Characterisation of human transcription factors binding mechanisms in human cell lines using a statistical thermodynamic framework

Alessandra Pisante

A dissertation submitted for the degree of Master of Science by Dissertation

School of Life Sciences

University of Essex

Submitted: 28 September 2021

1. Abstract

As key genome regulatory elements, transcription factors (TFs) bind to the DNA to control gene expression. High-throughput technologies such as ChIP-seq can experimentally determine TF binding; however, the raw data alone cannot fully answer questions such as whether a TF found in open chromatin is the one responsible for the relaxed state (thus displaying pioneering properties), or it came along and bound an already accessible site.

In this study we wanted to investigate the interplay between TFs and chromatin accessibility. I used ChIPanalyser, a Bioconductor tool, to model and predict ChIP-seq-like profiles in human cell lines, based on publicly available ChIP-seq and DNA accessibility datasets. I estimated the binding parameters of twenty TFs in IMR90 and HepG2 cell lines and used their profiles to evaluate their preference for (or lack of) open and dense chromatin.

Our analysis supports that there are a number of TFs (e.g. CTCF) that display the same properties whether we consider the DNA to be accessible or not, highlighting that some TFs are insensitive to chromatin accessibility, and this holds true for regions with weaker binding sites as well as strong. Our results also suggest there are subsets of TFs that have a preference for open chromatin (e.g. MAZ). Out of the 20 TFs analysed, 4 displayed pioneering functions when looking only at the strong-bound regions and 3 at both strong and medium-bound regions. These results are true when using different accessibility measure methods, like ATAC-seq, DNase-seq, MNase-seq and NOMe-seq, hence the results are not method-specific.

In short, ChIPanalyser can be used to model and predict ChIP-seq data and learn new biological insights, to predict TF binding events between cell lines, or for a screening process to understand a TF's behaviour.

2. Acknowledgements

First and foremost, my deepest gratitude goes to Dr. Nicolae Radu Zabet, who has guided me through the ups and downs of this project, every single day since September 2020. Without his knowhow, support and infinite help, this whole Master's thesis wouldn't have been possible.

I would also like to thank Dr. Antonio Marco, Dr. Francisco de Abreu e Lima and Dr. Jennifer Bromley, respectively my Essex supervisor and my former placement year tutors. You have taught me so much about Bioinformatics to give me the opportunity and the skills to complete a Bioinformatics Master's degree, as well as the encouragement I needed to have in order to know I could have succeeded. Thank you.

I would also like to thank my colleagues and friends from Dr. Zabet's lab whom I shared the past year with: Olivia Grant, Jareth Wolfe, Liudmila Mikheeva and Salma El-Sahhar. We kept each other company in times of stress and their support did not go unnoticed. Thanks to Dr. Patrick Martin who developed the R package ChIPanalyser which was extensively used throughout this project. You have all contributed to this project, one way or another, and for that I am most grateful.

Finally, to my family and all my friends, thank you for supporting me every day, no matter what. I am lucky to be surrounded by all of you. I hope this Master's degree will be the springboard to more success to come, to open up new doors and to welcome me into my future career, whatever that will be.

3. Table of contents

4. List of abbreviations

In alphabetical order:

