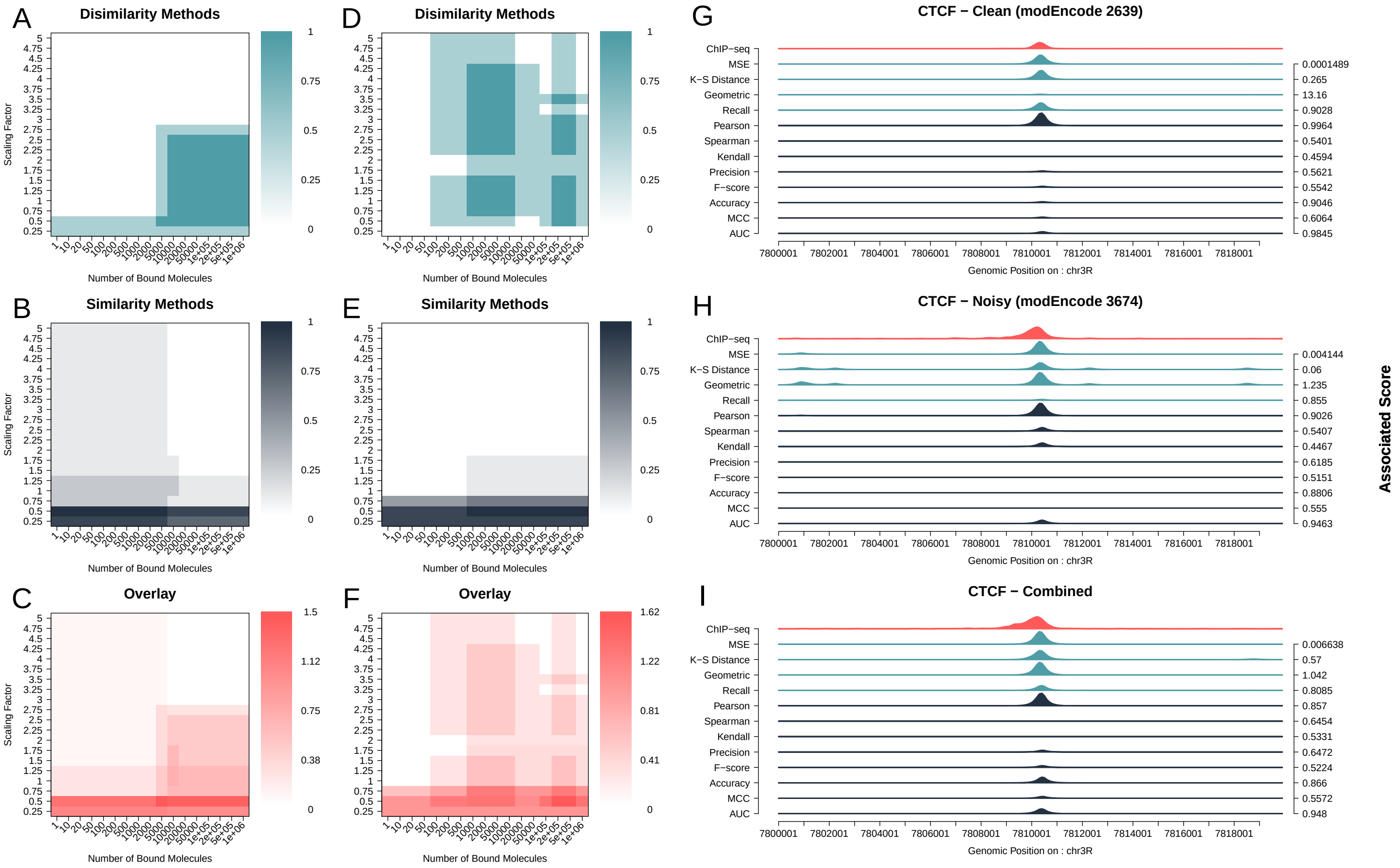


Figure 6: **Goodness of fit Methods are context dependant.** (**A-F**) Heatmaps show the overlap of best performing (top 10 %) combination of parameters for similarity and dissimilarity method as well as an overlay of all methods. I produced these heatmaps using the noisy (**D-F**) and clean (**A-C**) data sets. (**G**) ChIPanalyser correctly predicts CTCF peaks in a clean ChIP dataset (modEncode 2639) for the majority of metrics used. (**H**) For a noisier dataset (modEncode 3674) , dissimilarity metrics capture the height of the peak but also tend to show a high rate of False Positive peaks. In contrast, similarity metrics accurately predicted the location of the peak, but tend to fall short in terms on peak height. (**I**) Combining several ChIP replicates (all ChIP-seq datasets in S2 cells) does not reduce the rate of False Positive peaks for similarity metrics.

## ASSESSING PERFORMANCE OF THE MODEL

To evaluate the performance of the model, a chromosome withholding set-up was used. The principle behind this set up is to ensure that there are no biases between chromosomes and the model is not over-fitted to the data set used for training. The model was trained on the top ten region of chromosome 3R in a BEAF-32 data set (modEncode 922) and the optimal parameters were selected. The parameter combination was then directly plugged into the model and validated on the top ten regions of chromosome 3R (excluding regions used for training) as well as the top regions of chromosome 2R.

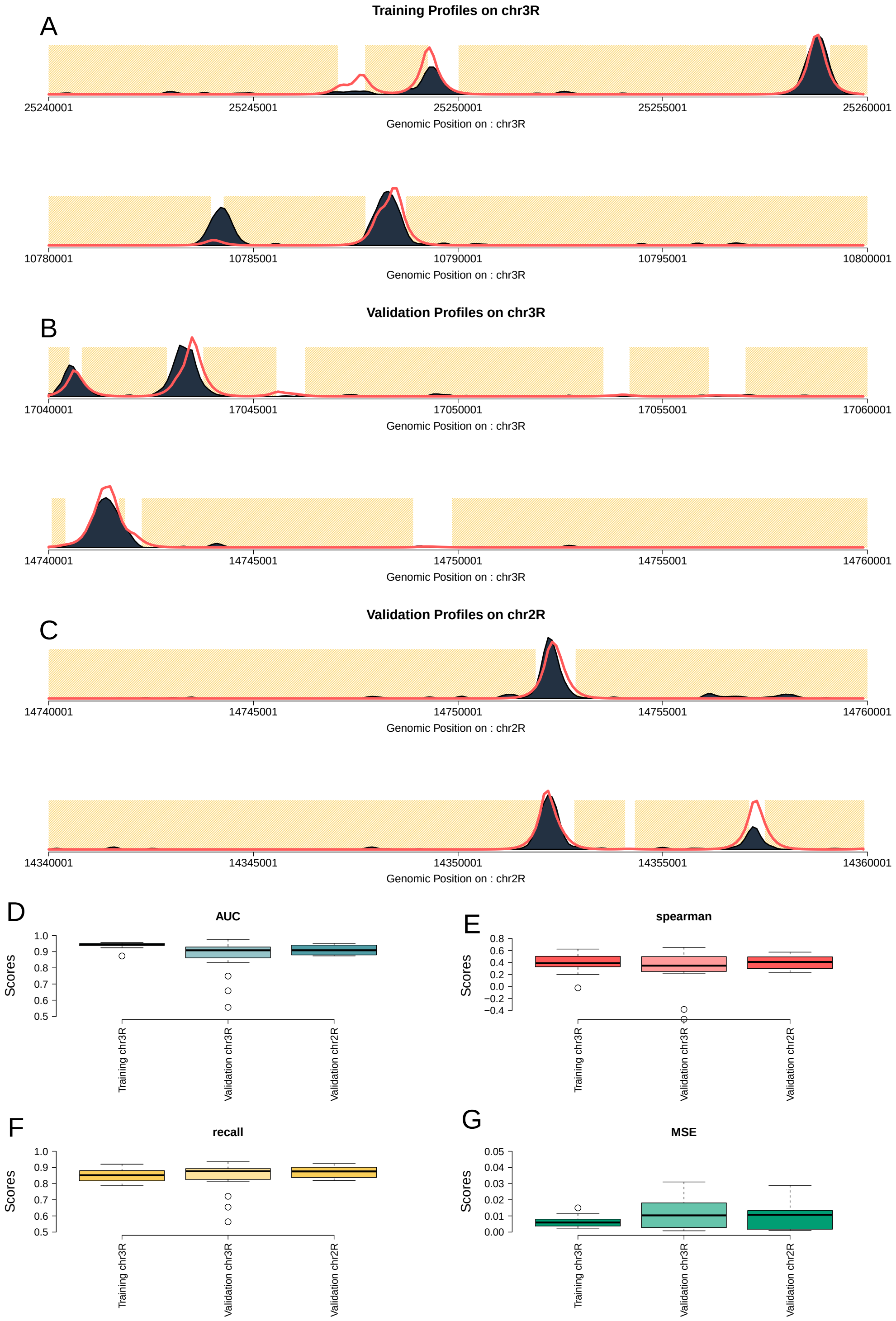


Figure 7: **Chromosome withholding setup for model validation.** I analysed BEAF-32 ChIP in S2 cells (modEncode 922) and I trained ChIPanlayser on the top 10 regions on chromosome 3R. I then validated the model on the top 20 regions on chromosome 2R and, for comparison, on top 10 regions on chromosome 3R that did not contain the training set. (**A**) shows example profiles obtained during training. (**B**) shows validation profiles obtained on chromosome 3R. (**C**) are profiles obtained during validation on chromosome 2R. Finally, (**D-G**) are the associated metrics for training and validation: AUC, Spearman correlation, recall and MSE respectively.

The model accurately recovers peaks in the training set as well in both validation sets (see Figure-7). As described previously, red lines represents the predicted ChIP-like profile, the dark blue shaded area represents experimental ChIP data and finally the yellow boxes are regions of inaccessible DNA. Figure-7 **A-C** are predicted profiles on the training set, same chromosome validation and different chromosome validation respectively.

When comparing different metrics between training and validation, the goodness of fit metrics remained similar between each set as seen in Figure-7 **D-G**. AUC, Spearman Correlation, recall and MSE were used as a point of comparison between metrics.

## COMPARING TO OTHER METHODS

In recent years, numerous tools for TF binding prediction have been developed. In 2016, a TF binding prediction competition was initiated by DREAM-challenges. Since, numerous tools have been developed and published. Some of the most popular tools include FactorNet, Anchor and Catchitt[Quang and Xie, 2019, Li et al., 2019, Keilwagen et al., 2019]. Previously, other tools have provided ways of predicting TF binding sites such as PIQ and msCENTIPEDE [Sherwood et al., 2014, Raj et al., 2015b]. These tools all have in common the use of DNA accessibility as a way of improving the predictions of TF binding sites. It should be noted that PIQ and msCENTIPEDE do not require genomic occupancy data but only requires a PWM and DNA accessibility data.

In order to asses the quality of the model, ChIPanalyser was compared to other tools and frameworks. For this aspect of the analysis, both FactorNet and Anchor were excluded. FactorNet was written using python2.7. Both python and FactorNet will be depreciated in the near future and numerous dependencies have moved away from python2.7 modules making FactorNet unstable and difficult to use. Furthermore, FactorNet as well as Anchor,Catchitt and msCENTIPEDE, would only recognise data from *Homo sapiens*. Anchor also suffered from challenging coding styles making the use of this tool extremely complicated. Both msCENTIPEDE and PIQ do not require genome occupancy data. Instead, binding sites are determined by using PWMs and DNA accessibility in the form of BAM files. These tools do not offer a training and validation set-up but rather a way of extracting TF binding sites based on DNA accessibility data. Catchitt offers a training and validation set-up where a data set can be trained on one chromosome and validated on another. I provided a break-down of each tool in Table-2.

To overcome these limitations, I selected a data set from the DREAM challenge series. Genome occupancy data for CTCF and DNA accessibility data from human astrocyte cells were selected. Chromosome 11 was used as the training chromosome and chromosome 18 was used as the validation chromosome. ChIPanalyser was trained on top ten regions of chromosome 11 and validated on the top ten regions of chromosome 18. Unfortunately, truncating data lead to computational issues or was simply infeasible with other tools. msCENTIPEDE and PIQ used the entire genome before outputting predicted TF binding sites. Catchitt on the other hand was trained on the entire chromosome (chromosome 11) and validated on chromosome 18. To demonstrate the packages ability to predict TF binding using low data input, I compared the performance of each tool by extracting prediction at the validation regions selected by ChIPanalyser.

| | ChIPanalyser | Catchitt | FactorNet | Anchor | PIQ | msCENTIPEDE | TRAP |
|---|---|---|---|---|---|---|---|
| **Language** | R | java | python 2.7 | python3.6/perl 5.1 | R | python 3.6 | C and web app |
| **Organisms** | All* | All | Human | Human | All* | Human | All |
| **Training & Validation** | Yes | Yes | Yes | Yes | No | No | No |
| **Plotting** | Yes | No | No | No | No | Limited | Yes |
| **Skill Level** | Low | Intermediate | High | High | Low | Intermediate | Low |
| **Availability** | Bioconductor | GitHub | GitHub | GitHub | bitbucket | GitHub | Dedicated Web page |

Table 2: **Tool and framework comparison.** I provide a breakdown of a few notable tools and frameworks for TF binding prediction. * Provided that these genomes are available through the Bsgenome package

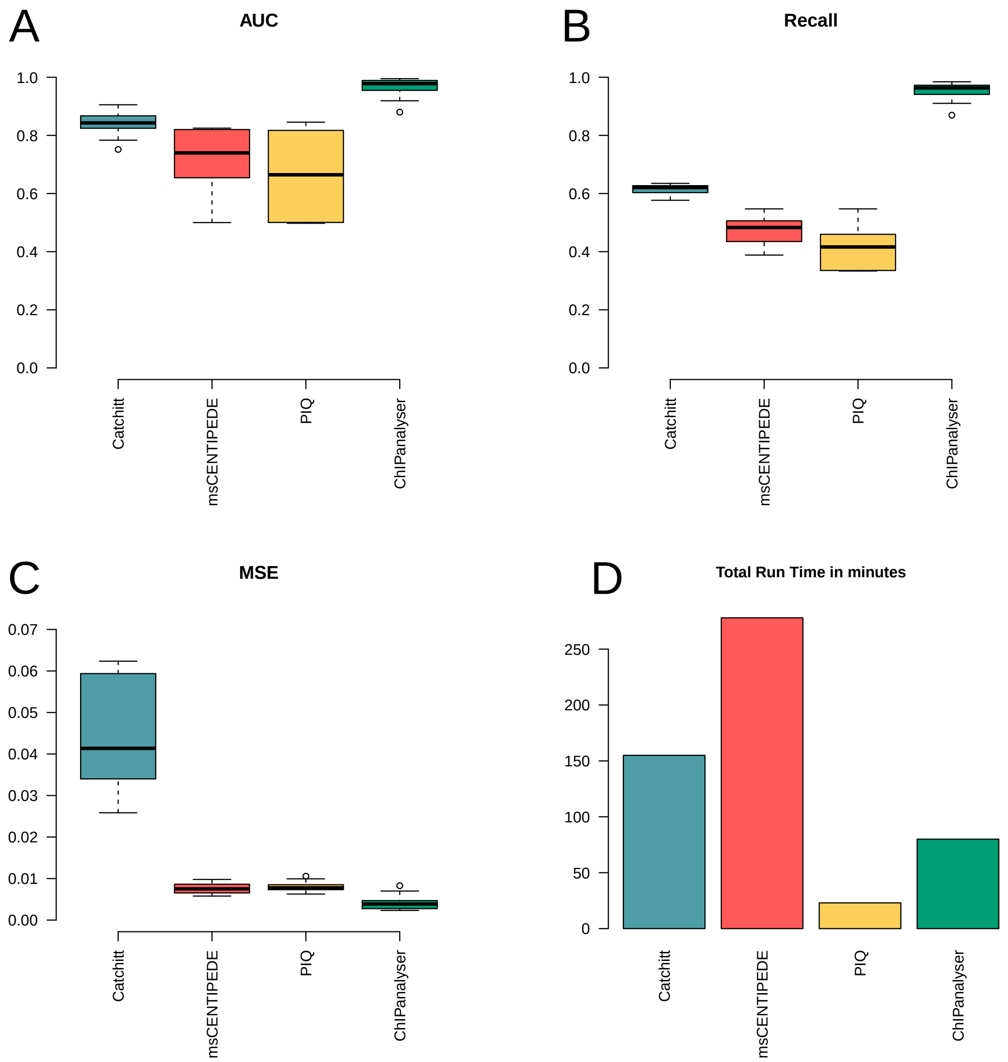


Figure 8: **Performance Comparison between TF binding predictions tools** ChIPanalyser performs well compared to other TF binding prediction tools using low input data for training. (**A**) shows AUC scores between Catchitt, msCENTIPEDE, PIQ ,and ChIPanalyser over the selected validation regions in chr18 on *Homo sapiens*.(**B** and **C**) are respectively recall and MSE over validation regions for each tool. Finally, (**D**) breaks down total run time for training and validation for each tool.

Figure-8 A-C show the overall performance of each tool over validation regions for AUC, recall and MSE. Figure-8 D compares total run time between each tool using. ChIPanalyser out-performed every tool in terms of predictive ability. It should be noted that ChIPanalyser uses an extremely stringent method to asses goodness of fit. Predicted profiles are compared to genome occupancy data by comparing predicted profiles to experimental profiles at a base pair level.

Tools developed through DREAM Challenges were scored only based on overlaps between predicted sites and ChIP peaks in bins of 200 bp. This approach negates the effect of background noise when scoring each model and disregards the local enrichment of the predicted peak. For this purpose, each tool were compared to each other using the approach developed in ChIPanalyser. When necessary, binding site probabilities were smoothed in order to produce a ChIP-like profile (100bp rolling mean window).

## DISCUSSION

### *Background noise and experimental artefacts remain a challenge in TF binding predictions*

I found that many ChIP datasets suffer from significant background noise that would reduce our ability to accurately assess the goodness of fit of the model. Despite my approaches to reduce background noise, it seems that ChIP data will always suffer from unspecific DNA pull-down [Teytelman et al., 2013]. Another possibility is that the noise in ChIP signal could be the result of unspecific binding of TFs to DNA followed by one-dimensional random walk along the genome [Zabet and Adryan, 2012, Hammar et al., 2012]. Nevertheless, the washing steps in the ChIP protocol would remove this non-specific binding from the final ChIP signal [Landt et al., 2012b]. Sequencing depth also plays a role in limiting background noise and false positive peaks [Sims et al., 2014]. In the context of ChIPanalyser, ChIP data sets with high sequencing depth would increase the goodness of fit of the model as it would be less affected by background signal. Furthermore, ChIP-seq protocols are affected by DNA accessibility as there is an over representation of DNA fragments within open chromatin [Auerbach et al., 2009]. This bias would affect the ability to predict the binding of TFs that preferentially bind to more restrictive or closed chromatin. While most data sets originate from modEncode and follow similar analysis pipelines, it should be noted that different mapping strategies will affect the downstream results and should be chosen with a specific biological question in mind [Fonseca, 2012]. These limitations may also apply to DNA accessibility data. DNase-seq relies on digestion of DNA fragments in nucleosome depleted regions [Song and Crawford, 2010]. DNA fragments are then digested again, attached to linker beads and finally sequenced. Similarly to ChIP, sequencing depth would also affect the quality of the data produced and potentially lead to mismatches between ChIP data and DNA accessibility data. The ChIPanalyser workflow and the subsequent interpretation of results could be affected by such mismatches.

I showed that choosing a goodness of fit method is context dependent. Interestingly, similarity methods (such as correlation, F-score or AUC) had the tendency to correctly call peak location but greatly underestimated the enrichment of the peak (see Figure-6). This behaviour results from the fact that these methods are highly penalised by false positive hits. The scaling factor can be described as how well a TF discriminates between a strong binding site over a weaker one. High values for the scaling factor translate to poorer ability for the TFs to discriminate between high and low affinity sites, which leads both to a higher number of false positive peaks and the model picking up smaller peaks. The number of bound molecules on the other hand, tend to affect the height of the peak (relative local enrichment). Similarity methods would avoid high values for N and $\lambda$ as this would penalise their goodness of fit score more severely as opposed to dissimilarity methods (see Figure 1). Choosing the right method will depend on the question at hand and similarity methods could be used to determine peak location, while dissimilarity metrics would be more appropriate to investigate the TF local enrichment.

*Model limitations and Core assumptions*

ChIPanalyser relies upon the statistical thermodynamic model presented by Zabet and Adryan [Zabet and Adryan, 2015]. Previously, it was demonstrated that using PWM threshold in order to select binding sites improved the performance of the model. While the performance of the model is improved, PWM thresholds might not display the complexity of the biology at hand. In the case of ChIPanalyser, only the top 30% of binding sites (based on their respective PWM scores and the default threshold value) are considered occupied by TFs. It has been suggested that lower affinity binding sites play a significant role in TF binding and in transcription itself [Farley et al., 2015]. The initiation of transcription could be the result of dose dependant binding of TFs to lower affinity binding sites [Spivakov, 2014]. Using ChIPanalyser's default PWM threshold value would fail to consider lower affinity binding sites. However, the entire spectrum of binding sites can be considered by setting the PWM threshold to 0. ChIPanalyser further increases this bias towards high affinity sites by using a PWM score scaling factor $\lambda$. The scaling factor describes how well a TF discriminates between low and high affinity sites. This scaling factor is only applied to sites above threshold, further increasing the "strength" of high affinity binding sites compared to lower affinity sites. While it would be interesting to test ChIPanalyser without a PWM threshold, it should be noted that the resulting profiles are likely to be of poorer quality. The model heavily penalises overestimation or underestimation of ChIP enrichment. Including all possible binding motifs will likely results in one of these scenarios.

Previous results also suggested that considering DNA as either open or closed was sufficient to explain TF binding. The same approach was used through out this chapter. However, a more in depth analysis of DNA accessibility is described in the next chapter.

### *Assessing ChIPanalyser as a tool*

ChIPanalyser was shown to be a powerful tool to predict TF binding. The performance of the model and the package out-performs other TF binding frameworks. However, it should be noted that the high performance of ChIPanalyser compared to other tools is a direct consequence of the way goodness of fit scores were computed. As described above, ChIPanalyser computes goodness of fit scores by comparing predicted profiles to experimental profiles in 100bp windows. This approach also accounts for signal enrichment. If ChIPanalyser predicts a peak at a given location but fails to recover signal enrichment or conversely overestimates signal enrichment, ChIPanalyser goodness of fit scores will be heavily penalised. This approach penalises other TF binding frameworks as they were not designed with signal enrichment prediction in mind. Catchitt for example was designed using DREAM challenge guide lines suggesting that a peak was correctly predicted if the predicted binding site was within 200 bp of that peak. Furthermore, ChIPanalyser performs at its best when only the strongest binding sites are considered. Validating the model on the top ten regions of chromosome 18 ensures that ChIPanalyser would perform with high accuracy and reduce the amount of background noise considered. However, this approach reduces the performance of other frameworks in favour of ChIPanalyser.

ChIPanalyser was designed to not only predict TF binding but also hopefully shed light on the mechanisms driving the binding of TFs to DNA. While I recognise that by using a different scoring method and validation set up, the performance of the other frameworks would be greatly improved, I selected this approach as it is better suited to assess the quality of the model underlying ChIPanalyser.

## CHAPTER CONCLUSION

In this chapter, I demonstrate that ChIPanalyser can accurately predict TF binding in both *Drosophila melanogaster* and *Homo sapiens*. Compared to other tools, ChIPanalyser performs well against other competing frameworks in both accuracy and run-time. This is possible thanks to the extremely low data input requirements in order to train the model. I show that goodness of fit metrics are context dependant and in the context of ChIPanalyser the question at hand should determine the choice of metric. Furthermore, experimental noise generated by genome occupancy profiling remains a recurring problem in TF binding prediction and suggests caution with the interpretation of TF binding predictions.

# 2

# CHIPANALYSER: INSIGHTS INTO BIOLOGY

## CHAPTER SUMMARY

In this chapter, I will show how ChIPanalyser can provide insight into the mechanisms driving TF binding. I will focus on two main aspects: DNA accessibility and TF abundance. I show that DNA accessibility is the main driver of differential binding of three architectural proteins ( CTCF, BEAF-32, and su(Hw) )in the three *Drosophila* cell lines. Using relative RNA levels, I show that for these proteins local concentration plays a lesser role in their respective binding. I also investigate the binding of three Hox TF (Dfd, Abd-b, and Ubx) and demonstrate that ChIPanalyser can recover known binding preferences of these Hox TFs with respect to chromatin accessibility.

## INTRODUCTION

The most fundamental aspect to consider concerning TF binding specificity is the DNA sequence itself. Most TFs exhibit a preferred binding motif. The most common way to describe this motif is in the form of a Position Weight Matrix (PWM); a measure of binding energy between TFs and DNA weighted by the genomic base pair frequency

[Ptashne and Gann, 1997, Spitz and Furlong, 2012, Berg and von Hippel, 1987b, Stormo and Zhao, 2010]. Nevertheless, TFs can have tens of thousands of putative binding sites within the genome, yet they only bind to a few hundred or thousand of them . Previous studies have shown that some TF binding events are concentration dependent [Chu et al., 2009, Kaplan et al., 2011b, Simicevic et al., 2013, Zabet and Adryan, 2015], where varying the concentration of the TF will drive the expression of different sets of genes. However, there are many more sites than binding sites where TF's could bind. This still begs the question: how do TFs distinguish between bound and unbound sites? One way to reduce the number of available sites is to consider DNA accessibility. Are these sites even available for binding in the first place? This assumes that TFs would bind only to sites that are accessible and cannot locate sites within dense chromatin [Klemm et al., 2019, Lamparter et al., 2017]. Nevertheless, there is a certain class of TFs known as pioneer TFs that would ignore accessibility restrictions [Soufi et al., 2015]. More specifically, pioneer TFs can bind sites in closed dense chromatin and subsequently open the chromatin. It was previously shown that statistical thermodynamics can be used to model TF binding to DNA with high accuracy. Considering only binding energy between TFs and DNA (estimated by the PWM and a scaling factor modulating the binding energy), the number of bound molecules to the DNA and DNA accessibility, Zabet and Adryan modelled binding of five TFs in Drosophila embryo [Zabet and Adryan, 2015]. Their results confirmed that, for some TFs, this model is sufficient to explain the majority of observed binding events in ChIP data and they were able to backwards infer number of bound molecules and specificity for five TFs in *Drosophila* embryo (bcd, cad, gt, hb and Kr). I used this model to describe the behaviour of several *Drosophila* TFs: CTCF, BEAF-32, su(Hw), Ubx, Abd-B and Dfd. The results provide a mechanistic interpretation of TF binding behaviour and propose a new classification of these TFs based on fine details of their binding mechanism. In particular, I found that DNA accessibility is the main driver that explains binding of CTCF, BEAF-32 and su(Hw) in three *Drosophila* cell lines (BG3, Kc167 and S2) and that relatively medium changes in the concentrations of these TFs lead to only negligible changes in their binding profiles. I also show that TF binding specificity can be achieved by their capacity to bind to regions with different levels of DNA accessibility. In particular, I show that Ubx, Abd-B and Dfd binding to DNA could be explained by their different capacity to bind dense chromatin, with Ubx binding only in highly accessible chromatin and Dfd and Abd-B binding in denser chromatin.

## DATA SET UP

As described in the previous chapter, ChIPanalyser requires a PWM and a reference DNA sequence as a minimal input. In order to validated the predicted profiles, a set of Genome wide assays are also required. I selected ChIP data sets for three architectural DNA binding proteins: CTCF, BEAF-32, and su(Hw). I also used ChIP data for three Hox TFs in *Drosophila*. A summary of the ChIP data used in this thesis can be found in Appendix A. In this chapter, I show the impact of DNA accessibility. DNA accessibility data was taken from DNase I hypersensitivity (DHS) data in three *Drosophila* cell lines (Kc167, BG3, and S2) [Kharchenko et al., 2010]. For Hox TFs, I used ATAC-seq data in Kc167 cells. Both ChIP-seq and ATAC-seq data sets were produced by the same lab [Porcelli et al., 2019]. The total amount of accessible DNA is similar in the three cell lines and *Drosophila* embryos (see Figure 9). In order to rescale TF abundance between cell lines, I used RNA-seq data from [Lee et al., 2014a]. RNA-seq relative abundance was used to rescale the estimated number of bound molecules from one cell line to another.

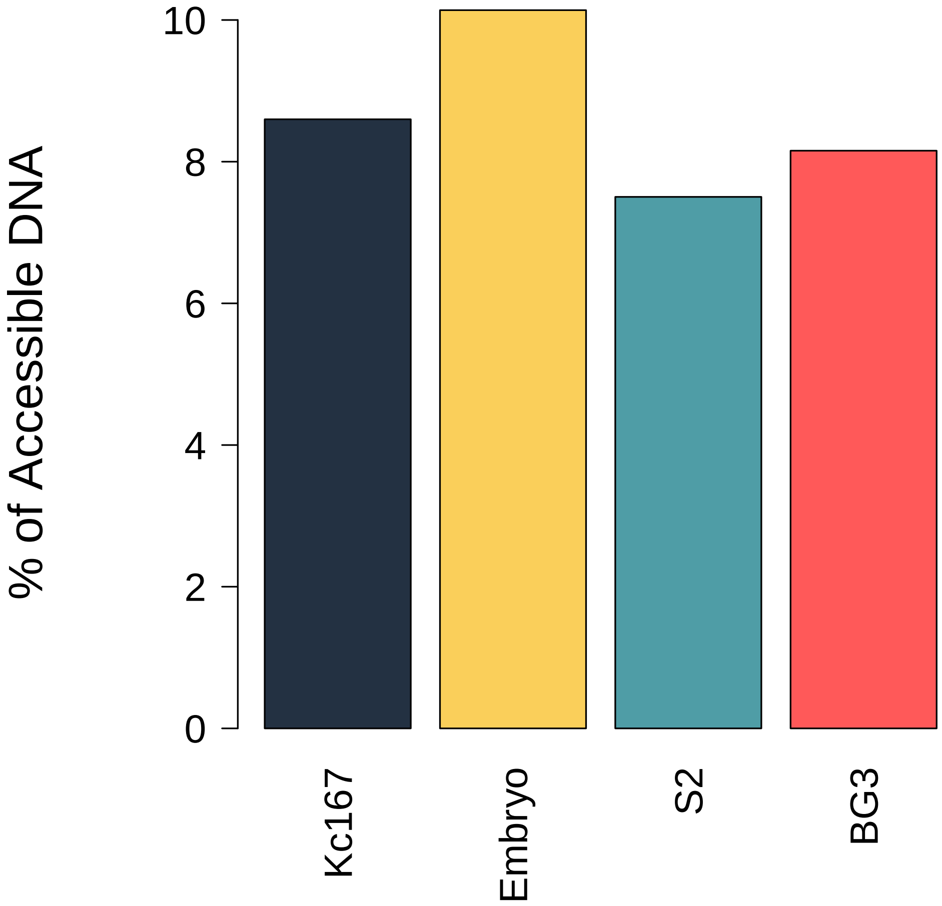


Figure 9: **Total accessibility between cell lines.** Accessibility for three cell lines was estimated from DNase Hypersensitivity Sites (DHS) by extracting DHS broad peaks . Kc167 cells displayed a slightly higher proportion of accessible DNA compared to S2 and BG3 cells. The proportions of accessible DNA remains similar to the proportion of accessible DNA in *Drosophila* embryos using the method proposed by [Kaplan et al., 2011a].

In order to ensure region consistency between each data set, I selected a total of 3293 20kbp regions that contained at least one peak of any architectural protein in any data set. Each region should also contain at least 100bp of accessible DNA and not contain any black listed regions as described by UCSC. This set of loci was used for all data sets and all architectural proteins. The same process was applied to Hox TFs and resulted in a total of 3838 regions.

For architectural proteins, ChIPanalyser was trained on the top ten regions for each data set as described in the previous chapter. In order to demonstrate the role of DNA accessibility, the top ten regions were trained with no accessibility data, continuous DNA accessibility and DHS only. Continuous accessibility was produced by using min/max normalised read pile-up scores.

## THE NUANCED ROLE OF DNA ACCESSIBILITY IN TRANSCRIPTION FACTOR BINDING

### *DNA accessibility influences the binding of Architectural Proteins*

Steric hindrance can influence the binding of some TFs to DNA, meaning that a TF molecule would only bind stretches of DNA if they are accessible. Any given genomic region can be considered either accessible or inaccessible and that is sufficient to explain the binding profiles of most TFs [Zabet and Adryan, 2015]. Here, I selected accessible DNA based on DNase Hypersensitivity Sites (DHS) in three Drosophila cell lines (Kc167, S2 and BG3) [Kharchenko et al., 2010]. In these circumstances, DNA was either considered accessible (score of 1) or inaccessible (score of 0). As a point of comparison, I also considered all DNA to be accessible (No Access - all regions are assigned a score of 1) and also used a min-max normalised DNase score as continuous DNA accessibility levels (a continuous value between 0 and 1). I focused my analysis on three TFs: CTCF, BEAF-32 and su(Hw). I trained the model on the top 10 regions for each data set. Then, I validated the results using the optimal parameters selected during training. Validation was carried out on the top 100 regions for each dataset (excluding the ones used for training). Figure-10 shows that for BEAF-32, the binding predictions were improved when considering DNA accessibility. Nevertheless, su(Hw) and CTCF displayed a different behaviour, as the mean AUC decreased when DNA accessibility was considered for most ChIP-seq datasets (Figure-10 **A-B**). This difference is especially striking in the case of su(Hw). The performance of the model drastically improves when all DNA was considered accessible or when I used continuous values for DNA accessibility. CTCF showed a similar trend although improvement was not as striking as in the case of su(Hw). This would indicate that only a small number of CTCF peaks are located in closed chromatin regions that display intermediary levels of accessibility. While DNA accessibility seems to play a role in the quality of our predictions, I also observed that the number of bound molecules ($N$) and scaling factor ($\lambda$) show a reduced influence when DNA accessibility is considered for CTCF (Figure-10). In particular, I observed less variation in MSE for different sets of parameters, when DNA accessibility was included, i.e., larger circles indicate that number of bound molecules and $\lambda$ have a more important role in TF binding, while smaller circles indicate that they have a less important role. This opposite trend is seen in the case of su(Hw) where N and $\lambda$ show an increased influence when DNA accessibility is considered. BEAF-32 on the other hand is negligibly influenced by N and $\lambda$ independently of whether or not I consider DNA accessibility.

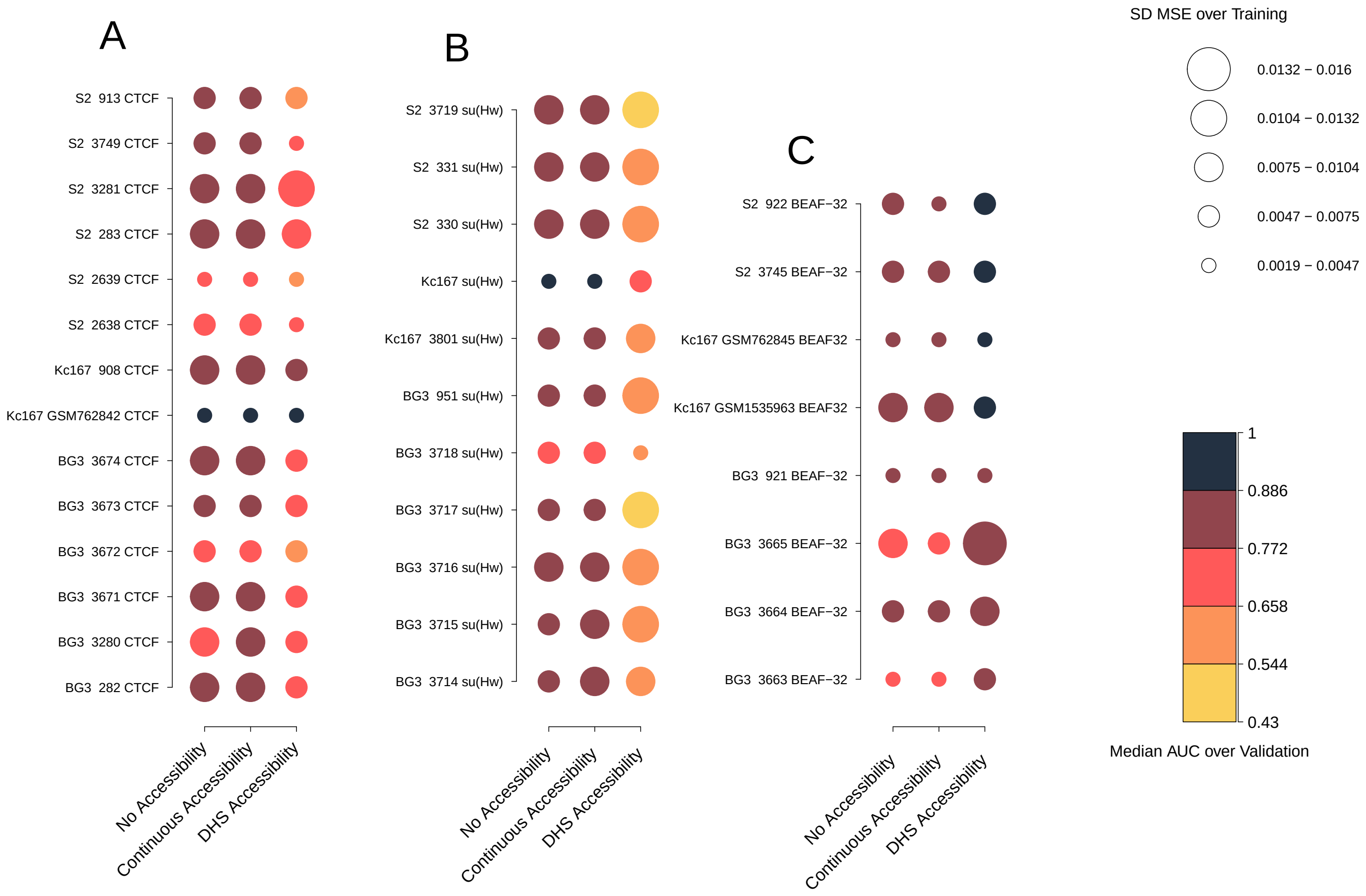


Figure 10: **DNA accessibility, number of molecules and binding energy have different roles in TF binding.** I selected optimal parameters by minimising MSE over the training set and then computed the median AUC scores over the top 100 regions in the validation set. I considered different ChIP replicates in S2, Kc167 and BG3 cells for: (A) CTCF, (B) su(Hw) and (C) BEAF-32. Darker colours indicate higher AUC scores, while lighter colours lower AUC scores. I also investigated the influence of number of bound molecules and scaling factor on TF binding by computing the standard deviation of MSE scores for all combination of parameters over the training set. Smaller circles indicate less variability in MSE when different parameters are used and larger circles more variability.

To factor in for potential differences in the capacity of the model to predict binding in regions with strong or weak ChIP signal, I trained ChIPanalyser on the top 10 regions for each data set and then selected the top 20, 50, 100, 150, 200, 500, 1000 and

3283 regions for validation (excluding regions used for training). I looked at how the median AUC scores (over all data sets) changes when regions with weaker binding are included in the analysis or when DNA accessibility is considered. For each number of regions selected for validation and for each data set, I subtracted the mean AUC score when no accessibility was considered from the AUC score with DHS accessibility (Delta mean AUC in Figure-11). First, I observed that CTCF exhibited a slightly lower AUC score when DNA accessibility was considered. The decrease in AUC scores observed upon considering more regions (see Figure-12 **A**,Figure13 **A** and Figure-14 **A** ) implies that CTCF binds preferentially to genome hotspots. CTCF binding is better explained at strong binding sites when all DNA is considered accessible. The effect of DNA accessibility decreases as the number of regions for validation increases. In contrast to CTCF, BEAF-32 displayed higher AUC scores when DNA accessibility was included, supporting the previous findings (see Figure-11 **B/E**) . BEAF-32 AUC scores were not affected by the increase in the number of regions (see Figure-12 **B**,Figure13 **B** and Figure-14 **B**), which means that BEAF-32 binding is not influenced by the number of regions selected. In other words, BEAF-32 would bind anywhere along the genome as long as it has an accessible site. Contrarily to CTCF, the binding of BEAF-32 is not susceptible to binding "strength" and there does not seem to be biological differences between strong binding sites and weaker binding sites. In this context, I propose BEAF-32 as a global binder and CTCF a hotspot TF. Furthermore, Figure-11 **C** and **F** shows that there is a strong and statistically significant ( $p < 0.05$ ) reduction in AUC score for su(Hw) when DNA accessibility is included, which indicates that su(Hw) would bind in less accessible DNA ( (see Figure-12 **C**,Figure13 **C** and Figure-14 **C**)). While, su(Hw) did not generally perform well when DNA accessibility is considered, the performance of the model increase when all DNA was considered accessible. Furthermore, an increase in number of regions selected for validation displayed a slight increase in Delta mean AUC score. These results suggest that the majority of su(Hw) binding sites are found in inaccessible DNA and that this tendency decreases with binding strength. It should be noted that the strong increase in Delta Mean AUC score when all regions are considered could be the results of many regions not containing any su(Hw) sites. In this case, DNA accessibility does not influence the performance of the model as there are no binding events to predict.