ADF – Accessibility dependent factor

AIF – Accessibility independent factor

ATAC-seq – Assay for transposase-accessible chromatin sequencing

AUC – Area under the curve

ChIP – Chromatin Immunoprecipitation

ChIP-seq – Chromatin Immunoprecipitation sequencing

DNA – Deoxyribonucleic acid

DNAse-seq – DNAse I hypersensitive sites sequencing

DNMT – DNA methyltransferase

FOXO – Forkhead box

hg38 – human genome assembly 38

IDF – Inaccessibility dependent factors

MACS – Model-Based analysis of ChIP-seq

mCpG – Methylated genomic island

MBD – Methyl CpG binding domain

MNase-seq – Micrococcal nuclease digestion with deep sequencing

MSE – Mean squared error

NCBI – National center for Biotechnology Information

NOMe-seq – Nucleosome occupancy and methylome sequencing

PWM – Position weight matrix

QC – Quality control

QDA – Quantised density accessibility

QDM – Quantile density methylation

Sam – Sequence alignment map

SRF – Serum response factor

TFBS – Transcription factor binding site

TF – Transcription factor

5. Introduction

5.1 An overview of transcription factors

Transcription factors (TFs) are specific regulatory proteins which, by binding to short DNA sequences, regulate the transcription of a target gene either positively or negatively in order to produce the observed effect on transcription. Indeed, the regulation of gene transcription is key to control both an organism's development by directing tissue-specific gene expression, as well as gene activity as a response to stimuli. The short DNA sequences whereby TFs bind are common to genes that exhibit the same pattern of expression. This translates into the fact that genes which transcription is induced as a response to, for example, an environmental stress such as heat or cold stress, will contain a common regulatory element which is absent from genes that do not show this pattern of transcription (Latchman, 1997). Given the pivotal role that these regulatory proteins play within the genome, their regulation is therefore vital. Their role is in fact essential in a wide variety of cellular processes including diseases such as developmental disorders, disorders of the hormone response, and cancer (Latchman, 1997). For example, Forkhead box (FOXO) transcription factor has a primary role in cancer and metastasis. FOXO TF regulates numerous biological processes including development and metabolism. Given its ability to regulate genes essential for cellular processes such as division, death, angiogenesis and metastasis, it has also been increasingly recognised as a tumour suppressor gene (Jiramongkol and Lam, 2020 and Zaret and Carroll, 2011). By clarifying the role and the regulation of FOXO protein in tumour initiation and progression, new therapeutic opportunities for tumourigenesis can be explored.

Indeed, disruptions of gene expression and gene regulation, some of which might be linked to the binding of the TFs in the first place, have been always associated to various diseases. Being able to understand where TFs bind, therefore, means gaining insights into the mechanisms that could lead to diseases such as cancer. Understanding TF binding can also be useful for targeting purposes in order to repress certain genes that, for instance, are involved with the cell growth of cancerous tissues. This targeting exercise could be used to inhibit specific TF:DNA binding as a potential therapeutic approach. Hence many efforts and numerous scientific resources have been – and are being – invested to delve in the complexion of TFs. This link in the chain is useful to understand the bigger picture of how cells work (fig. 1). By unravelling their binding mechanisms to the DNA one could get to the heart of gene regulatory networks (GNRs), where molecular regulators such as DNA, RNA and proteins interact with each other. All the positive and negative feedback interactions of many genes govern gene expression levels within the cell, eventually determining the function of the cell (Cussat-Blanc *et al*., 2019, and Latchman, 1996).

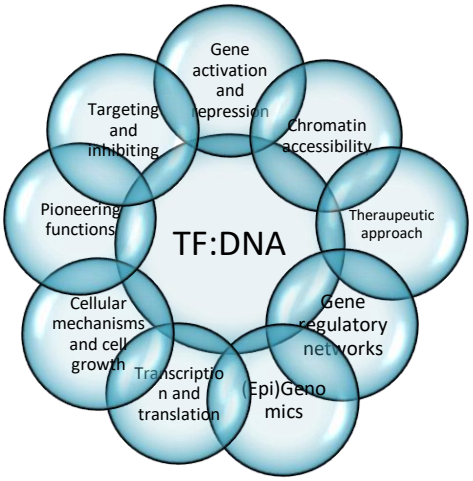

Figure 1: understanding the bigger picture of TF:DNA binding mechanisms. Understanding these mechanisms is pivotal to Genomics.

5.2 Transcription factor classification

Proteins bind to DNA through different types of interactions and they do so in a sequence-specific or non-sequence-specific manner (Wieczór and Czub, 2017). In order to understand the functions that these complexes play in regulating cellular processes, it is essential to classify these proteins based on their binding properties, e.g. the ability of binding to areas of dense chromatin, which would distinguish an accessibility dependent factor (such as a traditional TF) from an accessibility independent one (such as a pioneer TF), or again the ability to make DNA accessible to other proteins, which is the difference between a pioneer factor and a chromatin insulator (Wingender *et al*., 2014). The hypothesis that the distinguishing trait between the different classes of TFs was their potential role in chromatin binding and chromatin opening was investigated in a recent study by Ehsani *et al*. in 2016. They concluded that each subclass of TFs was enriched for properties that seemed to characterise the subclass relative to its role in gene regulation. This means that TF classification is given based on the TF's binding properties that subsequently give rise to a specific gene regulation pattern. Sherwood *et al*. in 2014 identified binding sites for over 700 TFs using a computational method to model the magnitude and shape of genome-wide DNase I hypersensitivity profiles used for identification of TF binding sites; based on the classification they proposed, in this project TFs have been grouped in three categories: AIFs (accessibility independent factors), ADFs (accessibility dependent factors) and IDFs (inaccessibility dependent factors).

1) **AIFs** include pioneers (and their co-factors) and chromatin remodelers (and their co-factors). They are capable of binding to DNA target sites – known as transcription factor binding sites or TFBSs – even in areas of inaccessible regions (heterochromatin, where chromatin is compacted and therefore inaccessible to most TFs). This event increases chromatin accessibility via nucleosome remodelling, therefore facilitating recruitment of other TFs that would otherwise be unable to access dense chromatin (Vanzan *et al*., 2021). They were described as proteins which can bind to compacted chromatin and can promote chromatin modification events required for gene activation by Cirillo *et al*. (2002). Moreover, AIFs have been found to be bound to these areas of dense chromatin even before activation of enhancers, which modulate gene expression (Pennacchio *et al*., 2013). Thus, AIFs can foster subsequent access to classical non-pioneer TFs to their target sites (Soufi *et al*., 2015).

   AIFs, compared to the other TF classes, are the ones that display the strongest and most specific DNA binding (Ehsani *et al*., 2016). The other characteristic that makes AIFs essential to gene expression regulation is their ability to relax the chromatin so as to make heterochromatin become accessible for non-pioneers to bind. AIFs are therefore proteins that can relax closed chromatin (Mayran *et al*., 2019) and can open up a pathway to recruit other factors, which will subsequently trigger gene transcription. An example of a known family of AIFs is the FOXO family (Zaret and Carroll, 2011). Table 1 shows a summary of most of the validated or predicted pioneer TFs in humans in the literature so far. In terms of ChIP-seq signals, AIFs' peaks will only ever be found in open chromatin, whether DNA is dense or not, because if there are no predicted peaks in dense chromatin, it means that the TF was able to open the chromatin wherever it bound.

2) **ADFs** include traditional factors (and their co-factors) and rely on the chromatin to be inheritably opened from a previous state (for example if a genomic region was accessible throughout the entire organism's development), or if the chromatin has been already

opened by an accessibility independent TF, as described in the above paragraph (Vandel *et al*., 2019). This is because they are not able to bind to inaccessible DNA regions, therefore the only way they can work is if an accessibility independent factor has previously made the chromatin accessible (Magnani, 2011). ADFs will therefore only bind to accessible DNA and will not change the chromatin accessibility state post binding. There is also a subclass of ADFs which binds just sporadically to the DNA, even when chromatin at their target sites is fully opened (Sherwood *et al*., 2014). ADFs, therefore, do not exhibit highly-specific DNA binding properties: they are the standard transcriptional machinery and will be there every time there is gene activation. Their ChIP peaks will be found mainly in open chromatin. However, if the DNA was considered dense, they could also bind to dense regions, if they were opened by an AIF.

3) Furthermore, recent studies have hypothesised other genomic elements that play a pivotal role in the establishment, maintenance and regulation of the specific chromatin states; (i.e. the three-dimensional chromatin organization within the nucleus of eukaryotic cells that takes place at the level of the 10 nm fibre, crucial for proper gene expression during development (Phillips and Corces 2009)), are the **IDFs**, which include chromatin insulators and their co-factors (Yang and Corces, 2011). IDFs will preferentially bind to dense chromatin, but they will also bind accessible regions in presence of an AIF, therefore they display high levels of binding in areas of dense chromatin and low to medium levels of binding in areas of open chromatin. The difference is that whilst AIFs will relax dense chromatin, IDFs will reinforce it, by blocking the interactions between promoters and enhancers, other than performing diverse regulatory functions (Song *et al*., 2011). A chromatin insulator consists of a DNA sequence that is capable of mediating intra- and inter-chromosomal interactions, thereby being partly responsible for the 3D folding of chromatin (Phillips-Cremins and Corces, 2013). This is obtained via two main mechanisms of action (Allison, 2012). The first is observed when the insulator is found between the enhancer and the promoter. In this case, enhancer-blocking IDFs work by interfering the enhancer's ability to interact with the gene promoter (West, 2002). The second case consists in barrier insulators stopping the spreading of heterochromatin, thereby preventing gene silencing and gene inactivation. They do so by recruiting histone modifying enzymes (Wang *et al*., 2014), which are genomic elements heavily involved in the control of gene expression, by catalyzing the modification, by addition or removal, of a number of histones and non-histone proteins (Marmorstein and Trievel, 2009).

   IDFs are also involved with diseases: as a consequence of a disruption in the extremely complex network of nuclear organization that chromatin insulators are a part of, which contributes to differential gene expression during cell differentiation, development of cancer or other diseases might be observed (Yang and Corces, 2011). An example of a known IDF is the CCCTC-binding factor, or CTCF (Phillips and Corces, 2009). This TF is involved in numerous regulatory functions, including transcriptional activation and repression. Among these, we find X chromosome inactivation, the chief and unique dosage compensation mechanism whereby, in a highly controlled and coordinated way, female of mammals transcriptionally silence one of their X chromosomes. This is achieved in order to correct the imbalance of having two gene-rich X chromosomes compared to males who carry one gene-poor (Y) and one gene-rich (X) chromosome (Ahn and Lee, 2008). IDFs' preferentially bind dense chromatin however, if the DNA is considered accessible, there could potentially be some binding there as well. IDFs bind chromatin in three different scenarios (fig. 2b): (i)

dense chromatin is bound and maintained; (ii) dense chromatin is bound and reinforced, thus the number of nucleosomes in that genomic region is increased; and lastly (iii) open chromatin is bound and then becomes compacted. Is it difficult to distinguish between the three cases from ChIP data alone since what we observe is that, after binding, the TF is located in dense chromatin and we do not know if the chromatin was already closed and the TF is there to maintain it (first case) or it is responsible for the strengthening or compaction of the chromatin (second and third scenario). An example of a IDF that would behave like example iii) is CTCF: due to its diverse functions, CTCF does present binding regions in open chromatin. However, by binding with target sequences, CTCF can act as an IDF by both blocking interactions between enhancers and promoters as well as preventing heterochromatin expansion, effectively acting as a chromatin barrier (Rojano, Seoane, Ranea and Perkins, 2018).