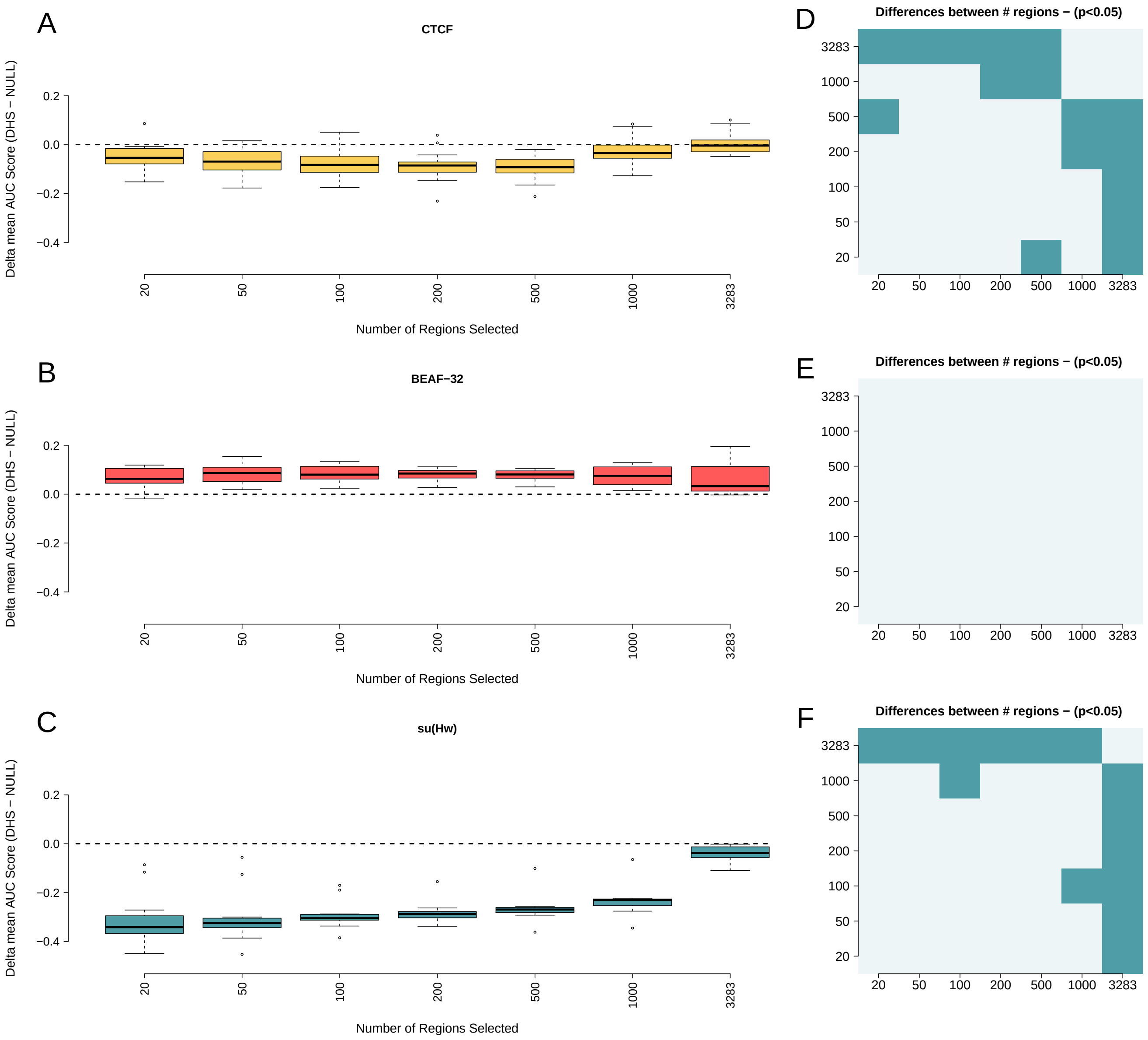


Figure 11: **Number of selected regions sheds light on TF behaviour.** (**A-C**) Boxplot representing the difference in AUC (over validation sets) between the model with and without DNA accessibility for several biological replicates and different number of selected bins. (**D-F**) T-test to assess whether the differences are statistically significant (blue indicates statistically significant differences, while light grey represent non significant combinations). (**A** and **D**). Predictions of CTCF binding shows CTCF's ability to bind to less accessible DNA. The effect of DNA accessibility decreases as the number of regions used for selection increases. Predictions of BEAF-32 binding are improved by DNA accessibility and are not affected by number of regions selected.(**B** and **E**). su(Hw) performs better when all DNA is considered accessible (**C** and **D**)

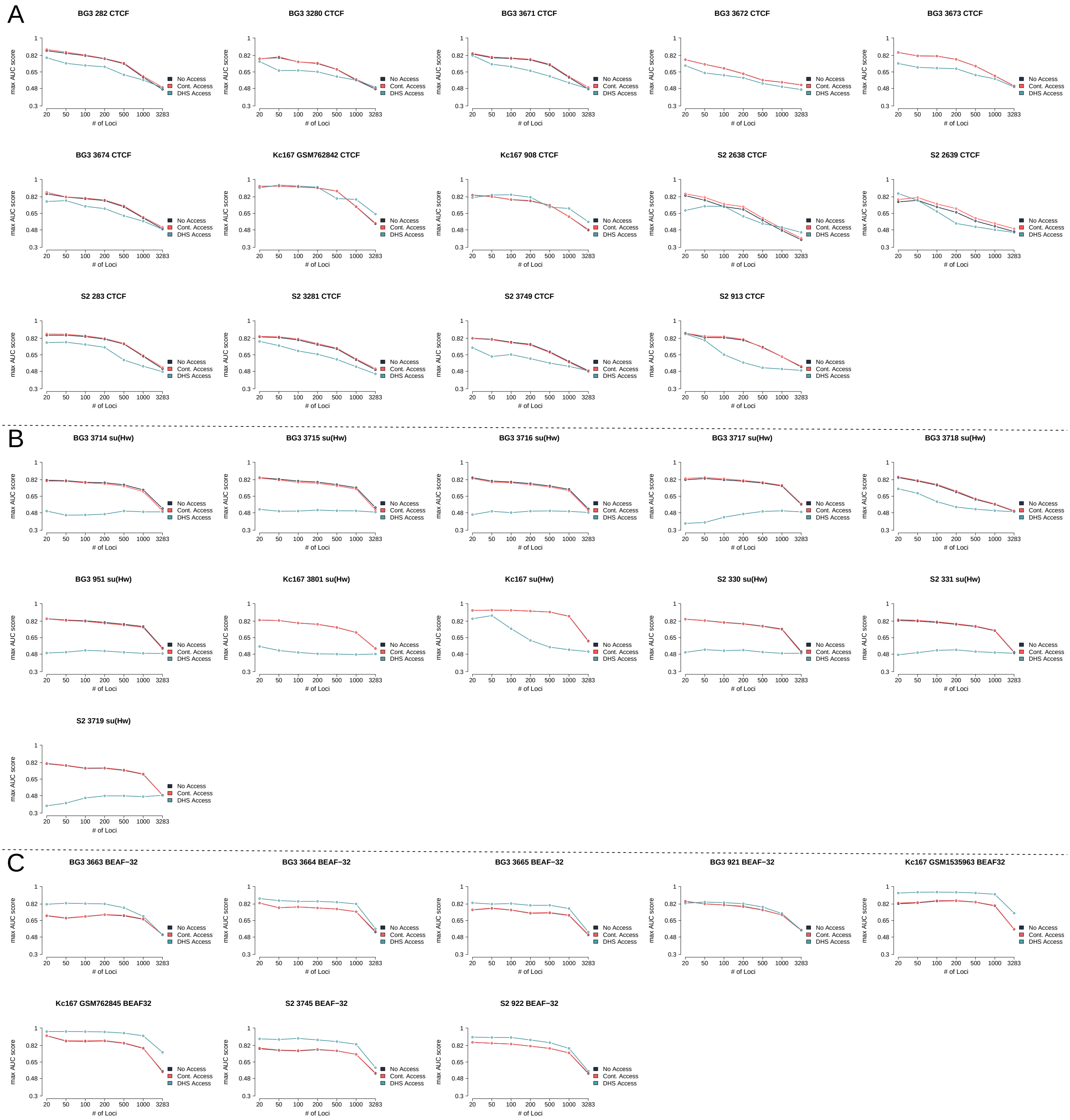


Figure 12: **Increasing the number of regions used during validation sheds light of TF behaviour and binding preferences.** (**A**), (**B**) and (**C**) describe the maximum AUC for each data set and each TF as the number of regions used for validation increase. CTCF decreases in predictability as the number of regions increase while BEAF-32 remain consistent. su(Hw) show a slight decrease in performance only when all DNA is considered accessible.

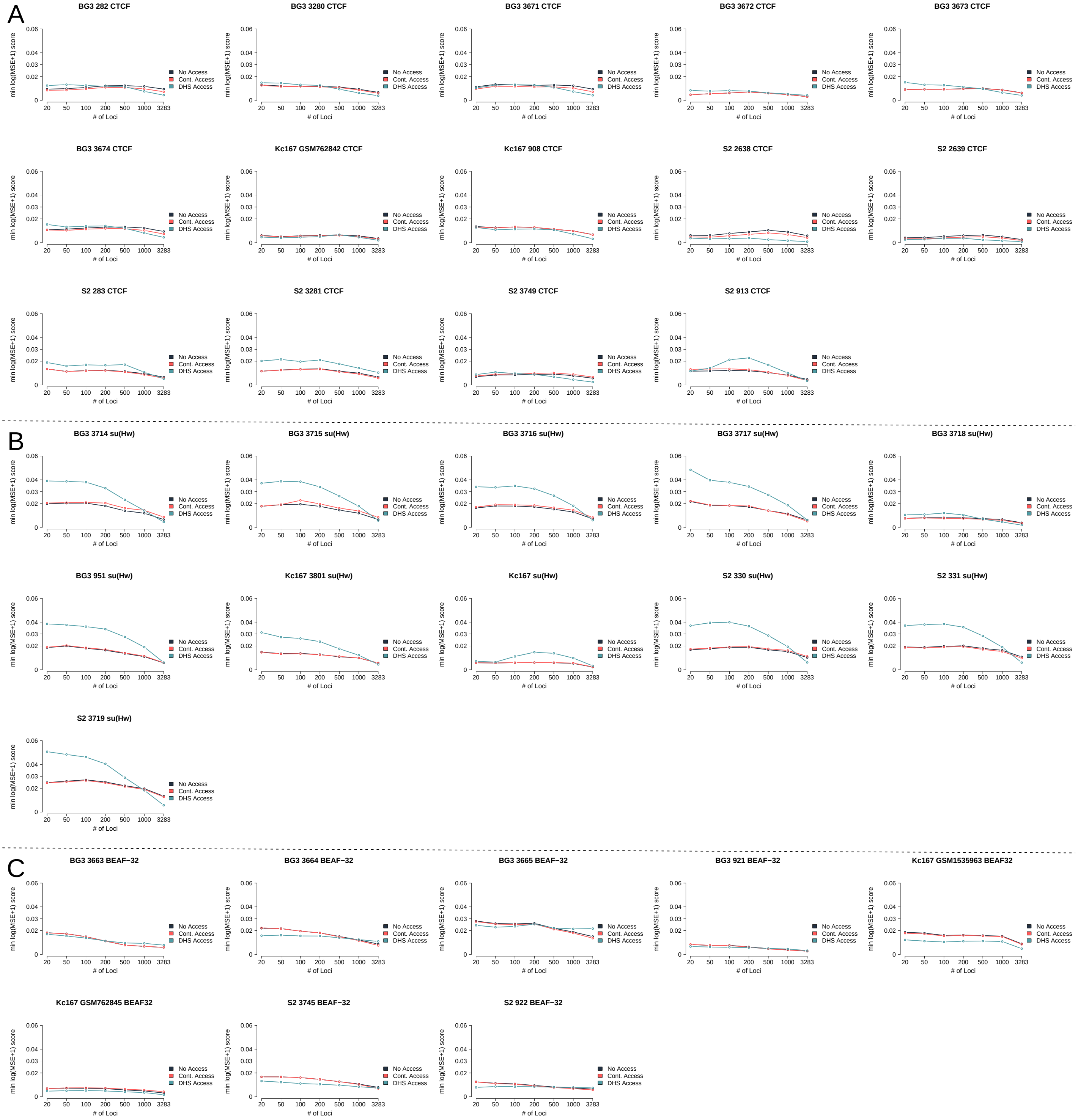


Figure 13: **Increasing the number of regions used during validation sheds light of TF behaviour and binding preferences.** (**A**), (**B**) and (**C**) describe the minimum MSE for each data set and each TF as the number of regions used for validation increase. CTCF decreases in predictability as the number of regions increase while BEAF-32 remain consistent. su(Hw) show a slight decrease in performance only when all DNA is considered accessible.

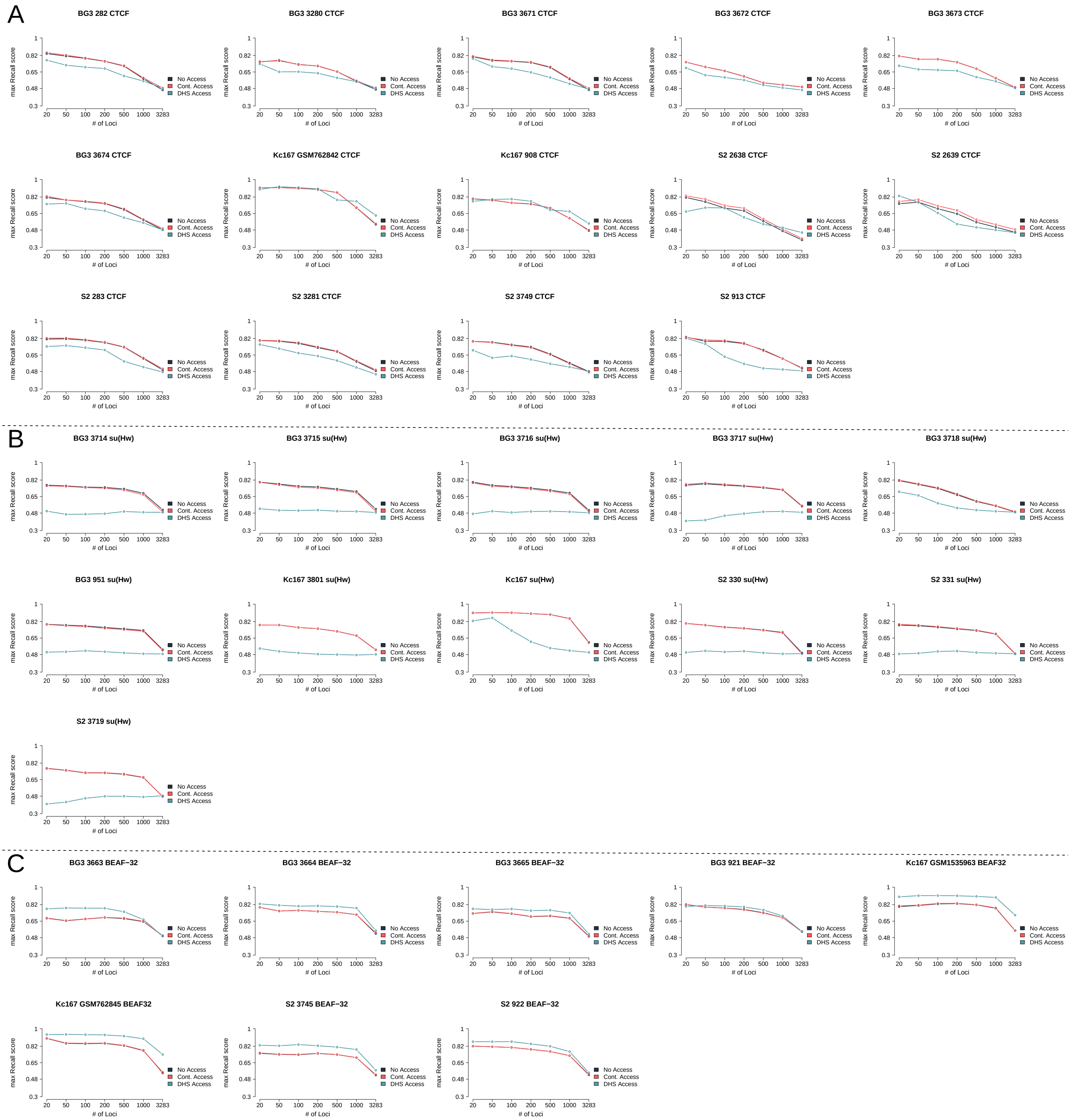


Figure 14: **Increasing the number of regions used during validation sheds light of TF behaviour and binding preferences.** (**A**), (**B**) and (**C**) describe the maximum Recall for each data set and each TF as the number of regions used for validation increase. CTCF decreases in predictability as the number of regions increase while BEAF-32 remain consistent. su(Hw) show a slight decrease in performance only when all DNA is considered accessible.

### *DNA accessibility driving HOX Transcription factor binding*

Hox proteins are key players during development. Recently it has been suggested that Hox proteins show different binding preferences with respect to DNA accessibility [Porcelli et al., 2019] . Most notably, Ubx and Abd-A would bind predominately in open chromatin, while other Hox TF (Lab, Pg, Dfd, Scr and Abd-B) would prefer closed chromatin. I selected three Hox TFs (Ubx, Dfd and Abd-B) and ran the model using different levels of DNA accessibility. DNA accessibility levels were selected based on quantile distribution of ATAC-seq scores (QDA - Quantized Distribution Accessibility). This means that higher QDA scores lead to fewer regions being marked as accessible. I trained the model on the top ten regions selected from the 3838 selected loci for the Hox analysis for each QDA. The results show that Ubx exhibits a preference towards open chromatin. In Figure-15, the maximum AUC score for Ubx increases with the increase of the QDA score. Dfd and Abd-B on the other hand were not strongly influenced by QDA. This means that these TFs can bind in inaccessible DNA. According to the model, Ubx performed best with 0.99 QDA (top 1% ATAC-seq scores - AUC 0.928), while Abd-B and Dfd with 0.95 QDA (top 5% ATAC-seq scores) and 0.8 QDA (top 20% ATAC-seq scores) respectively (see Figure-15 **B**). It should be noted that these scores are extracted from the training set as the goal was to understand how QDA would effect the training of the model. I then validated our model on the top 100 regions (excluding the ones used for training) using the optimal set of parameters inferred during training and plotted the predicted profiles for Hox TF (see Figure-15 **C**,**D**, and **E**).The model recovers the position of peaks accurately especially for Ubx (see Figure-15 **C**). While for Dfd and Abd-B most of the peaks are detected, their height is not always an accurate representation of the strength of the ChIP-seq signal (see Figure-15 **D**/**E**). Hox TFs are known to display cooperative interactions and there are reports that both Dfd and Abd-B have a higher number of sites in the bound peaks, suggesting they bind cooperatively to open the chromatin. Furthermore, Table-3 show the optimal parameters selected for each TF with its optimal QDA. Generally, the model overestimates the number of bound molecules. This would suggest that the model is overcompensating for the lack of co-factors by increasing the number of estimated bound molecules. The model does not include cooperative interactions and this could explain the reduced performance for Dfd and Abd-B.

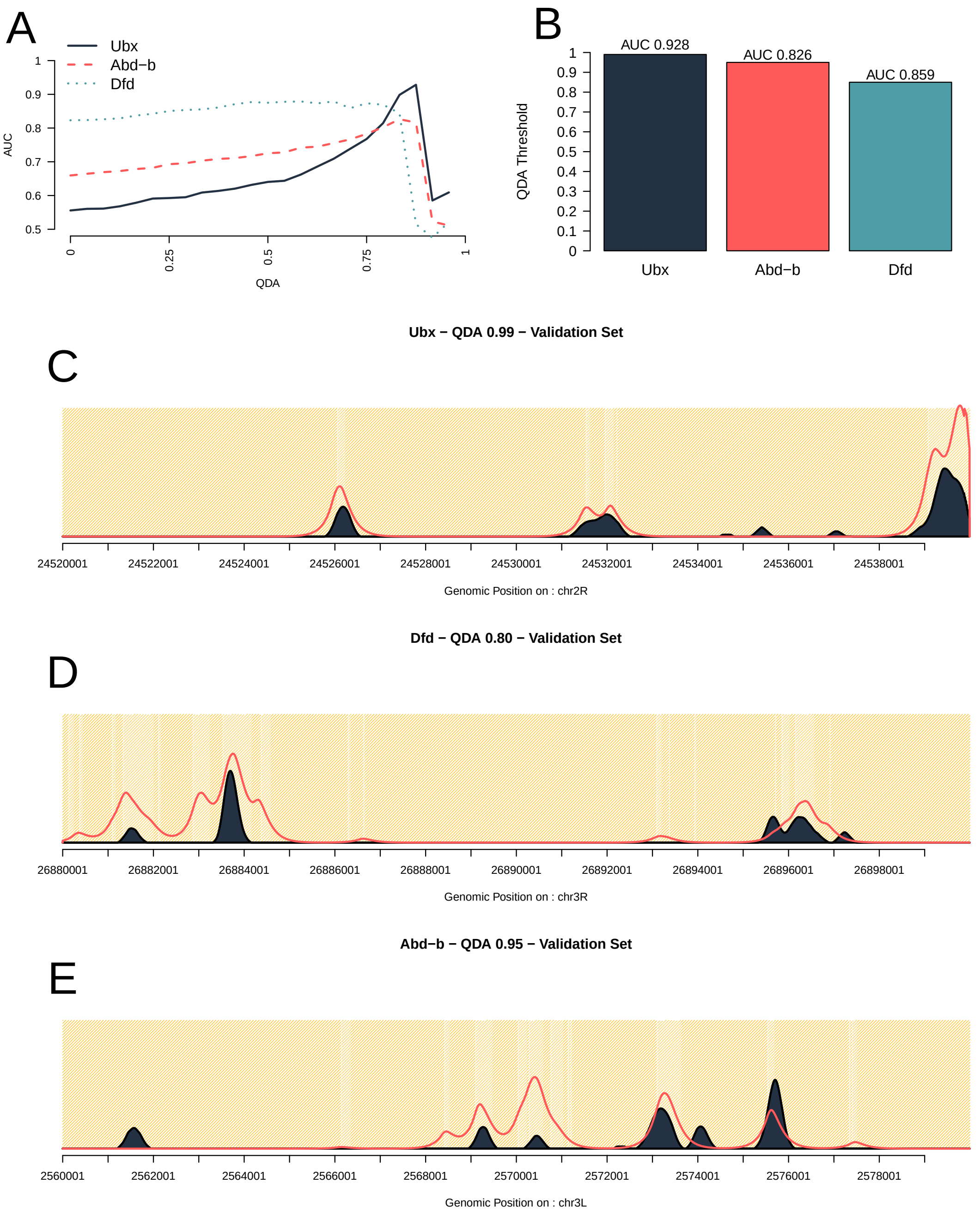


Figure 15: **Hox genes show binding preferences towards DNA accessibility.** I tested the model using different DNA accessibility stringencies. (**A**) Maximum AUC score as a function of stringency of DNA accessibility (the higher the QDA value the less DNA is called accessible) for three Hox TFs: Ubx, Dfd and Abd-B. (**B**) The best performing QDA accessibility in terms of AUC. (**C, D** and **E**) Binding profiles and prediction of the ChIP data at individual loci taken from the validation set for the three TFs

| TF | N | lambda | MSE | N | lambda | AUC | N | lambda | recall |
|---|---|---|---|---|---|---|---|---|---|
| Ubx QDA 0.99 | 2.00e+05 | 3.75 | 0.007 | 1.00e+06 | 3.75 | 0.928 | 1.00e+06 | 3.75 | 0.795 |
| Dfd QDA 0.8 | 1e+04 | 0.5 | 0.006 | 1e+05 | 0.5 | 0.859 | 1e+05 | 0.5 | 0.808 |
| Abd-b QDA 0.95 | 2e+05 | 3 | 0.008 | 2e+04 | 0.75 | 0.826 | 1e+06 | 2 | 0.748 |

Table 3: **Optimal set of Parameters for Hox TFs.** The following table shows the optimal parameters for best performing QDA for each Hox TF. MSE, AUC and recall are included.

## THE ROLE OF NUMBER OF BOUND MOLECULES IN TRANSCRIPTION FACTOR BINDING

### *Number of bound molecules and TF specificity plays a limited role in the binding of architectural proteins*

To investigate the robustness of the estimated parameters, I computed the optimal parameters for different biological replicates. Despite strong variations between experimental data, I show that the predicted optimal set of parameters when using MSE remained similar between biological replicates (see Figure-16). This suggests that despite biological and technical variation between replicates performed by different labs using different protocols, the model robustly infers a similar number of bound molecules and scaling factor for a given TF. The optimal parameters estimated over the training set can be found in Table-4, Table-5, Table-6 and Table-7 for MSE, AUC, recall and Spearman correlation coefficient.

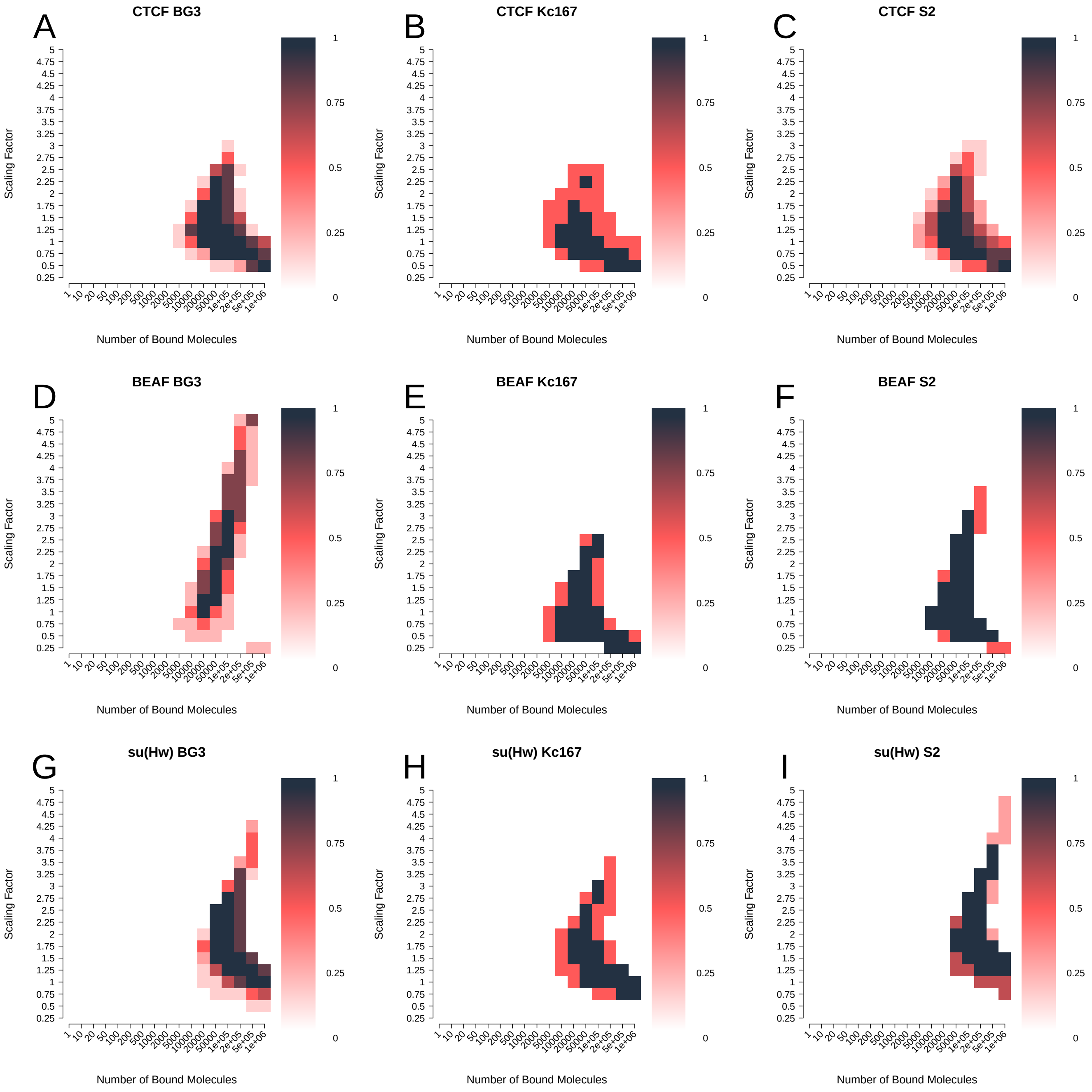


Figure 16: **Optimal parameters consistency among biological replicates for MSE using DHS accessibility.** Heat maps show an overlay of the top 10 % combinations of parameters when minimising MSE for: (**A-C**) CTCF, (**D-F**) BEAF-32 and (**G-I**) su(Hw). I plotted the following cell lines: (**A**, **D** and **G**) BG3, (**B**, **E** and **H**) Kc167 and (**C**, **F** and **I**) BG3.

To investigate the influence of these parameters, I assumed that a high variation of goodness of fit score for each combination of parameters would suggest a strong influence of these parameters on TF binding. If a goodness of fit scores varied little between parameter combinations, I can then conclude that they do not strongly influence our predicted profiles. This means that DNA accessibility would be the strongest driver towards predicting TF binding of these architectural proteins. Restricting the amount of available binding motifs would be more influential than TF copy number and the ability of a TF to discriminate between high and low affinity sites. Interestingly, this is still true in the case of su(Hw); I show that su(Hw) binding sites are most likely found in less accessible DNA. These results suggest that relative TF abundance only plays a role on binding sites found in accessible DNA (see Figure-10).

| TF | # Bound | Lambda | MSE | # Bound | Lambda | MSE | # Bound | Lambda | MSE |
|---|---|---|---|---|---|---|---|---|---|
| BG3 modEncode 282 CTCF | 1e+06 | 0.75 | 0.009 | 1e+06 | 0.5 | 0.009 | 5e+05 | 0.75 | 0.009 |
| BG3 modEncode 3280 CTCF | 5e+05 | 0.75 | 0.013 | 5e+05 | 0.5 | 0.013 | 20000 | 1.5 | 0.018 |
| BG3 modEncode 3671 CTCF | 1e+06 | 0.75 | 0.008 | 1e+06 | 0.5 | 0.008 | 5e+05 | 0.75 | 0.007 |
| BG3 modEncode 3672 CTCF | 1e+06 | 0.5 | 0.006 | 50000 | 0.5 | 0.006 | 1e+05 | 1 | 0.007 |
| BG3 modEncode 3673 CTCF | 5e+05 | 0.75 | 0.01 | 20000 | 0.75 | 0.01 | 5e+05 | 0.75 | 0.008 |
| BG3 modEncode 3674 CTCF | 1e+06 | 0.75 | 0.01 | 1e+06 | 0.5 | 0.01 | 5e+05 | 0.75 | 0.011 |
| Kc167 GSM762842 CTCF | 50000 | 1 | 0.007 | 500 | 1 | 0.007 | 10000 | 1.5 | 0.006 |
| Kc167 modEncode 908 CTCF | 5e+05 | 0.75 | 0.011 | 50000 | 0.5 | 0.01 | 5e+05 | 0.75 | 0.009 |
| S2 modEncode 2638 CTCF | 50000 | 1.25 | 0.002 | 500 | 1.25 | 0.002 | 5000 | 1.5 | 0.002 |
| S2 modEncode 2639 CTCF | 20000 | 1.25 | 0.002 | 200 | 1.25 | 0.002 | 5000 | 2 | 0.002 |
| S2 modENCODE 283 CTCF | 5e+05 | 0.75 | 0.014 | 20000 | 0.5 | 0.013 | 1e+06 | 0.75 | 0.013 |
| S2 modEncode 3281 CTCF | 5e+05 | 0.75 | 0.017 | 20000 | 0.5 | 0.017 | 50000 | 2 | 0.019 |
| S2 modEncode 3749 CTCF | 1e+05 | 1 | 0.009 | 1000 | 1 | 0.009 | 20000 | 1.25 | 0.01 |
| S2 modEncode 913 CTCF | 2e+05 | 0.75 | 0.01 | 1e+06 | 0.25 | 0.01 | 5e+05 | 0.5 | 0.014 |
| BG3 modEncode 3714 Su(Hw) | 1e+05 | 2 | 0.023 | 5000 | 2 | 0.023 | 5000 | 2 | 0.045 |
| BG3 modEncode 3715 Su(Hw) | 1e+05 | 2 | 0.024 | 5000 | 2 | 0.024 | 10000 | 2.75 | 0.046 |
| BG3 modEncode 3716 Su(Hw) | 2e+05 | 1.5 | 0.016 | 5000 | 1.25 | 0.016 | 10000 | 2 | 0.031 |
| BG3 modEncode 3717 Su(Hw) | 1e+05 | 2 | 0.016 | 2000 | 1.75 | 0.016 | 10000 | 2.25 | 0.032 |
| BG3 modEncode 3718 Su(Hw) | 50000 | 1.75 | 0.012 | 1000 | 1.75 | 0.012 | 5e+05 | 1 | 0.019 |
| BG3 modEncode 951 Su(Hw) | 1e+05 | 1.75 | 0.017 | 2000 | 1.5 | 0.017 | 2e+05 | 0.75 | 0.041 |
| Kc167 modEncode 3801 Su(Hw) | 1e+05 | 2.25 | 0.017 | 5000 | 2.25 | 0.017 | 2e+05 | 1.25 | 0.03 |
| Kc167 Su(Hw) | 50000 | 1.25 | 0.004 | 200 | 1.25 | 0.004 | 10000 | 1.25 | 0.005 |
| S2 modEncode 330 Su(Hw) | 2e+05 | 1.75 | 0.022 | 5000 | 1.75 | 0.022 | 5e+05 | 0.75 | 0.05 |
| S2 modEncode 331 Su(Hw) | 1e+06 | 1.25 | 0.017 | 5000 | 1.25 | 0.017 | 5e+05 | 0.75 | 0.037 |
| S2 modEncode 3719 Su(Hw) | 2e+05 | 2.25 | 0.033 | 10000 | 2.25 | 0.032 | 2e+05 | 0.75 | 0.066 |
| BG3 modEncode 3663 BEAF-32 | 1e+05 | 2.75 | 0.026 | 10000 | 2.5 | 0.026 | 50000 | 3 | 0.023 |
| BG3 modEncode 3664 BEAF-32 | 1e+05 | 2.5 | 0.027 | 10000 | 2.5 | 0.027 | 1e+05 | 4 | 0.019 |
| BG3 modEncode 3665 BEAF-32 | 2e+05 | 3 | 0.039 | 20000 | 2.75 | 0.04 | 2e+05 | 2.5 | 0.034 |
| BG3 modEncode 921 BEAF-32 | 20000 | 1.25 | 0.012 | 2000 | 1.75 | 0.012 | 20000 | 2.5 | 0.008 |
| Kc167 GSM1535963 BEAF32 | 50000 | 0.75 | 0.012 | 1000 | 1 | 0.012 | 50000 | 0.75 | 0.008 |
| Kc167 GSM762845 BEAF32 | 20000 | 1 | 0.006 | 1000 | 1.25 | 0.006 | 10000 | 1.25 | 0.005 |
| S2 modEncode 3745 BEAF-32 | 50000 | 1.25 | 0.017 | 1000 | 1 | 0.017 | 50000 | 3.25 | 0.014 |
| S2 modEncode 922 BEAF-32 | 50000 | 1.5 | 0.014 | 2000 | 1.5 | 0.014 | 50000 | 3 | 0.009 |

Table 4: **Optimal set of parameters after training on top ten regions and minimising MSE.** White columns show optimal parameters when No Accessibility was considered. Light grey columns show optimal parameters for continuous accessibility. Dark grey columns show optimal parameters for DHS accessibility.

| TF | # Bound | Lambda | AUC | # Bound | Lambda | AUC | # Bound | Lambda | AUC |
|---|---|---|---|---|---|---|---|---|---|
| BG3 modEncode 282 CTCF | 5e+05 | 0.25 | 0.944 | 20000 | 0.25 | 0.943 | 1 | 0.25 | 0.885 |
| BG3 modEncode 3280 CTCF | 20000 | 0.25 | 0.895 | 2000 | 0.25 | 0.895 | 5e+05 | 0.25 | 0.812 |
| BG3 modEncode 3671 CTCF | 2e+05 | 0.5 | 0.936 | 5e+05 | 0.25 | 0.937 | 10000 | 1 | 0.946 |
| BG3 modEncode 3672 CTCF | 10 | 1 | 0.922 | 1 | 1 | 0.922 | 2000 | 1.25 | 0.832 |
| BG3 modEncode 3673 CTCF | 20000 | 0.5 | 0.903 | 1000 | 0.5 | 0.903 | 20 | 1.25 | 0.853 |
| BG3 modEncode 3674 CTCF | 1e+06 | 0.25 | 0.929 | 5e+05 | 0.25 | 0.929 | 500 | 0.75 | 0.847 |
| Kc167 GSM762842 CTCF | 1e+06 | 0.5 | 0.954 | 10000 | 0.5 | 0.955 | 1e+06 | 0.5 | 0.887 |
| Kc167 modEncode 908 CTCF | 1e+05 | 0.25 | 0.952 | 50 | 0.25 | 0.952 | 1 | 1.25 | 0.963 |
| S2 modEncode 2638 CTCF | 50000 | 0.5 | 0.99 | 100 | 0.5 | 0.99 | 5000 | 0.5 | 0.963 |
| S2 modEncode 2639 CTCF | 100 | 2 | 0.915 | 1 | 2 | 0.916 | 2e+05 | 0.25 | 0.857 |
| S2 modENCODE 283 CTCF | 1e+06 | 0.25 | 0.926 | 200 | 0.25 | 0.926 | 10 | 1.25 | 0.886 |
| S2 modEncode 3281 CTCF | 1e+06 | 0.25 | 0.882 | 10000 | 0.25 | 0.882 | 50 | 1 | 0.864 |
| S2 modEncode 3749 CTCF | 50000 | 0.25 | 0.857 | 10 | 0.25 | 0.857 | 1e+06 | 0.25 | 0.848 |
| S2 modEncode 913 CTCF | 2000 | 0.75 | 0.938 | 1 | 0.75 | 0.938 | 2000 | 0.5 | 0.872 |
| BG3 modEncode 3714 Su(Hw) | 20000 | 0.75 | 0.928 | 100 | 0.75 | 0.928 | 1 | 0.25 | 0.458 |
| BG3 modEncode 3715 Su(Hw) | 50 | 0.75 | 0.927 | 1 | 0.75 | 0.927 | 1 | 0.25 | 0.606 |
| BG3 modEncode 3716 Su(Hw) | 10000 | 0.75 | 0.929 | 50 | 0.75 | 0.93 | 1 | 0.25 | 0.677 |
| BG3 modEncode 3717 Su(Hw) | 5000 | 0.75 | 0.927 | 20 | 0.75 | 0.927 | 10000 | 1.25 | 0.662 |
| BG3 modEncode 3718 Su(Hw) | 1 | 0.75 | 0.93 | 10 | 0.75 | 0.93 | 20 | 1.5 | 0.766 |
| BG3 modEncode 951 Su(Hw) | 2000 | 0.75 | 0.931 | 10 | 0.75 | 0.931 | 200 | 0.25 | 0.583 |
| Kc167 modEncode 3801 Su(Hw) | 50000 | 0.75 | 0.925 | 20 | 0.75 | 0.925 | 20000 | 0.75 | 0.548 |
| Kc167 Su(Hw) | 1e+06 | 0.75 | 0.962 | 500 | 0.75 | 0.962 | 20000 | 1.25 | 0.915 |
| S2 modEncode 330 Su(Hw) | 50000 | 0.5 | 0.931 | 50 | 0.5 | 0.931 | 1 | 0.25 | 0.565 |
| S2 modEncode 331 Su(Hw) | 1e+05 | 0.5 | 0.937 | 100 | 0.75 | 0.937 | 5000 | 0.25 | 0.643 |
| S2 modEncode 3719 Su(Hw) | 10000 | 1 | 0.917 | 50 | 1 | 0.916 | 1 | 0.25 | 0.43 |
| BG3 modEncode 3663 BEAF-32 | 20000 | 0.25 | 0.66 | 1000 | 0.25 | 0.661 | 1 | 2.25 | 0.81 |
| BG3 modEncode 3664 BEAF-32 | 5000 | 4 | 0.789 | 1000 | 4 | 0.789 | 100 | 3.5 | 0.859 |
| BG3 modEncode 3665 BEAF-32 | 500 | 1.5 | 0.726 | 1 | 1.5 | 0.726 | 10000 | 0.75 | 0.816 |
| BG3 modEncode 921 BEAF-32 | 5000 | 4.25 | 0.831 | 1000 | 4 | 0.831 | 1e+05 | 3 | 0.897 |
| Kc167 GSM1535963 BEAF32 | 2e+05 | 0.25 | 0.847 | 5000 | 0.25 | 0.847 | 1000 | 3.75 | 0.963 |
| Kc167 GSM762845 BEAF32 | 10000 | 4 | 0.937 | 1000 | 4 | 0.937 | 50 | 4.25 | 0.986 |
| S2 modEncode 3745 BEAF-32 | 2e+05 | 0.25 | 0.857 | 1000 | 0.25 | 0.857 | 200 | 3.75 | 0.869 |
| S2 modEncode 922 BEAF-32 | 10000 | 4 | 0.825 | 2000 | 4.25 | 0.826 | 1 | 3 | 0.928 |

Table 5: **Optimal set of parameters after training on top ten regions and maximising AUC.** White columns show optimal parameters when No Accessibility was considered. Light grey columns show optimal parameters for continuous accessibility. Dark grey columns show optimal parameters for DHS accessibility.