In conclusion, the three main classes of TFs are AIFs, ADFs and IDFs (fig. 2a), each with their respective binding properties (fig. 2b). Characterisation is given by how differently they bind to chromatin. Based on the TF class, prediction of gene expression patterns can be attempted. Each class presents its own unique features and understanding them could bring us closer to predict gene expression.

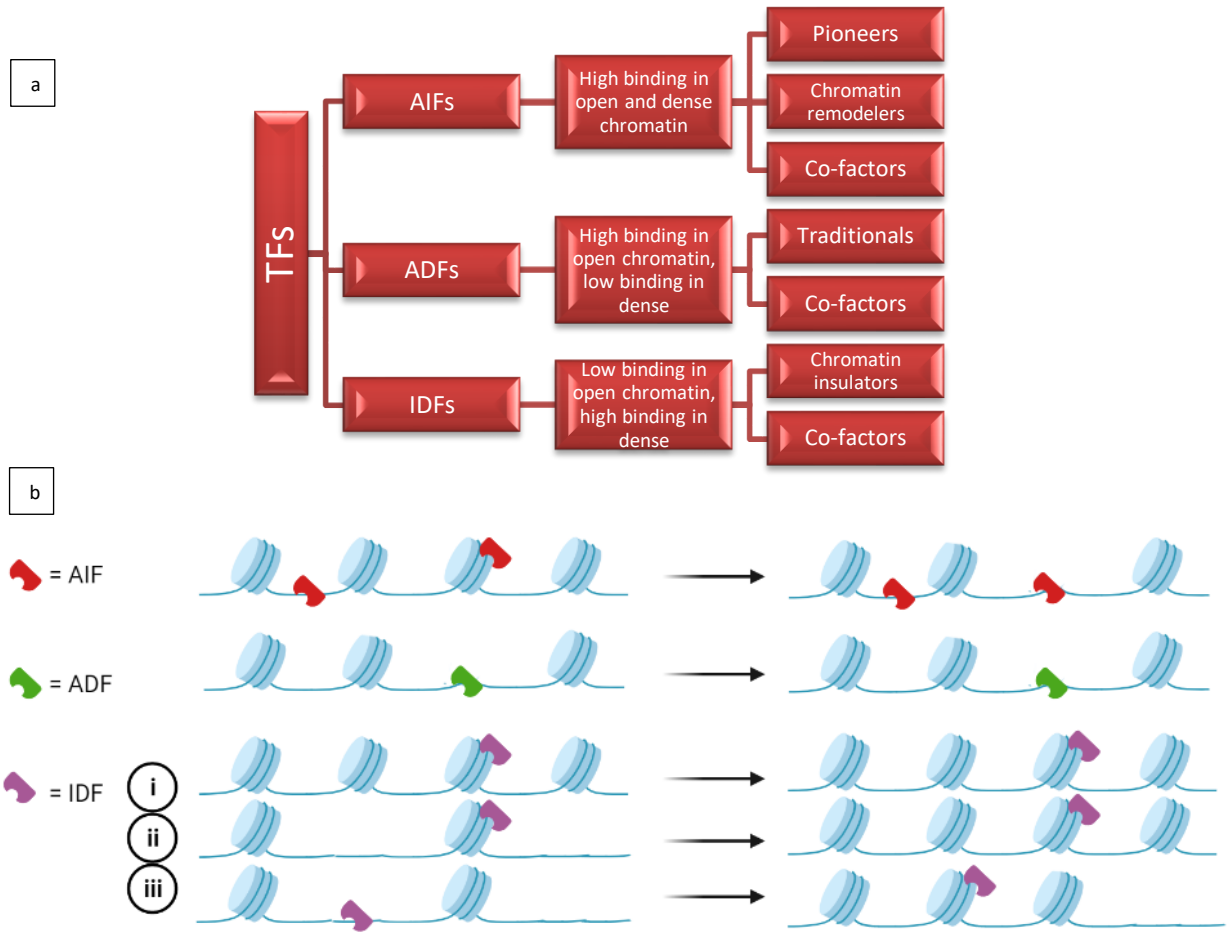

Figure 2: the 3 main classes of TFs are AIFs, ADFs and IDFs. Classification of TFs based on their binding properties. Graphical representation of: (a) what type of TF each class includes and (b) what are the binding properties of each class and how each class affects chromatin after binding. AIFs bind open or dense chromatin without preference, but when nucleosomes are bound, they are displaced. ADFs bind open chromatin only and no changes are observed post binding. Finally, IDFs bind chromatin in three different scenarios: (i) IDF binds dense chromatin and maintains it (no changes are observed); (ii) IDF binds dense chromatin and reinforces it, thus changes are observed in the increased number of nucleosomes; and lastly (iii) IDF binds open chromatin which becomes compacted.

Table 1. Summary of most of the predicted/validated pioneer factors in humans in the literature. The table has been expanded from the one published by Iwafuchi-Doi and Zaret in 2014.

TF name	Predicted/validated pioneering activity	Reference
FoxA	Inducing transdifferentiation	Huang et al. 2011; Sekiya and Suzuki 2011
	Establishment of the competence for liver development	Lee et al. 2005
	Occupying silent liver enhancer in endoderm progenitors	Gualdi et al. 1996
	Recruitment of histone variant H2A.Z	Updike and Mango 2006
	Chromatin decompaction	Fakhouri et al. 2010
	Binding to mitotic chromatin	Caravaca et al. 2013
	Recruitment of hormone receptor	Zaret and Carroll 2011; Jozwik and Carroll 2012
	Binding to mononucleosomes, dinucleosomes, and nucleosome array independent of other factors in vitro	Cirillo and Zaret 1999; Cirillo et al. 2002
	Increasing DNA accessibility in compacted nucleosome array independent of other factors in vitro	Cirillo and Zaret 1999; Cirillo et al. 2002
Oct3/4 Pou5f3	Inducing zygotic genome activation	Foygel et al. 2008; Lee et al. 2013
	Binding to target sites before zygotic genome activation	Leichsenring et al. 2013
	Inducing reprogramming	Takahashi and Yamanaka 2006
	Binding to chromatin that is not DNase-sensitive and not premarked by common histone modifications in vivo	Soufi et al. 2012
Sox2	Inducing zygotic genome activation	Pan and Schultz 2011; Lee et al. 2013
	Inducing reprogramming	Takahashi and Yamanaka 2006
	Binding to chromatin that is nonhypersensitive and not premarked by common histone modifications in vivo	Soufi et al. 2012
	Inducing tumor formation	Bass et al. 2009; Boumahdi et al. 2014
Klf4	Inducing reprogramming	Takahashi and Yamanaka 2006
	Binding to chromatin that is nonhypersensitive and not premarked by common histone modifications in vivo	Soufi et al. 2012
Ascl1	Inducing transdifferentiation	Vierbuchen et al. 2010
	Binding to nonhypersensitive chromatin in vivo	Caiazzo et al. 2011; Son et al. 2011; Wapinski et al. 2013
Pax7	Establishment of the competence for melanotrope development	Budry et al. 2012; Drouin 2014
	Increasing DNA accessibility in native chromatin	Budry et al. 2012; Drouin 2014
PU.1	Inducing transdifferentiation	Feng et al. 2008
	Increasing DNA accessibility in native chromatin	Ghisletti et al. 2010; Heinz et al. 2010; Barozzi et al. 2014
GATA4	Inducing transdifferentiation	Huang et al. 2011
	Occupying silent liver enhancer in endoderm progenitors	Bossard and Zaret 1998
	Recruitment of hormone receptor	Zaret and Carroll 2011; Jozwik and Carroll 2012
	Binding to nucleosome array independent of other factors in vitro (not as effective as FoxA)	Cirillo et al. 2002
	Increasing DNA accessibility in compacted nucleosome array independent of other factors in vitro (not as effective as FoxA	Cirillo et al. 2002
GATA1	Binding to mitotic chromatin	Kadauke et al. 2012
CLOCK:BMAL1	Increasing DNA accessibility in native chromatin	Menet et al. 2014
P53	Increasing DNA accessibility in native chromatin	Laptenko et al. 2011
	Binding to chromatin that encodes high intrinsic nucleosome occupancy	Lidor Nili et al. 2010
	Binding to nucleosome independent of other factors in vitro	Espinosa and Emerson 2001
	Recruitment of histone acetyltransferase p300	Espinosa and Emerson 2001
MyoD	Structure organizer of 3D genome architecture in muscle cells	Chen et al. 2020
Pbx1	Skeletal muscle lineage specification	Berkes et al., Maves et al.; Magnani et al.; Thiaville et al. Grebbin et al.
	Breast cancer (regulation of the estrogen response)	Berkes et al., Maves et al.; Magnani et al.; Thiaville et al. Grebbin et al.
	Adult neurogenesis and neuronal lineage specification	Berkes et al., Maves et al.; Magnani et al.; Thiaville et al. Grebbin et al.
Gro/TLE/Grg	Meditate embryonic segmentation, dorsal-ventral patterning, neurogenesis, and Notch and Wnt signaling	Sekiya and Zaret, 2007
AP-1	Nucleosome displacement at enhancers in murine mammary epithelial cells	Atak et al., 2021
CREB1	Induces transcription of genes in response to hormonal stimulation of the cAMP pathway	Sherwood et al., 2014