| TF | # Bound | Lambda | Recall | # Bound | Lambda | Recall | # Bound | Lambda | Recall |
|---|---|---|---|---|---|---|---|---|---|
| BG3 modEncode 282 CTCF | 20000 | 0.75 | 0.887 | 1000 | 0.75 | 0.887 | 1 | 1.5 | 0.778 |
| BG3 modEncode 3280 CTCF | 10000 | 0.75 | 0.816 | 500 | 0.75 | 0.818 | 1 | 1.25 | 0.676 |
| BG3 modEncode 3671 CTCF | 10000 | 0.75 | 0.884 | 500 | 0.75 | 0.884 | 1000 | 1.5 | 0.859 |
| BG3 modEncode 3672 CTCF | 1 | 1.25 | 0.883 | 1 | 1.25 | 0.883 | 5000 | 1.5 | 0.693 |
| BG3 modEncode 3673 CTCF | 20000 | 0.75 | 0.855 | 500 | 0.75 | 0.854 | 100 | 1.5 | 0.758 |
| BG3 modEncode 3674 CTCF | 10000 | 0.75 | 0.865 | 2000 | 0.75 | 0.865 | 100 | 1.5 | 0.735 |
| Kc167 GSM762842 CTCF | 5e+05 | 0.75 | 0.932 | 5000 | 0.75 | 0.932 | 10000 | 1.75 | 0.807 |
| Kc167 modEncode 908 CTCF | 10000 | 0.75 | 0.88 | 50 | 0.75 | 0.88 | 1 | 1.25 | 0.831 |
| S2 modEncode 2638 CTCF | 10000 | 1 | 0.962 | 100 | 1 | 0.962 | 2000 | 1.5 | 0.859 |
| S2 modEncode 2639 CTCF | 100 | 2 | 0.891 | 1 | 2.25 | 0.892 | 2000 | 1.75 | 0.727 |
| S2 modENCODE 283 CTCF | 10000 | 0.75 | 0.86 | 50 | 0.75 | 0.86 | 1 | 1.5 | 0.744 |
| S2 modEncode 3281 CTCF | 20000 | 0.75 | 0.805 | 50 | 0.75 | 0.805 | 500 | 1.25 | 0.71 |
| S2 modEncode 3749 CTCF | 20000 | 1 | 0.816 | 20 | 1 | 0.814 | 2000 | 1.25 | 0.682 |
| S2 modEncode 913 CTCF | 2000 | 1 | 0.875 | 20 | 1 | 0.875 | 1 | 1.75 | 0.711 |
| BG3 modEncode 3714 Su(Hw) | 500 | 1.25 | 0.85 | 20 | 1.25 | 0.85 | 10000 | 1.25 | 0.208 |
| BG3 modEncode 3715 Su(Hw) | 1000 | 1.25 | 0.848 | 10 | 1.25 | 0.848 | 10 | 2.25 | 0.397 |
| BG3 modEncode 3716 Su(Hw) | 20000 | 1.25 | 0.86 | 50 | 1.25 | 0.86 | 10000 | 1.5 | 0.443 |
| BG3 modEncode 3717 Su(Hw) | 2000 | 1.25 | 0.855 | 20 | 1.25 | 0.855 | 10000 | 1.25 | 0.419 |
| BG3 modEncode 3718 Su(Hw) | 5000 | 1.25 | 0.869 | 50 | 1.25 | 0.868 | 1 | 2 | 0.547 |
| BG3 modEncode 951 Su(Hw) | 5000 | 1.25 | 0.846 | 50 | 1.25 | 0.846 | 1 | 4.25 | 0.346 |
| Kc167 modEncode 3801 Su(Hw) | 10 | 1.25 | 0.851 | 1 | 1.25 | 0.851 | 20 | 1.5 | 0.225 |
| Kc167 Su(Hw) | 50000 | 1 | 0.936 | 50 | 1 | 0.935 | 20000 | 1.25 | 0.833 |
| S2 modEncode 330 Su(Hw) | 5000 | 1.25 | 0.841 | 20 | 1.25 | 0.841 | 5000 | 1.5 | 0.286 |
| S2 modEncode 331 Su(Hw) | 5000 | 1.25 | 0.863 | 20 | 1.25 | 0.862 | 10000 | 1.25 | 0.406 |
| S2 modEncode 3719 Su(Hw) | 5000 | 1.25 | 0.813 | 10 | 1.25 | 0.812 | 1e+05 | 1 | 0.132 |
| BG3 modEncode 3663 BEAF-32 | 10 | 1 | 0.627 | 20 | 1 | 0.627 | 1 | 2.75 | 0.69 |
| BG3 modEncode 3664 BEAF-32 | 5000 | 4 | 0.762 | 1000 | 4 | 0.761 | 5000 | 4 | 0.732 |
| BG3 modEncode 3665 BEAF-32 | 2000 | 2 | 0.687 | 200 | 2 | 0.687 | 2e+05 | 1.75 | 0.663 |
| BG3 modEncode 921 BEAF-32 | 5000 | 4.25 | 0.807 | 1000 | 4.25 | 0.806 | 1e+05 | 3 | 0.829 |
| Kc167 GSM1535963 BEAF32 | 1 | 1.5 | 0.823 | 1 | 1.5 | 0.823 | 1000 | 3.75 | 0.902 |
| Kc167 GSM762845 BEAF32 | 10000 | 4 | 0.918 | 1000 | 4 | 0.918 | 50 | 4.25 | 0.95 |
| S2 modEncode 3745 BEAF-32 | 1000 | 0.75 | 0.806 | 20 | 0.75 | 0.806 | 1 | 3.25 | 0.745 |
| S2 modEncode 922 BEAF-32 | 10000 | 4 | 0.799 | 2000 | 4.25 | 0.801 | 2000 | 3.25 | 0.856 |

Table 6: **Optimal set of parameters after training on top ten regions and maximising recall.** White columns show optimal parameters when No Accessibility was considered. Light grey columns show optimal parameters for continuous accessibility. Dark grey columns show optimal parameters for DHS accessibility.

| TF | # Bound | Lambda | spearman | # Bound | Lambda | spearman | # Bound | Lambda | spearman |
|---|---|---|---|---|---|---|---|---|---|
| BG3 modEncode 282 CTCF | 1e+06 | 0.25 | 0.584 | 2e+05 | 0.25 | 0.584 | 5e+05 | 0.25 | 0.523 |
| BG3 modEncode 3280 CTCF | 1e+06 | 0.25 | 0.512 | 2e+05 | 0.25 | 0.513 | 5e+05 | 0.25 | 0.461 |
| BG3 modEncode 3671 CTCF | 1e+06 | 0.25 | 0.582 | 1e+06 | 0.25 | 0.583 | 1 | 0.25 | 0.581 |
| BG3 modEncode 3672 CTCF | 1e+05 | 0.5 | 0.502 | 5000 | 0.5 | 0.503 | 1 | 0.25 | 0.404 |
| BG3 modEncode 3673 CTCF | 1e+06 | 0.25 | 0.463 | 1e+06 | 0.25 | 0.464 | 1 | 0.25 | 0.462 |
| BG3 modEncode 3674 CTCF | 1e+06 | 0.25 | 0.483 | 50000 | 0.25 | 0.483 | 2000 | 0.5 | 0.471 |
| Kc167 GSM762842 CTCF | 1 | 0.25 | 0.299 | 5e+05 | 0.25 | 0.312 | 20 | 3.25 | 0.367 |
| Kc167 modEncode 908 CTCF | 5e+05 | 0.25 | 0.551 | 100 | 0.25 | 0.551 | 10 | 1.25 | 0.638 |
| S2 modEncode 2638 CTCF | 1e+06 | 0.25 | 0.608 | 10000 | 0.25 | 0.608 | 5e+05 | 0.25 | 0.544 |
| S2 modEncode 2639 CTCF | 1 | 0.25 | 0.438 | 10000 | 0.25 | 0.438 | 1 | 0.25 | 0.323 |
| S2 modENCODE 283 CTCF | 1 | 0.25 | 0.596 | 20000 | 0.25 | 0.597 | 1 | 0.25 | 0.517 |
| S2 modEncode 3281 CTCF | 1e+06 | 0.25 | 0.496 | 500 | 0.25 | 0.496 | 5e+05 | 0.25 | 0.448 |
| S2 modEncode 3749 CTCF | 1e+06 | 0.25 | 0.446 | 10000 | 0.25 | 0.448 | 5e+05 | 0.25 | 0.467 |
| S2 modEncode 913 CTCF | 5e+05 | 0.25 | 0.613 | 10000 | 0.25 | 0.614 | 200 | 1.25 | 0.556 |
| BG3 modEncode 3714 Su(Hw) | 5e+05 | 0.5 | 0.689 | 1000 | 0.5 | 0.689 | 1 | 0.25 | -0.109 |
| BG3 modEncode 3715 Su(Hw) | 50 | 0.75 | 0.704 | 1 | 0.75 | 0.704 | 1 | 0.25 | 0.1 |
| BG3 modEncode 3716 Su(Hw) | 1000 | 0.75 | 0.705 | 1 | 0.75 | 0.705 | 1 | 0.25 | 0.22 |
| BG3 modEncode 3717 Su(Hw) | 2000 | 0.75 | 0.682 | 1 | 0.75 | 0.682 | 5e+05 | 1 | 0.238 |
| BG3 modEncode 3718 Su(Hw) | 1 | 0.5 | 0.647 | 50 | 0.5 | 0.645 | 1 | 0.75 | 0.462 |
| BG3 modEncode 951 Su(Hw) | 5e+05 | 0.5 | 0.718 | 20000 | 0.5 | 0.721 | 1 | 0.25 | 0.056 |
| Kc167 modEncode 3801 Su(Hw) | 1e+06 | 0.5 | 0.674 | 200 | 0.5 | 0.674 | 1e+05 | 0.75 | 0.062 |
| Kc167 Su(Hw) | 50000 | 0.75 | 0.421 | 20 | 0.75 | 0.428 | 200 | 1.5 | 0.361 |
| S2 modEncode 330 Su(Hw) | 1e+06 | 0.5 | 0.748 | 1e+06 | 0.25 | 0.757 | 1 | 0.25 | 0.037 |
| S2 modEncode 331 Su(Hw) | 5e+05 | 0.5 | 0.675 | 1e+06 | 0.25 | 0.691 | 1 | 0.25 | 0.194 |
| S2 modEncode 3719 Su(Hw) | 50000 | 0.75 | 0.743 | 100 | 0.75 | 0.746 | 1 | 0.25 | -0.266 |
| BG3 modEncode 3663 BEAF-32 | 100 | 0.5 | 0.16 | 1 | 0.5 | 0.16 | 500 | 0.5 | 0.503 |
| BG3 modEncode 3664 BEAF-32 | 500 | 0.25 | 0.317 | 20 | 0.25 | 0.317 | 1000 | 1.75 | 0.507 |
| BG3 modEncode 3665 BEAF-32 | 50000 | 0.25 | 0.341 | 1 | 0.75 | 0.34 | 50000 | 0.75 | 0.512 |
| BG3 modEncode 921 BEAF-32 | 1e+06 | 0.25 | 0.252 | 1e+05 | 0.25 | 0.253 | 1000 | 0.5 | 0.31 |
| Kc167 GSM1535963 BEAF32 | 2e+05 | 0.25 | 0.385 | 2000 | 0.25 | 0.385 | 1000 | 3 | 0.575 |
| Kc167 GSM762845 BEAF32 | 1e+06 | 0.25 | 0.297 | 10000 | 0.25 | 0.296 | 200 | 4 | 0.552 |
| S2 modEncode 3745 BEAF-32 | 50000 | 0.25 | 0.385 | 200 | 0.25 | 0.385 | 50 | 2.25 | 0.517 |
| S2 modEncode 922 BEAF-32 | 5e+05 | 0.25 | 0.238 | 5000 | 0.25 | 0.237 | 5000 | 1.75 | 0.448 |

Table 7: **Optimal set of parameters after training on top ten regions and maximising Spearman correlation.** White columns show optimal parameters when No Accessibility was considered. Light grey columns show optimal parameters for continuous accessibility. Dark grey columns show optimal parameters for DHS accessibility.

### *ChIPanalyser predicts TF binding in different cell lines by considering relative mRNA abundance*

I wanted to further investigate the predictive capabilities of the model and also demonstrate its mechanistic soundness for CTCF, BEAF-32 and su(Hw) in the three selected cell lines. For that, I estimated the optimal set of parameters in one cell line and aimed to predict TF binding in a different cell line taking into account changes in DNA accessibility (DHS) and changes in number of bound molecules using relative changes in RNA abundance. For example, I estimated the optimal set of parameters for CTCF in Kc167 cells that would minimise MSE as $\lambda = 1.5$ and $N = 10^4$ over the top 10 regions. By rescaling N based on relative RNA-seq levels of CTCF in the two cell lines, I could approximate the number of CTCF molecules bound to DNA in BG3 cells ($N \approx 1.6 \times 10^4$ ). This together with BG3-specific DNA accessibility data is capable of predicting with high accuracy the ChIP-seq profile in BG3 cells (see Figure-17). RNA rescaling seems to recover both the number of peaks and their location with high accuracy. Moreover, the rescaling of number of bound molecules did not lead to any difference in terms of MSE variation between estimated and rescaled (Figure17 **G**). The same analysis was performed for BEAF-32 (Figure-17 **C, D** and **H**), where I estimated parameters in BG3 cells ($\lambda = 2.5$ and $N = 2 \times 10^4$ ) and rescaled the number of molecules in S2 cells ($N \approx 1.2 \times 10^4$ ). Once again, the model correctly predicts ChIP profiles in both location and relative enrichment. Finally, for su(Hw) (Figure-17 **E, F** and **I**) I estimated parameters in Kc167 cells ($\lambda = 1.25$ and $N = 10^4$ ) and rescaled the number of molecules in S2 cells ($N \approx 6 \times 10^3$ ). Again, the predictions of the model are accurate. The results show that changes between cell lines in DNA accessibility and number of bound molecules seem sufficient to explain the changes in TF binding profiles. Nevertheless, I still do not know which of the two is the more important factor or whether both have similar contributions.

To address this, I also assumed that there is (i) no change, (ii) a 10 fold reduction and (iii) one 100 fold reduction in the number of bound molecules and repeated the analysis. Figure-17 shows that using the same TF abundance as in the original cell line did not change the prediction's quality at all. In fact, I observed a significant reduction in the predicted local enrichment only when reducing the number of bound molecules by 100 (for CTCF and su(Hw)) or 10-fold (for BEAF-32). These results show that cell differences in binding profiles of TFs, at their strong binding regions, would mainly come from differences in DNA accessibility and not relatively small changes in TF abundance. The only way that TF abundance could impact the binding profile (and, consequently, lead to changes in gene regulation) is when the expression of these TFs are strongly down-regulated.

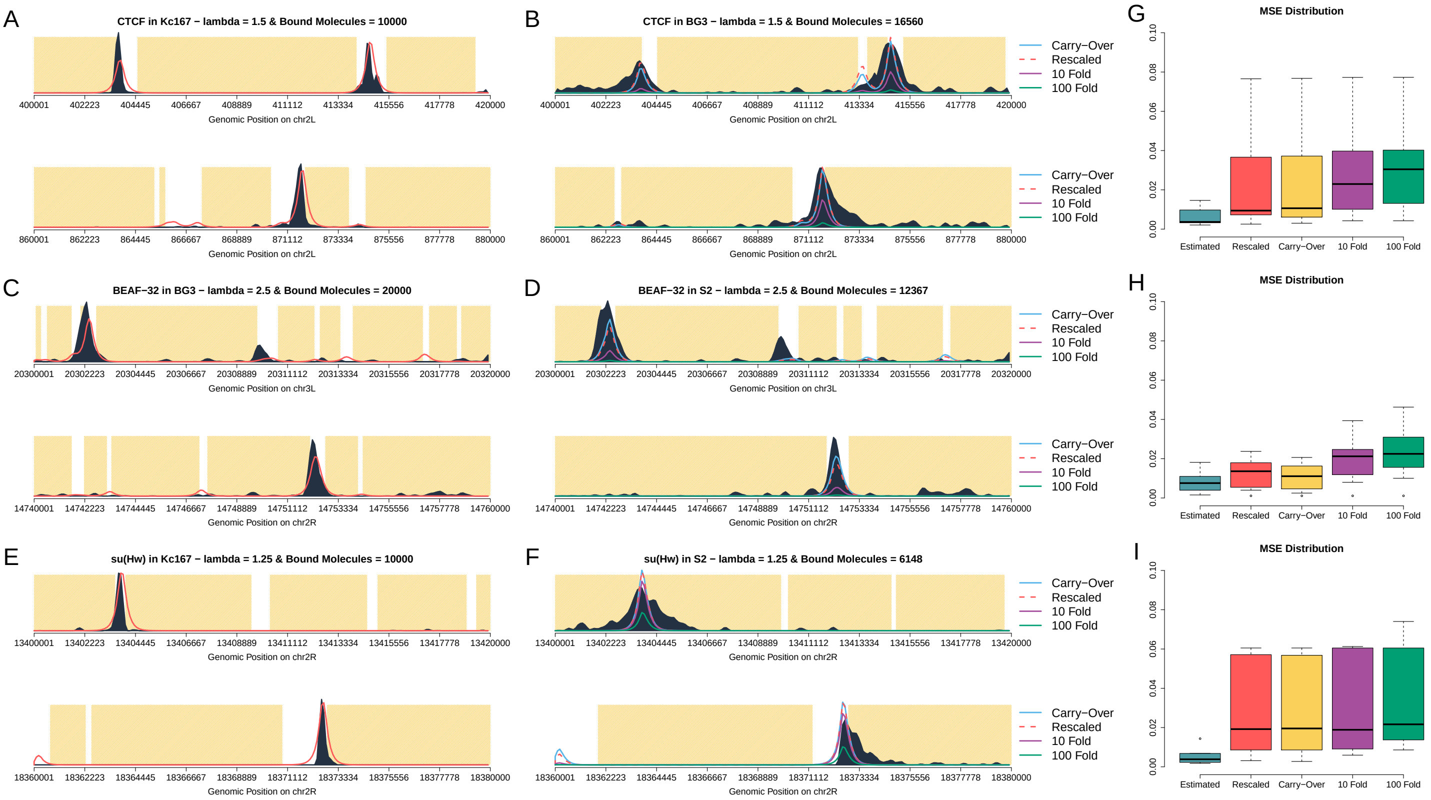


Figure 17: **TF abundance remains stable between different cell lines when considering relative mRNA abundance. A-F** show predicted ChIP-seq profiles with the TF abundance estimated based on RNA-seq. The yellow area represents inaccessible DNA, the dark area represents experimental ChIP signal and the red lines are the predicted profiles. I estimated the number of bound molecules in one cell line (**A, C** and **E**) and rescaled our estimate using relative mRNA abundance in an other cell line (**B, D** and **F**). (**B, D** and **F**) The dashed red line represents the rescaled value of number of bound molecules based on relative RNA-seq abundance, the light blue the original value estimated in (**A, C** and **E**).The purple line and the green line represent the original estimated value reduced 10 and 100 times respectively. (**G , H** and **I**) Boxplots with MSE for all cases in the estimated and predicted profiles at top 10 regions for both training and validation.

## DISCUSSION

### *TFs use different binding mechanisms*

In this analysis, I focused my attention on three DNA binding proteins: CTCF, BEAF-32 and su(Hw). All three TFs are known architectural proteins in *Drosophila* but also play roles in transcription regulation and insulation [Van Bortle et al., 2014, Chathoth and Zabet, 2019]. Moreover, it was shown that these three TFs have distinct binding behaviours and were classified into three subclasses with respect to chromatin architecture[Bushey et al., 2009, Vogelmann et al., 2014] . In this analysis, I show that they all exhibit different behaviours with respect to DNA binding.

CTCF has been shown to play a role in loop formation and participating in Topologically Associated Domains (TADs) boundary maintenance [Chathoth and Zabet, 2019]. However, only a subset of CTCF sites are involved in these structures and that many CTCF sites do not conform to this rule [Guo et al., 2015, Tang et al., 2015]. In my analysis, CTCF displayed strong sensitivity to DNA accessibility but reduced sensitivity to the number of bound molecules and scaling factor when DNA accessibility was considered (see Figures 10 and 11). My findings suggest that CTCF binds to hotspots along the genome and this could be explained by the observation that the strongest peaks are in fact highly conserved binding sites [VietriÂăRudan et al., 2015]. As the number of sites increase, the conservation of binding sites decreases, as does the goodness fit. Thus, CTCF binding to highly conserved sites can be explained by the model, but something else is is responsible for the reduced binding at less conserved sites (i.e. cell specific CTCF binding).

BEAF-32 is a Drosophila specific genetic insulator [Schoborg and Labrador, 2010] that shows preferential binding towards TAD boundaries, but also is involved in transcription itself. More specifically, BEAF-32 was identified as a cis-regulatory element separating close head-to-head genes with different transcription regulation modes [Jiang et al., 2009]. In Drosophila, there is a high density of these genes through out the genome and BEAF-32 tends to bind closely to the TSS [Rennie et al., 2018]. This is further confirmed by studies showing that BEAF-32 has uniform binding along the entire genome [Bushey et al., 2009]. TSSs are generally considered open chromatin and, if BEAF-32 binds in close proximity of the TSSs, it comes to no surprise that BEAF-32 would show a high sensitivity towards DNA accessibility. My results confirm that BEAF-32 shows a strong preference towards DNA accessibility and, to a lesser extent to local abundance (see Figures 10 and 11).

Furthermore, I show that su(Hw) binds in both open and closed chromatin and also displays a high sensitivity towards number of bound molecules and scaling factor when DNA accessibility is considered. There is a significant body of work showing the role su(Hw) plays in chromatin insulation and remodelling [Kurshakova et al., 2007, Kuhn-Parnell et al., 2008, Soshnev et al., 2013, Vorobyeva et al., 2013]. It had been suggested that the role of insulator is only possible when paired with other DNA binding proteins such as CP190 and mdg4. su(Hw) is also a primary actor in the interaction between the genome and nuclear lamina (also know as Lamina Associated Domains) [van Bemmel et al., 2010, van Steensel and Belmont, 2017]. Both chromatin insulation and LADs would induce closed chromatin in order to maintain chromosomal structure and this would explain why su(Hw) can bind in both open and closed chromatin. In this context, ChIP-seq peaks might not overlap well with DNase hyper sensitivity data (see Figures 10 and 11).

It has been shown that su(Hw) binding sites tend to cluster together (with varying number of sites) and that these sites are constitutively bound by su(Hw) [Parnell et al., 2006, Adryan et al., 2007]. Interestingly, it seems that only isolated high affinity sites had a role in transcriptional regulation and the clustered sites were more involved in chromatin architecture.

*DNA accessibility is the main driver of binding to DNA for some TFs*

These results show that DNA accessibility and number of bound molecules control the binding profiles of TFs (see Figure-10 and Figure-17). When I estimated the binding parameters ($\lambda$ and $N$ ) in one cell line and then predicted TF binding profiles in a different cell line based on changes in DNA accessibility and number of TF molecules (using changes in mRNA), I found a good agreement between the predictions and the actual ChIP-seq dataset (see Figure-17). Nevertheless, the changes in number of TF molecules between the two cell lines did not seem to make any difference to the predicted profiles at strong binding sites (compare blue and dashed red line in Figure-17 **B**, **D** and **F**). This means that biologically relevant fluctuations in TF numbers between different cell lines would have little effect on the differences in binding profiles of TFs, which would be mainly driven by changes in DNA accessibility. Furthermore, only very strong knock-downs would decrease or deplete ChIP peaks. It should be noted that CTCF, BEAF-32 and su(Hw) are highly expressed architectural and insulator proteins and, thus, they would be expected to saturate their binding sites. Why would changes in concentration of the TF have such a limited effect on their binding? One potential explanation is that these TFs control the expression of essential genes that should be tightly regulated to buffer fluctuations in number of molecules that affect the cell [Schoech and Zabet, 2014, Nicolas et al., 2017]. Finally, I also investigate the capacity of our model to differentiate between TFs that can bind only in open chromatin or also partially opened chromatin. The results showed that while Ubx displays a strong sensitivity to open chromatin and binds in the top 1% accessible sites, the binding of Abd-B and Dfd is less influenced by DNA accessibility (with Abd-B and Dfd binding in top 5% and 20% respectively accessible regions); see Figure-15. Hox TFs are known for having a similar motif, but displaying differences in their binding profiles [Chauvet et al., 2000a, Gehring et al., 1994, Pellerin et al., 1994]. It was hypothesised that binding cooperativity could explain the difference in binding profiles coupled with protein sequence changes [Hayashi and Scott, 1990, Joshi et al., 2010, Rezsohazy et al., 2015]. Here, I showed that DNA accessibility could also be responsible for the difference in binding profiles of Hox TFs (see Figure-15). Interestingly, our results support a model where Hox TFs would be able to bind to regions of DNA showing different level of accessibility and the DNA accessibility would be sufficient to explain these differences in the binding profiles of Hox TFs. Nevertheless, we also observed a poorer quality in modelling the binding profiles of TFs that can bind in dense chromatin (e.g., Abd-B or Dfd), which suggests that cooperative binding would be required to explain their binding. Due to the fact that our model does not include cooperativity, the predictions for these TFs would not be as accurate as in the case of TFs that preferentially bind to open chromatin.

### *Modelling DNA accessibility*

In this chapter, I described the role of DNA accessibility in TF binding. I compared the performance of ChIPanalyser without DNA accessibility data (No Access - all DNA is considered accessible), with continuous DNA accessibility data and with DHS DNA accessibility ( DNA is either accessible or inaccessible). As described above, the role of DNA accessibility is nuanced as certain TFs such as su(Hw) and CTCF are capable of binding to seemingly inaccessible DNA. This could indicate that their binding sites are located within less permissive chromatin. DNA would find itself in an intermediate state between closed chromatin and open chromatin. Under these circumstance, one would assume that continuous DNA accessibility values would be better suited to explain TF binding. However, the results described in Figure-10 seem to show that DHS DNA accessibility improves the performance of ChIPanalyser rather than continuous DNA accessibility.

One explanation for these unintuitive results could be the way continuous DNA accessibility was modelled. Continuous DNA accessibility scores were computed by min/max normalising DNase I hypersensitivity signal. This approach ensures that scores are bound between 0 and 1. However, this approach comes with limitations. Firstly, most DNase I hypersensitivity scores are non-zero and often slightly over the minimum score. In the context of ChIPanalyser, this will produce occupancy scores in regions of closed chromatin. The resulting predicted profiles would be much more closely related to the profiles produced with all DNA considered accessible than profiles produced with DHS accessibility. Furthermore, using a min/max normalisation approach to modelling DNA accessibility also compresses the "distance" between DNase I hypersensitivity scores. This increases the weight of low DNase I hypersensitivity scores compared to high DNase I hypersensitivity scores. The importance of open chromatin (or at least permissive chromatin ) would be minimised in favour closed chromatin. One approach to overcome this issue would be to model accessibility using an exponential. This would induce an *inflation* of scores associated with more accessible DNA and potentially correct the bias described above. A similar approach was described by Teif [Teif et al., 2014] for nucleosome occupancy based on the works by [Goh et al., 2010].

## CONCLUSION

I show that ChIPanalyser can shed light on the mechanisms driving TF binding. For architectural proteins, DNA accessibility is the main drivers towards differential binding between cell lines. Furthermore, all three architectural binding proteins show different binding mechanisms with respect to DNA accessibility. CTCF binds to genome hotspots as well as binding sites located in lesser accessible DNA. BEAF-32 bind anywhere along the genome as long as its binding sites is found in open chromatin while su(Hw) binds preferentially in closed chromatin. Finally, I describe how the model can recover the binding preferences of three Hox TF with respect to

DNA accessibility. While Ubx binds to open chromatin, Dfd and Abd-b are more permissive in terms of DNA accessibility.

# 3

# THE ROLE OF CHROMATIN STATES ON TF BINDING

## CHAPTER SUMMARY

The following chapter will describe the role of chromatin states on TF binding. First, I describe the development of a genetic algorithm to assess chromatin state affinity scores and the overall performance of such an algorithm. Second, I demonstrate that binding of both architectural proteins and Hox TFs are better explained with the inclusion of chromatin states. Finally, as describe previously, architectural proteins show a low sensitivity towards changes in protein abundance. Differential binding between cell lines is driven by changes in chromatin states rather than changes in abundance. Overall, chromatin states improves the performance of the model but most importantly increase the understanding of TF binding mechanisms.

## INTRODUCTION

There are many factors that influence the binding of TFs to DNA. Some of the most notable factors are binding motifs, TF abundance and DNA accessibility. As described previously, TF binding motifs are overly abundant across the genome. Most of the binding sites (characterised by PWMs) do not correlate well with genome wide occupancy assay peaks such as ChIP-seq or ChIP-on-chip. A common approach to decrease the number of false positive binding motif is to consider DNA accessibility. In this circumstance, one needs to assume that TFs will only bind to accessible DNA. Binding motifs located in closed chromatin will be "masked" by compacted DNA and thus unavailable for TF binding. In the previous chapters, I described ChIPanalyser, a Bioconductor package that predicts and models TF binding by using a statistical thermodynamic framework [Martin, 2017]. Using the package, I detailed the significant role of DNA accessibility in the binding of three *Drosophila* architectural proteins (CTCF, BEAF-32, and su(Hw)). For example, the binding of BEAF-32 is strongly driven by DNA accessibility and to a lesser extent protein abundance. By including DNA accessibility into the model, the package accurately recovers the binding of BEAF-32 on a genome wide scale. I demonstrated the lesser role of protein abundance by training the model in one cell line and rescaling the estimated number of bound molecules in another cell line using relative RNA levels. The model accurately recovered BEAF-32 peaks between cell lines. Rescaling the number of bound molecules had little effect on the binding predictions. These results suggest that BEAF-32 only bind in open chromatin and that changes in DNA accessibility are the main drivers behind differential binding between cell lines.

In the case of su(Hw), including DNA accessibility told a very different story. The model performed better when all DNA was considered accessible. This implies that many su(Hw) peaks are located in regions of closed chromatin. Interestingly, despite DNA accessibility playing a different role in the binding of su(Hw), protein abundance also played a lesser role in differential binding of su(Hw) between cell lines.

Including DNA accessibility into the model might not improve the predictions but still delivers valuable insight into the mechanisms of TF binding. In the case of su(Hw), binding is driven by something else than simply open or closed chromatin. In *Drosophila*, most cell lines have approximately 10% of open chromatin. This leaves close to 90% of "uncharted" DNA landscape. This increases up to 98 % in *Homo sapiens*.

Although extremely useful, DNA accessibility on its own shows some limitations.

One approach to overcome these limitations is to consider chromatin states instead of considering chromatin as either open or closed. DNA is wrapped around octameric complexes of histones called nuclosomes. Histones can undergo post translational modifications such as methlyation or acetylation. The addition or removal of chemical groups modifies the functional properties of a stretch of DNA [Bannister and Kouzarides, 2011]. These changes can occur either by directly modifying the physical bond between DNA and histone or by enabling the recruitment of chromatin re-modeller proteins. Chromatin states are defined by a specific combination of histone modifications associated with a genomic function. As an example, enhancers are generally associated with increased levels of H3K4me1, H3K27ac and H3K56ac [Andersson and Sandelin, 2019] . Interestingly, histone acetlylation is associated with more open chromatin as acytly groups neutralise the positive charge on lysine residues [Zhang and Presgraves, 2017, Shogren-Knaak et al., 2006]. This decreases the strength of the bond between DNA and nucleosomes. Genome wide chromatin state maps are generated by using histone ChIP-seq data and computational methods such as hidden markov models. Recently, chromatin maps for *Drosophila* were produced by using 24 histone modifications and by including DNAse I hypersensitvity data [Skalska et al., 2015]. This resulted in 11 chromatin states across the *Drosophila* genome. Chromatin states provide a more accurate representation of chromatin. Open chromatin can be described by multiple states with varying levels of accessibility while closed chromatin can be described by various heterochromatic states.

In this chapter, I demonstrate how chromatin states can be integrated to the statistical thermodynamic framework. More specifically, I developed a genetic algorithm that optimise chromatin state preferences of TFs by using ChIPanalyser at its core. Using this updated approach, I further investigate the binding of three architectural proteins

(CTCF, BEAF-32 and su(Hw)) and three Hox TFs (Ubx, DfD and Abd-b) in two *Drosophila* cell lines (Kc167 and BG3). I show that architectural proteins show clear preferences towards certain chromatin states. Unsurprisingly, BEAF-32 preferentially bind in states related to open chromatin. CTCF showed preferences towards open chromatin states but also intermediate states. On the other hand, predicting binding profiles of su(Hw) were drastically increased when including chromatin states. su(Hw) did not show a clear preference towards closed chromatin states but rather varying affinities for numerous chromatin states. Furthermore, I demonstrate that differential binding of architectural proteins between cell line is driven by chromatin states and that proteins abundance plays a lesser role. These results further support the idea that DNA accessibility plays a nuanced role in TF binding. Finally, Hox TFs preferences towards varying levels of accessibility can accurately be described by varying affinities towards chromatin states.

## DATA SETS

To carry out this analysis, data was downloaded from various sources. DNA sequences, PFMs and TF ChIP-seq data are described in the previous chapters (see chapter-1 and Appendix A). Chromatin state maps for Kc167 and BG3 cells were previously published by [Skalska et al., 2015]. As chromatin state maps were only available for BG3 and Kc167, the following analysis will focus on these two cell lines. The distribution of chromatin states shows slight difference between cell lines (see Figure-18). BG3 cells displayed an increase in heterochromatin state while Kc167 cells showed an increase in the "Basal" state. However the distribution of binding sites ( as described by PWM scores) does not indicate any bias towards a specific chromatin state. This suggests that binding motifs are evenly distributed along the genome.

To insure that all selected regions (see chapter 1 and chapter 2) contained chromatin states, I overlapped the selected regions with chromatin state maps. This resulted in a slightly reduced number of regions (3171). The process by which top regions were selected remained the same as the one previously described. The model was trained on the top 100 regions and validated on the remaining regions (3071). The number of regions used for training was increased to 100 in order to include every chromatin state at least once.

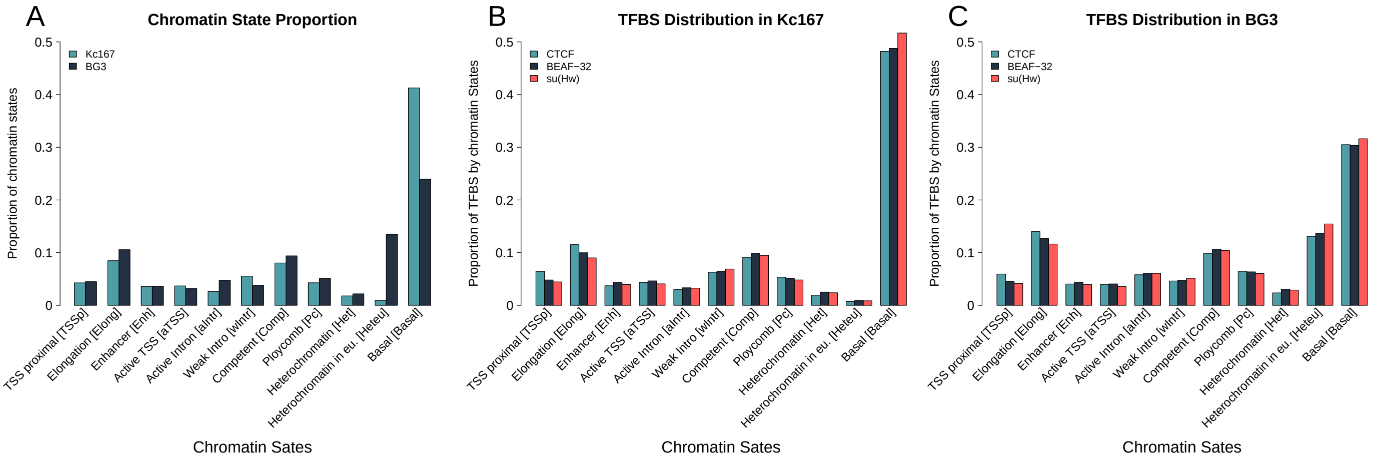


Figure 18: **Chromatin state maps show slight difference between cell lines.** While the Basal state is strongly increased in Kc167 cells compared to BG3 cells, BG3 cells display a higher proportion of heterochromatin in euchromatin. I investigated the distribution of binding sites between each chromatin state in both cell lines. Unsurprisingly, there does not seem to be any bias towards a specific chromatin state. Binding sites ( as described by a PWM) are evenly distributed across the genome.

## WIELDING NATURAL SELECTION

### *Genetic Algorithms*

To investigate chromatin states and their potential role in TF binding, the first step was to adapt the model. Previously, the model assumed that the four main factors driving the binding of TFs were: binding energy (as a PWM score), the number of bound molecules, a PWM scaling factor and finally DNA accessibility. Incorporating chromatin states only required to consider DNA affinity in the master equation. Similarly to varying an affinity towards accessible DNA, the new model considers varying affinities towards chromatin states.

$$P(N, a, \lambda, \omega)_j = \frac{N \cdot a_j \cdot e^{(\frac{1}{\lambda} \cdot \omega_j)}}{N \cdot a_j \cdot e^{(\frac{1}{\lambda} \cdot \omega_j)} + L \cdot n \cdot [a_i \cdot e^{(\frac{1}{\lambda} \cdot \omega_j)}]_i} \tag{3}$$

with:

- $N$ , the average number of bound molecules
- $a_j$ , chromatin state affinity at site j
- $\omega$ , the binding energy required for a TF to bind to site $j$ - in the form of a Position Weight Matrix Score
- $\lambda$ , a scaling factor for the Position Weight Matrix score
- $L$ , the length of the genome of interest
- $n$ , the ploidy level of the organism

Chromatin state affinity is defined as the following:

$$a_j = \sum_k \alpha_k \cdot c_j^k \tag{4}$$

with $\alpha$ the chromatin state affinity score and $c$ the chromatin state at site $j$. Conceptually, $\alpha$ represents the inferred chromatin state affinity for a given TF. This remains constant over the entire genome. $c$ represents the chromatin state at site $j$. Both $\alpha$ and $c$ are of length $k$, the number of different chromatin states. While $\alpha$ is characterised by a vector of length $k$ with inferred affinity scores for each chromatin state, $c$ is a vector of length $k$ describing the presence or absence of a chromatin state at site $j$ (0 for absent and 1 for present). One and only one chromatin state can be present at site $j$.

To uncover potential TF affinities for chromatin states, I developed a genetic algorithm with at its heart the core functionalities offered by ChIPanalyser. The main aim was to infer chromatin state affinities as well as number of bound molecules and $\lambda$ using the model described in equation (3). A genetic algorithm can be described as a machine learning algorithm that mimics natural selection. A population evolves over time and only the fittest individuals make it to the next generation carrying over their distinct traits. The starting population will fluctuate by undergoing both cross-over events and mutations for every trait. In this case, the starting population is characterized by a set of 14 traits: Number of bound molecules, a scaling factor, a PWM threshold ( as described in Chapter 1), and a starting affinity towards each of the 11 states. Traits are contained in a so called "chromosome". Chromosomes contain the values assigned to each trait. For the purpose of this analysis, all traits would select a random value within a predefined set of values. The pre-defined values for each trait are described as following:

- $N$ : 1,10,20,50,100,200,500,1000,2000,5000,10000,20000,50000,100000,200000,500000,1000000
- $\lambda$: 0 to 5 by 0.25 increments
- PWM Threshold: 0 to 1 by 0.1 increments
- Chromatin States: 0 to 1 by 0.2 increments

The starting population contained 100 individuals. Only 10 of these individuals are carried over to the next generation. Between each generation there is a 0.2 probability of mutation. To stay true to sexual reproduction, cross-over events occurred for every individual with a randomised extent of "chromosome" cross-over. The genetic algorithm ran for a total of 50 generations. Fitness of each individual was computed based maximising or minimising goodness of fit metrics. While ChIPanalyser offers 12 different metrics to asses goodness of fit, I elected to only use MSE, AUC and recall. These three metrics have shown to be either the most reliable at describing the goodness of fit of the model or are commonly used metrics in machine learning approaches.

It should be noted that a more straight forward approach could have been considered by overlapping ChIP experimental peaks and chromatin states. If a given TF would have an increased affinity for a set of chromatin states, there would be a higher number of ChIP peaks within that chromatin state. One could calculate an affinity score based on the proportion of peaks within each chromatin state taking into account the extent of each chromatin state. This approach could be taken a step further by also considering peak local enrichment. A TF's affinity score would not only consider the proportion of peaks within each chromatin state but also the "strength" of peaks within each range. It is conceivable that the strength of TF binding events would differ between chromatin states. Interestingly, computing affinity scores using the aforementioned method would provided a good starting point for the model. The affinity scores could be directly used within equation (1). The results presented in this chapter assume no prior knowledge of chromatin state affinities are provided.

*How fit can you get?*

After running the genetic algorithm on each data set for 50 generations, the best performing "individual" of each generation was extracted and plotted. Most data sets converged to an optimal solution after 20 generations. Figure-20 shows AUC, MSE, and recall scores over 50 generations for each data set in Kc167 cells. Figure-19 shows AUC, MSE, and recall scores over 50 generations for each data set in BG3 cells. Finally, Figure-21 shows AUC, MSE, and recall scores over 50 generations for Hox data sets in Kc167 cells. Unsurprisingly, the performance of BEAF-32 remains similar to the score obtained when using only DNA accessibility (see Table-8). If BEAF-32 only binds to accessible DNA, it would be expected for BEAF-32 to show a higher affinity towards states of open chromatin and low affinity towards other states. CTCF showed a slight improvement compared to only using DNA accessibility (see Table-8). CTCF would bind in chromatin states defined as open chromatin or closely related states. Finally, su(Hw) showed the highest score improvement compared to only using DNA accessibility (see Table-8). In this case, it is expected for su(Hw) to show a higher affinity towards chromatin states associated with heterochromatin.

| Data set | N | lambda | MSE | N | lambda | AUC | N | lambda | recall |
|---|---|---|---|---|---|---|---|---|---|
| BG3 modEncode 282 CTCF | 5e+05 | 0.75 | 0.008 | 500 | 0.25 | 0.901 | 20 | 0.25 | 0.86 |
| BG3 modEncode 3280 CTCF | 50000 | 1 | 0.01 | 500 | 0.25 | 0.86 | 500 | 0.25 | 0.808 |
| BG3 modEncode 3671 CTCF | 2e+05 | 0.75 | 0.009 | 500 | 0.5 | 0.892 | 2000 | 0.5 | 0.847 |
| BG3 modEncode 3672 CTCF | 10000 | 1.75 | 0.003 | 200 | 0.25 | 0.8 | 10000 | 1.25 | 0.762 |
| BG3 modEncode 3673 CTCF | 1e+05 | 1 | 0.008 | 500 | 0.25 | 0.892 | 500 | 0.25 | 0.846 |
| BG3 modEncode 3674 CTCF | 50000 | 1.25 | 0.01 | 500 | 0.25 | 0.889 | 500 | 0.25 | 0.843 |
| BG3 modEncode 3663 BEAF-32 | 2e+05 | 3 | 0.011 | 5000 | 3.5 | 0.908 | 1e+06 | 3.5 | 0.85 |
| BG3 modEncode 3664 BEAF-32 | 2e+05 | 2.5 | 0.012 | 50 | 4.5 | 0.949 | 10 | 4.5 | 0.894 |
| BG3 modEncode 3665 BEAF-32 | 5e+05 | 3 | 0.014 | 1 | 3.5 | 0.938 | 1 | 3.5 | 0.867 |
| BG3 modEncode 921 BEAF-32 | 50000 | 2 | 0.006 | 200 | 4.5 | 0.902 | 200 | 4.5 | 0.867 |
| BG3 modEncode 3714 Su(Hw) | 1e+05 | 2.5 | 0.02 | 5000 | 0.75 | 0.918 | 5000 | 0.75 | 0.831 |
| BG3 modEncode 3715 Su(Hw) | 1e+05 | 2.5 | 0.019 | 1000 | 0.5 | 0.93 | 5000 | 0.75 | 0.837 |
| BG3 modEncode 3716 Su(Hw) | 1e+05 | 2 | 0.015 | 1000 | 0.75 | 0.928 | 1000 | 0.75 | 0.846 |
| BG3 modEncode 3717 Su(Hw) | 1e+05 | 2 | 0.017 | 1000 | 0.75 | 0.926 | 1000 | 0.75 | 0.841 |
| BG3 modEncode 3718 Su(Hw) | 20000 | 2 | 0.007 | 100 | 0.75 | 0.905 | 200 | 0.75 | 0.838 |
| BG3 modEncode 951 Su(Hw) | 1e+05 | 2 | 0.017 | 5000 | 0.75 | 0.921 | 5000 | 1 | 0.839 |
| Kc167 GSM762842 CTCF | 20000 | 1.25 | 0.006 | 200 | 2.5 | 0.939 | 5000 | 2 | 0.921 |
| Kc167 modEncode 908 CTCF | 1e+05 | 1 | 0.011 | 100 | 0.75 | 0.891 | 1000 | 0.75 | 0.841 |
| Kc167 GSM1535963 BEAF32 | 2e+05 | 2 | 0.009 | 1 | 5 | 0.972 | 10 | 5 | 0.942 |
| Kc167 GSM762845 BEAF32 | 50000 | 1.75 | 0.005 | 1000 | 5 | 0.957 | 1000 | 5 | 0.934 |
| Kc167 modEncode 3801 Su(Hw) | 1e+05 | 3 | 0.012 | 200 | 0.75 | 0.903 | 1000 | 0.75 | 0.829 |
| Kc167 Su(Hw) | 1e+05 | 1.75 | 0.005 | 2e+05 | 2.5 | 0.952 | 2e+05 | 2.5 | 0.928 |
| Kc167 Ubx | 1e+06 | 1.5 | 0.004 | 20000 | 0.5 | 0.83 | 5e+05 | 0.5 | 0.773 |
| Kc167 Dfd | 2e+05 | 0.75 | 0.003 | 5e+05 | 0.5 | 0.854 | 5e+05 | 0.5 | 0.806 |
| Kc167 Abdb | 1e+06 | 1.5 | 0.003 | 10 | 1 | 0.832 | 5e+05 | 0.75 | 0.767 |

Table 8: **Optimal Parameters obtained with a genetic algorithms after 50 generations.**White columns are N and $\lambda$ minimising MSE. Light grey are N and $\lambda$ maximising AUC and dark grey columns are N and $\lambda$ maximising recall. Architectural proteins as well as Hox TFs are included.