5.3 Tools and models to investigate this relationship: experimental and computational methods are complementary

To this day, the most used **experimental** technique that has been developed to study protein:DNA interactions is chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Landt *et al*., 2012). ChIP-seq techniques use Next Generation Sequencing to map proteins' binding sites within the genome (Barski and Zhao, 2009), and are used to assess whether TF-specific binding sites are occupied (Héberlé and Bardet, 2019). When a genomic area is identified to be enriched in the number of aligned reads for a specific TF, those enriched regions represent the likely locations of where that TF interacts with the DNA. By overlapping the ChIP-seq peaks with genomic regions of different DNA accessibility (i.e. open VS dense chromatin, methylated VS unmethylated DNA), one could pinpoint where those TFs are bound in the genome and therefore classify that TF based on its binding properties. For instance, if a TF has bound to inaccessible DNA, it could be a potential IDF, as the steady state of the cell showed that the chromatin was kept dense post TF binding. If it is the other way around and the TF is found in open chromatin, one would not know if the TF had opened the chromatin (like an AIF would do) or had just bound there because the chromatin was already accessible (like an ADF, as its binding is restricted to certain areas of open DNA). Experimentally, potential pioneers can be validated by TF overexpression analysis to observe gain in accessibility in areas of increased binding (Iwafuchi-Doi and Zaret, 2014). These experimental techniques therefore allow to come to conclusions to certain extents; however they also present limitations such as difficulties to adapt for high-throughput screening (Park, 2009). For example, just by looking at the overlaps from the raw data (e.g. intersecting ChIP-seq with ATAC-seq to know where peaks are found within the genome) one would not necessarily understand the relationship and the information on the behaviour of that TF. Hence why the need to use **computational** methods too.

Areas of enrichment can also be identified computationally using methods such as peak calling (Taleb *et al*., 2019), which rely on algorithms capable of defining sites of protein:DNA binding. For my project I will be using a Bioconductor package called ChIPanalyser which algorithm is based on a statistical thermodynamic framework. The field of statistical thermodynamics is the study of how one can organize all the different combinations of particles in a system. ChIPanalyser, available on Bioconductor as an R package, models and predicts the binding of TFs to the DNA (Zabet and Adryan, 2015; Gentleman *et al*., 2015; Martin and Zabet, 2020). The question that ChIPanalyser aims to answer is "what describes the probability of a specific DNA region to be bound by a TF?". ChIPanalyser will output ChIP-seq-like profiles based on:

(i) a Position Weight Matrix (PWM) which is a representation of patterns in a biological sequence, called motif, which TFs will preferentially bind to. This informs the model how likely the protein is to bind at a specific genomic sequence. A PWM will indicate, for each base pair, how likely this motif is going to show a binding event. One issue associated to models which solely rely on PWMs as markers for TF:DNA binding is that there are more motifs that fit the high probability of having a binding event than actual binding events. Therefore, it is necessary to understand what other factors contribute to TF binding;

(ii) the number of molecules bound to the DNA indicated by the letter N. It states how many TFs are bound to the DNA, on average, at any time point. TFs can also unbind; ChIPanalyser does not look at the dynamics of N to the DNA, therefore N indicates the steady-state population average;

(iii) a scaling factor, which is a modulator for the binding specificity indicated by the symbol λ. This controls how well a TF can discriminate between high and low affinity binding sites. High values for λ translate into poorer ability for the TFs to discriminate between high and low affinity sites (Martin and Zabet, 2020);

(iv) DNA-accessibility of that specific site, which indicates whether chromatin is compacted or not.

Equal weight of importance is given to these four parameters. The established relationship between these four parameters allows to come to conclusions regarding TF binding patterns: if the prediction is accurate, those parameters are sufficient to explain the relationship between that TF and DNA. On the other hand, if the model fails to predict TF:DNA binding mechanisms, that is a sign that there is something else that needs to be taken into account, and those parameters are not sufficient to explain the binding event. ChIPanalyser therefore, models TF binding based on real ChIP-seq data and provides a predictive tool to extract features of the TF binding mechanisms, eventually allowing to discriminate between TFs that show different binding properties and therefore to shortlist potential AIFs. With partial experimental data ChIPanalyser can show where to expect TF binding events and allows to highlight what the main force that drives binding is. For example, is TF binding dependent upon whether chromatin is accessible, or does it depend on the number of TFs bound to that stretch of DNA? To summarise, when one overlaps ChIP peaks with DNA accessibility for both accessibility dependent and independent factors, it will be clear that they both will bind to accessible DNA, but while classical TFs bind only in open chromatin, accessibility independent TFs bind in dense chromatin and open it (Lamparter *et al*., 2017). There are two possibilities to distinguish between the two: one experimental and one computational. Both wet and dry lab approaches allow to come to these conclusions: as a matter of fact, computational and experimental designs are complementary (Ding *et al*., 2010).

5.5 Chromatin accessibility

Computational and experimental models which determine the interactions between DNA and TFs require incorporation of DNA accessibility measures. There are several methods that can be used to map DNA accessibility, including ATAC-seq, DNAse-seq, MNase-seq and NOMe-seq. **ATAC-seq** (Assay for Transposase Accessible Chromatin using Sequencing) works by probing DNA transposases to DNA sequences at a genome-wide level. The transposases will insert themselves into open chromatin rather than dense chromatin, which translates into higher gene expression because more TFs can bind to the DNA sites. The outputted reads reflect genomic accessible regions that correspond to lack of nucleosomes, therefore map the chromatin accessibility landscape at a genome-wide level (Sun *et al*., 2019).

Identification of active gene regulatory elements can also be assayed via **DNase-seq**, which stands for DNase I hypersensitive sites sequencing (Song and Crawford, 2010). DNase-seq is a high-resolution technique to map active gene regulatory elements genome-wide, and is therefore a useful tool to identify promoters, enhancers and other regulatory regions. DNase-seq uses the restriction enzyme DNase I to selectively digest accessible DNA regions. Instead, the regions where chromatin is inaccessible are not digested. By capturing and sequencing in high throughput the DNase-digested fragments across the whole genome, DNase-seq allows to identify the most active regulatory regions within the genome (Song and Crawford, 2010).

MNase-seq, short for Micrococcal Nuclease digestion with deep sequencing, is another technique used to assay chromatin accessibility. MNase-seq is used as probe of DNA sequence organization and chromatin structure (Keene, 1981). When a region of DNA is occupied by histones or other proteins such as TFs, MNase, which is an enzyme that digests DNA in regions that are not stably bound by proteins (Cuatrecasas *et al*.,1967), is unable to bind to the DNA. The sequenced fragments output nucleosome location information, therefore revealing structural information about chromatin. In this case absence of nucleosomes denotes accessible regions, whilst presence of nucleosomes indicates dense chromatin. MNase-seq was used by Teif *et al*. in 2012 to map nucleosome positions by genome-wide sequencing of nucleosomal DNA from mouse embryonic stem cells, neural progenitor cells and mouse embryonic fibroblasts after digesting the linker DNA between nucleosomes with MNase. They observed that pluripotency master regulators such as Sox2, Oct4 and Nanog as well as chromatin remodelers Chd7 and Brg1 preferentially bound to DNA regions that coincided with well-positioned nucleosomes. They concluded that these factors had pioneering properties as they could efficiently bind while the DNA target site interacted with a histone octamer (Teif *et al*., 2012). According to these findings, from the MNase-seq data it can be concluded that a factor displays pioneering properties as its peaks coincide with nucleosomes. However, this approach would not allow to distinguish between TFs that bind nucleosomes and displace them, like AIFs, and TFs that bind nucleosomes and maintain or even reinforce the closed chromatin state, like IDFs.

The last technique to assay chromatin accessibility that I will discuss is **NOMe-seq**. Short for Nucleosome Occupancy and Methylome sequencing, NOMe-seq measures the relationship between DNA methylation and nucleosome occupancy and generates a map of these two signals showing how these epigenetic components correlate with each other (Lay *et al*., 2017). The genome is engineered to introduce a GpC methyltransferase enzyme that artificially methylates GpC dinucleotides that are not protected by nucleosomes or protein binding sites, i.e. everything that is accessible. In order to measure the increase in accessibility, the methylation level is measured, and it is found that everything that gained methylation was accessible for the methyltransferase enzyme to add methylation (NOMe-Seq - Nucleosome Occupancy and Methylome Sequencing, 2021).