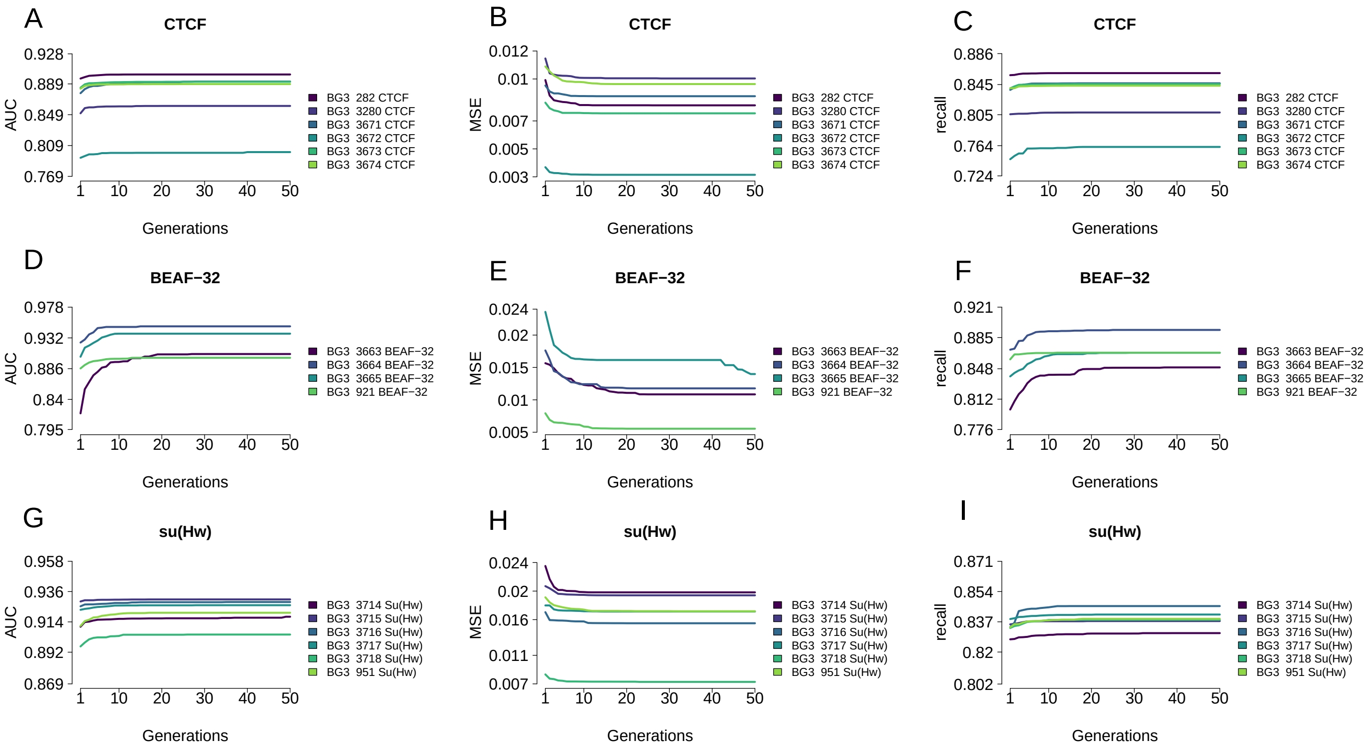


Figure 19: **Genetic Algorithm performance over 50 generations in BG3 cells.** After running the genetic algorithm for 50 generations, the best performing "individual" at each generation was extracted based on three goodness of fit metrics in BG3c cells. Most datasets converge to an optimal set of parameters after 20 generations. **A-C** shows AUC, MSE and recall scores for CTCF over 50 generations. **D-F** shows AUC, MSE and recall scores for BEAF-32 over 50 generations. **G-I** shows AUC, MSE and recall scores for su(Hw) over 50 generations.

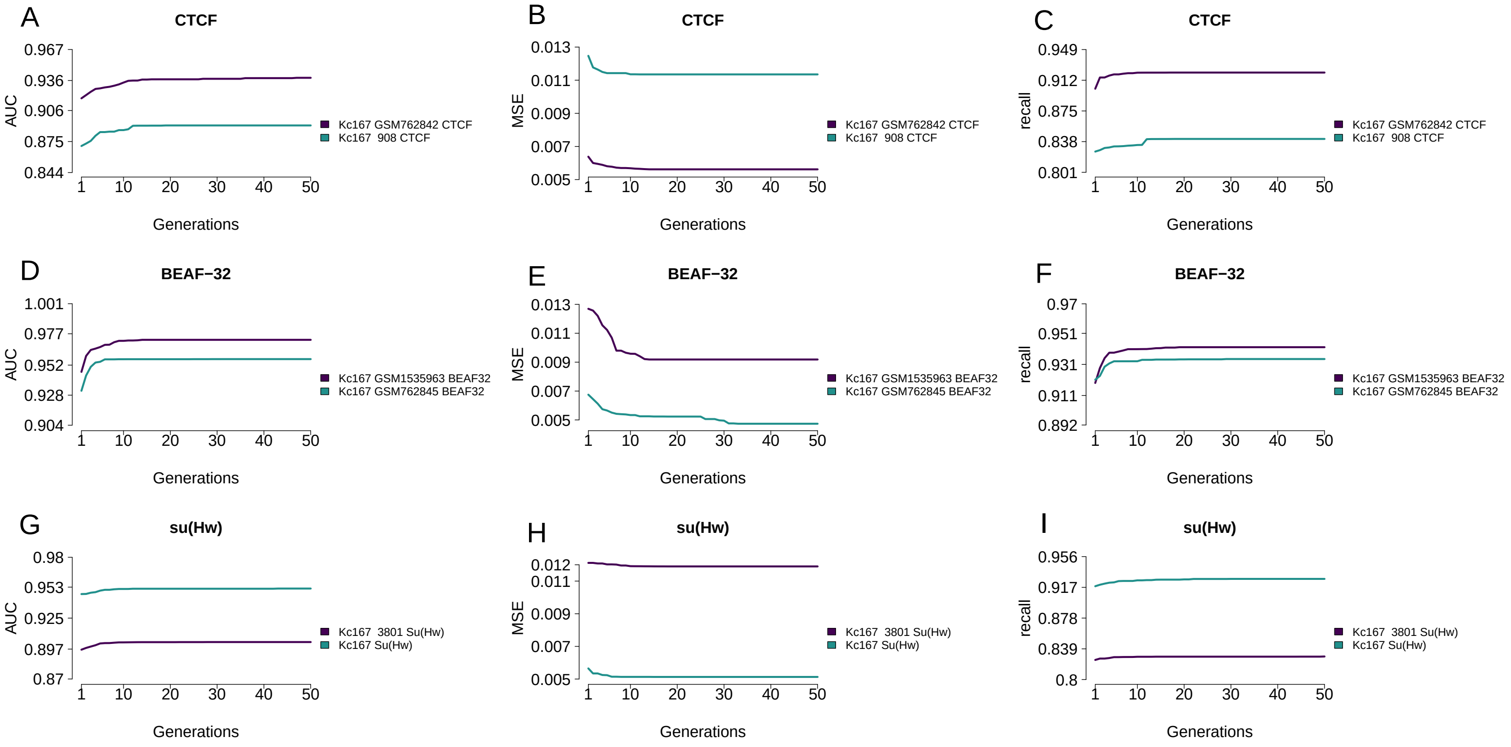


Figure 20: **Genetic Algorithm performance over 50 generations in Kc167 cells.** After running the genetic algorithm for 50 generations, the best performing "individual" at each generation was extracted based on three goodness of fit metric in Kc167 cells. Most datasets converge to an optimal set of parameters after 20 generations.**A-C** shows AUC, MSE and recall scores for CTCF over 50 generations. **D-F** shows AUC, MSE and recall scores for BEAF-32 over 50 generations. **G-I** shows AUC, MSE and recall scores for su(Hw) over 50 generations.

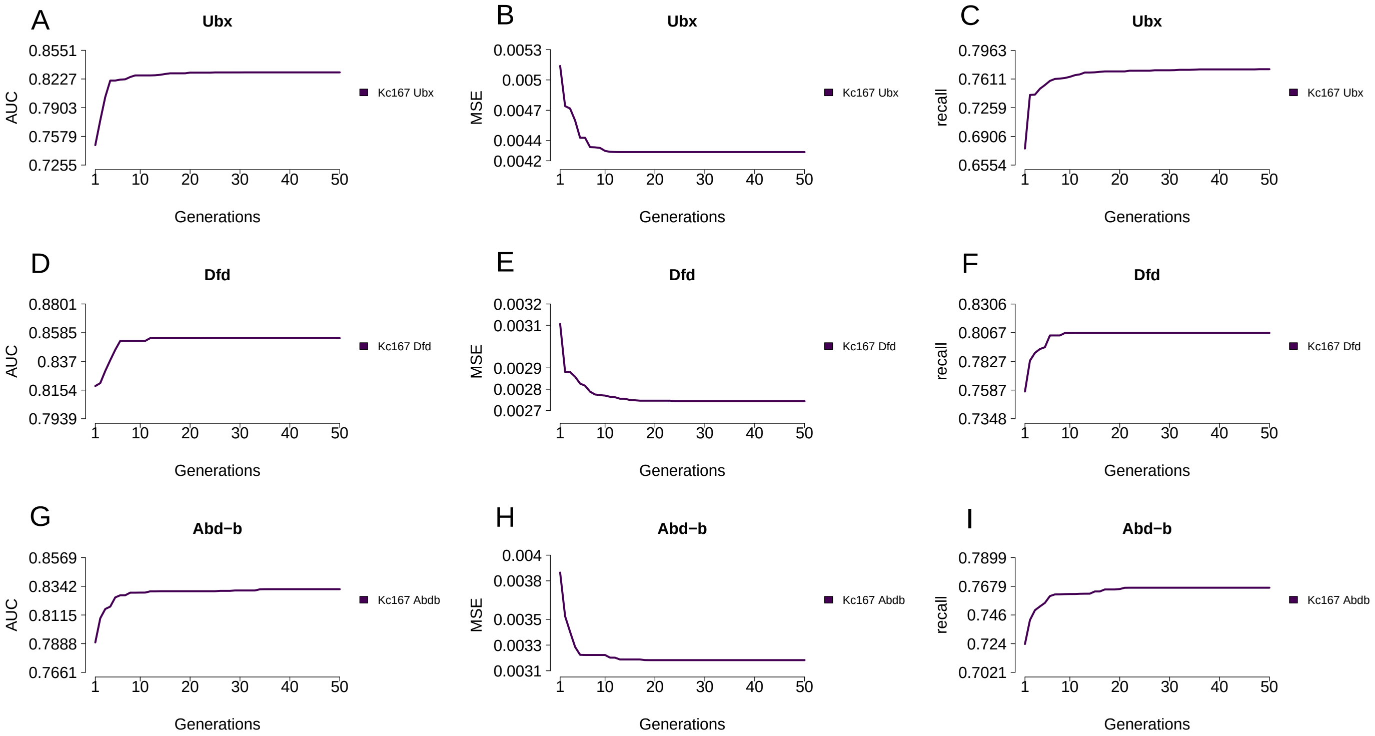


Figure 21: **Genetic Algorithm performance over 50 generations in Kc167 cells for Hox TFs .** After running the genetic algorithm for 50 generations, the best performing "individual" at each generation was extracted based on three goodness of fit metrics in Kc167 cells for Hox TFs. Datasets converge to an optimal set of parameters after 20 generations.**A-C** shows AUC, MSE and recall scores for Ubx over 50 generations. **D-F** shows AUC, MSE and recall scores for Dfd over 50 generations. **G-I** shows AUC, MSE and recall scores for Abd-b over 50 generations.

## CHROMATIN STATE AFFINITY

Architectural DNA binding proteins point in the direction of having varying affinities towards different chromatin states. In order to verify if I could recover the expected chromatin state affinities, I extracted the best performing "individuals" of the last generation by maximising or minimising a goodness of fit metric (MSE, AUC and recall). Chromatin states affinities for the selected individuals in each data set were combined and averaged. Variance in chromatin state affinities was also extracted and computed from selected individuals. It should be noted that a single data set can produce numerous individuals with the same fitness score while displaying varying affinity scores. Variance of affinity scores show the fluctuations in affinity scores over replicates as well as within a single data sets. Even when only one replicate is available, variance scores can still be extracted based on top performing individuals.

### *Architectural Proteins show preferences towards different chromatin states*

For each TF, chromatin state affinity and variance is summarised in Figure-22. BEAF-32 preferentially binds to active transcription start sites (aTSS) and to a lesser extent to enhancer regions. This holds true between both cell lines and between goodness of fit metrics (see Figure-22 **B/E**). In BG3 cells, Figure-22 **E** shows an increased affinity for heterochromatin. However, the variance between affinity scores is also increased. Increased variance decreases confidence in chromatin state affinity scores. Furthermore, High variance in affinity scores suggests that these chromatin states could play a lesser role in TF binding events.

CTCF binds to both aTSS and enhancers. Figure-22 **A** and **D** show a high affinity for both of these states in both cell lines. Interestingly, CTCF also displays a intermediate affinity towards the competent state. The Competent state is considered as an enhancer regions in the process of being fully activated or conversely enhancers being repressed. Despite slight variation, all three metrics indicate intermediate to high affinity towards the competent state.

Finally, su(Hw) tells a more complex and nuanced story. su(Hw) binds with an intermediate to low affinity towards Elongation states and active introns. Unsurprisingly, su(Hw) exhibited an increased affinity (intermediate to high) towards heterochromatin, heterochromatin in euchromatin and basal. However, the higher variance observed for these states suggest that something else is afoot.

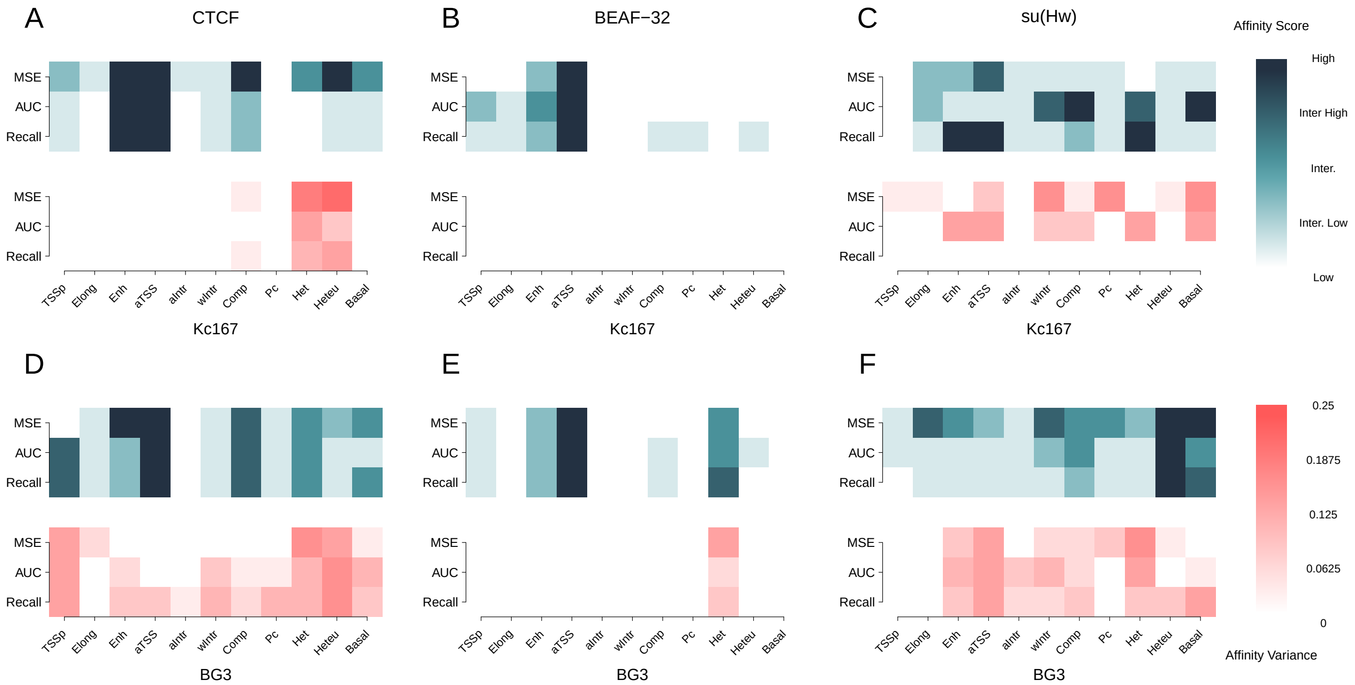


Figure 22: **Chromatin state affinity scores for Architectural proteins.** In order to determine chromatin state affinity for architectural proteins, I extracted and averaged the best performing "individuals" for each data set and averaged their respective affinity scores. To ensure, the robustness of affinity scores, the variance in affinity scores between top "individuals" was also computed. CTCF displays a high affinity towards enhancer and active TSS regions but also intermediate affinity towards the competent state (**A-D**). BEAF-32 is illustrated by a clear affinity towards active TSS and to a lesser extent enhancer regions (**B-E**). Finally, su(Hw) displays complex and nuanced patterns towards many chromatin states (**C-F**). Nevertheless, su(hw) shows an increased affinity towards hetreochromatin and closed chromatin states.

After 50 generations, the optimal set of parameters and affinities were then applied to the validation set. The validation set contained all other regions that were not contained in the training set. Figure-23 shows predicted profiles in the validation set for the three architectural proteins in BG3 cells. The red lines represents the predicted profiles. The dark blue area shows genome occupancy. Underneath each profile, the multi-coloured rectangles are the different chromatin states and their respective ranges. The colour code can be found on the right hand side of the plot. Predicted profiles accurately recover experimental data both in location and in local enrichment for CTCF (Figure-23 **A**), BEAF-32 (Figure-23 **B**) and su(Hw) (Figure-23 **C**). Figure-24 shows predicted and experimental data in Kc167 cells. In both cell lines, the prediction over the validation set remains similar. These results demonstrate ChIPanalyser's ability to accurately predict binding using chromatin states.

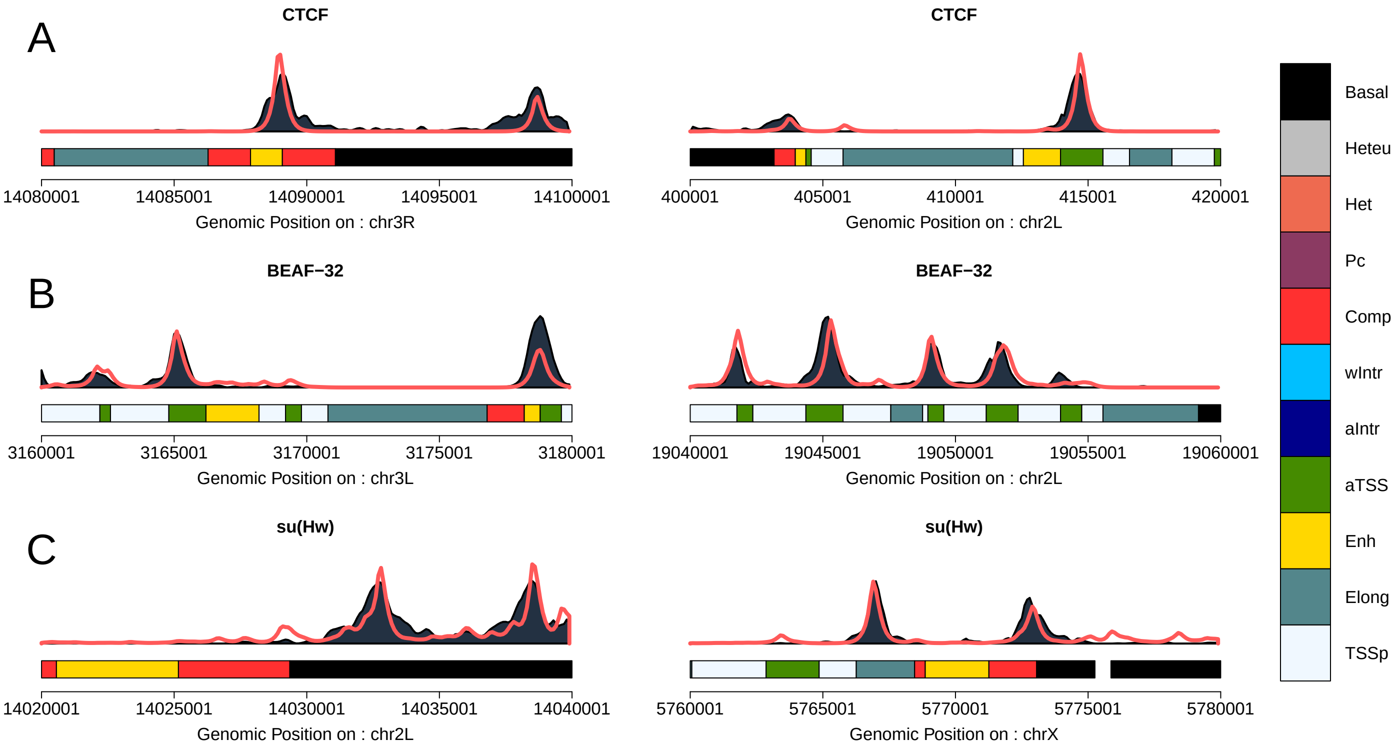


Figure 23: **Predicted Profiles for architectural proteins in BG3 cells.** After training the model and the genetic algorithm on the top 100 regions, the optimal set of parameters were then applied to a validation set in BG3 cells. The resulting predicted profiles for CTCF are described in (**A**). (**B**) shows predicted profiles for BEAF-32 and (**C**) displays the predicted profiles for su(Hw). The red line represents the prediction while the dark blue area describes ChIP data. Coloured rectangles described the chromatin state and their extent.

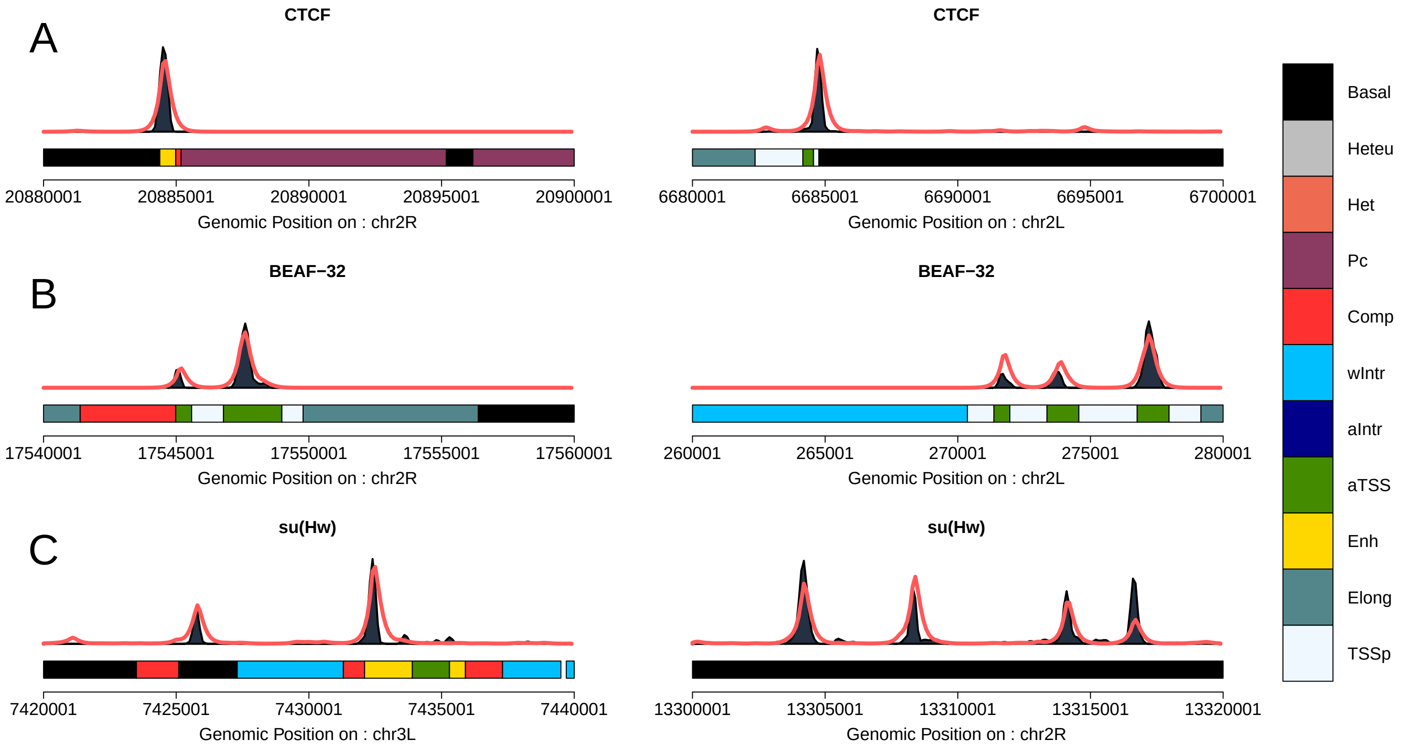


Figure 24: **Predicted Profiles for architectural proteins in Kc167 cells.** After training the model and the genetic algorithm on the top 100 regions, the optimal set of parameters were then applied to a validation set in Kc167 cells. The resulting predicted profiles for CTCF are described in (**A**). (**B**) shows predicted profiles for BEAF-32 and (**C**) displays the predicted profiles for su(Hw). The red line represents the prediction while the dark blue area describes ChIP data. Coloured rectangles described the chromatin state and their extent.

### *Hox TFs show preference towards different chromatin states*

In the previous chapter, Hox TFs were shown to have differential preferences towards DNA accessibility. Ubx binds in open chromatin whereas Dfd and Abd-b would bind in less accessible DNA. By including chromatin states into the model, would the same binding preferences arise? I ran the genetic algorithm over 50 generations and extracted the best performing "individuals" as described above. The chromatin state affinity scores for top performing "individuals" were combined and averaged. Score variance was also extracted following the same protocol as for architectural proteins. Figure-25 shows the affinity scores for each TF. Ubx clearly displays a strong preference towards Enhancer marks. Interestingly, when using MSE as goodness of fit metric, active TSS was also picked up as preferred chromatin state. These results recover Ubx's preferences towards open chromatin as described in the previous chapter. Both Dfd and Abd-b reveal a preference towards a wider range of chromatin states. Unsurprisingly, both have a high affinity towards enhancer marks. Hox TFs are involved in development and thus play an active role in gene expression. Both TFs show a high affinity towards active TSS as well an intermediate affinity towards introns (active and weak),and competent states. Dfd distinguishes itself by also showing an intermediate to high affinity towards heterochromatin for at least two of the metrics while Abd-b displayed intermediate to low affinity towards polycomb states. Many of these states show varying levels of DNA accessibility or gene activity. This demonstrates that the binding of Dfd and Abd-b in varying levels of DNA accessibility can in fact be explained by their affinity towards certain chromatin states. In all case the low variance suggests that these states are stable between high ranking individuals. As described above, single replicates may yield multiple individuals with the same fitness score while displaying varying affinity scores. However, the lack of biological replicates could also imply that these affinity scores are the results of over fitting. Affinity score should then be considered with caution.

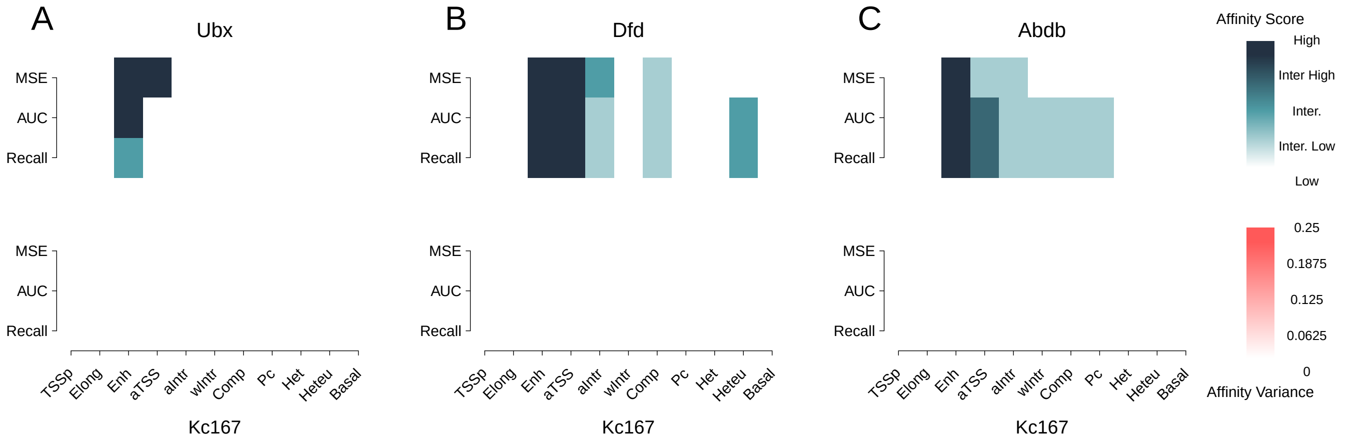


Figure 25: **Chromatin state affinity scores for Hox TFs .** In order to determine chromatin state affinity for architectural proteins, I extracted and averaged the best performing "individuals" for each data set and average their respective affinity scores. To ensure the robustness of affinity scores, the variance in affinity scores between top "individuals" was also computed. Ubx demonstrated a clear preference towards enhancer states (**A**) while Dfd and Abd-b displayed high affinity towards enhancer states as well as states related to repressive DNA (**B-C**).

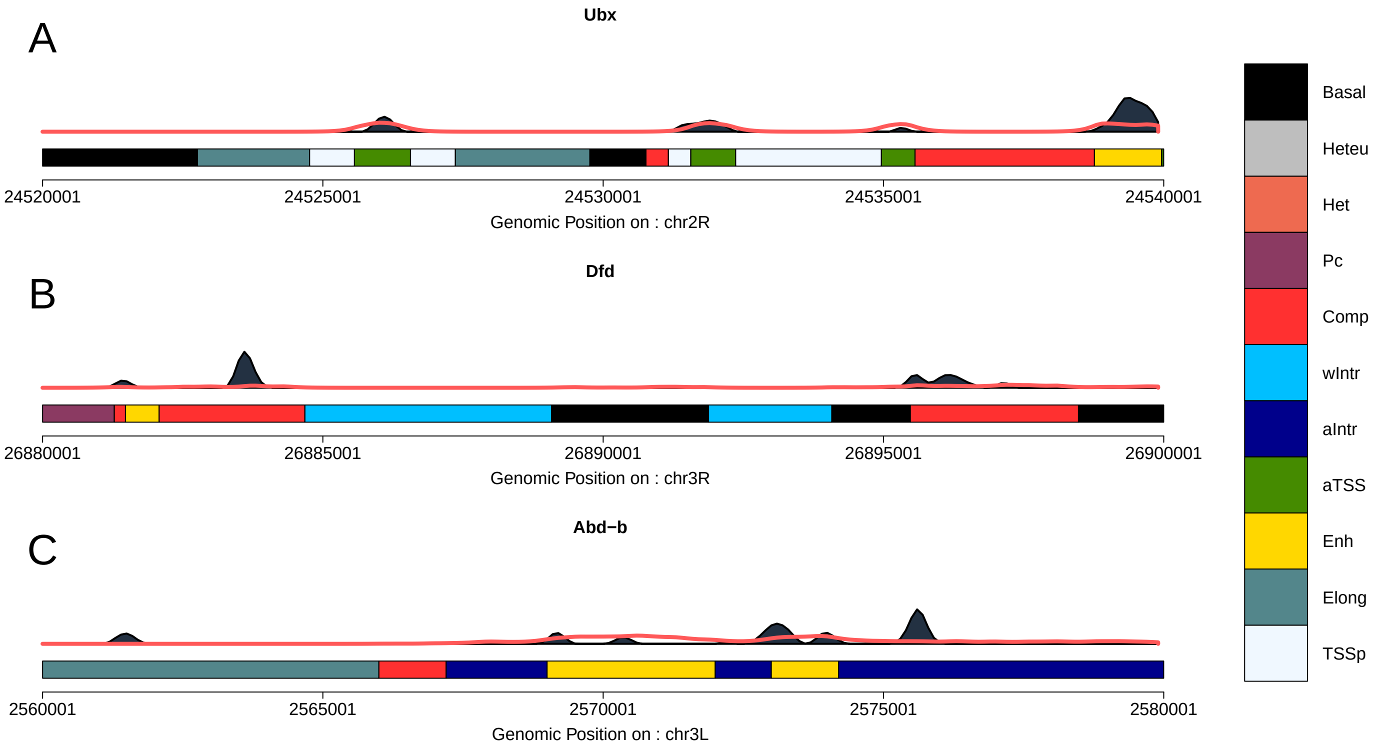


Figure 26: **Predicted Profiles for Hox TFs in Kc167 cells.** After training the model and the genetic algorithm on the top 100 regions, the optimal set of parameters were then applied to a validation set in Kc167 cells for Hox TFs. The resulting predicted profiles for Ubx are described in (**A**). (**B**) shows predicted profiles for Dfd and (**C**) show display the predicted profiles for Abd-b. The red line represents the prediction while the dark blue area describes ChIP data. Coloured rectangles described the chromatin state and their extent.

Once the optimal parameters and chromatin state preferences were computed and extracted, the model was applied to the validation set. Figure-26 describes the predicted profiles compared to ChIP-seq data. Interestingly, the profiles are the opposite of what was found when using varying levels of DNA accessibility. All profiles are fairly flat and seem to underestimate the height of the peaks. Previously, the profiles strongly over estimated the height of peaks. However, Hox TF data sets showed low peak enrichment. Only a few peaks were associated with a strong signal and this lower signal could affect the models ability to correctly assess parameters. Flatter predicted profiles could also be a clear indication of cooperative binding. Despite increasing the information content of genomic DNA, there are other factors ( or co-factors should I say) that drive the binding of Hox TFs.

*Binding of Architectural proteins is not disrupted by low changes in protein abundance*

The genetic algorithm also optimised for number of bound molecules and $\lambda$. The selected optimal values were stable across top individuals for each data set. Both parameters were extracted from the best performing individuals and averaged. The optimal parameters for each cell line are shown in Table-8. The estimated number of bound molecules remains for the most part within biologically acceptable boundaries. Optimal parameters remains fairly consistent between data sets and cell lines further demonstrating that the models abundance estimates are robust. The estimated number of bound molecules for Hox TF is shown in Table-8. In this case, the model over estimated the number of bound molecules for all Hox TFs.

Previously, I demonstrated that protein abundance plays a lesser role in the binding of architectural proteins. Differential binding between cell lines is driven by changes in DNA accessibility between cell lines and to a much lesser extent by changes in proteins abundance. Here, I demonstrated that chromatin states significantly improves the predictions of the model and can explain the nuanced role of chromatin states in TF binding. Based on these results, changes in chromatin states would be sufficient to explain deferential binding between cell lines. In order to test this hypothesis, the model was trained on the top 100 region in Kc167 cells. Then, the optimal parameters and affinity scores were applied to BG3 cells. The number of bound molecules was rescaled based on relative RNA levels as well as divided by a factor of ten and one hundred. Figure-27 **A-C** provides predicted profiles for CTCF, BEAF-32 and su(Hw) in Kc167 cells. Figure-27 **D-F** shows the predicted profiles in BG3 cells after carrying over the number of bound molecules and rescaling. All three architectural proteins still remain unaffected by small to medium changes in protein abundance. Predicted profiles only displayed a decrease in enrichment after a strong reduction in number of bound molecules (10 Fold or 100 Fold decrease). The effect on model performance was fairly low and is most likely due to changes in dataset quality as seen in Figure-27 **G-I**. These results further confirm that differential binding between cell lines is driven by changes in chromatin states rather than changes in abundance.

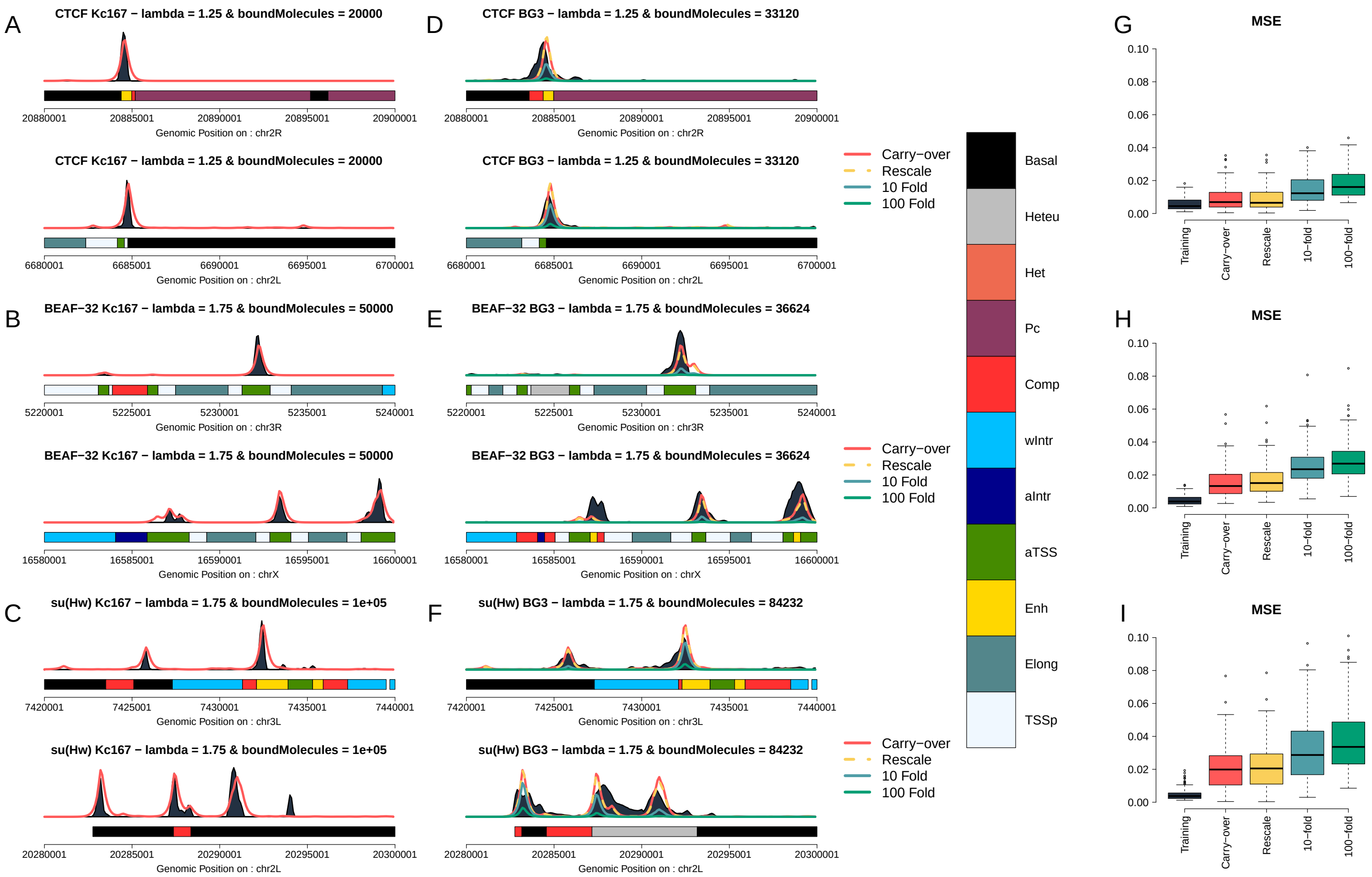


Figure 27: **Differential binding between cell lines is mainly driven by changes in chromatin states.** In order to investigate the role of protein abundance, the model and the genetic algorithm was trained on the top 100 regions in Kc167 cells. The resulting parameters where then applied to BG3 cells. The number of bound molecules was rescaled using relative mRNA levels as well as divided by a factor of 10 and 100. Predicted profiles in Kc167 cells are shown in **A,B** and **C** for CTCF, BEAF-32 and su(Hw) respectively. Estimated profiles in BG3 cells are illustrated in **D,E** and **F** for CTCF, BEAF-32 and su(Hw) respectively. Architectural proteins are robust to small to medium changes in proteins abundance between cell lines. Only strong changes in protein abundance leads to changes in predicted profiles. Furthermore, the performance of the model remains similar between cell lines after RNA rescaling (**G-I**). Changes in MSE are likely due to differences in dataset quality.

*Chromatin states enable low affinity binding site recognition*

Architectural binding proteins demonstrated different behaviours with respect to the number of regions selected for validation. CTCF was less well predicted when the number of validation regions increased. This suggested that CTCF preferentially binds to genome hotspots or highly conserved binding motifs. BEAF-32 on the other hand displayed little variation with the increased number of regions. BEAF-32 would bind anywhere along the genome as long as it is accessible. Finally, su(Hw) showed bias towards increase number of regions only when all DNA was considered accessible. The effect of DNA accessibility was the strongest driver towards su(hw) binding. Based on these results, I wanted to investigate if the same behaviour would be observed with the addition of chromatin states. I extracted MSE, AUC and recall after increasing the number of regions included in the validation set. Then, I averaged the score over selected regions. Figure-28 and Figure-29 show the change in goodness of fit score for all data sets combined in Kc167 and BG3 cells respectively.

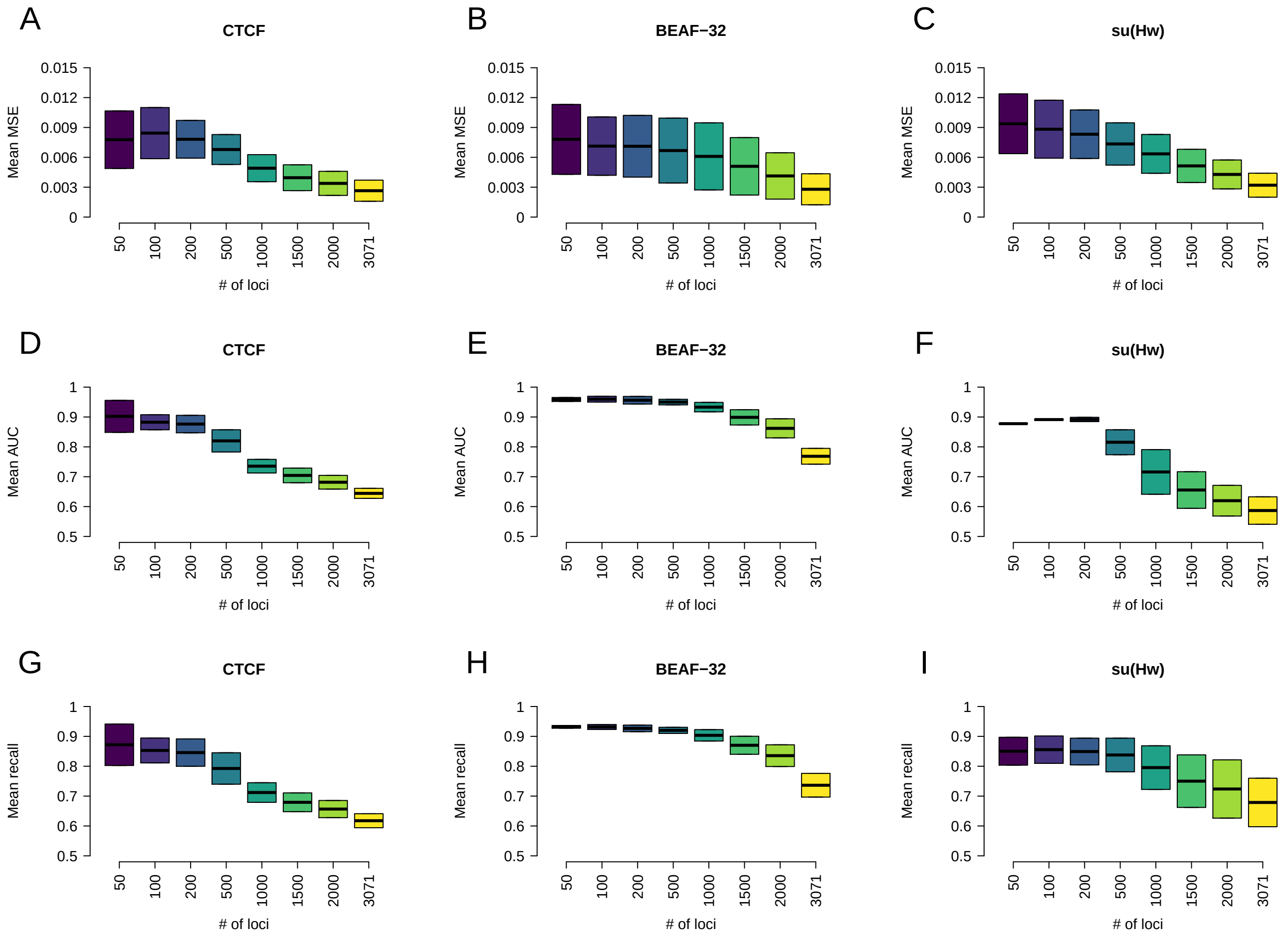


Figure 28: **Increased number of validation regions tells tales of binding preferences in Kc167 cells.** The performance of CTCF dropped after 500 regions were included in the validation set (**A**,**D** and **G**. BEAF-32 remained unaffected by the increased number of regions (**B**,**E** and **H**) The slight drop around 2000 regions is likely due to the lack of BEAF-32 peaks. Finally, su(Hw) was less well predicted after 500 regions were included in the validation set (**C**,**F** and **I**).

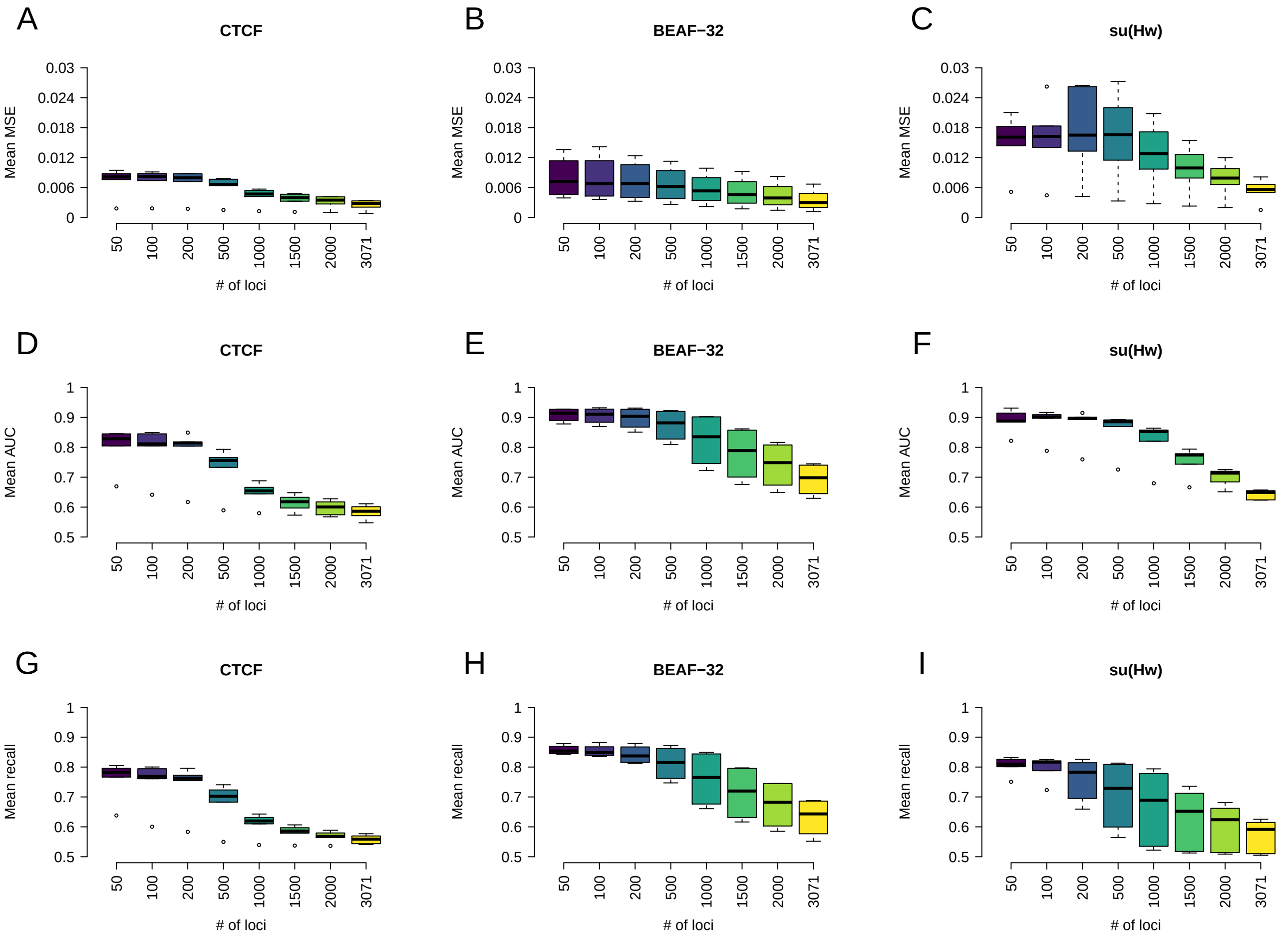


Figure 29: **Increased number of validation regions tells tales of binding preferences in BG3 cells.** The performance of CTCF dropped after 500 regions were included in the validation set (**A**,**D** and **G**. BEAF-32 showed a slight decrease in performance when including more regions for validation.(**B**,**E** and **H**) The slight drop around 1500 regions is likely due to the lack of BEAF-32 peaks. Finally, su(Hw) was less well predicted after 500 regions were included in the validation set (**C**,**F** and **I**).

The inclusion of chromatin states shows a slightly different tale. In both cell lines, the ability to predict CTCF binding drops after 500 regions are included in the validation set. Interestingly, this drop occurs much later than when only DNA accessibility was included. CTCF still bind to hotspots (or stronger binding sites) but cell specific binding events are driven by changes in chromatin states rather than accessibility on its own. These results go hand in hand with CTCF's predicted affinity towards chromatin states. CTCF showed the ability to bind in states that would not always be considered accessible (i.e competent state ). In both cell lines, increasing the number of regions slightly reduced the overall ability of the model to predict BEAF-32 binding. However, performance decay occurs at a much slower rate. BEAF-32 still binds the entire genome in regions of open chromatin ( as described by BEAF-32's chromatin state affinity scores - see Figure-23 and Figure-24). The performance drop could be the direct consequence of the absence of BEAF-32 peaks in those regions. Finally, the model's ability to predict su(Hw) binding decreased as the number of regions increased. These results are similar to the ones found when all DNA was considered accessible. su(Hw) would preferentially bind to high affinity sites but in closed chromatin. Conversely to CTCF, differential binding between cell lines might not only be driven by changes in chromatin states. The difficulty to clearly pinpoint su(Hw)'s binding mechanisms could imply that su(Hw) requires cofactors in order to specifically bind to its target sites.

## DISCUSSION

The addition of chromatin states to the model introduces a more precise perspective on the binding of architectural proteins and developmental TFs. Moreover, the overall performance of the model is improved by introducing chromatin states.

The results presented in this chapter suggest that the binding of architectural proteins is strongly driven by chromatin states. As expected, BEAF-32 displayed a high affinity towards active TSS and to a lesser extent towards enhancers further confirming its preferences towards open chromatin. As the introduction of chromatin states do not change the mechanistic interpretation of BEAF-32 binding, BEAF-32 will not be further discussed in this chapter. CTCF was preferentially bound to enhancers, active TSS but also competent states. This demonstrates CTCF's ability to bind into both open chromatin and into intermediate states. Finally, su(Hw) showed varying affinities towards many different states but most noticeably towards heterochromatin. Despite displaying an increased affinity towards repressive chromatin states, the binding affinities of su(Hw) are challenging to assess.

### *Chromatin affinity scores and affinity variance*

As described above, I extracted the variance associated with each affinity score. Multiple affinity scores can be associated with the same goodness of fit score. The variance in affinity score describes the "spread" of these scores over the top performing "individuals". All top performing "individuals" are described by the same goodness of fit score but not necessarily the same values associated to each parameter. For example, let a TF *A* have an AUC score of 0.9 and mean affinity score associated with heterochromatin of 0.7 but with a 0.35 variance ( values described are just examples and do not represent real data). One could describe TF *A* as having an intermediate to high affinity towards heterochromatin based on the mean affinity score. However, TF *A* is also characterised by a high variance for heterochromatin. This suggests that regardless of the affinity score associated with heterochromatin, it would not impact the model's performance. In this case, the binding of TF *A* is driven by something else other than heterochromatin on its own. Multiple and drastically different affinity scores can be associated with heterochromatin while still maintaining the same model performance. In the case of low variance for a given affinity score, it would demonstrate that changing the value of this affinity score would affect model performance quite significantly. In the example given above, if TF *A* now displayed a low variance towards heterochromatin while maintaining the same affinity score, one could conclude that heterochromatin could drive the binding of TF *A* and that TF *A* is capable of binding to heterochromatin.
Changes in affinity scores demonstrate how certain TFs show preferential binding in certain chromatin states other others. The purpose of the genetic algorithm is to

optimise these values and uncover which chromatin state is the main driver of binding for a given TF. Theoretically, it would be possible to assign near equal chromatin state affinities across all states. However, this would yield similar results as considering all DNA to be accessible as described in the previous chapter. Affinity scores modulate the probability of a TF binding to a given site. If all affinities are given then same value, then chromatin states are not technically considered within the model. Affinity scores would become a factor multiplying the probability of binding uniformly over the entire genome. This approach does not conform with the heterogeneous nature of chromatin *in vivo*.

*The role of chromatin states in CTCF binding*

CTCF is known to play the role of a chromatin insulator. It can block the interaction between enhancers and target genes [Kim et al., 2015, Nichols and Corces, 2015]. It has also been shown to act as a barrier against hetereochromatin spreading [Guelen et al., 2008, Van Bortle et al., 2014]. Finally, there is evidence suggesting that CTCF plays a direct role in transcriptional regulation [Tang et al., 2015, Smith et al., 2009]. Taken together, recovered CTCF chromatin affinity supports the chromatin state model. High affinity towards active TSS and enhancer demonstrates CTCF involvement in direct transcriptional regulation. The increased affinity towards the competent state could be an indication of the model picking up on CTCFs role as a chromatin insulator. Despite the models ability to correctly recover CTCF binding mechanisms, there are still some unanswered questions. The model's ability to predict CTCF drops when increasing the number of regions used for validation. These results suggest that CTCF binds to genome hotspots, generally correlated with highly conserved binding sites [Nakahashi et al., 2013] or that CTCF is better explained at stronger binding sites. However, the drop in performance could also imply that there is an increased number of false positive peaks appearing in these regions. As described above, regions selected for analysis should contain at least one peaks of any of the three architectural proteins from any dataset. Many of these regions will not contain any CTCF peaks but the model will still sometimes predict CTCF binding. This increase in false positives will irremediably lead to decrease in goodness of fit. Interestingly, the drop in model performance occurred much later when chromatin states were included into the model suggesting that chromatin states do play a strong role in low affinity CTCF binding. Changes in chromatin states would be responsible for cell specific CTCF binding. However, there are many binding motifs that could theoretically be bound but are not. Many of these sites could be inaccessible for CTCF binding from a structural perspective. The complex folding of DNA within cells could create pockets of DNA that are not accessible ( or less accessible ) for TF binding despite being in an advantageous chromatin state. Furthermore, it would seem that chromatin structure would play a role in guiding CTCF to its target sites [Hansen et al., 2019]. Finally, lower affinity sites could be bound by other DNA binding proteins inhibiting the binding of CTCF.

### *The role of chromatin states in su(Hw) binding*

The binding of su(Hw) is just as complex to unravel. The results presented in this chapter describe su(Hw) as being able to bind to many chromatin states with varying affinities. As expected, su(hw) binds with high affinity in heterochromatin (and heterochromatin in euchromatin) but also intermediate to high affinity towards Polycomb. Both of these states are associated with repressive chromatin. Affinity towards these states can easily be explained by su(Hw)'s role as chromatin insulator and association with lamina associated domains (LAD) [Kurshakova et al., 2007, Kuhn-Parnell et al., 2008, van Bemmel et al., 2010]. Curiously, su(Hw) also displayed intermediate affinity towards active marks such enhancer or active TSS suggesting a direct role in transcription. However, recent studies have shown that su(Hw) is directly involved in transcriptional repression by blocking enhancer to gene interactions [Adryan et al., 2007, Melnikova et al., 2019]. Increased affinity towards enhancer marks could be reflective of such a role. It should be noted that many chromatin state scores were also accompanied with affinity score variance. High affinity score variance is indicative that changes in affinity scores do not affect fitness scores significantly. Resulting affinity scores should be taken with caution with respect to biological significance. The mechanism of su(Hw) binding remain difficult to explain. There have been some studies suggesting that su(Hw) binding involves co-factors [Baxley et al., 2017, Glenn and Geyer, 2019].

### *The role of chromatin states in Hox TF binding*

I had previously shown that Hox TFs display different binding preferences with respect to DNA. In this chapter, I investigated if this preference could be explained by changes in chromatin state affinities. The results presented here suggest that Hox TFs show different affinities towards chromatin states. Ubx exhibited a clear preference towards enhancers, a chromatin state associated with open chromatin. On the other hand, both Dfd and Abd-b were illustrated by high affinity towards enhancers and active TSS but also intermediate affinity towards competent state and active introns. Dfd was also characterised by a intermediate affinity for hetrochromatin while Abd-b displayed a intermediate affinity towards weak intron and Polycomb states. Generally, these states are associated with repressive chromatin or at least less permissive chromatin. Unfortunately, the predicted binding profiles showed a low agreement with Hox ChIP data. Predicted profiles greatly underestimated the binding of Hox TFs while still estimating a high number of bound molecules.
Hox TF rely on their homeodomain in order to bind to DNA [Kuziora and McGinnis, 1989]. It has been suggested that homeodomains form a loose bond with DNA and the binding of Hox TF requires multiple binding sites in close proximity [Pellerin et al., 1994, Chauvet et al., 2000b, Rezsohazy et al., 2015]. Most of the peaks observe in the pre-

dicted binding profiles are wide despite lacking height. This suggests that the model picks up on the multiple sites but fails to asses the importance of site clustering in Hox binding. Furthermore, Hox TFs bind to DNA cooperatively [Moens and Selleri, 2006, Mann et al., 2009, Joshi et al., 2010]. In order to recover peak enrichment, the model overestimated the number of bound molecules (see Table-8).The binding of Hox TFs would require Hox TFs and associated co-factors to be present in the right abundance in order to trigger gene expression [Petkova et al., 2019]. Mechanistically, the model predicts a higher number of bound TFs to compensate for the fact that co-factors are not included. The current model does not include cooperative binding and thus can explain why it does not pick up on Hox TF enrichment in predicted profiles. While the model description of chromatin state is plausible, the contribution of all parameters should be taken into consideration in order to understand the binding of Hox TFs.

*Binding of Architectural proteins between cell lines is by driven chromatin states*

Previously, I demonstrated that the binding of CTCF, BEAF-32 and su(Hw) was driven by changes in DNA accessibility rather than changes in protein abundance. Along those lines, I hypothesised that the same results would hold true with the addition of chromatin states into the model. The results presented in this chapter demonstrate that cell specific binding of architectural proteins is not driven by changes in proteins abundance. Changes in predicted profiles were only observed after a 10 fold decrease in estimated number of bound molecules. Architectural proteins are involved both in genome architecture maintenance but also regulation of essential genes [Van Bortle et al., 2014]. The maintenance of these structure and regulatory mechanisms are crucial to cellular homoeostasis hence the robustness in the face of concentration fluctuations.

*Genetic algorithms as interpretable machine learning*

The prediction of TF binding has been a hot topic for many years and many machine learning algorithms have done an amazing job a predicting TF binding. However, in many case, what machine learning has in predictive power, it sometimes lack in explainability. The creation of explainable machine learning models has become a crucial aspect of modern genomics. Genetic algorithms are one of the many proposed solutions for interpretable machine learning. The strength of genetic algorithms resides in the fact that all parameters are known in advance. A user will know exactly which values go into the algorithm and which values are selected by the algorithm as main driver towards accurate predictions. This is especially important in the case of ChIPanalyser. The model underlying ChIPanalyser describes TF binding using a set of predetermined parameters such as chromatin state affinity or number of bound molecules. This makes genetic algorithms particularly suited for being used in concert with ChIPanalyser. While other method and tools have successfully predicted and explained TF binding [Tareen and Kinney, 2019, Avsec et al., 2019, Salekin et al., 2017], genetic algorithms are the best suited for the work presented in this thesis. However, despite the success , there are a few limitations that should be mentioned. First, the problem related to choice of goodness of fit metric as discussed in chapter 1 still remains an issue. Each metric will tend to penalise different aspects. While MSE will recover peak enrichment but produces an increased amount of false positive hits, metrics such as AUC and recall will accurately predict peak location but generally miss peak height. Second, the case of su(Hw) is reflective of issues arising with an increased feature space. There are close to 200 million possible parameter combinations in the current model. The genetic algorithm displayed a complex binding behaviour for su(Hw) with respect to chromatin statse. This could be the result of hitting a local minima or maxima. However given the increase in goodness of fit scores and that the model converges consistently after 20 generations, this scenario remains unlikely.

## CONCLUSION

In this chapter, I described the addition of chromatin states to the existing ChIPanalyser model. In order to ascertain chromatin state affinities for architectural proteins and Hox TFs, I developed a genetic algorithm using the core functionalities offered by ChIPanalyser. Chromatin states are a strong driver towards the differential binding of TFs. Binding preferences towards chromatin states and accessibility can accurately be recovered by the model. Furthermore, the addition of chromatin states confirms that the cell specific binding of architectural proteins is driven by changes in chromatin states rather than changes in TF abundance. Overall, ChIPanalyser recovers known binding mechanisms. This further demonstrates the packages ability to gain insight into the mechanisms of TF binding.

# 4

# INVESTIGATING SU(H)

## CHAPTER SUMMARY

The following chapter describes the binding mechanisms of Su(H). Thanks to the wealth of data available, Su(H) is an ideal case study to further investigate TF binding mechanisms with ChIPanalyser. I demonstrate that Su(H) preferentially binds to open chromatin and that DNA accessibility is sufficient to explain its binding. I show that Notch activation increases Su(H) binding to DNA using Notch induced and Non induced ChIP data sets. Finally, I show that Su(H) binding is affected by changes in protein abundance.

## INTRODUCTION

The Notch signalling pathway is one of the most conserved signalling pathways in the animal kingdom. On top of its high conservation, Notch signalling operates in numerous cell types and varying stages of development [Weinmaster et al., 1992, Bray, 2016]. Simply put, NOTCH signalling is triggered by the activation of NOTCH receptors at the cytoplasmic membrane by NOTCH ligands. Activated receptors initiate the proteolytic cleavage of receptors and the release of the Notch Intracellular domain (NICD). The NICD enters the nucleus to interact with CBF1/Su(H)/LAG1 family of TFs. The interaction between the NICD and nuclear factors will trigger the expression of target genes [Wang et al., 2014, Yashiro-Ohtani et al., 2014, Wang et al., 2015].

Su(H) (fondly known as suppressor of hairless) is part of the *Drosophila* DNA binding proteins in the Notch pathway. Interestingly, Su(H) is constitutively present within the nucleus and it still binds to its targets at a lower rate [Bray, 2016]. The induction of the Notch signalling pathway translocates the NICD to the nucleus and the interaction of Su(H) and the NICD increases Su(H) binding [Krejčí and Bray, 2007, Gomez-Lamarca et al., 2018]. However, increased binding is thought to be reflective of increased binding at Su(H) target sites rather than an increase in Su(H) abundance. The role of Su(H) abundance is not to be neglected. Changes in Su(H) abundance affects its ability to trigger gene regulation at target genes [Wang et al., 2015]

Su(H) preferentially binds in open chromatin [Lake et al., 2014]. Recent work has suggested that Su(H) not only binds in accessible DNA but directly affects chromatin accessibility by increasing H3K56ac around binding sites. Chromatin opening would be the result of both chromatin state recognition and co-factors recruitment such as Lz/Runx [Skalska et al., 2015]. Notch activation would induce a higher binding rate of Su(H) consequently increasing H3K56ac spreading. Interestingly, while the majority of binding motifs were found in the Basal state, Su(H) peaks were shown to be mainly located in enhancer states and active TSS. These results suggests that chromatin environment plays a significant role in Su(H) binding specificity.

Given the wealth of data available, Su(H) is the ideal candidate to thoroughly test the ability of ChIPanalyser to describe TF binding mechanisms. The binding mechanisms of Su(H) have been thoroughly studied experimentally but the questions remains: Can the model recover those mechanisms *in silico*? First, I will show that the binding of Su(H) is indeed better described when DNA accessibility is considered. Second, I will demonstrate that Notch induction does not seem to be accompanied by in increase in Su(H) abundance but rather an increase in DNA occupancy. Third, the model accurately recovers chromatin state preferences towards open chromatin. Finally, I will demonstrate Su(H)'s sensitivity towards TF abundance by training the model in Wild Type ChIP-seq and comparing rescaled predictions in partial RNAi knock-downs.

## MATERIALS AND METHODS

### *DNA Sequence*

For *Drosophila melanogaster*, the dm6 version of the genome was used. References genomes are available in R within the BSgenome suit of packages [Pages, 2018]. When required, data was aligned to the dm6 version of the genome.

### *Binding Motifs*

The binding motif for Su(H) was downloaded from the JASPAR database [Mathelier et al., 2014]. PFM and PPM are converted to PWM by ChIPanalyser using the method described by Stormo [Stormo and Zhao, 2010].

### *Genome binding profiling*

ChIP-seq data sets for Su(H) were produced by Zabet, Skalska and Bray (Unpublished). This includes constitutive Su(H), Notch-Induced Su(H) as well as Su(H) RNAi partial knock-downs. Notch-Induced data sets represent data sets where the NOTCH pathway was chemically induced by cleaving NOTCH receptors and subsequent translocation of the NICD. The experimental method is described by [Skalska et al., 2015]. Datasets were aligned to the (dm6) genome using bowtie-2 (–non deterministic) [Langmead and Salzberg, 2012]. *SAM* files were converted to *BAM* files using smatools [Li et al., 2009]. Peaks and pile-up signal were called using macs2 with a 0.01 FDR [Zhang et al., 2008].

### *DNA Accessibility and chromatin states*

DNase I hypersensitivity data was generated by modEncode for *Drosophila* cell line, BG3. Fastq files were aligned to the dm6 genome using bowtie-2 (–non-deterministic) [Langmead and Salzberg, 2012]. *SAM* files were converted to *BAM* files using samtools [Li et al., 2009]. Peaks and read pile-up were called using macs2 (-broad-call -cutoff 0.05 -q 0.05) [Zhang et al., 2008]. The role of DNA accessibility on Su(H) binding was assessed three ways: (i) All DNA considered accessible (No Access), (ii) only DHS sites are considered accessible (DHS), (iii) varying a DNA accessibility threshold based on quantized pile-up scores (QDA). DNase I hypersensitivity pile up scores were split using varying threshold. Only regions with scores above a given threshold were considered accessible. Chromatin state maps were produced by [Skalska et al., 2015]. Chromatin state maps were produced for both pre and post Notch activation.

### *Analysis workflow*

For the purpose of this analysis, I analysed the behaviour of Su(H) at 10 known functional loci, namely drm, Mtk, trio, Rgl, W, CG32425, Irc, E(spl), kirre, and N. Selecting these regions for analysis ensured that only the strongest Su(H) peaks are considered. Contrarily to architectural proteins that display close to 30 000 peaks across the genome, Su(H) displays less than 200. Previously, I demonstrated that ChIPanalyser performs well using only the top ten regions. The same principle is applied for this analysis.

When investigating the role of DNA accessibility on Su(H) binding, ChIPanalyser was used on its own. Chromatin state affinity was investigate using the same genetic algorithm as described in the previous chapter.

## RESULTS

### *Su(H) preferentially binds to open chromatin*

Work of the binding of Su(H) has shown that Su(H) preferentially binds in open chromatin [Lake et al., 2014]. Based on these results, it would be safe to assume that ChIPanalyser would better predict the binding of Su(H) when DNA accessibility is included. To test this hypothesis, the model was trained on the selected loci with and without DNA accessibility data. In this case, only DHS peaks were considered.

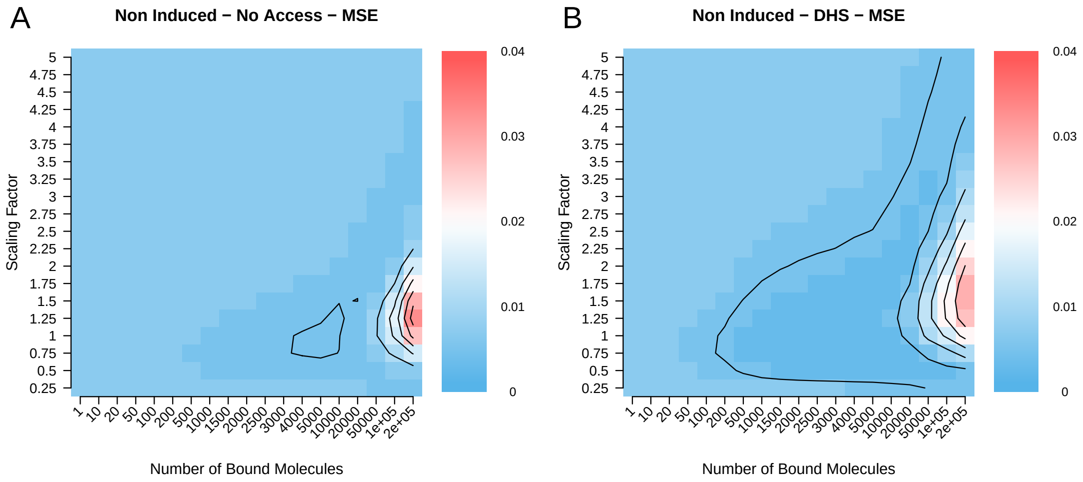


Figure 30: **SuH preferentially binds in open chromatin.** The optimal set of parameters were computed with and without accessibility Data by minimising MSE. (**A**) shows optimal parameters when all DNA is considered accessible. (**B**) shows optimal parameters when DHS accessibility is considered. MSE drops with the inclusion of DNA accessibility. Furthermore, the estimated bound molecules also seem to decrease (from 5000 to 3000)

The optimal set of parameters were computed between the two conditions (No Access and DHS) in non-induced Notch. The overall performance of the model is improved by the addition of DNA accessibility in the model (see Figure-30). MSE displays a slight decrease with the inclusion of DNA accessibility data. Predicted profiles show a better agreement with ChIP-seq data when DNA accessibility is included. Figure-31 shows a comparison in predicted profiles between the two conditions. Figure-31 **A** shows predicted profiles without accessibility. As expected, there is an increased level of false positive hits. Figure-31 **B** shows how DNA accessibility can successfully reduce the number of available binding sites and increase the predictive power of the model. These results show that Su(H) prefers binding in open chromatin and ChIPanalyser can recover this behaviour.

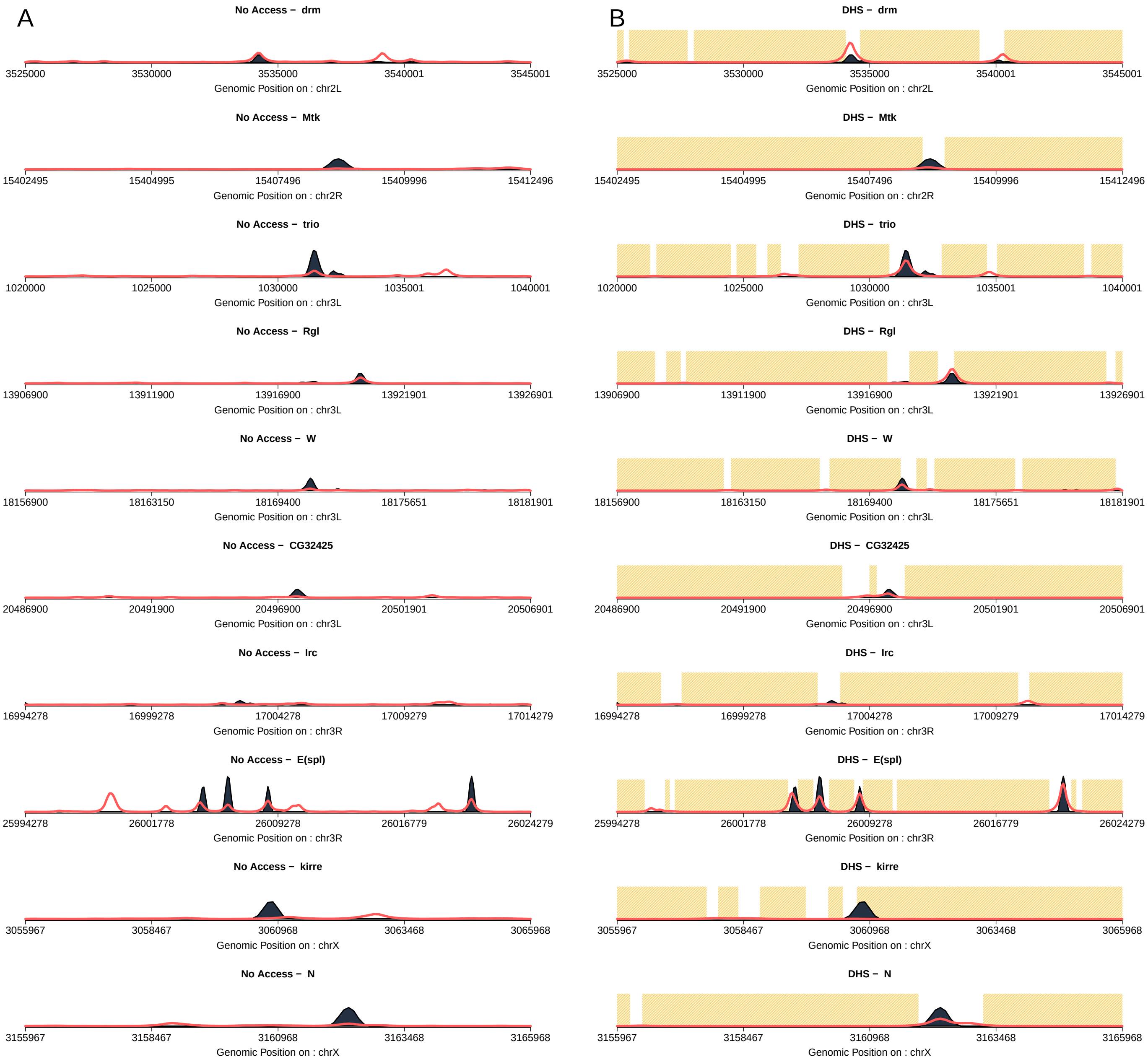


Figure 31: **Su(H) preferentially binds in open chromatin.** (**A**) shows predicted profiles without accessibility data. The model recovers some Su(H) peaks however this is also accompanied with an increase in false positive binding. (**B**) shows the predicted profiles with DNA accessibility data included. DNA accessibility reduces the number of available binding and in turn reduces the number of false positive predictions. The red line represents predicted profiles while the dark blue area represents ChIP data. Finally, yellow boxes are regions of inaccessible DNA.

The induction of the Notch signalling pathway is accompanied by an opening of chromatin at Su(H) binding sites. To investigate if the model could accurately recover this mechanisms, I trained the model using varying levels of open chromatin. More specifically, 17 QDA threshold between 0.9 and 0.999 were selected. For the purpose this analysis, I elected to use QDA thresholds that would closely mimic DHS peaks and by extension biologically relevant DNA accessibility. Furthermore, the model was trained on Notch induced and Non-Induced ChIP-seq data. Notch induced should perform better with slightly relaxed QDA thresholds compared to Non-Induced Notch. Unfortunately, I was unable to recover chromatin opening by Su(H) binding (see Figure-32). Interestingly, Figure-32 **A** and **B** illustrates changes in AUC scores and MSE with varying QDA levels. AUC scores peaks around 0.96 QDA in both induced and non-induced condition while MSE is at its lowest with 0.99 QDA in both conditions. The resulting profiles in induced and non-induced are extremely similar (see Figure-32 **C** and **D**). The changes in DNA accessibility induced by Notch binding are likely to be too minute for the model to pick up on this mechanisms. Furthermore, slight variation in DNA accessibility would be offset by changes in ChIP-seq quality. A slight increase in open chromatin windows are also accompanied by varying levels of ChIP noise. Together, these opposing factors negate the ability of the model to pick up on chromatin opening by Su(H) binding.

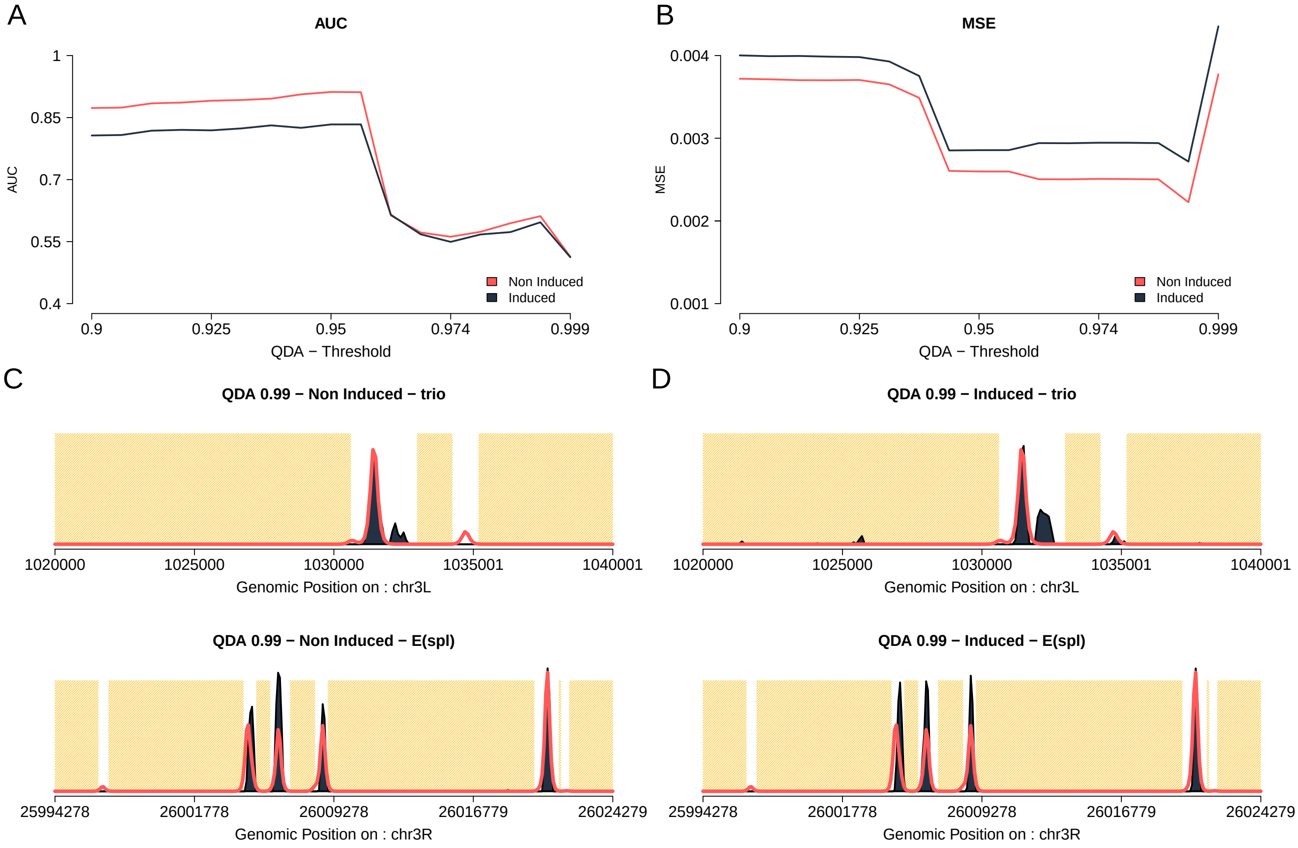


Figure 32: **Su(H) shows little to no effect on DNA accessibility around binding sites.** Using varying levels of DNA accessibility, I attempted to recover chromatin opening around Su(H) binding sites after Notch induction. (**A**) and (**B**) show no indication that Notch induction increases DNA accessibility at Su(H) binding sites. Furthermore, (**C**) and (**D**) display no difference in predicted profiles between the two conditions.

*Notch Induction increases Su(H) to DNA binding*

The induction of the Notch signalling pathway drives the Notch Intra-cellular domain to translocate to the nucleus. The NICD along with Su(H) will result in an increased binding to DNA. However, experimentally, the translocation of the NCID does not seem to increase Su(H) levels but rather modulates binding specificity. This would suggest that the number of bound molecules should slightly increase after Notch activation. To test if ChIPanalyser could recover this behaviour, the optimal set of parameters were

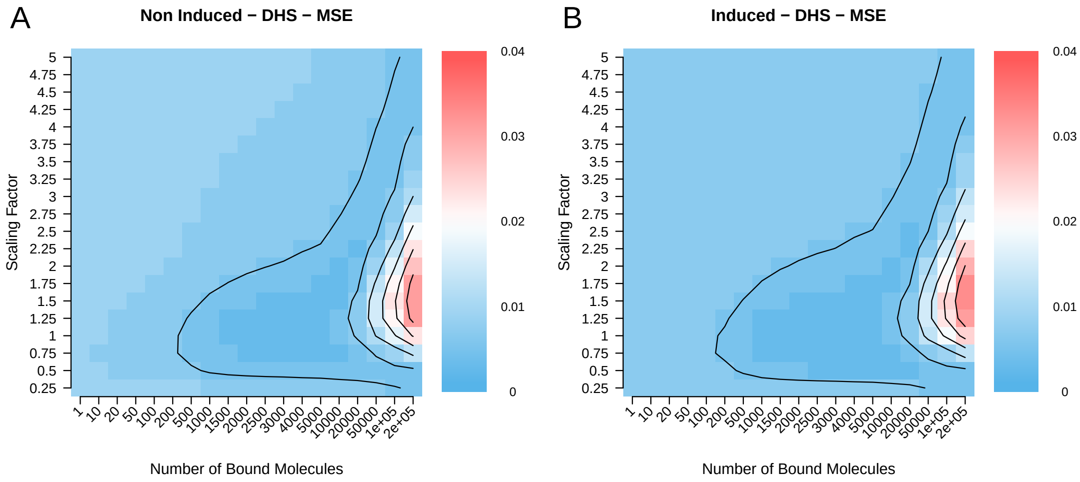


Figure 33: **Notch Induction increases Su(H) binding to DNA.** The model was trained on Notch Induced and Non-induced data sets and the optimal set of parameters were extracted. The number of bound molecules slightly increases after Notch Induction (**B**) (N=4000) compared to the Notch Non-Induced system (**A**) (N=3000)

computed by maximising or minimising a goodness of fit metrics (MSE, AUC, and recall) over selected regions between each condition. As DNA accessibility improves the binding predictions of Su(H), DHS was included in the model at this stage.

As reported previously, goodness of fit metrics are context dependant. MSE is the most apt metric at determining the optimal set of parameters (see Figure-33).The optimal number of bound molecules in slightly increased after Notch activation (N=4000) as compared to Non Induced Notch (N=3000). Optimal parameters are summarised in Table-9. Increased DNA occupancy by Su(H) after Notch activation is translated in a slight increase in peak enrichment (see Figure-32 **C**). ChIPanalyser accurately recovers the increase DNA occupancy by Su(H).

*Su(H) binding can be recovered by considering chromatin states*

Not only does Su(H) preferentially bind to open chromatin but shows preferences towards certain chromatin states. Su(H) peaks were mainly located in enhancers and

| Data set | N | lambda | MSE | N | lambda | AUC | N | lambda | recall |
|---|---|---|---|---|---|---|---|---|---|
| DHS Non-Induced | 3000 | 1.25 | 0.003 | 100 | 1.5 | 0.963 | 100 | 1.5 | 0.908 |
| DHS Induced | 4000 | 1.25 | 0.003 | 1e+05 | 2.5 | 0.920 | 1e+05 | 2.5 | 0.844 |
| CS Non-Induced | 20000 | 1 | 0.005 | 1e+05 | 1.5 | 0.924 | 2e+05 | 2 | 0.895 |
| CS Induced | 20000 | 0.25 | 0.006 | 1e+05 | 1.25 | 0.86 | 20000 | 1 | 0.82 |

Table 9: **Optimal parameters for Su(H) after minimising MSE and maximising AUC and recall.** The table includes optimal parameters inferred using DHS in both Notch systems (Induced and Non Induced). CS data sets refers to the optimal parameters selected after running the genetic algorithm for 50 generations. Both Notch systems are included (Induced and Non-Induced).

active TSS as well as active Introns [Skalska et al., 2015]. In the previous chapter, I described how using genetic algorithms in coordination with ChIPanalyser can uncover chromatin state binding preferences. Here, I investigated if the chromatin state affinities described experimentally can be recovered *in silico*. The genetic algorithm was trained in selected regions for 50 generations on both Notch Induced and Non-Induced. At every generation, only 10 out of 100 individuals made their way to the next generations. The selected few would reproduce and undergo both cross-over and mutations events. As described previously, mutations occurred with a 0.2 probability. MSE , AUC and recall were used as fitness scores.

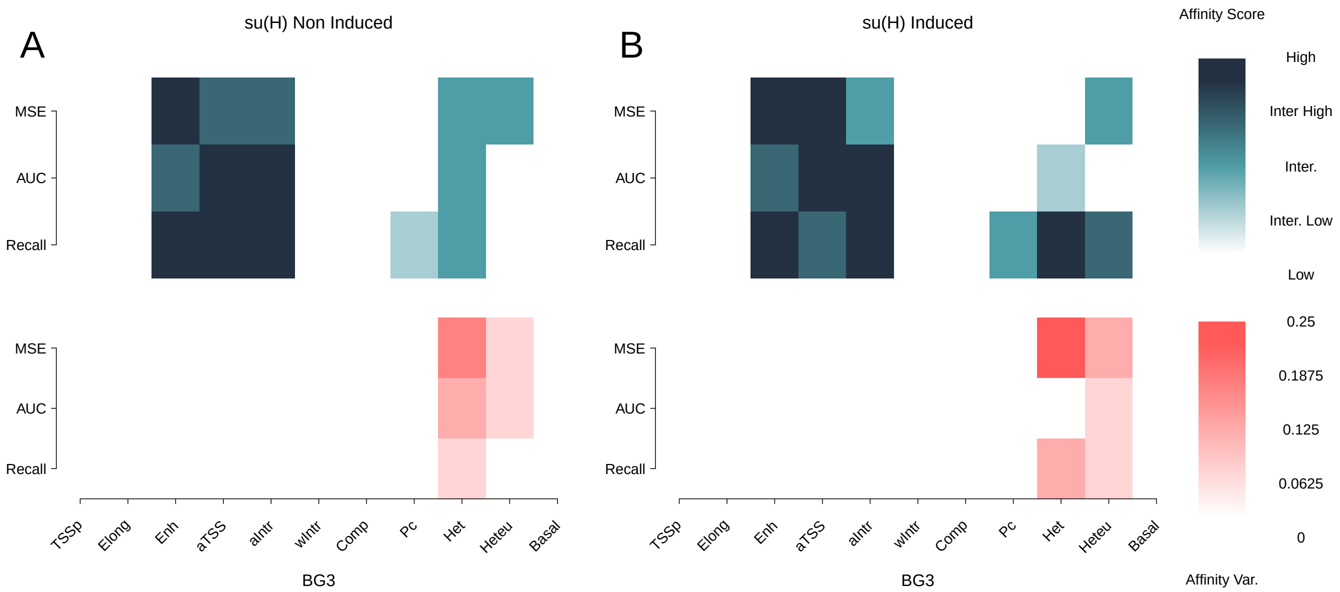
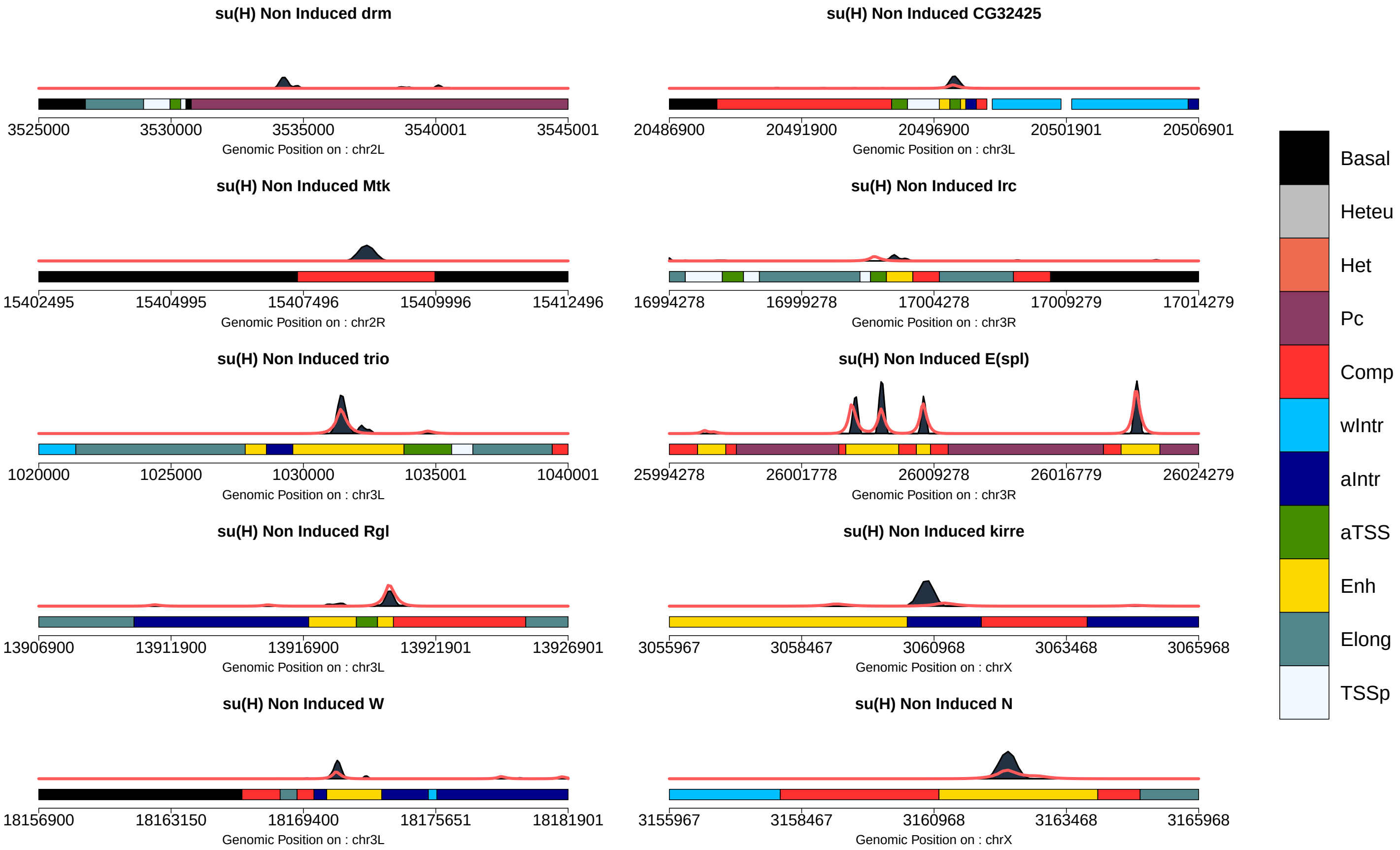


Figure 34: **Chromatin state affinity for Su(H) in BG3 cells.** After 50 generations, Su(H) displayed a clear preferences towards enhancers, active TSS and active introns in both Non-Induced (**A**) and Induced (**B**) conditions. Su(H) showed an intermediate affinity towards heterochromatin. However, the variability associated to these states suggests that they do not represent true biological affinities but rather an artefact of the genetic algorithm.

To describe chromatin state affinities, the fitness scores of the top "individuals" after 50 generations were extracted, combined and averaged. The variance of fitness scores was also computed in order to determine the robustness of affinity scores. High variance for a certain chromatin states illustrates a reduced role of that chromatin state in explaining the binding of Su(H). Regardless of the affinity score assigned to that chromatin state, the fitness score is not affected.

Figure-34 **A** shows the chromatin state affinity of Su(H) in Non-Induced Notch. Su(H) displays a clear preference towards Enhancers, active TSS and active introns. The same affinities are described in Figure-34 **B** in the Notch Induced system. In both case there also seems to be an intermediate to high affinity for Polycomb and various heterochromatin states. However, the variability associated to these states suggests that they do not represent true biological affinities but rather an artefact of the genetic algorithms. The chromatin state affinities recovered by ChIPanalyser accurately recovers the chromatin preferences determined experimentally [Skalska et al., 2015].

To exemplify the performance of the model, Figure-35 illustrates predicted profiles with respect to Su(H) ChIP-seq data in Notch Non-Induced BG3 cells. Interestingly, predicted profiles are similar to profiles presented in Figure-31 **B**. This similarity is unsurprising as Su(H) preferentially binds to accessible DNA. Enhancers, active TSS and active Introns are all chromatin states associated with open chromatin. Curiously, not all peaks are recovered by the model. This suggests that something else could drive the binding of Su(H).



su(H) Non Induced drm

su(H) Non Induced CG32425

su(H) Non Induced Mtk

su(H) Non Induced lrc

su(H) Non Induced trio

su(H) Non Induced E(spl)

su(H) Non Induced Rgl

su(H) Non Induced kirre

su(H) Non Induced W

su(H) Non Induced N

Figure 35: **Chromatin state illustrate Su(H) binding preferences in BG3 cells.** Predicted profiles after training the genetic algorithm for 50 generations. The red line represents the model prediction. The dark blue represents ChIP data. The coloured rectangle beneath the profiles show chromatin states and their respective range. Chromatin colour code is described on the right hand side. Predicted profiles generally recover Su(H) binding well after the inclusion of chromatin state affinity scores.

Translocation of the NICD to the nucleus is correlated with changes in DNA accessibility, more specifically with spreading of H3K56ac. Increase in histone acetylation would also imply changes in chromatin states. [Skalska et al., 2015] produced chromatin state maps before and after Notch Induction. Figure-34 illustrates that in both conditions chromatin state affinities remain similar if not identical. However, to ensure that changes in chromatin states between condition were effectively recovered by the model, I trained the genetic algorithm over 50 generations on Notch Non-Induced ChIP-seq data. Then, I selected the best performing parameters and validated chromatin state affinities ( as well as number of bound molecules and $\lambda$) on Notch Induced ChIP-seq data. The model accurately recovers the binding of Su(H) in the validation set. Figure-36 **A** shows the training profiles and Figure-36 **B** exhibits the validation profiles. Changes in chromatin states remain subtle and the most notable differences is found at the Esp(l) locus. The Competent states shifts and extends itself towards the enhancer state. This change is unsurprising as the competent state is often described as an enhancer that has not yet fully been activated. Increase in H3K56ac drives an increase in chromatin accessibility accompanied with a change in chromatin state. Interestingly, these results are also incredibly similar with the profiles describes when only DHS is used (see Figure-31 **B**). Despite recovering chromatin state affinities, these results would suggest that DNA accessibility is sufficient to explain the binding of Su(H) at the selected loci.

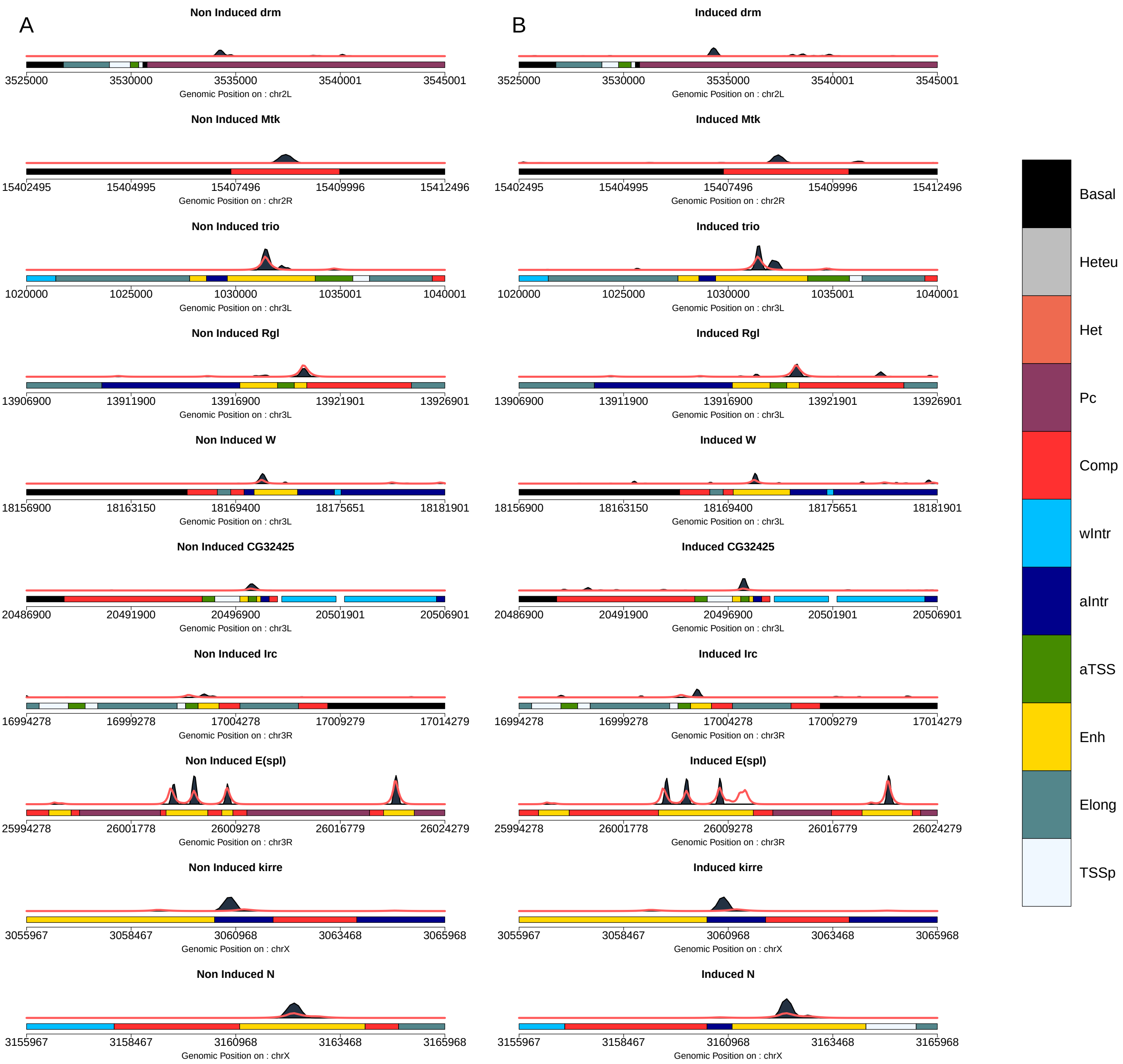


Figure 36: **Notch Induction induces changes in chromatin states.** The activation of the Notch pathway and the translocation of the NICD induces increased levels of H3k56ac around Su(H) binding sites. The model was trained on Non-Induced Notch Su(H) ChIP-seq data (**A**) and the optimal parameters where validated on Notch Induced ChIP data (**B**). Changes in chromatin states are minute. Nevertheless, the model accurately recovers binding profiles pre and post Notch activation. The red line represents the predicted profile. The dark blue shaded area represents ChIP data. The coloured rectangle show the various chromatin states and their extent.

*RNAi partial knock-down demonstrate Su(H) sensitivity towards changes in proteins abundance*

One question that might arise is how quantitative is the prediction of number of bound molecules. Can the model predict loss of binding in a knock-down experiment? Partial Knock-downs of Su(H) using RNAi constructs demonstrated that Su(H) binding is sensitive to changes in TF abundance. It was measured that RNAi knock-downs would lead to a 70% decrease in Su(H) abundance and that this decrease lead to a reduction in Su(H) peak enrichment. I investigated the model's ability to predict changes in Su(H) binding after RNAi treatment. First, ChIP scores of control and knock-downs were cross normalised in order to ensure that I would recover changes in read density. Then, I trained the model on wild type Non-Induced Su(H) ChIP data and simply multiplied the estimated number of bound molecules by a factor of 0.3. Using these parameters, I predicted changes in binding profiles of the RNAi knock-downs. For the purpose of this analysis, I opted to use DNA accessibility instead of chromatin states. Despite accurately recovering chromatin state preferences, Su(H) binding was mainly driven by open chromatin. DNA accessibility on its own is sufficient to explain Su(H) binding. Furthermore, this drastically reduced computational time as the feature space is significantly smaller in the case of DNA accessibility only.

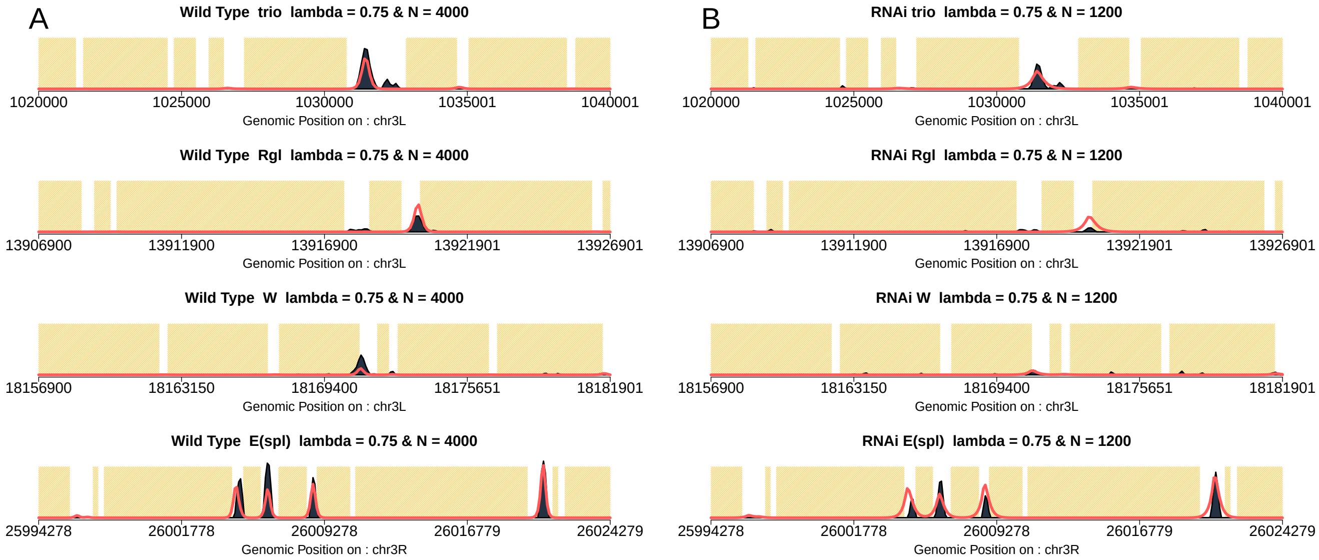


Figure 37: **ChIPanalyser recovers partial Su(H) knock-downs in BG3 cells.** The model was trained in wild type Su(H) and the optimal parameters ( N and $\lambda$) were estimated by minimising MSE (**A**). The number of bound molecules were then rescaled to mimic the extent of the RNAi knock-downs ( 70% reduction in abundance). The model recovers the drop in Su(H) enrichment (**B**). The red line represents the predicted profiles. The dark blue shaded area represents ChIP data. Finally, yellow boxes are regions of inaccessible DNA.

The model was able to accurately recover the binding profiles between wild type and RNAi knock-down (see Figure-37). After RNAi knock-down, Su(H) ChIP profiles show a decrease in enrichment. By simply dividing the estimate Su(H) abundance, ChIPanalyser recovers the decrease in ChIP enrichment and the loss of peaks.

## DISCUSSION

The well studied case of Su(H) provides an interesting case study for ChIPanalyser. Numerous data sets are available describing the binding of Su(H) in various conditions. By using the wealth of data available, I sought to demonstrate that ChIPanalyser could recover the binding mechanisms of Su(H). The results presented here further demonstrate the ability of the model to describe TF binding with respect to nature of chromatin and TF abundance. ChIPanalyser has shown to be a useful tool in predicting and understanding TF binding.

*DNA accessibility is sufficient to explain Su(H) binding*

One of the key factors of the model is DNA accessibility. Su(H) has been shown to preferentially bind to open chromatin and increase chromatin accessibility post binding [Lake et al., 2014, Skalska et al., 2015]. In this chapter, I was able to recover Su(H) preferences towards open chromatin. Figure-30 shows that the inclusion of DNA accessibility increase the predictive power of the model. In Figure-31, I illustrate how the inclusion of DNA accessibility reduces the number of available binding sites. In turn, this translates to a better agreement between the predicted profiles and ChIP data. Unfortunately, the model was not able to recover the opening of chromatin resulting from Su(H) binding. The opening of the chromatin is further increased after Notch activation[Skalska et al., 2015]. I attempted to uncover this mechanisms by varying chromatin accessibility during model training. However, no clear differences can be observed between Notch Induced and Non-Induced data sets. The results demonstrate that in both cases Su(H) binding is better predicted when DNA accessibility is considered but not that Su(H) induced changes in chromatin after binding. This is likely due to the fact that changes in chromatin accessibility after Su(H) binding are likely minute. Small changes in DNA accessibility at the loci of interest would be offset by an increase in ChIP noise. Together, these opposing factors negate the ability of the model to pick up on chromatin opening by Su(H) binding. One approach to overcome this issue would be to extend DHS peaks instead of using QDA. The slightly extended peaks would be more apt to mimic the extension of open chromatin.

The use of ChIPanalyser together with a genetic algorithm demonstrated that the model can accurately describe Su(H)'s chromatin state affinities. Su(H) preferentially binds to enhancers, active TSS and active introns [Skalska et al., 2015]. Chromatin state affinity scores show that Su(H) preferentially binds to open chromatin. Despite being able to recover chromatin state affinities, it would seem that DNA accessibility is sufficient to predict the binding of Su(H). The use of chromatin states increases the ability to understand the intricacies of Su(H) binding but in terms of predictive capabilities are unnecessary. Curiously, chromatin state affinity scores also suggests an affinity towards heterochromatin. However, the variance of the affinity score was also increased thus it is unlikely that this is biologically relevant. Furthermore, [Skalska et al., 2015] observed Su(H) peaks in Polycomb and heterochromatic states but attributed this to ChIP-seq noise.

Finally, despite DNA accessibility being a strong driver in Su(H) binding, certain ChIP peaks were completely missed by ChIPanalyser. This could be due to two main factors. First, the binding motif used to predict binding has been put into question as to its ability to accurately describe Su(H)'s preferred binding motif. Other binding motifs are available and result in different binding predictions. After testing both binding motifs, I selected the JASPAR motif as overall it seemed to recover more peaks. Second, recent work has suggested that Su(H) binding operates with high efficiency thanks to the aid of co-factors such as Lz/Runx [Skalska et al., 2015]. The current model does not include cooperative interactions thus it is likely that ChIPanalyser is unable to capture all Su(H) binding events.

### *Notch Induction does not significantly induce an increased in Su(H) abundance*

The activation of the Notch pathway is accompanied by the translocation of the NICD. The NICD translocation is translated into an increase of Su(H) binding but not an increase in Su(H) abundance [Krejčí and Bray, 2007, Bray, 2016, Gomez-Lamarca et al., 2018]. In an effort to recover this behaviour, the model was trained in both conditions using

DNA accessibility only. The estimated number of bound molecules was slightly increased after Notch Induction (N=4000) compared to Non-Induced Notch (N=3000). The nuclear concentration of Su(H) might not change but there will be an increased number of Su(H) molecules bound to DNA. The increase in peak enrichment at certain loci would be the consequence of increased Su(H) binding but not necessarily increase in Su(H) nuclear concentration. Recent studies on Su(H) mutants have revealed that Su(H) abundance could be linked to NICD translocation but also to Su(H) alter-ego Hairless (H) [Praxenthaler et al., 2017]. Nevertheless, the optimal set of parameters described in Figure-33 suggest that the number of bound molecules remains similar. The true role of Notch Induction on Su(H) abundance is still up to debate.

*RNAi knock-downs reduce Su(H) abundance*

The binding of Su(H) is affected by changes in nuclear abundance [Wang et al., 2015]. Experimentally, this was demonstrated by the use of RNAi Su(H) partial knock-downs. Su(H) abundance was reduced by 70% after knock-downs. Using the ability of ChIPanalyser to estimate number of bound molecules, I investigated the role of protein abundance on Su(H) binding. To do so, I trained the model on wild type Su(H) and rescaled the estimated number of bound molecules to reflect the effect of the RNAi knock-down. The optimal parameters were then applied to RNAi ChIP data. The model accurately recovered the reduction in Su(H) binding between wild type and treatment. Interestingly, at the selected loci the change in ChIP signal was not as striking as one would expect. Nevertheless, ChIPanalyser demonstrates its ability to accurately recover changes in TF abundance. The abundance of Su(H) is an integral part of its biological function as it is believe that changes in Su(H) abundance will adjust its ability to locally modify chromatin. Furthermore, Su(H) depletion compromises its ability to regulate target genes [Yuan et al., 2016].

## CONCLUSION

The binding of Su(H) has been well studied and has provided wealth of data describing the binding mechanisms of Su(H). By using this well studied system, I demonstrated ChIPanalyser's ability to recover numerous known binding mechanisms. Despite displaying affinities towards chromatin states, DNA accessibility is sufficient to explain the binding of Su(H) with respect to the nature of chromatin. Finally, change in Su(H) abundance affects the binding of Su(H) to its target genes. Overall, ChIPanalyser proved to be a powerful tool to not only predict TF binding but also understand the mechanisms behind the binding of TFs.

# 5

# DISCUSSION

TFs play a central role in gene regulation. It comes to no surprise that TFs have been studied for decades and there is still a vibrant field dedicated to understanding how TFs function in the grand scale of a genome. For the most part, TFs recognise a sequence of DNA and bind to a preferred binding motif. TFs induce or repress the recruitment of the transcriptional machinery. This form of regulation will lead to the activation ( or repression ) of gene transcription. The main issue is that binding sites are fairly short (8 to 20bp) and occur at a high rate throughout the genome. Despite numerous target motifs, genome occupancy profiling methods ( the gold standard of *in vivo* TF binding determination) seem to show that only a subset of these sites are bound by TFs. Furthermore, the argument could be made that *in vivo* binding to the genome does not constitute functional binding. Functional binding would require some sort of regulation of transcription. It has been suggests that only a subset of genes bound by TFs were differentially expressed after TF knock-down. Most TF-DNA interaction would not results in change in gene expression [Cusanovich et al., 2014].

Over millions of years of evolution, cells have found many tricks to ensure that TFs would bind when and where they are required. Binding motifs can be made unavailable for TF binding by decreasing the accessibility of DNA around those sites [Klemm et al., 2019, Lamparter et al., 2017]. TF would only bind to sites in open chro-

matin. This assumption comes with a few limitations. First, some TFs can bind into regions of closed chromatin. Pioneer TFs have been shown to bind closed chromatin and induce the opening of said chromatin [Soufi et al., 2015, Mayran et al., 2018, Donaghey et al., 2018, Zaret and Carroll, 2011]. Second, other DNA binding proteins have shown varying affinities to DNA accessibility without being a pioneer transcription factor [Van Bortle et al., 2012, Porcelli et al., 2019].
Instead of considering DNA as either open or closed, chromatin can be described with respect to its local state. The chromatin state of a loci is defined by a specific combination of histone modifications and/or histone variants. Chromatin states are correlated with a "genomic" function such as enhancers or active TSS [Baker, 2011].

In recent years, the scientific community has seen a burst in computational based methods to understand and predict biological phenomena. This is possible thanks to the formidable increase in available data but also the computational means to analyse large scale data sets. This lead to the rise of machine learning and Artificial intelligence in the field of genomics. Machine learning methods have shown great promise thanks to their powerful ability to recognise patterns in vast seas of data. However, what they have in predictive capabilities, they lack in interpretability. Machines can easily annihilate any human in a game of chess or Go but it sadly cannot explain why a strategy is better than an other. The same principle can be applied to their use in genomics. Deep learning algorithms are most certainly some of the strongest predictors of TF binding [Salekin et al., 2017, Salekin et al., 2018, Alipanahi et al., 2015, Quang and Xie, 2019, Keilwagen et al., 2019, Li et al., 2019]. However, they lack the ability to explain why they predict binding at at specific location. Recently, there has been a strong push towards creating explainable machine learning. Methods that we humans could understand. One such method takes inspiration directly from nature and biology: Genetic algorithms.

In biology, understanding mechanisms often trumps predictive power. It is in this context that biology turned towards physics. The tools used in physics can be applied to biological questions. Thanks to explicit modelling of parameters, we can gain insight into the mechanisms driving biological phenomena. TF binding was no exception. Statistical thermodynamics has brought an interesting perspective on the mechanisms of TF binding. Zabet and Adryan described the binding of TF as the results of four main factors: binding energy, a specificity scaling factor, number of bound molecules and, DNA accessibility [Zabet and Adryan, 2015].
It is in this context that the work in this thesis should be described. Hopefully, the work presented in this thesis would have shed light on the mechanisms driving TF binding.

## THE ROLE OF DNA ACCESSIBILITY AND CHROMATIN STATES

A large part of this thesis was dedicated to understanding the role of DNA accessibility in TF binding. To do so, I analysed among others three architectural proteins: namely CTCF, BEAF-32 and su(Hw). All three proteins are known as *Drosophila* architectural proteins [Van Bortle et al., 2014, Chathoth and Zabet, 2019]. This suggests that not only do they have a role in transcription but they also play important roles in genome architecture maintenance and insulation. It is important to note that none of these DNA binding proteins are considered pioneer transcription factors. At least, they have never been described as such. In theory, we would expect these proteins to be better explained by the model when DNA accessibility is considered. In practice , the answer was not clear cut.

BEAF-32 was better predicted when DNA accessibility was considered. By masking binding motifs in closed chromatin, the cell is able to drive the binding of this TF to its target motifs. BEAF-32 binding occurs at a genome wide scale: any available target site will be bound by BEAF-32. This preference towards open chromatin was further confirmed when chromatin states were included. Chromatin state affinity scores suggest that BEAF-32 preferentially binds to active TSS and to a lesser extent to enhancers. Both of these chromatin states are unsurprisingly considered open and active chromatin.

The role of accessibility in the binding of CTCF was unclear. The predictions did not seem to improve when DNA accessibility was included. This would suggest that at least certain peaks are located in less accessible DNA. Interestingly, CTCF has been described as a chromatin insulator. CTCF plays the role of a road block stopping the spreading of heterochromatin [Guelen et al., 2008, Van Bortle et al., 2014]. This could explain why some peaks are partially covered by inaccessible DNA. Using chromatin states demonstrated that CTCF had indeed a more nuanced role with respect to open chromatin. CTCF showed increased affinity towards enhancer regions, active TSS and competent states. Interestingly, competent states are often considered as enhancers that have not yet been fully activated [Kamakaka and Thomas, 1990, Skalska et al., 2015]. CTCF has shown to be involved in insulating enhancers from their target gene [Kim et al., 2015, Nichols and Corces, 2015]. The chromatin state affinities described in this thesis seem to recover some of CTCF's biological functions. CTCF displayed an intermediate affinity towards heterochromatin that would confirm its role in chromatin insulation. The overlap between CTCF ChIP data and chromatin states ( or DNA accessibility for that matter) would not be clear cut. Many CTCF peaks would be located in less accessible DNA. However, these results should be taken with caution as affinity scores were also accompanied by increased affinity variance. Over all replicates and all top performing "individuals", there was no clear consensus on which affinity score produced the best fitness. Furthermore, CTCF's performance tapered off when more regions were included in the validation set. CTCF preferentially binds to genome hotspots and lower affinity sites ( cell specific sites) would be controlled by changes in chromatin states.

The last architectural protein I analysed was su(Hw). This protein performed poorly when DNA accessibility was included into the model. *Drosophila* displayed around 10% of accessible DNA. The question remains as to what drives the binding of su(Hw). Thankfully, the predictive power of the model was greatly increased when chromatin states were included. The picture drawn by su(hw) still remained fairly unclear. While it did show increased affinity for heterochromatic states it also seem to exhibit intermediate affinity towards transcriptionally active states. su(Hw) has been suggested to be involved in many biological functions such as LAD stabilisation, TAD borders and repression of gene expression [Kurshakova et al., 2007, Kuhn-Parnell et al., 2008, van Bemmel et al., 2010, Adryan et al., 2007, Melnikova et al., 2019]. These biological functions tend to display increased heterochromatin around TF binding sites. The nuanced behaviour of su(Hw) can be partially explained by chromatin state affinity. However, su(Hw) binding specificity could also occur with the help of co-factors [Baxley et al., 2017, Glenn and Geyer, 2019].

Varying the levels of DNA accessibility by using quantized thresholds proved to yield some interesting results. First, it showed that Hox TFs have varying affinities towards DNA accessibility. While Ubx preferentially bind in open chromatin, both Dfd and Abd-b show a more permissive behaviour with respect to DNA accessibility. This method was also applied to Su(H) but in a different context. The aim was to describe the opening of chromatin after Notch Induction. However, the model did not recover this mechanism. Changes in DNA accessibility were to minute too be recovered and moreover offset by an increase in ChIP seq noise.

With the addition of chromatin state affinity scores, Ubx and Su(H) displayed similar results as to the ones obtained with DNA accessibility. Unsurprisingly, both TFs showed high affinity scores towards highly open chromatin states. In this case, DNA accessibility is sufficient to explain their binding mechanisms. Abd-b and Dfd on the other hand were better explained with the addition of chromatin states. Being able to bind in lesser accessible DNA suggests that these TFs recognise certain histone modification. Furthermore, chromatin affinities imply that Hox TFs may not bind in heterochromatin but rather in regions of an intermediate accessibility (e.g. competent state).

Taken together, these results show that the role of DNA accessibility is not clear cut. Assuming that TFs can only bind to accessible DNA is only true for a subset of TFs. For these TFs, increasing the level of information with the addition of chromatin states did not significantly improve the model's performance. DNA accessibility on its own is sufficient to predict their binding. The other DNA binding proteins showed varying affinities towards both DNA accessibility and chromatin states. Unfortunately, there is no clear indication that a certain type of TF would bind to open or closed chromatin. Their direct involvement in gene expression does not warrant the assumption that they will bind open chromatin. A non-pioneer TF could also bind to less permissive chromatin with the help of chromatin re-modellers. This was the case for Abd-b and Dfd for example. Although it is tempting to make this assumption, understanding the binding mechanisms of TFs requires more than DNA accessibility. The assumption that only pioneer transcription factors bind to heterochromatin could be an over simplification of the biology at hand.

## THE ROLE OF PROTEIN ABUNDANCE AND BINDING SITE SPECIFICITY

On top of DNA accessibility and DNA affinity, the model also includes three other factors: binding energy (as a PWM), a scaling factor modulating PWM specificity and the number of bound molecules.
Binding motifs are often described as Position weight matrices. PWMs are convenient ways to describe weighted binding sequences. However, there are some limitations to using PWMs. First, from a technical perspective, all base pairs in the binding motif are considered independent from each other (see Models). Yet , in certain cases, the binding of TF requires binding motif to act as unit, as a multimeric entity [Zhou et al., 2015]. Second, PWMs fail to demonstrate DNA shape feature that in certain cases have been shown to influence TF binding [Zhou et al., 2015]. Finally, it seems that binding motifs can also occur in clusters and that lower affinity binding sites can be preferred over higher affinity sites [Farley et al., 2016].

Interestingly, clustering of lower affinity binding sites will be considered within the ChIPanalyser workflow. ChIPanalyser selects sites above a certain threshold and these sites will be considered as potential targets. Modulating this threshold value will include more or less potential binding motifs. Occupancy scores are computed for each of these sites and ChIP-like profiles are generated by applying a gamma distribution to the occupancy scores. The gamma distribution will both consider and amplify clusters of lower affinity binding sites. Despite being technically included, ChIPanalyser does not explicitly model low affinity binding motif clustering. This limitation was demonstrated with Hox TFs and chromatin states. The predicted profiles generated wide yet flat peaks. Multiple peaks would be present under the curve but as they represent low affinity sites the model does not capitalise on their importance in TF binding. Furthermore, the inclusion of the scaling factor will increase the strength of high affinity sites over lower affinity sites despite being above the PWM threshold. It would be interesting to test the model without using a PWM threshold. This would yield a much higher number of binding sites and potentially lead to a clearer picture of TF binding. On the other hand, this could also reduce the ability of the model to predict ChIP peaks at it is likely the model would optimise against noise rather than ChIP peaks. The resulting profiles might end up being flat as this would affect the scoring method the least or conversely overestimating TF binding leading to a reduced ability to recover ChIP peaks and their enrichment. ChIPanalyser uses ChIP enrichment scores for goodness of fit assessment and heavily penalises predicted profiles that overestimate or underestimate ChIP enrichment. While the best goodness of fit score is not the goal of the package, recovering ChIP peak location and enrichment is one of them as this serves as the basis for understanding the mechanisms of TF binding.

The number of bound molecules is assumed to be a proxy for TF concentration. However, I believe it is important to recognise that bound molecules and concentration are different concepts. First, as demonstrated by Notch activation described in chapter 4, the number of bound molecules slightly increases after the activation of the Notch pathway. It is believed that Notch induction does not change the abundance of Su(H) but rather increases the binding of Su(H) to DNA. In this context, Su(H) is present at a higher rate bound to DNA. The model recovers this mechanisms but in this case number of bound molecules can not be used a proxy for concentration.

Second, concentration is a function of volume. In the context of cells and DNA, local concentration can be increased by decreasing the volume in which proteins find themselves. Numerous studies on phase separation have suggested that chromatin confirmation plays an active role in TF binding by modulating local concentration [Hnisz et al., 2017, Boeynaems et al., 2018]. Furthermore, high local concentration would increase the probability of binding to low affinity sites [Lickwar et al., 2012, Erbas and Marko, 2019].

Third, concentration of a single TF does not consider the interaction between a TF and its cofactors. Hox TFs are a prime example of TF binding driven by concentration. However, Hox TFs also show cooperative binding and the correct binding to Hox target genes seems to be driven by complex relationships between each cofactor [Petkova et al., 2019]. Despite being in high concentration, certain TFs would require the binding of co-factors in order to induce favourable binding conditions [Iwafuchi-Doi and Zaret, 2014]. Furthermore, many TF posse protein-protein interaction domains [Sammak and Zinzalla, 2015, Shokri et al., 2019]. The presence of these domains and the known role of co-factors suggests that protein concentration describes an incomplete picture of TF binding.

Forth, I described that architectural proteins saturate their binding sites and only strong knock-downs can reduce the enrichment of strong peaks. This suggests that in many cases, DNA binding proteins are at saturating levels in order to ensure that protein abundance fluctuations would not affect their biological function.

Finally, the number of bound molecules are inferred by maximising the goodness of fit between predicted ChIP like profiles and ChIP profiles. Simply, the higher the peak, the higher the expected number of bound molecules. However, ChIP-seq is not a fully quantitative method. While peaks with a high enrichment score are considered to be sites highly bound by TFs, peak enrichment scores are also dependant on number of cells used, sequencing depth , anti-body specificity and the ChIP protocol used [Chen et al., 2016]. It should be noted that ChIPanalyser uses normalised ChIP scores to ensure that data sets are as comparable as possible. While there has been an ongoing effort to make ChIP data quantitative [Bonhoure et al., 2014, Egan et al., 2016, Orlando et al., 2014], these protocols and methods had not been applied to the data sets used in this thesis. The estimated number of bound molecules shown in this thesis tend to fall into biologically acceptable ranges but should be taken with caution. The same caution should be applied to experimental ChIP peaks as they might not fully represent stronger TF binding events.

## THE LIMITATIONS OF GENOME OCCUPANCY PROFILING

ChIP-seq has become the gold standard of defining *in vivo* protein binding to DNA. Despite over a decade of maturation, ChIP-seq and the subsequent analysis comes with caveats.

First, the choice of anti-body can affect the specificity of DNA binding pull down. This is further increased by unspecific DNA pull down [Teytelman et al., 2013]. The increase of unwanted DNA will lead to an increase of false positive peaks.

Second, the ChIP-seq protocol is affected by DNA accessibility [Auerbach et al., 2009]. The fragmentation of the genome occurs at a higher rate in open chromatin. This bias will likely lead to an over representation of DNA fragments in accessible DNA. Understanding the binding of DNA binding proteins such as su(Hw) would be hindered by this bias.

Third, the production of false positive peaks is also related to sequencing depth [Sims et al., 2014]. Increasing sequencing depth will limit the number of false positive peaks. It has been reported that many publicly available ChIP data sets display low sequencing depth. The modEncode and Encode consortium have developed guidelines [Landt et al., 2012a] to increase the overall quality of publicly available data however there is no guarantee that these guidelines will be correctly followed. It should be noted that spike in controls and negative controls tend to limit false positive peaks. Nevertheless, all peaks observed in ChIP data are not always reflective of true *in vivo* binding [Wreczycka et al., 2017].

Finally, the end product of ChIP analysis is often the production of ChIP peak coordinates and/or signal enrichment pile-up. The ability of peak calling algorithms to distinguish true peaks from background noise are limited by both the quality of the data used but also the statistical method used. Peak calling limitations was reviewed by [Nakato and Shirahige, 2017] and benchmarked by [Thomas et al., 2016].

The quality of ChIP data has a strong influence on the prediction quality of ChIPanalyser. This is especially true when considering the scoring method used by ChIPanalyser. Many competing tools provide a prediction of TF binding based on the overlap between the predicted peak location and actual ChIP peak in 200bp windows. ChIPanalyser on the other hand uses enrichment scores to assess goodness of fit. This approach attempts to recover peak height as well as peak location. As described above, ChIP enrichment scores are normalised prior to analysis in order to maximise the "comparability" of each data set. Moreover, the prediction resolution can be increased to base pair level rather than 200bp windows. The trade-off of using this method resides in the fact that ChIP noise or missing enrichment will be heavily penalised. In many cases, I observed data sets containing wide and noisy peaks and the model was not able to recover experimental fluctuations. The goodness of fit score would reflect this despite accurately describing peak location. Noise filtering methods were able to slightly reduce noise however the filtering occurred on low level peaks. If strong peaks were still characterised by a messy signal, ChIPanalyser would still be strongly penalised.
One way to overcome these issues is by using alternative methods such as CUT&run, CUT &tag, or ChIP-exo [Skene and Henikoff, 2017, Kaya-Okur et al., 2019, Serandour et al., 2013]. These methods offer sharper peaks and reduced background noise. As described above,

both back ground noise and peak "sharpness" strongly affect the goodness of fit of the model. It should be noted that the goal of decreasing background noise and unspecific binding should not be to improve goodness of fit of the model but rather displaying a fairer assessment of the model's performance. Reducing background noise and increasing peak sharpness would mitigate over or under representation of model performance. Furthermore, methods such as CUT &run or CUT &tag are both *in situ* method using few cells to determine TF binding. This lends a more quantitative nature to these methods. As described above, estimating the number of bound molecules based on ChIP data should be taken with caution as ChIP-seq is not always a quantitative. While not perfect, this could potentially lead to clearer estimations of bound molecules.

A question that one might ask is whether low level peaks are actual noise or rather the consequence of low affinity binding. It has been suggested that TFs can bind to DNA and perform 1D walks before finding their target sites [Zabet and Adryan, 2012, Hammar et al., 2012]. This could also transpire at strong peaks. The noise surrounding a peak could be the illustration of such a mechanism. The vast majority of ChIP data is based on a large population of cells. Cell population ChIP would recover the spectrum of binding locations thus giving the impression of noisy peaks. Furthermore, lower affinity peaks could actually play a biological role. Cohesin has been shown to act as a transcriptional regulator independently of CTCF at peaks of lower enrichment in *Homo sapiens* [Schmidt et al., 2010]. It could be conceivable that a similar mechanism exists in *Drosophila*.

Finally, ChIP-seq binding illustrates *in vivo* binding of TFs to DNA but does not describe functional TF binding. Generally, TFs are involved in transcriptional regulation. The binding of a TF will activate or repress the transcription of target genes. However, ChIP-seq on its own does not give any indication wherever TF binding peaks will lead to transcriptional regulation. Furthermore, the role of a DNA binding proteins might not directly lead to transcriptional regulation. Architectural proteins for example could stabilise genome structures that will then lead to gene regulation by other regulatory TFs. It is important to recognise the limitations of ChIP data in order to fully understand the mechanisms of gene regulation.

## dissecting the mechanisms of tf binding

Work on machine learning and artificial intelligence has been an ongoing field since the 1950s. Interestingly, the game of chess has been at the heart of the development of artificial intelligence. The 1950s saw the first human to loose a game of chess against these smart algorithms. In 1996, Deep Blue was famous for being the first machine capable of beating a chess grand master (non other than Garry Kasparov). Unfortunately, the enthusiasms for artificial intelligence dropped in the 1970s and 1980s as it seemed they did not deliver on the promise of understanding human consciousness and intelligence. However, machine learning was not forgotten but rather picked up by other fields of study. The field of finance was especially keen on developing algorithms capable of predicting changes in the stock market. This push came at a cost. Most of these new methods were developed with predictive power in mind and not explainability. Decades later, when machine learning came back to world of pure science, the crisis of explainable AI started. Many algorithms have shown to be powerful predictive tools. Recently, image recognition algorithm has demonstrated their ability to recognise cancerous growths at a higher rate than human specialists [Wu et al., 2019, McKinney et al., 2020]. The main issue is that despite following the rules of mathematics, machine learning algorithms are inherently black boxes. What happens inside, what happens to trigger differential weighting of neural network layers remains largely unknown.

Overcoming these limitations has been at the center of AI research. In fact, genetic algorithms are one of the ways to overcome the black box problem. All factors are explicitly given before hand and only these factors are optimised in order to improve predictive power. Recently, there was an approach to incorporate thermodynamic parameters to neural networks to explain gene expression [Tareen and Kinney, 2019]. The goal of this thesis was to demonstrate the ability of statistical thermodynamics to describe the mechanisms of TF binding. ChIPanalyser has shown to be a valuable tool to predict and understand TF binding. However, it some case, ChIPanalyser predicted TF binding quite poorly. The binding of su(Hw) using only DNA accessibility proved to be limited.

The strength of statistical thermodynamics and by extension biophysical models resides in interpreting the mechanisms driving TF binding. Biophysics explicitly models a relationship between factors thought to contribute to TF binding. Even when predictions fall short, biophysical models still give us the ability to speculate on other potential contributing factors. ChIPanalyser exemplifies how biophysics can be used to understand biological mechanisms even when predictions fall short. While the package showed a limited ability to predict the binding of Su(Hw), it was still possible to understand or at least have an educated guess onto which factors contribute to its binding. While predicting biological phenomena is a crucial aspect of modern science, the goal of this thesis and ChIPanalyser was to understand how they happen.

Part III

# CONCLUDING REMARKS

# 1

# CONCLUSION

The work in the thesis explores the mechanisms of TF binding using a statistical thermodynamic framework. The model proposed by Zabet and Adryan suggests that TF binding to DNA is the results of four main factors: binding energy in the form of PWMs, a PWM scaling factor, the number of bound molecules, and finally DNA accessibility. I aimed to recover known binding mechanisms and potentially uncover unknown factors driving the binding of certain TFs.

To do so, I developed ChIPanalyser, a user-friendly R package available on Bioconductor. I demonstrated the package's ability to accurately predict the binding of TFs in both *Drosophila melanogaster* and *Homo sapiens*. Furthermore, ChIPanalyser can infer the number of bound molecules and scaling factor by maximising or minimising a goodness of fit metric. The estimated number of bound molecules remain within biologically acceptable boundaries. The development of ChIPanlyser lead to interesting insight into the choice of goodness of fit metrics. I showed that goodness of metrics could be classified into two classes (similarity and dissimilarity metrics). While similarity metrics correctly predict peak location, they often fall short in terms of peak enrichment. The opposite was observed for dissimilarity metrics. They correctly recovered local enrichment but were often subject to a higher number of false positive peaks. Choosing the right metric will depend on both the question and the data at hand.

Then, I demonstrated that DNA accessibility plays a more nuanced role in TF binding than previously thought. I investigated the binding of three architectural proteins (CTCF , BEAF-32 and su(Hw)) in three *Drosophila* cell lines (Kc167, BG3, and S2). While the inclusion of DNA accessibility greatly increased the binding prediction of BEAF-32, CTCF and su(Hw) were less well predicted. Both CTCF and su(Hw) displayed peaks in inaccessible DNA suggesting that open chromatin was not sufficient to explain their binding. I investigated the role of TF abundance on the binding of architectural proteins only to demonstrate that TF abundance plays a minor role in their binding. Differential binding preferences towards DNA accessibility was further demonstrated by investigating the binding of three Hox TFs. Ubx was preferentially bound to open chromatin while Dfd and Abd-b were more permissive with DNA accessibility in terms of binding. These results suggests that DNA accessibility can not only be considered as either open or closed. The role of chromatin on TF binding stems from more complex relationships.

To further investigate the role of chromatin, I developed a genetic algorithm to demonstrate that TFs display different binding affinities towards chromatin states. Chromatin states affinity scores were able to recover known binding affinities and give a more nuanced picture of TF binding mechanisms. The binding of CTCF and su(Hw) revealed to be driven by chromatin states more than DNA accessibility on its own. Binding of TFs to DNA would be better explained by the specific nature of chromatin rather than being in an open or closed state.

Finally, I explored the model's ability to recover the binding mechanisms of Su(H). The wealth of data available for Su(H) made this TF an ideal case study. While Su(H) displayed clear affinities towards chromatin states, DNA accessibility seemed to be sufficient to predict Su(H) binding. The model also accurately recovered changes in TF abundance after partial RNAi Su(H) knock downs.

Despite being able to recover many known binding mechanisms, the model showed some limitations. The current model does not include cooperative binding. Many TFs examined in this thesis have been shown to bind to DNA with the help or in collaboration with co-factors. The model also fails to capitalise on low affinity binding sites where site clustering would induce binding. Finally, the role of 3D chromatin structure should not be neglected.

Overall, ChIPanalyser provides useful insight into the binding mechanisms of TFs. Chromatin states would drive the binding of TFs to DNA rather than simply open or closed chromatin. Furthermore, protein abundance does not always explain why TFs bind the way they do. In some instance, TFs would bind at saturating levels as a mechanisms against unwanted protein level fluctuations. Assumptions of TF binding have often been over-simplified with respect to the nature of chromatin.

# 2

# LOOKING FORWARD

The mechanisms of TF binding and gene regulation are complex and poorly understood. ChIPanalyser has provided insights into the binding of TFs to DNA. That being said, I recognise that many improvements could be made in order to further our understanding of TF binding.

First, the bulk of the analysis presented in this thesis focused on *Drosophila*. It would be interesting to investigate the performance of the model of other organisms. To do so, it would be wise to test a large number of TFs. As shown in this thesis, the assumption that non-pioneer TFs bind to open chromatin is somewhat an over-simplification of the problem at hand. To capture a wide range of binding mechanisms, I would suggest testing the model in multiple organisms with as many data sets as possible.

Second, as described in this thesis, the model could be improved by the inclusion of cooperative binding. Preliminary work has already been done with respect to cooperative binding inclusion. However, the underlying master equation would need to be drastically modified. This would also require a complete re-factoring of ChIPanalyser as a package.

Third, including 3D chromatin structure will provide a more in depth understanding of TF binding. In its current form, ChIPanalyser provides the option to select the optimal parameters ( $N$ and $\lambda$) based on frequency of occurrence of parameter combinations for every region selected for analysis. Simply put, every region selected ( in the case of this thesis - over 3000 regions of 20kbp) would return an optimal combination of parameters. By ranking the combinations of parameters with respect to frequency of occurrence, it would be possible to select optimal parameters on a region to region basis. By including chromosome conformation capture contact maps, it would be interesting to investigate if regions that display a high contact frequency with each other would also share similar predicted number of bound molecules and PWM scaling factor at TF binding events.

An experimental validation of the results obtained in this thesis would be a welcome addition. More specifically, the work in this thesis would benefit by experimentally determining protein abundance. One of the main parameters of the model presented in this thesis is the number of bound molecules. While the inferred number of bound molecules often seemed to be in biologically acceptable boundaries, experimentally quantifying the number of molecules within a cell would greatly increase the credence of the model presented within this thesis. The model should not only be able to predict the binding of TF to DNA but also describe the mechanisms driving TF binding. Furthermore, experimentally quantifying protein abundance is possible thanks to methods such as Fluorescence Correlation Spectroscopy, Flow Cytometry, or in-gel fluorescence. While explicit measurement of binding energies to random binding sequences *in vitro* could give an insight into $\lambda$, *in vivo* binding specificity remain extremely challenging to determine.

Finally, while I have shown that differential binding between cell lines could be explained by change in chromatin accessibility, it would be interesting to alter chromatin states to further demonstrate the role of chromatin accessibility and chromatin states.

Global alteration of chromatin accessibility using degron fusions of core chromatin proteins could provided modified chromatin accessibility/ chromatin state maps to be used as input to the model. It would be interesting to investigate how strong changes in chromatin states would affect the models ability to predict and explain TF binding. The same approach could be applied to specific regions using (CRISPRi/CRISPRa). For example, in the case of Su(H), changes to DNA accessibility or chromatin states at known binding loci ( e.g. E(spl) locus) could further our understanding of the binding mechanisms driving its binding. While it was shown that Su(H) seems to open chromatin around its strong binding sites, what would happen if the chromatin had already been opened?

# A

# APPENDIX

**Appendix-A: Data sources for all TFs.** Data sets have been separated into modEncode ChIP data , GEO ChIP data, Accessibility Data, and ENCODE data for *Homo sapiens*

| Cell Line | TF | modEncode Accession # |
|---|---|---|
| S2 | CTCF | 2638 |
| S2 | CTCF | 2639 |
| S2 | CTCF | 283 |
| S2 | CTCF | 3281 |
| S2 | CTCF | 3749 |
| S2 | CTCF | 913 |
| S2 | BEAF-32 | 922 |
| S2 | BEAF-32 | 3745 |
| S2 | BEAF-32 | 274 |
| S2 | Su(Hw) | 3719 |
| S2 | Su(Hw) | 330 |
| S2 | Su(Hw) | 331 |
| BG3 | CTCF | 282 |
| BG3 | CTCF | 3280 |
| BG3 | BEAF-32 | 3663 |
| BG3 | BEAF-32 | 3664 |
| BG3 | BEAF-32 | 3665 |
| BG3 | BEAF-32 | 921 |
| BG3 | CTCF | 3671 |
| BG3 | CTCF | 3672 |
| BG3 | CTCF | 3673 |
| BG3 | CTCF | 3674 |
| BG3 | Su(Hw) | 3714 |
| BG3 | Su(Hw) | 3715 |
| BG3 | Su(Hw) | 3716 |
| BG3 | Su(Hw) | 3717 |
| BG3 | Su(Hw) | 3718 |
| BG3 | Su(Hw) | 951 |
| Kc167 | CTCF | 908 |
| Kc167 | Su(Hw) | 3801 |

| Cell Line | TF | GEO Accession # |
|---|---|---|
| Kc167 | CTCF | GSM762842 |
| Kc167 | Su(Hw) | GSM762839 |
| Kc167 | BEAF-32 | GSM762845 |
| Kc167 | BEAF-32 | GSM1535963 |

| Cell Line | Method | GEO Accession # |
|---|---|---|
| Kc167 | DNase | Kharchenko et al. 2011 |
| S2 | DNase | Kharchenko et al. 2011 |
| BG3 | DNase | Kharchenko et al. 2011 |
| Kc167 | ATAC-seq | GSE122575 |

| Cell Line | Method | File Type | ENCODE # |
|---|---|---|---|
| Astrocyte | DNase | BAM | ENCFF384CCQ |
| Astrocyte | DNase | BAM | ENCFF885IAD |
| Astrocyte | DNase | BigWig | ENCFF901UBX |
| Astrocyte | DNase | bed | ENCFF021SAS |
| Astrocyte | DNase | bed | ENCFF352LYZ |
| Astrocyte | ChIP-seq | bigWig | ENCFF424JNY |
| Astrocyte | ChIP-seq | bed | ENCFF600CYD |
| Astrocyte | ChIP-seq | bed | ENCFF183YLB |

# BIBLIOGRAPHY

[Abe et al., 2005] Abe, M., Kobayashi, Y., Yamamoto, S., Daimon, Y., Yamaguchi, A., Ikeda, Y., Ichinoki, H., Notaguchi, M., Goto, K., and Araki, T. (2005). FD, a bZIP Protein Mediating Signals from the Floral Pathway Integrator FT at the Shoot Apex. *Science (80-. ).*, 309(5737):1052–1056.

[Abe et al., 2015] Abe, N., Dror, I., Yang, L., Slattery, M., Zhou, T., Bussemaker, H., Rohs, R., and Mann, R. (2015). Deconvolving the Recognition of DNA Shape from Sequence. *Cell*, 161(2):307–318.

[Adryan et al., 2007] Adryan, B., Woerfel, G., Birch-Machin, I., Gao, S., Quick, M., Meadows, L., Russell, S., and White, R. (2007). Genomic mapping of Suppressor of Hairy-wing binding sites in Drosophila. *Genome Biol.*, 8(8):R167.

[Aguilar-Arnal and Sassone-Corsi, 2015] Aguilar-Arnal, L. and Sassone-Corsi, P. (2015). Chromatin landscape and circadian dynamics: Spatial and temporal organization of clock transcription. *Proc. Natl. Acad. Sci. U. S. A.*, 112(22):6863–70.

[Akiyama-Oda et al., 1998] Akiyama-Oda, Y., Hosoya, T., and Hotta, Y. (1998). Alteration of cell fate by ectopic expression of Drosophila glial cells missing in non-neural cells. *Dev. Genes Evol.*, 208(10):578–585.

[Alexander et al., 2009] Alexander, T., Nolte, C., and Krumlauf, R. (2009). <i>Hox</i> Genes and Segmentation of the Hindbrain and Axial Skeleton. *Annu. Rev. Cell Dev. Biol.*, 25(1):431–456.

[Alipanahi et al., 2015] 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. *Nat. Biotechnol.*, 33(8).

[Andersson and Sandelin, 2019] Andersson, R. and Sandelin, A. (2019). Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.*, pages 1–17.

[Angermueller et al., 2016] Angermueller, C., Pärnamaa, T., Parts, L., Stegle, O., Albert, F., Treusch, S., Shockley, A., Bloom, J., Kruglyak, L., Alipanahi, B., Delong, A., Weirauch, M., Frey, B., Angermueller, C., Lee, H., Reik, W., Stegle, O., Asgari, E., Mofrad, M., Bach, S., Binder, A., Montavon, G., Klauschen, F., Muller, K., Samek, W., Battle, A., Khan, Z., Wang, S., Mitrano, A., Ford, M., Pritchard, J., Gilad, Y., Bell, J., Pai, A., Pickrell, J., Gaffney, D., PiqueâĂŘRegi, R., Degner, J., Gilad, Y., Pritchard, J., Bengio, Y., Courville, A., Vincent, P., Cheng, C., Yan, K., Yip, K., Rozowsky, J., Alexander, R., Shou, C., Gerstein, M., Deng, L., Eduati, F., Mangravite, L., Wang, T., Tang, H., Bare, J., Huang, R., Norman, T., Kellen, M., Menden, M., Yang, J., Zhan, X., Zhong, R., Xiao, G., Xia, M., Abdo, N., Kosyk, O., Collaboration, N., Friend, S., Dearry, A., Simeonov, A., Eickholt, J., Cheng, J., Eickholt, J., Cheng, J., Farley, B., Clark, W., Gawehn, E., Hiss, J., Schneider, G., Gibbs, J., van der Brug, M., Hernandez, D., Traynor, B., Nalls, M., Lai, S., Arepalli, S., Dillman, A., Rafferty, I., Troncoso, J., Grubert, F., Zaugg, J., Kasowski, M., Ursu, O., Spacek, D., Martin, A., Greenside, P., Srivas, R., Phanstiel, D., Pekowska, A., Heidari, N., Euskirchen, G., Huber, W., Pritchard, J., Bustamante, C., Steinmetz, L., Kundaje, A., Snyder, M., Hastie, T., Tibshirani, R., Friedman, J., Franklin, J., Hinton, G., Salakhutdinov, R., Hinton, G., Osindero, S., Teh, Y., Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., Hubel, D., Wiesel, T., Hubel, D., Wiesel, T., Hutter, F., Hoos, H., LeytonâĂŘBrown, K., Kang, H., Ye, C., Eskin, E., Karlic, R., Chung, H., Lasserre, J., Vlahovicek, K., Vingron, M., Kell, D., Kelley, D., Snoek, J., Rinn, J., Koh, P., Pierson, E., Kundaje, A., LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L., LeCun, Y., Bengio, Y., Hinton, G., Leung, M., Xiong, H., Lee, L., Frey, B., Li, J., Ching, T., Huang, S., Garmire, L., Libbrecht, M., Noble, W., Lyons, J., Dehzangi, A., Heffernan, R., Sharma, A., Paliwal, K., Sattar, A., Zhou, Y., Yang, Y., Mamoshina, P., Vieira, A., Putin, E., Zhavoronkov, A., Märtens, K., Hallin, J., Warringer, J., Liti, G., Parts, L., McCulloch, W., Pitts, W., Menden, M., Iorio, F., Garnett, M., McDermott, U., Benes, C., Ballester, P., SaezâĂŘRodriguez, J., Montgomery, S., Sammeth, M., GutierrezâĂŘArcelus, M., Lach, R., Ingle, C., Nisbett, J., Guigo, R., Dermitzakis, E., Nesterov, Y., Ning, F., Delhomme, D., LeCun, Y., Piano, F., Bottou, L., Barbano, P., Park, Y., Kellis, M., Pärnamaa, T., Parts, L., Parts, L., Stegle, O., Winn, J., Durbin, R., Parts, L., Liu, Y., Tekkedil, M., Steinmetz, L., Caudy, A., Fraser, A., Boone, C., Andrews, B., Rosebrock, A., Pickrell, J., Marioni, J., Pai, A., Degner, J., Engelhardt, B., Nkadori, E., Veyrieras, J., Stephens, M., Gilad, Y., Pritchard, J., Rakitsch, B., Stegle, O., Rampasek, L., Goldenberg, A., Rosenblatt, F., Rumelhart, D., Hinton, G., Williams, R., Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Salakhutdinov, R., Hinton, G., Schmidhuber, J., Spencer, M., Eickholt, J., Cheng, J., Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., Stegle, O., Parts, L., Durbin, R., Winn, J., Stormo, G., Schneider, T., Gold, L., Ehrenfeucht, A., Swan, A., Mobasheri, A., Allaway, D., Liddell, S., Bacardit, J., Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P., Waszak, S., Delaneau, O., Gschwind, A., Kilpinen, H., Raghav, S., Witwicki, R., Orioli, A., Wiederkehr, M., Panousis, N., Yurovsky, A., RomanoâĂŘPalumbo, L., Planchon, A., Bielser, D., Padioleau, I., Udin, G., Thurnheer, S., Hacker, D., Hernandez, N., Reymond, A., Deplancke, B., Xiong, H., Alipanahi, B., Lee, L., Bretschneider, H., Merico, D., Yuen, R., Hua, Y., Gueroussov, S., Najafabadi, H., Hughes, T., Morris, Q., Barash, Y., Krainer, A., Jojic, N., Scherer, S., Blencowe, B., Frey, B., Xu, R., Wunsch, D., Frank, R., Zhou, J., and Troyanskaya, O. (2016). Deep learning for computational biology. *Mol. Syst. Biol.*, 12(7):878.

[Auerbach et al., 2009] Auerbach, R. K., Euskirchen, G., Rozowsky, J., Lamarre-Vincent, N., Moqtaderi, Z., Lefrançois, P., Struhl, K., Gerstein, M., and Snyder, M. (2009). Mapping accessible chromatin regions using Sono-Seq. *Proc. Natl. Acad. Sci. U. S. A.*, 106(35):14926–14931.

[Avsec et al., 2019] Avsec, Ž., Weilert, M., Shrikumar, A., Alexandari, A., Krueger, S., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., and Zeitlinger, J. (2019). Deep learning at base-resolution reveals motif syntax of the cis-regulatory code. *bioRxiv*, page 737981.

[Baker, 2011] Baker, M. (2011). Making sense of chromatin states. *Nat. Methods*, 8(9):717–722.

[Bannister and Kouzarides, 2011] Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Res.*, 21(3):381–395.

[Baxley et al., 2017] Baxley, R. M., Bullard, J. D., Klein, M. W., Fell, A. G., Morales-Rosado, J. A., Duan, T., and Geyer, P. K. (2017). Deciphering the DNA code for the function of the Drosophila polydactyl zinc finger protein Suppressor of Hairy-wing. *Nucleic Acids Res.*, 45(8):4463–4478.

[Beagan et al., 2016] Beagan, J. A., Gilgenast, T. G., Kim, J., Plona, Z., Norton, H. K., Hu, G., Hsu, S. C., Shields, E. J., Lyu, X., Apostolou, E., Hochedlinger, K., Corces, V. G., Dekker, J., and Phillips-Cremins, J. E. (2016). Local genome topology can exhibit an incompletely rewired 3D-folding state during somatic cell reprogramming. *Cell Stem Cell*, 18(5).

[Benos et al., 2002] Benos, P. V., Lapedes, A. S., and Stormo, G. D. (2002). Is there a code for protein?DNA recognition? Probab(ilistical)ly? *BioEssays*, 24(5):466–475.

[Berg and von Hippel, 1987a] Berg, O. G. and von Hippel, P. H. (1987a). Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, 193(4):723–50.

[Berg and von Hippel, 1987b] Berg, O. G. and von Hippel, P. H. (1987b). Selection of DNA binding sites by regulatory proteins statistical-mechanical theory and application to operators and promoters. *Journal of Molecular Biology*, 193(4):723–750.

[Boeynaems et al., 2018] Boeynaems, S., Alberti, S., Fawzi, N. L., Mittag, T., Polymenidou, M., Rousseau, F., Schymkowitz, J., Shorter, J., Wolozin, B., Van Den Bosch, L., Tompa, P., and Fuxreiter, M. (2018). Protein Phase Separation: A New Phase in Cell Biology. *Trends Cell Biol.*, 28(6):420–435.

[Bogliotti and Ross, 2012] Bogliotti, Y. S. and Ross, P. J. (2012). Mechanisms of histone H3 lysine 27 trimethylation remodeling during early mammalian development. *Epigenetics*, 7(9):976–981.

[Boija et al., 2018] Boija, A., Klein, I. A., Sabari, B. R., Dall'Agnese, A., Coffey, E. L., Zamudio, A. V., Li, C. H., Shrinivas, K., Manteiga, J. C., Hannett, N. M., Abraham, B. J., Afeyan, L. K., Guo, Y. E., Rimel, J. K., Fant, C. B., Schuijers, J., Lee, T. I., Taatjes, D. J., and Young, R. A. (2018). Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell*, 175(7):1842–1855.e16.

[Bonhoure et al., 2014] Bonhoure, N., Bounova, G., Bernasconi, D., Praz, V., Lammers, F., Canella, D., Willis, I. M., Herr, W., Hernandez, N., Delorenzi, M., Hernandez, N., Delorenzi, M., Deplancke, B., Desvergne, B., Guex, N., Herr, W., Naef, F., Rougemont, J., Schibler, U., Andersin, T., Cousin, P., Gilardi, F., Gos, P., Lammers, F., Raghav, S., Villeneuve, D., Fabbretti, R., Vlegel, V., Xenarios, I., Migliavacca, E., Praz, V., David, F., Jarosz, Y., Kuznetsov, D., Liechti, R., Martin, O., Delafontaine, J., Cajan, J., Gustafson, K., Krier, I., Leleu, M., Molina, N., Naldi, A., Rib, L., Symul, L., Bounova, G., and Bounova, G. (2014). Quantifying ChIP-seq data: a spiking method providing an internal reference for sample-to-sample normalization. *Genome Res.*, 24(7):1157–1168.

[Bray, 2016] Bray, S. J. (2016). Notch signalling in context. *Nat. Rev. Mol. Cell Biol.*, 17(11):722–735.

[Buenrostro et al., 2015] Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523(7561).

[Bushey et al., 2009] Bushey, A. M., Ramos, E., and Corces, V. G. (2009). Three subclasses of a Drosophila insulator show distinct and cell type-specific genomic distributions. *Genes Dev.*, 23(11):1338–50.

[Casey et al., 2018] Casey, B. H., Kollipara, R. K., Pozo, K., and Johnson, J. E. (2018). Intrinsic DNA binding properties demonstrated for lineage-specifying basic helix-loop-helix transcription factors. *Genome Res.*, 28(4).

[Champagne and Kutateladze, 2009] Champagne, K. and Kutateladze, T. (2009). Structural Insight Into Histone Recognition by the ING PHD Fingers. *Curr. Drug Targets*, 10(5):432–441.

[Chathoth and Zabet, 2019] Chathoth, K. T. and Zabet, N. R. (2019). Chromatin architecture reorganisation during neuronal cell differentiation in drosophila genome. *Genome Res.*, 29:613–625.

[Chauvet et al., 2000a] Chauvet, S., Merabet, S., Bilder, D., Scott, M. P., Pradel, J., and Graba, Y. (2000a). Distinct Hox protein sequences determine specificity in different tissues. *Proc. Natl. Acad. Sci.*, 97(8):4064–4069.

[Chauvet et al., 2000b] Chauvet, S., Merabet, S., Bilder, D., Scott, M. P., Pradel, J., and Graba, Y. (2000b). Distinct hox protein sequences determine specificity in different tissues. *Proc. Natl. Acad. Sci. U. S. A.*, 97(8):4064–9.

[Chen et al., 2016] Chen, K., Hu, Z., Xia, Z., Zhao, D., Li, W., and Tyler, J. K. (2016). The Overlooked Fact: Fundamental Need for Spike-In Control for Virtually All Genome-Wide Analyses. *Molecular and Cellular Biology*, 36(5):662–667.

[Chereji et al., 2016] Chereji, R. V., Kan, T.-W., Grudniewska, M. K., Romashchenko, A. V., Berezikov, E., Zhimulev, I. F., Guryev, V., Morozov, A. V., and Moshkin, Y. M. (2016). Genome-wide profiling of nucleosome sensitivity and chromatin accessibility in Drosophila melanogaster. *Nucleic Acids Res.*, 44(3):1036–51.

[Chronis et al., 2017] Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., and Plath, K. (2017). Cooperative Binding of Transcription Factors Orchestrates Reprogramming. *Cell*, 168(3).

[Chu et al., 2009] Chu, D., Zabet, N., and Mitavskiy, B. (2009). Models of transcription factor binding: Sensitivity of activation functions to model assumptions. *Journal of Theoretical Biology*, 257(3).

[Cubeñas-Potts et al., 2017] Cubeñas-Potts, C., Rowley, M. J., Lyu, X., Li, G., Lei, E. P., and Corces, V. G. (2017). Different enhancer classes in Drosophila bind distinct architectural proteins and mediate unique chromatin interactions and 3D architecture. *Nucleic Acids Res.*, 45(4):1714–1730.

[Cusanovich et al., 2014] Cusanovich, D. A., Pavlovic, B., Pritchard, J. K., and Gilad, Y. (2014). The Functional Consequences of Variation in Transcription Factor Binding. *PLoS Genet.*, 10(3):e1004226.

[Davis et al., 2018] Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J., Jolanki, O., Tanaka, F. Y., and Cherry, J. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, 46(D1):D794–D801.

[De Laat and Duboule, 2013] De Laat, W. and Duboule, D. (2013). Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature*, 502(7472):499–506.

[Dekker and Heard, 2015] Dekker, J. and Heard, E. (2015). Structural and functional diversity of Topologically Associating Domains. *FEBS Lett.*, 589(20).

[Dekker et al., 2013] Dekker, J., Marti-Renom, M. A., and Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.*, 14(6):390–403.

[Djordjevic et al., 2003] Djordjevic, M., Sengupta, A. M., and Shraiman, B. I. (2003). A biophysical approach to transcription factor binding site discovery. *Genome Res.*, 13(11):2381–2390.

[Domcke et al., 2015] 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(7583):575–579.

[Donaghey et al., 2018] Donaghey, J., Thakurela, S., Charlton, J., Chen, J. S., Smith, Z. D., Gu, H., Pop, R., Clement, K., Stamenova, E. K., Karnik, R., Kelley, D. R., Gifford, C. A., Cacchiarelli, D., Rinn, J. L., Gnirke, A., Ziller, M. J., and Meissner, A. (2018). Genetic determinants and epigenetic effects of pioneer-factor occupancy. *Nat. Genet.*, page 1.

[Dubuis et al., 2013a] Dubuis, J. O., Samanta, R., and Gregor, T. (2013a). Accurate measurements of dynamics and reproducibility in small genetic networks. *Mol. Syst. Biol.*, 9.

[Dubuis et al., 2013b] Dubuis, J. O., Tkacik, G., Wieschaus, E. F., Gregor, T., and Bialek, W. (2013b). Quantifying positional information during early embryonic development. *Phd thesis*, 110(41):16301–16308.

[Egan et al., 2016] Egan, B., Yuan, C. C., Craske, M. L., Labhart, P., Guler, G. D., Arnott, D., Maile, T. M., Busby, J., Henry, C., Kelly, T. K., Tindell, C. A., Jhunjhunwala, S., Zhao, F., Hatton, C., Bryant, B. M., Classon, M., and Trojer, P. (2016). An alternative approach to ChIP-Seq normalization enables detection of genome-wide changes in histone H3 lysine 27 trimethylation upon EZH2 inhibition. *PLoS ONE*, 11(11).

[El Khattabi et al., 2019] El Khattabi, L., Zhao, H., Kalchschmidt, J., Young, N., Jung, S., Van Blerkom, P., Kieffer-Kwon, P., Kieffer-Kwon, K.-R., Park, S., Wang, X., Krebs, J., Tripathi, S., Sakabe, N., Sobreira, D. R., Huang, S.-C., Rao, S. S. P., Pruett, N., Chauss, D., Sadler, E., Lopez, A., Nóbrega, M. A., Aiden, E. L., Asturias, F. J., and Casellas, R. (2019). A Pliable Mediator Acts as a Functional Rather Than an Architectural Bridge between Promoters and Enhancers. *Cell*, 178(5):1145–1158.e20.

[Elemento and Tavazoie, 2005] Elemento, O. and Tavazoie, S. (2005). Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol.*, 6(2):R18.

[ENCODE Project Consortium, 2012] ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74.

[Erbas and Marko, 2019] Erbas, A. and Marko, J. F. (2019). How do DNA-bound proteins leave their binding sites? The role of facilitated dissociation. *Curr. Opin. Chem. Biol.*, 53:118–124.

[Farley et al., 2015] Farley, E. K., Olson, K. M., Zhang, W., Brandt, A. J., Rokhsar, D. S., and Levine, M. S. (2015). Suboptimization of developmental enhancers. *Science (80-. ).*, 350(6258):325–328.

[Farley et al., 2016] Farley, E. K., Olson, K. M., Zhang, W., Rokhsar, D. S., and Levine, M. S. (2016). Syntax compensates for poor binding sites to encode tissue specificity of developmental enhancers. *Proc. Natl. Acad. Sci. U. S. A.*, 113(23):6508–13.

[Farnham, 2009] Farnham, P. J. (2009). Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, 10(9):605–616.

[Fonseca, 2012] Fonseca, N. A. (2012). Tools for mapping high-throughput sequencing data . *Bioinfromatics*, 28(24):3169–3177.

[Gehring et al., 1994] Gehring, W. J., Qian, Y. Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A. F., Resendez-Perez, D., Affolter, M., Otting, G., and Wüthrich, K. (1994). Homeodomain-DNA recognition. *Cell*, 78(2):211–23.

[Gentleman et al., 2004] Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80.

[Ghavi-Helm et al., 2019] Ghavi-Helm, Y., Jankowski, A., Meiers, S., Viales, R. R., Korbel, J. O., and Furlong, E. E. M. (2019). Highly rearranged chromosomes reveal uncoupling between genome topology and gene expression. *Nat. Genet.*, 51(8).

[Gibcus and Dekker, 2013] Gibcus, J. H. and Dekker, J. (2013). The Hierarchy of the 3D Genome. *Mol. Cell*, 49(5).

[Glenn and Geyer, 2019] Glenn, S. E. and Geyer, P. K. (2019). Investigation of the developmental requirements of drosophila HP1 and insulator protein partner, HIPP1. *G3 Genes, Genomes, Genet.*, 9(2):345–357.

[Goh et al., 2010] Goh, W. S., Orlov, Y., Li, J., and Clarke, N. D. (2010). Blurring of high-resolution data shows that the effect of intrinsic nucleosome occupancy on transcription factor binding is mostly regional, not local. *PLoS Computational Biology*, 6(1).

[Gomez-Lamarca et al., 2018] Gomez-Lamarca, M. J., Falo-Sanjuan, J., Stojnic, R., Abdul Rehman, S., Muresan, L., Jones, M. L., Pillidge, Z., Cerda-Moya, G., Yuan, Z., Baloul, S., Valenti, P., Bystricky, K., Payre, F., O'Holleran, K., Kovall, R., and Bray,

S. J. (2018). Activation of the Notch Signaling Pathway InÂăVivo Elicits Changes in CSL Nuclear Dynamics. *Dev. Cell*, 44(5):611–623.e7.

[Granek and Clarke, 2005] Granek, J. A. and Clarke, N. D. (2005). Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol.*, 6(10):R87.

[Grant et al., 2011] Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017.

[Guelen et al., 2008] Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M. B., Talhout, W., Eussen, B. H., De Klein, A., Wessels, L., De Laat, W., and Van Steensel, B. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453(7197):948–951.

[Guo et al., 2015] Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D. U., Jung, I., Wu, H., Zhai, Y., Tang, Y., Lu, Y., Wu, Y., Jia, Z., Li, W., Zhang, M. Q., Ren, B., Krainer, A. R., Maniatis, T., and Wu, Q. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell*, 162(4).

[Halder et al., 1995] Halder, G., Callaerts, P., and Gehring, W. J. (1995). Induction of ectopic eyes by targeted expression of the eyeless gene in Drosophila. *Science*, 267(5205):1788–92.

[Hammar et al., 2012] Hammar, P., Leroy, P., Mahmutovic, A., Marklund, E. G., Berg, O. G., and Elf, J. (2012). The lac repressor displays facilitated diffusion in living cells. *Science*, 336(6088):1595–1598.

[Hansen et al., 2019] Hansen, A. S., Amitai, A., Cattoglio, C., Tjian, R., and Darzacq, X. (2019). Guided nuclear exploration increases CTCF target search efficiency. *Nat. Chem. Biol.*

[Hansen et al., 2018] Hansen, A. S., Cattoglio, C., Darzacq, X., and Tjian, R. (2018). Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus*, 9(1).

[Harr et al., 1983] Harr, R., Häggström, M., and Gustafsson, P. (1983). Search algorithm for pattern match analysis of nucleic acid sequences. *Nucleic Acids Res.*, 11(9):2943–57.

[Hayashi and Scott, 1990] Hayashi, S. and Scott, M. P. (1990). What determines the specificity of action of Drosophila homeodomain proteins? *Cell*, 63(5):883–894.

[Hayes and Wolffe, 1992] Hayes, J. J. and Wolffe, A. P. (1992). The interaction of transcription factors with nucleosomal DNA. *BioEssays*, 14(9):597–603.

[He et al., 2010] He, X., Samee, M. A. H., Blatti, C., Sinha, S., Casamayor, A., Lebrecht, D., Foehr, M., Smith, E., Lopes, F., Vanario-Alonso, C., Arnosti, D., Barolo, S., Levine, M., Small, S., Fakhouri, W., Ay, A., Sayal, R., Dresch, J., Dayringer, E., Shea, M., Ackers, G., Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., Gaul, U., Buchler, N., Gerland, U., Hwa, T., Joung, J., Le, L., Hochschild, A., Berg, O., von Hippel, P., Stormo, G., Fields, D., Bauer, D., Bailey, T., Gertz, J., Siggia, E., Cohen, B., Janssens, H., Hou, S., Jaeger, J., Kim, A., Myasnikova, E., Zinzen, R., Papatsenko, D., Tanay, A., Lifanov, A., Makeev, V., Nazina, A., Papatsenko, D., Gray, S., Levine, M., Kulkarni, M., Arnosti, D., Makeev, V., Lifanov, A., Nazina, A., Papatsenko, D., Nibu, Y., Zhang, H., Bajor, E., Barolo, S., Small, S., Sauer, F., Fondell, J., Ohkuma, Y., Roeder, R., Jackle, H., Green, M., Carey, M., Struhl, K., Veitia, R., Sauer, F., Hansen, S., Tjian, R., Ma, X., Yuan, D., Diepold, K., Scarborough, T., Ma, J., Hoch, M., Gerwin, N., Taubert, H., Jackle, H., Moses, A., Pollard, D., Nix, D., Iyer, V. V. V., Li, X., Dermitzakis, E., Bergman, C., Clark, A., Birney, E., Stamatoyannopoulos, J., Dutta, A., Guigo, R., Gingeras, T., Hu, Z., Killion, P., Iyer, V. V. V., Reinitz, J., Hou, S., Sharp, D., Beer, M., Tavazoie, S., Zinzen, R., Girardot, C., Gagneur, J., Braun, M., Furlong, E., Benos, P., Bulyk, M., Stormo, G., Maerkl, S., Quake, S., Keller, S., Mao, Y., Struffi, P., Margulies, C., Yurk, C., Hermsen, R., Tans, S., ten Wolde, P., Teif, V., Andrioli, L., Vasisht, V., Theodosopoulou, E., Oberstein, A., Small, S., Jimenez, G., Guichet, A., Ephrussi, A., Casanova, J., Noyes, M., Meng, X., Wakabayashi, A., Sinha, S., Brodsky, M., Bergman, C., Carlson, J., Celniker, S., Homsi, D., Gupta, V., Stormo, G., Ray, P., Shringarpure, S., Kolar, M., Xing, E., Borneman, A., Gianoulis, T., Zhang, Z., Yu, H., Rozowsky, J., Ludwig, M., Gao, F., Foat, B., Bussemaker, H., Burz, D., Rivera-Pomar, R., Jackle, H., Hanes, S., Chi, T., Lieberman, P., Ellwood, K., Carey, M., Gray, S., Szymanski, P., Levine, M., Small, S., Arnosti, D., Levine, M., Rosee-Borggreve, A. L., Hader, T., Wainwright, D., Sauer, F., Jackle, H., Small, S., Blair, A., Levine, M., Yan, R., Small, S., Desplan, C., Dearolf, C., Jr, J. D., Krumm, A., Hickey, L., Groudine, M., Nakanishi, H., Mitarai, N., Sneppen, K., Zenklusen, D., Larson, D., Singer, R., Fowlkes, C., Hendriks, C., Keranen, S., Weber, G., Rubel, O., Morozov, A., Fortney, K., Gaykalova, D., Studitsky, V., Widom, J., Wasson, T., Hartemink, A., Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Lusk, R., and Eisen, M. (2010). Thermodynamics-Based Models of Transcriptional Regulation by Enhancers: The Roles of Synergistic Activation, Cooperative Binding and Short-Range Repression. *PLoS Comput. Biol.*, 6(9):e1000935.

[Hernandez et al., 2000] Hernandez, A., Smith, F., Wang, Q., Wang, X., and Evers, B. M. (2000). Assessment of differential gene expression patterns in human colon cancers. *Ann. Surg.*, 232(4):576–85.

[Heumann et al., 1994] Heumann, J. M., Lapedes, A. S., and Stormo, G. D. (1994). Neural networks for determining protein specificity and multiple alignment of binding sites. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 2:188–194.

[Hnisz et al., 2017] Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K., and Sharp, P. A. (2017). A Phase Separation Model for Transcriptional Control. *Cell*, 169(1):13–23.

[Honrado et al., 2006] Honrado, E., Osorio, A., Palacios, J., and Benitez, J. (2006). Pathology and gene expression of hereditary breast tumors associated with BRCA1, BRCA2 and CHEK2 gene mutations. *Oncogene*, 25(43):5837–5845.

[Inukai et al., 2017] Inukai, S., Kock, K. H., and Bulyk, M. L. (2017). Transcription factorâĂŞDNA binding: beyond binding site motifs. *Curr. Opin. Genet. Dev.*, 43:110–119.

[Iwafuchi-Doi, 2019] Iwafuchi-Doi, M. (2019). The mechanistic basis for chromatin regulation by pioneer transcription factors. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 11(1):e1427.

[Iwafuchi-Doi and Zaret, 2014] Iwafuchi-Doi, M. and Zaret, K. S. (2014). Pioneer transcription factors in cell reprogramming. *Genes Dev.*, 28(24):2679–92.

[Jiang et al., 2009] Jiang, N., Emberly, E., Cuvier, O., and Hart, C. M. (2009). Genome-wide mapping of boundary element-associated factor (BEAF) binding sites in Drosophila melanogaster links BEAF to transcription. *Mol. Cell. Biol.*, 29(13):3556–68.

[Jolma et al., 2013] 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(1-2):327–339.

[Joshi et al., 2010] Joshi, R., Sun, L., and Mann, R. (2010). Dissecting the functional specificities of two Hox proteins. *Genes Dev.*, 24(14):1533–45.

[Kamakaka and Thomas, 1990] Kamakaka, R. T. and Thomas, J. O. (1990). Chromatin structure of transcriptionally competent and repressed genes. *EMBO J.*, 9(12):3997–4006.

[Kaplan et al., 2011a] Kaplan, T., Li, X.-Y., Sabo, P. J., Thomas, S., Stamatoyannopoulos, J. A., Biggin, M. D., and Eisen, M. B. (2011a). Quantitative Models of the Mechanisms That Control Genome-Wide Patterns of Transcription Factor Binding during Early

Drosophila Development. *PLoS Genet*, 7(2):e1001290.

[Kaplan et al., 2011b] Kaplan, T., Li, X.-Y., Sabo, P. J., Thomas, S., Stamatoyannopoulos, J. A., Biggin, M. D., and Eisen, M. B. (2011b). Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early *Drosophila* development. *PLoS Genetics*, 7(2):e1001290.

[Kasahara et al., 2017] Kasahara, K., Shiina, M., Fukuda, I., Ogata, K., and Nakamura, H. (2017). Molecular mechanisms of cooperative binding of transcription factors Runx1-CBF$\beta$-Ets1 on the TCR$\alpha$ gene enhancer. *PLoS One*, 12(2).

[Kaya-Okur et al., 2019] Kaya-Okur, H. S., Wu, S. J., Codomo, C. A., Pledger, E. S., Bryson, T. D., Henikoff, J. G., Ahmad, K., and Henikoff, S. (2019). CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications*, 10(1):1–10.

[Keilwagen et al., 2019] Keilwagen, J., Posch, S., and Grau, J. (2019). Accurate prediction of cell type-specific transcription factor binding. *Genome Biol.*, 20(1):9.

[Kharchenko et al., 2010] Kharchenko, P. V., Alekseyenko, A. A., Schwartz, Y. B., Minoda, A., Riddle, N. C., Ernst, J., Sabo, P. J., Larschan, E., Gorchakov, A. A., Gu, T., Linder-Basso, D., Plachetka, A., Shanower, G., Tolstorukov, M. Y., Luquette, L. J., Xi, R., Jung, Y. L., Park, R. W., Bishop, E. P., Canfield, T. P., Sandstrom, R., Thurman, R. E., MacAlpine, D. M., Stamatoyannopoulos, J. A., Kellis, M., Elgin, S. C. R., Kuroda, M. I., Pirrotta, V., H., G., and Park, P. J. (2010). Comprehensive analysis of the chromatin landscape in *Drosophila* melanogaster. *Nature*.

[Kim et al., 2015] Kim, S., Yu, N. K., and Kaang, B. K. (2015). CTCF as a multifunctional protein in genome regulation and gene expression. *Exp. Mol. Med.*, 47:e166.

[Klemm et al., 2019] Klemm, S. L., Shipony, Z., and Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.*, page 1.

[Krejčí and Bray, 2007] Krejčí, A. and Bray, S. (2007). Notch activation stimulates transient and selective binding of Su(H)/CSL to target enhancers. *Genes Dev.*, 21(11):1322–1327.

[Kuhn-Parnell et al., 2008] Kuhn-Parnell, E. J., Helou, C., Marion, D. J., Gilmore, B. L., Parnell, T. J., Wold, M. S., and Geyer, P. K. (2008). Investigation of the Properties of Non- <i>gypsy</i> Suppressor of Hairy-wing-Binding Sites. *Genetics*, 179(3):1263–1273.

[Kurshakova et al., 2007] Kurshakova, M., Maksimenko, O., Golovnin, A., Pulina, M., Georgieva, S., Georgiev, P., and Krasnov, A. (2007). Evolutionarily Conserved E(y)2/Sus1 Protein Is Essential for the Barrier Activity of Su(Hw)-Dependent Insulators in Drosophila. *Mol. Cell*, 27(2):332–338.

[Kuziora and McGinnis, 1989] Kuziora, M. A. and McGinnis, W. (1989). A homeodomain substitution changes the regulatory specificity of the Deformed protein in drosophila embryos. *Cell*, 59(3):563–571.

[Lai et al., 2018] Lai, X., Stigliani, A., Vachon, G., Carles, C., Smaczniak, C., Zubieta, C., Kaufmann, K., and Parcy, F. (2018). Building Transcription Factor Binding Site Models to Understand Gene Regulation in Plants. *Mol. Plant*.

[Lake et al., 2014] Lake, R. J., Tsai, P. F., Choi, I., Won, K. J., and Fan, H. Y. (2014). RBPJ, the Major Transcriptional Effector of Notch Signaling, Remains Associated with Chromatin throughout Mitosis, Suggesting a Role in Mitotic Bookmarking. *PLoS Genet.*, 10(3).

[Lambert et al., 2018] Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The Human Transcription Factors. *Cell*, 172(4).

[Lamparter et al., 2017] Lamparter, D., Marbach, D., Rueedi, R., Bergmann, S., and Kutalik, Z. (2017). Genome-Wide Association between Transcription Factor Expression and Chromatin Accessibility Reveals Regulators of Chromatin Accessibility. *PLOS Comput. Biol.*, 13(1):e1005311.

[Landt et al., 2012a] Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J., and Snyder, M. (2012a). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, 22(9):1813–1831.

[Landt et al., 2012b] Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E., Bickel, P., Brown, J. B., Cayting, P., Chen, Y., DeSalvo, G., Epstein, C., Fisher-Aylor, K. I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A. J., Hoffman, M. M., Iyer, V. R., Jung, Y. L., Karmakar, S., Kellis, M., Kharchenko, P. V., Li, Q., Liu, T., Liu, X. S., Ma, L., Milosavljevic, A., Myers, R. M., Park, P. J., Pazin, M. J., Perry, M. D., Raha, D., Reddy, T. E., Rozowsky, J., Shoresh, N., Sidow, A., Slattery, M., Stamatoyannopoulos, J. A., Tolstorukov, M. Y., White, K. P., Xi, S., Farnham, P. J., Lieb, J. D., Wold, B. J., and Snyder, M. (2012b). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, 22(9):1813–1831.

[Langmead and Salzberg, 2012] Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359.

[Latchman, 2001] Latchman, D. S. (2001). Transcription factors: bound to activate or repress. *Trends Biochem. Sci.*, 26(4):211–3.

[Lee et al., 2014a] Lee, H., McManus, C. J., Cho, D.-Y., Eaton, M., Renda, F., Somma, M. P., Cherbas, L., May, G., Powell, S., Zhang, D., Zhan, L., Resch, A., Andrews, J., Celniker, S. E., Cherbas, P., Przytycka, T. M., Gatti, M., Oliver, B., Graveley, B., and MacAlpine, D. (2014a). Dna copy number evolution in drosophila cell lines. *Genome Biology*, 15(8):R70.

[Lee et al., 2014b] Lee, Y.-H., Kim, K.-S., Jang, Y.-S., Hwang, J.-H., Lee, D.-H., and Choi, I.-H. (2014b). Global gene expression responses to waterlogging in leaves of rape seedlings. *Plant Cell Rep.*, 33(2):289–299.

[Li, 2002] Li, E. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. *Nat. Rev. Genet.*, 3(9):662–673.

[Li et al., 2009] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, . G. P. D. P. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

[Li et al., 2019] Li, H., Quang, D., and Guan, Y. (2019). Anchor: trans-cell type prediction of transcription factor binding sites. *Genome Res.*, 29(2):281–292.

[Li et al., 2017] Li, J., Sagendorf, J. M., Chiu, T. P., Pasi, M., Perez, A., and Rohs, R. (2017). Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic Acids Res.*, 45(22):12877–12887.

[Li et al., 2015] Li, L., Lyu, X., Hou, C., Takenaka, N., Nguyen, H. Q., Ong, C.-T., CubeÃśas-Potts, C., Hu, M., Lei, E. P., Bosco, G., Qin, Z. S., and Corces, V. G. (2015). Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. *Molecular Cell*, 58(2):216–231.

[Li et al., 2008] Li, X.-y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D. A., Iyer, V. N., Hechmer, A., Simirenko, L., Stapleton, M., Hendriks, C. L. L., Chu, H. C., Ogawa, N., Inwood, W., Sementchenko, V., Beaton, A., Weiszmann, R., Celniker, S. E., Knowles, D. W., Gingeras, T., Speed, T. P., Eisen, M. B., and Biggin, M. D.

(2008). Transcription Factors Bind Thousands of Active and Inactive Regions in the Drosophila Blastoderm. *PLoS Biol.*, 6(2):e27.

[Li et al., 2011] Li, X.-Y., Thomas, S., Sabo, P. J., Eisen, M. B., Stamatoyannopoulos, J. A., and Biggin, M. D. (2011). The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biol.*, 12(4):R34.

[Liang and Pardee, 2003] Liang, P. and Pardee, A. B. (2003). Timeline: Analysing differential gene expression in cancer. *Nat. Rev. Cancer*, 3(11):869–876.

[Lickwar et al., 2012] Lickwar, C. R., Mueller, F., Hanlon, S. E., McNally, J. G., and Lieb, J. D. (2012). Genome-wide proteinâĂŞDNA binding dynamics suggest a molecular clutch for transcription factor function. *Nature*, 484(7393):251–255.

[Lieberman-Aiden et al., 2009] Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., and Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–93.

[Liu et al., 2016] Liu, X., Wang, C., Liu, W., Li, J., Li, C., Kou, X., Chen, J., Zhao, Y., Gao, H., Wang, H., Zhang, Y., Gao, Y., and Gao, S. (2016). Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos. *Nature*, 537(7621):558–562.

[Lupiáñez et al., 2015] Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., and Mundlos, S. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025.

[Lyko et al., 2000] Lyko, F., Ramsahoye, B. H., and Jaenisch, R. (2000). DNA methylation in Drosophila melanogaster. *Nature*, 408(6812):538–540.

[Mallo and Alonso, 2013] Mallo, M. and Alonso, C. R. (2013). The regulation of Hox gene expression during animal development. *Development*, 140(19):3951–3963.

[Mann et al., 2009] Mann, R. S., Lelli, K. M., and Joshi, R. (2009). Hox specificity unique roles for cofactors and collaborators. *Curr. Top. Dev. Biol.*, 88:63–101.

[Martin, 2017] Martin, P. C. (2017). *ChIPanalyser: Predicting Transcription Factor Binding Sites*. R package version 1.8.

[Mathelier et al., 2014] Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C.-y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., and Wasserman, W. W. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 42(D1):D142–D147.

[Mayran et al., 2018] Mayran, A., Khetchoumian, K., Hariri, F., Pastinen, T., Gauthier, Y., Balsalobre, A., and Drouin, J. (2018). Pioneer factor Pax7 deploys a stable enhancer repertoire for specification of cell fate. *Nat. Genet.*, page 1.

[McKinney et al., 2020] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. C., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., Ledsam, J. R., Melnick, D., Mostofi, H., Peng, L., Reicher, J. J., Romera-Paredes, B., Sidebottom, R., Suleyman, M., Tse, D., Young, K. C., De Fauw, J., and Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94.

[Melnikova et al., 2019] Melnikova, L., Elizar'ev, P., Erokhin, M., Molodina, V., Chetverina, D., Kostyuchenko, M., Georgiev, P., and Golovnin, A. (2019). The same domain of Su(Hw) is required for enhancer blocking and direct promoter repression. *Sci. Rep.*, 9(1).

[Miller and Grant, 2013] Miller, J. L. and Grant, P. A. (2013). The role of DNA methylation and histone modifications in transcriptional regulation in humans. *Subcell. Biochem.*, 61:289–317.

[Mirny, 2010] Mirny, L. A. (2010). Nucleosome-mediated cooperativity between transcription factors. *Proceedings of the National Academy of Sciences of the United States of America*, 107(52):22534–22539.

[modENCODE Consortium et al., 2010] modENCODE Consortium, T., Roy, S., Ernst, J., Kharchenko, P. V., Kheradpour, P., Negre, N., Eaton, M. L., Landolin, J. M., Bristow, C. A., Ma, L., Lin, M. F., Washietl, S., Arshinoff, B. I., Ay, F., Meyer, P. E., Robine, N., Washington, N. L., Di Stefano, L., Berezikov, E., Brown, C. D., Candeias, R., Carlson, J. W., Carr, A., Jungreis, I., Marbach, D., Sealfon, R., Tolstorukov, M. Y., Will, S., Alekseyenko, A. A., Artieri, C., Booth, B. W., Brooks, A. N., Dai, Q., Davis, C. A., Duff, M. O., Feng, X., Gorchakov, A. A., Gu, T., Henikoff, J. G., Kapranov, P., Li, R., MacAlpine, H. K., Malone, J., Minoda, A., Nordman, J., Okamura, K., Perry, M., Powell, S. K., Riddle, N. C., Sakai, A., Samsonova, A., Sandler, J. E., Schwartz, Y. B., Sher, N., Spokony, R., Sturgill, D., van Baren, M., Wan, K. H., Yang, L., Yu, C., Feingold, E., Good, P., Guyer, M., Lowdon, R., Ahmad, K., Andrews, J., Berger, B., Brenner, S. E., Brent, M. R., Cherbas, L., Elgin, S. C. R., Gingeras, T. R., Grossman, R., Hoskins, R. A., Kaufman, T. C., Kent, W., Kuroda, M. I., Orr-Weaver, T., Perrimon, N., Pirrotta, V., Posakony, J. W., Ren, B., Russell, S., Cherbas, P., Graveley, B. R., Lewis, S., Micklem, G., Oliver, B., Park, P. J., Celniker, S. E., Henikoff, S., Karpen, G. H., Lai, E. C., MacAlpine, D. M., Stein, L. D., White, K. P., and Kellis, M. (2010). Identification of functional elements and regulatory circuits by drosophila modencode. *Science*, 330(6012):1787–1797.

[Moens and Selleri, 2006] Moens, C. B. and Selleri, L. (2006). Hox cofactors in vertebrate development. *Dev. Biol.*, 291(2):193–206.

[Nakahashi et al., 2013] Nakahashi, H., Kwon, K.-R., Resch, W., Vian, L., Dose, M., Stavreva, D., Hakim, O., Pruett, N., Nelson, S., Yamane, A., Qian, J., Dubois, W., Welsh, S., Phair, R., Pugh, B., Lobanenkov, V., Hager, G., and Casellas, R. (2013). A Genome-wide Map of CTCF Multivalency Redefines the CTCF Code. *Cell Rep.*, 3(5):1678–1689.

[Nakato and Shirahige, 2017] Nakato, R. and Shirahige, K. (2017). Recent advances in ChIP-seq analysis: From quality management to whole-genome annotation. *Brief. Bioinform.*, 18(2):279–290.

[Nichols and Corces, 2015] Nichols, M. H. and Corces, V. G. (2015). A CTCF Code for 3D Genome Architecture. *Cell*, 162(4).

[Nicolas et al., 2017] Nicolas, D., Phillips, N. E., and Naef, F. (2017). What shapes eukaryotic transcriptional bursting? *Mol. BioSyst.*, 13:1280–1290.

[Ohnishi et al., 2014] Ohnishi, Y., Huber, W., Tsumura, A., Kang, M., Xenopoulos, P., Kurimoto, K., Oleå, A. K., Araúzo-Bravo, M. J., Saitou, M., Hadjantonakis, A. K., and Hiiragi, T. (2014). Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nat. Cell Biol.*, 16(1):27–37.

[Orlando et al., 2014] Orlando, D. A., Chen, M. W., Brown, V. E., Solanki, S., Choi, Y. J., Olson, E. R., Fritz, C. C., Bradner, J. E., and Guenther, M. G. (2014). Quantitative ChIP-Seq normalization reveals global modulation of the epigenome. *Cell Reports*, 9(3):1163–1170.

[Osório, 2015] Osório, J. (2015). Gene regulation: Landscape and mechanisms of transcription factor cooperativity. *Nat. Rev. Genet.*, 17(1):5–5.

[Osório, 2016] Osório, J. (2016). Gene regulation: Landscape and mechanisms of transcription factor cooperativity. *Nat. Rev. Genet.*, 17(1):5.

[Pages, 2018] Pages, H. (2018). *BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs*. R package version 1.49.5.

[Papadopoulos et al., 2019] Papadopoulos, D. K., Skouloudaki, K., Engström, Y., Terenius, L., Rigler, R., Zechner, C., Vukojević, V., and Tomancak, P. (2019). Control of hox transcription factor concentration and cell-to-cell variability by an auto-regulatory switch. *Dev.*, 146(12).

[Park, 2009] Park, P. J. (2009). Chip-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680.

[Parnell et al., 2006] Parnell, T. J., Kuhn, E. J., Gilmore, B. L., Helou, C., Wold, M. S., and Geyer, P. K. (2006). Identification of Genomic Sites That Bind the Drosophila Suppressor of Hairy-wing Insulator Protein. *Mol. Cell. Biol.*, 26(16):5983–5993.

[Passegué et al., 2005] Passegué, E., Wagers, A. J., Giuriato, S., Anderson, W. C., and Weissman, I. L. (2005). Global analysis of proliferation and cell cycle gene expression in the regulation of hematopoietic stem and progenitor cell fates. *J. Exp. Med.*, 202(11):1599–611.

[Pellerin et al., 1994] Pellerin, I., Schnabel, C., Catron, K. M., and Abate, C. (1994). Hox proteins have different affinities for a consensus DNA site that correlate with the positions of their genes on the hox cluster. *Mol. Cell. Biol.*, 14(7):4532–45.

[Petkova et al., 2019] Petkova, M. D., Tkačik, G., Bialek, W., Wieschaus, E. F., and Gregor, T. (2019). Optimal Decoding of Cellular Identities in a Genetic Network. *Cell*, 176(4):844–855.e15.

[Phelps and Brand, 1998] Phelps, C. B. and Brand, A. H. (1998). Ectopic Gene Expression inDrosophilaUsing GAL4 System. *Methods*, 14(4):367–379.

[Pique-Regi et al., 2011] Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., and Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, 21(3):447–55.

[Pope et al., 2014] Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D. L., Wang, Y., Hansen, R. S., Canfield, T. K., Thurman, R. E., Cheng, Y., Gülsoy, G., Dennis, J. H., Snyder, M. P., Stamatoyannopoulos, J. A., Taylor, J., Hardison, R. C., Kahveci, T., Ren, B., and Gilbert, D. M. (2014). Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527).

[Porcelli et al., 2019] Porcelli, D., Fischer, B., Russell, S., and White, R. (2019). Chromatin accessibility plays a key role in selective targeting of Hox proteins. *Genome Biol.*, 20(1):115.

[Praxenthaler et al., 2017] Praxenthaler, H., Nagel, A. C., Schulz, A., Zimmermann, M., Meier, M., Schmid, H., Preiss, A., and Maier, D. (2017). Hairless-binding deficient Suppressor of Hairless alleles reveal Su(H) protein levels are dependent on complex formation with Hairless. *PLoS Genet.*, 13(5).

[Ptashne and Gann, 1997] Ptashne, M. and Gann, A. (1997). Transcriptional activation by recruitment. *Nature*, 386(6625):569–577.

[Qi and Zhang, 2018] Qi, Y. and Zhang, B. (2018). Predicting three-dimensional genome organization with chromatin states. *bioRxiv*, page 282095.

[Quang and Xie, 2019] Quang, D. and Xie, X. (2019). FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Methods*, 166:40–47.

[Rae and Steele, 1979] Rae, P. M. and Steele, R. E. (1979). Absence of cytosine methylation at C-C-G-G and G-C-G-C sites in the rDNA coding regions and intervening sequences of Drosophila and the rDNA of other higher insects. *Nucleic Acids Res.*, 6(9):2987–2995.

[Raj et al., 2015a] Raj, A., Shim, H., Gilad, Y., Pritchard, J. K., and Stephens, M. (2015a). msCentipede: Modeling Heterogeneity across Genomic Sites and Replicates Improves Accuracy in the Inference of Transcription Factor Binding. *PLoS One*, 10(9):e0138030.

[Raj et al., 2015b] Raj, A., Shim, H., Gilad, Y., Pritchard, J. K., and Stephens, M. (2015b). msCentipede: Modeling Heterogeneity across Genomic Sites and Replicates Improves Accuracy in the Inference of Transcription Factor Binding. *PLoS One*, 10(9):e0138030.

[Rao et al., 2014] Rao, S., Huntley, M., Durand, N., Stamenova, E., Bochkov, I., Robinson, J., Sanborn, A., Machol, I., Omer, A., Lander, E., and Aiden, E. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159(7):1665–1680.

[Rennie et al., 2018] Rennie, S., Dalby, M., Lloret-Llinares, M., Bakoulis, S., DalagerÂă-VaagensÃÿ, C., HeickÂăJensen, T., and Andersson, R. (2018). Transcription start site analysis reveals widespread divergent transcription in d. melanogaster and core promoter-encoded enhancer activities. *Nucleic Acids Research*, 46(11):5455–5469.

[Rezsohazy et al., 2015] Rezsohazy, R., Saurin, A. J., Maurel-Zaffran, C., and Graba, Y. (2015). Cellular and molecular insights into Hox protein action. *Development*, 142(7):1212–1227.

[Roider et al., 2007] Roider, H. G., Kanhere, A., Manke, T., and Vingron, M. (2007). Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, 23(2):134–141.

[Roscher et al., 2019] Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2019). Explainable machine learning for scientific insights and discoveries. Technical report.

[Rouzina and Bloomfield, 1997] Rouzina, I. and Bloomfield, V. A. (1997). Competitive electrostatic binding of charged ligands to polyelectrolytes: Practical approach using the non-linear Poisson-Boltzmann equation. *Biophysical Chemistry*, 64(1-3):139–155.

[Rowicka et al., 2007] Rowicka, M., Kudlicki, A., Tu, B. P., and Otwinowski, Z. (2007). High-resolution timing of cell cycle-regulated gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, 104(43):16892–7.

[Salekin et al., 2017] Salekin, S., Zhang, J. M., and Huang, Y. (2017). A deep learning model for predicting transcription factor binding location at single nucleotide resolution. *2017 IEEE EMBS Int. Conf. Biomed. Heal. Informatics*, pages 57–60.

[Salekin et al., 2018] Salekin, S., Zhang, J. M., and Huang, Y. (2018). Base-pair resolution detection of transcription factor binding site by deep deconvolutional network. *bioRxiv*, page 254508.

[Sammak and Zinzalla, 2015] Sammak, S. and Zinzalla, G. (2015). Targeting protein-protein interactions (PPIs) of transcription factors: Challenges of intrinsically disordered proteins (IDPs) and regions (IDRs). *Prog. Biophys. Mol. Biol.*, 119(1):41–46.

[Schmidt et al., 2010] Schmidt, D., Schwalie, P. C., Ross-Innes, C. S., Hurtado, A., Brown, G. D., Carroll, J. S., Flicek, P., and Odom, D. T. (2010). A CTCF-independent role for cohesin in tissue-specific transcription. *Genome Res.*, 20(5):578–588.

[Schneuwly et al., 1987] Schneuwly, S., Klemenz, R., and Gehring, W. J. (1987). Redesigning the body plan of Drosophilaby ectopic expression of the homoeotic gene Antennapedia. *Nature*, 325(6107):816–818.

[Schoborg and Labrador, 2010] Schoborg, T. A. and Labrador, M. (2010). The Phylogenetic Distribution of Non-CTCF Insulator Proteins Is Limited to Insects and Reveals that BEAF-32 Is Drosophila Lineage Specific. *J. Mol. Evol.*, 70(1):74–84.

[Schoech and Zabet, 2014] Schoech, A. P. and Zabet, N. R. (2014). Facilitated diffusion buffers noise in gene expression. *Phys. Rev. E*, 90(3):32701.

[Schwarzer et al., 2017] Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N. A., Huber, W., Haering, C. H., Mirny, L., and Spitz, F. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, 551(7678).

[Sekar et al., 2015] Sekar, S., McDonald, J., Cuyugan, L., Aldrich, J., Kurdoglu, A., Adkins, J., Serrano, G., Beach, T. G., Craig, D. W., Valla, J., Reiman, E. M., and Liang, W. S. (2015). Alzheimer's disease is associated with altered expression of genes involved in immune response and mitochondrial processes in astrocytes. *Neurobiol. Aging*, 36(2):583–591.

[Serandour et al., 2013] Serandour, A. A., Brown, G. D., Cohen, J. D., and Carroll, J. S. (2013). Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. *Genome Biology*, 14(12):R147.

[Sexton and Cavalli, 2015] Sexton, T. and Cavalli, G. (2015). The role of chromosome domains in shaping the functional genome. *Cell*, 160(6).

[Shannon and Richards, 2018] Shannon, P. and Richards, M. (2018). *MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs*. R package version 1.24.1.

[Sherwood et al., 2014] Sherwood, R. I., Hashimoto, T., O'Donnell, C. W., Lewis, S., Barkal, A. A., van Hoff, J. P., Karun, V., Jaakkola, T., and Gifford, D. K. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.*, 32(2):171–178.

[Shogren-Knaak et al., 2006] Shogren-Knaak, M., Ishii, H., Sun, J. M., Pazin, M. J., Davie, J. R., and Peterson, C. L. (2006). Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science (80-. ).*, 311(5762):844–847.

[Shokri et al., 2019] Shokri, L., Inukai, S., Hafner, A., Basler, K., Deplancke, B., and Correspondence, M. L. B. (2019). A Comprehensive Drosophila melanogaster Transcription Factor Interactome. *Cell Reports*.

[Simicevic et al., 2013] Simicevic, J., Schmid, A. W., Gilardoni, P. A., Zoller, B., Raghav, S. K., Krier, I., Gubelmann, C., Lisacek, F., Naef, F., Moniatte, M., and Deplancke, B. (2013). Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics. *Nature Methods*, 10:570–576.

[Sims et al., 2014] Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: Key considerations in genomic analyses. *Nat. Rev. Genet.*, 15(2):121–132.

[Skalska et al., 2015] Skalska, L., Stojnic, R., Li, J., Fischer, B., Cerda-Moya, G., Sakai, H., Tajbakhsh, S., Russell, S., Adryan, B., and Bray, S. J. (2015). Chromatin signatures at Notch-regulated enhancers reveal large-scale changes in H3K56ac upon activation. *EMBO J.*, 34(14):1889–904.

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

[Slattery et al., 2014] Slattery, M., Ma, L., Spokony, R. F., Arthur, R. K., Kheradpour, P., Kundaje, A., Nègre, N., Crofts, A., Ptashkin, R., Zieba, J., Ostapenko, A., Suchy, S., Victorsen, A., Jameel, N., Grundstad, A. J., Gao, W., Moran, J. R., Rehm, E. J., Grossman, R. L., Kellis, M., and White, K. P. (2014). Diverse patterns of genomic targeting by transcriptional regulators in Drosophila melanogaster. *Genome Res.*, 24(7):1224–35.

[Smith et al., 2009] Smith, S. T., Wickramasinghe, P., Olson, A., Loukinov, D., Lin, L., Deng, J., Xiong, Y., Rux, J., Sachidanandam, R., Sun, H., Lobanenkov, V., and Zhou, J. (2009). Genome wide ChIP-chip analyses reveal important roles for CTCF in Drosophila genome organization. *Dev. Biol.*, 328(2):518–528.

[Sokolik et al., 2015] Sokolik, C., Liu, Y., Bauer, D., McPherson, J., Broeker, M., Heimberg, G., Qi, L. S., Sivak, D. A., and Thomson, M. (2015). Transcription Factor Competition Allows Embryonic Stem Cells to Distinguish Authentic Signals from

Noise. *Cell Syst.*, 1(2):117–129.

[Solomon et al., 1988] Solomon, M. J., Larsen, P. L., and Varshavsky, A. (1988). Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. *Cell*, 53(6):937–947.

[Song and Crawford, 2010] Song, L. and Crawford, G. E. (2010). DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.*, 5(2).

[Soshnev et al., 2013] Soshnev, A. A., Baxley, R. M., Manak, J. R., Tan, K., and Geyer, P. K. (2013). The insulator protein Suppressor of Hairy-wing is an essential transcriptional repressor in the Drosophila ovary. *Development*, 140(17):3613–3623.

[Soufi et al., 2015] Soufi, A., Garcia, M., Jaroszewicz, A., Osman, N., Pellegrini, M., and Zaret, K. (2015). Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming. *Cell*, 161(3):555–568.

[Spitz and Furlong, 2012] Spitz, F. and Furlong, E. E. M. (2012). Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, 13(9):613–626.

[Spivakov, 2014] Spivakov, M. (2014). Spurious transcription factor binding: Non-functional or genetically redundant? *BioEssays*, 36(8):798–806.

[Stadhouders et al., 2019] Stadhouders, R., Filion, G. J., and Graf, T. (2019). Transcription factors and 3D genome conformation in cell-fate decisions. *Nature*, 569(7756):345–354.

[Stormo, 1982] Stormo, G. (1982). Use of the'Perceptron'Algorithm to Distinguish Translational Initiation Sites In E. Coli. *Nucleic Acids ...*.

[Stormo, 2000] Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23.

[Stormo and Zhao, 2010] Stormo, G. D. and Zhao, Y. (2010). Determining the specificity of proteinâĂŞDNA interactions. *Nat. Rev. Genet.*, 11(11):751.

[Tang et al., 2015] Tang, Z., Luo, O. J., Li, X., Zheng, M., Zhu, J. J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S. Z., Penrad-Mobayed, M., Sachs, L. M., Ruan, X., Wei, C.-L., Liu, E. T., Wilczynski, G. M., Plewczynski, D., Li, G., and Ruan, Y. (2015). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell*, 163(7):1611–27.

[Tareen and Kinney, 2019] Tareen, A. and Kinney, J. B. (2019). Biophysical models of cis-regulation as interpretable neural networks. *bioRxiv*, page 835942.

[Teif et al., 2014] Teif, V. B., Beshnova, D. A., Vainshtein, Y., Marth, C., Mallm, J. P., and Rippe, T. H. (2014). Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. *Genome Research*, 24(8):1285–1295.

[Teytelman et al., 2013] 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. *PNAS*, 110(46):18602âĂŽÃĎÃň18607.

[Thomas et al., 2016] Thomas, R., Thomas, S., Holloway, A. K., and Pollard, K. S. (2016). Features that define the best ChIP-seq peak calling algorithms. *Brief. Bioinform.*, page bbw035.

[Tillo et al., 2010] Tillo, D., Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Field, Y., Lieb, J. D., Widom, J., Segal, E., and Hughes, T. R. (2010). High Nucleosome Occupancy Is Encoded at Human Regulatory Sequences. *PLoS One*, 5(2):e9129.

[Tompa et al., 2005] Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., Van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, 23(1):137–144.

[Tsodikov et al., 2001] Tsodikov, O. V., Holbrook, J. A., Shkel, I. A., and Record, M. T. (2001). Analytic binding isotherms describing competitive interactions of a protein ligand with specific and nonspecific sites on the same DNA oligomer. *Biophysical Journal*, 81(4):1960–1969.

[Urieli-Shoval et al., 1982] Urieli-Shoval, S., Gruenbaum, Y., Sedat, J., and Razin, A. (1982). The absence of detectable methylated bases in Drosophila melanogaster DNA. *FEBS Lett.*, 146(1):148–152.

[van Bemmel et al., 2010] van Bemmel, J. G., Pagie, L., Braunschweig, U., Brugman, W., Meuleman, W., Kerkhoven, R. M., and van Steensel, B. (2010). The insulator protein SU(HW) fine-tunes nuclear lamina interactions of the Drosophila genome. *PLoS One*, 5(11):e15013.

[Van Bortle et al., 2014] Van Bortle, K., Nichols, M. H., Li, L., Ong, C.-T., Takenaka, N., Qin, Z. S., and Corces, V. G. (2014). Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biology*, 15(5):R82.

[Van Bortle et al., 2012] Van Bortle, K., Ramos, E., Takenaka, N., Yang, J., Wahi, J. E., Corces, V. G., Bortle, K. V., Ramos, E., Takenaka, N., Yang, J., Wahi, J. E., and Corces, V. G. (2012). Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. *Genome Res.*, 22(11):2176–2187.

[van Steensel and Belmont, 2017] van Steensel, B. and Belmont, A. S. (2017). Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell*, 169(5):780–791.

[van Steensel and Furlong, 2019] van Steensel, B. and Furlong, E. E. (2019). The role of transcription in shaping the spatial organization of the genome. *Nat. Rev. Mol. Cell Biol.*, 20(6).

[VietriÂăRudan et al., 2015] VietriÂăRudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D., Tanay, A., and Hadjur, S. (2015). Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Rep.*, 10(8):1297–1309.

[Vogelmann et al., 2014] Vogelmann, J., Le Gall, A., Dejardin, S., Allemand, F., Gamot, A., Labesse, G., Cuvier, O., Nègre, N., Cohen-Gonsaud, M., Margeat, E., and Nöllmann, M. (2014). Chromatin Insulator Factors Involved in Long-Range DNA Interactions and Their Role in the Folding of the Drosophila Genome. *PLoS Genet.*, 10(8).

[Vorobyeva et al., 2013] Vorobyeva, N. E., Mazina, M. U., Golovnin, A. K., Kopytova, D. V., Gurskiy, D. Y., Nabirochkina, E. N., Georgieva, S. G., Georgiev, P. G., and Krasnov, A. N. (2013). Insulator protein Su(Hw) recruits SAGA and Brahma complexes and constitutes part of Origin Recognition Complex-binding sites in the Drosophila genome. *Nucleic Acids Res.*, 41(11):5717–5730.

[Voss et al., 2011] Voss, T. C., Schiltz, R. L., Sung, M.-H., Yen, P. M., Stamatoyannopoulos, J. A., Biddie, S. C., Johnson, T. A., Miranda, T. B., John, S., Hager, G. L., and et Al. (2011). Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism. *Cell*, 146(4):544–54.

[Wang et al., 2015] Wang, H., Zang, C., Liu, X. S., and Aster, J. C. (2015). The role of Notch receptors in transcriptional regulation. *J. Cell. Physiol.*, 230(5):982–8.

[Wang et al., 2014] Wang, H., Zang, C., Taing, L., Arnett, K. L., Wong, Y. J., Pear, W. S., Blacklow, S. C., Liu, X. S., and Aster, J. C. (2014). NOTCH1-RBPJ complexes drive target gene expression through dynamic interactions with superenhancers. *Proc.*

*Natl. Acad. Sci. U. S. A.*, 111(2):705–10.

[Wang et al., 2009] Wang, Y., Guo, L., Golding, I., Cox, E. C., and Ong, N. (2009). Quantitative Transcription Factor Binding Kinetics at the Single-Molecule Level. *Biophys. J.*, 96(2):609–620.

[Weinmaster et al., 1992] Weinmaster, G., Roberts, V. J., Lemke, G., Yan, Q., Sassoon, D., Kitajewski, J., Okazaki, S., Kawaichi, M., Shiota, K., Mak, T. W., and Honjo, T. (1992). Notch2: a second mammalian Notch gene. *Development*, 116(4):931–41.

[Whitfield et al., 2002] Whitfield, M. L., Sherlock, G., Saldanha, A. J., Murray, J. I., Ball, C. A., Alexander, K. E., Matese, J. C., Perou, C. M., Hurt, M. M., Brown, P. O., and Botstein, D. (2002). Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors. *Mol. Biol. Cell*, 13(6):1977–2000.

[Won et al., 2010] Won, K.-J., Ren, B., and Wang, W. (2010). Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biol.*, 11(1):R7.

[Wreczycka et al., 2017] Wreczycka, K., Franke, V., Uyar, B., Wurmus, R., and Akalin, A. (2017). HOT or not: Examining the basis of high-occupancy target regions. *bioRxiv*, page 107680.

[Wu et al., 2019] Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzebski, S., Fevry, T., Katsnelson, J., Kim, E., Wolfson, S., Parikh, U., Gaddam, S., Lin, L. L. Y., Ho, K., Weinstein, J. D., Reig, B., Gao, Y., Pysarenko, H. T. K., Lewin, A., Lee, J., Airola, K., Mema, E., Chung, S., Hwang, E., Samreen, N., Kim, S. G., Heacock, L., Moy, L., Cho, K., and Geras, K. J. (2019). Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Trans. Med. Imaging*, pages 1–1.

[Xin and Rohs, 2018] Xin, B. and Rohs, R. (2018). Relationship between histone modifications and transcription factor binding is protein family specific. *Genome Res.*, page gr.220079.116.

[Yashiro-Ohtani et al., 2014] Yashiro-Ohtani, Y., Wang, H., Zang, C., Arnett, K. L., Bailis, W., Ho, Y., Knoechel, B., Lanauze, C., Louis, L., Forsyth, K. S., Chen, S., Chung, Y., Schug, J., Blobel, G. A., Liebhaber, S. A., Bernstein, B. E., Blacklow, S. C., Liu, X. S., Aster, J. C., and Pear, W. S. (2014). Long-range enhancer activity determines Myc sensitivity to Notch inhibitors in T cell leukemia. *Proc. Natl. Acad. Sci. U. S. A.*, 111(46):E4946–53.

[Yuan et al., 2016] Yuan, Z., Praxenthaler, H., Tabaja, N., Torella, R., Preiss, A., Maier, D., and Kovall, R. A. (2016). Structure and Function of the Su(H)-Hairless Repressor Complex, the Major Antagonist of Notch Signaling in Drosophila melanogaster. *PLOS Biol.*, 14(7):e1002509.

[Zabet and Adryan, 2012] Zabet, N. R. and Adryan, B. (2012). A comprehensive computational model of facilitated diffusion in prokaryotes. *Bioinformatics*, 28(11):1517–1524.

[Zabet and Adryan, 2015] Zabet, N. R. and Adryan, B. (2015). Estimating binding properties of transcription factors from genome-wide binding profiles. *Nucleic Acids Research*, 43(1):84–94.

[Zabet et al., 2013] Zabet, N. R., Foy, R., and Adryan, B. (2013). The Influence of Transcription Factor Competition on the Relationship between Occupancy and Affinity. *PLoS One*, 8(9).

[Zaret and Carroll, 2011] Zaret, K. S. and Carroll, J. S. (2011). Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.*, 25(21):2227–41.

[Zeng et al., 2010] Zeng, W., Ball, A. R., and Yokomori, K. (2010). HP1: Heterochromatin binding proteins working the genome. *Epigenetics*, 5(4):287–292.

[Zhan et al., 2017] Zhan, Y., Mariani, L., Barozzi, I., Schulz, E. G., Bl??thgen, N., Stadler, M., Tiana, G., and Giorgetti, L. (2017). Reciprocal insulation analysis of Hi-C data shows that TADs represent a functionally but not structurally privileged scale in the hierarchical folding of chromosomes. *Genome Res.*, 27(3).

[Zhang and Marr, 1993] Zhang, M. Q. and Marr, T. G. (1993). A weight array method for splicing signal analysis. *Comput. Appl. Biosci.*, 9(5):499–509.

[Zhang et al., 2005] Zhang, X., Odom, D. T., Koo, S.-H., Conkright, M. D., Canettieri, G., Best, J., Chen, H., Jenner, R., Herbolsheimer, E., Jacobsen, E., Kadam, S., Ecker, J. R., Emerson, B., Hogenesch, J. B., Unterman, T., Young, R. A., and Montminy, M. (2005). Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target gene activation in human tissues. *Proc. Natl. Acad. Sci.*, 102(12):4459–4464.

[Zhang et al., 2008] Zhang, Y., Liu, T., Meyer, C., Eeckhoute, J., Johnson, D., Bernstein, B., Nusbaum, C., Myers, R., Brown, M., Li, W., and Liu, X. S. (2008). Model-based Analysis of {ChIP-Seq (MACS)}. *Genome Biology*, 9(9):R137.

[Zhang and Presgraves, 2017] Zhang, Z. and Presgraves, D. C. (2017). Translational compensation of gene copy number alterations by aneuploidy in Drosophila melanogaster. *Nucleic Acids Res.*, 45(6):2986–2993.

[Zhou et al., 2015] Zhou, T., Shen, N., Yang, L., Abe, N., Horton, J., Mann, R. S., Bussemaker, H. J., Gordân, R., and Rohs, R. (2015). Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci.*, 112(15):4654–4659.

[Zhou et al., 2019] Zhou, Y., Gerrard, D. L., Wang, J., Li, T., Yang, Y., Fritz, A. J., Rajendran, M., Fu, X., Schiff, R., Lin, S., Frietze, S., and Jin, V. X. (2019). Temporal dynamic reorganization of 3D chromatin architecture in hormone-induced breast cancer and endocrine resistance. *Nat. Commun. 2019 101*, 10(1):1–14.

[Zhu et al., 2018] Zhu, F., Farnung, L., Kaasinen, E., Sahu, B., Yin, Y., Wei, B., Dodonova, S. O., Nitta, K. R., Morgunova, E., Taipale, M., Cramer, P., and Taipale, J. (2018). The interaction landscape between transcription factors and the nucleosome.

*Nature*, page 1.