5.6 Why using ChIPanalyser

What ChIPanalyser adds to the equation is that once the model has been trained with different scenarios (e.g. varying the DNA accessibility percentage), one can better dissect what happens to proteins:DNA interaction. Instead, by just studying the overlaps of the experimental data, we could be missing out additional insights that led to that final result (perhaps the presence of a co-factor of a pioneer TF that previously made chromatin accessible). Metaphorically, this is like the “chicken and the egg” problem: if one observes a TF binding open chromatin, it will not be possible to know whether it was that TF’s presence that had opened the chromatin, or if that genomic area was already accessible for TFs to bind – which is why the presence of co-factors is not to exclude. By calculating TF binding based on a biophysical model, one can filter out potential AIFs. Using ChIPanalyser is therefore *complementary* to running ChIP-seq experiments, and will still answer questions such as “is that specific TF bound in that specific region (e.g. open chromatin), because it is restricted to binding to certain regions only, or could that same TF bind to other areas as well (e.g. dense chromatin), maybe because DNA accessibility does not affect its binding?”. In that case, a TF might be found in open chromatin according to experimental data following ChIP-seq, but that chromatin, which is now accessible, might have been opened by a cofactor or a different TF which could be the real pioneer TF. This means that that specific TF is a traditional one, which is normally

affected by DNA accessibility levels, whilst the TF that had previously made chromatin open is an accessibility independent factor. At the same time, one cannot just look at the binding motifs and take for granted that *motif* equals *binding event*, because there are many more motifs than there are actual binding events. For example, some sites are concentration-dependent and, simply, there might not be enough of a TF to bind every single motif.

In conclusion, in order to run ChIPanalyser one still needs experimental data for ChIP-seq, as well as some measure of DNA accessibility like ATAC-seq or DNase-seq (even DNA methylation can be used, when taken in the context of its correlation with closed chromatin in many cases). However, experimentally one cannot easily tweak these parameters. For instance, binding in accessible chromatin does not mean that the TF cannot bind dense chromatin too, and this can be explored with ChIPanalyser by varying the percentage of accessible genome and studying the binding pattern. In two cases there will not be peaks in dense chromatin: either the TF has pioneering properties, which means whatever site wes there has been opened, or there are binding sites, but the TF does not have pioneering properties, and the binding sites are masked by dense chromatin. Testing different hypothesis allows dissecting the binding properties of different TFs and classifying them based on their properties.

5.7 Aims of project

Previous analyses within Dr. Zabet's laboratory included *D. melanogaster*, *M. Musculus* and *H. sapiens* genomes studies. Several TFs from human and mouse cell lines have already been characterised, thereby TF classification into different groups based on their chromatin accessibility preference has been possible. As a result of these analyses, 48 potential AIFs, including known pioneers such as CTCF and MYC, have been identified. These results suggest that there might be more pioneers and co-factors of pioneers than it had been previously hypothesised, in human and mouse data. I will be using ChIPanalyser on human ChIP-seq data hoping to further classify TF binding properties using different metrics to assess DNA accessibility. It will be interesting to not only shed light on newly discovered potential AIFs, IDFs and other TF classes, but also to compare those results with similar mouse genome-wide studies. The overarching aim of the project will therefore be to have a comprehensive view of TF binding models in mammalian systems and to shortlist potential AIFs in mammalian cell lines. The main question that this project aims to tackle is: can ChIPanalyser, in conjunction with ChIP-seq data and several methods to map DNA accessibility, allow us to identify whether TFs display pioneering functions? And therefore, are there different mechanisms for the binding of TFs? I.e. do all classes bind similarly to the DNA, regardless of their ability to bind to euchromatin vs heterochromatin? Fig. 3 shows ChIPanalyser's workflow in a nutshell.

The novel aspects compared to previous analyses are:

(i) comparing the impact of different DNA accessibility measures to see if results are consistent. These include DNase, ATAC-seq, MNase-seq and NOMe-seq (Meyer and Liu, 2014);

(ii) instead of scaling the data within known genomic regions (e.g. 10 to 60, i.e. the stronger-bound regions), it could be interesting to check how scalable this is by taking 50 medium-bound regions (e.g. 500 to 550) and 50 weak-bound regions (regions that have lower levels of binding, e.g. 1000 to 1050) to understand how well the weaker-bound regions perform. The training set is made of the regions at the top where the ratio between true

positive and true negative signals is balanced. Validating on the training set would yield a biased optimal model fit as those regions were used to teach the model what to predict. In fact, it is known that biophysical models perform better for stronger bound regions (Kaplan *et al*., 2011), but the regions displaying weaker binding will be used to validate the model which will have been previously trained on the top 10 regions. Checking validation on the weaker-bound regions too is important because by doing so one can understand whether a TF can still bind to DNA without taking into account DNA accessibility; this is so that we can also understand the Biology behind TF binding – does a TF display pioneering properties only at strong sites, or is it capable to open chromatin at weaker-bound ones too? – other than testing the reliability and the accuracy of the model.

ChIPanalyser cannot be trained on weaker-bound regions as the ChIP-seq data presents so many true negatives (regions that are not bound), that if the model was trained on poor quality data it would predict everything as negative. There would not be enough peaks on the training dataset for the model to learn the real TF's binding patterns, thus the model would behave as if predicting no peaks would always be accurate. Since this study is conducted genome wide, selecting the top 10 regions on a chromosome translates into the fact that, when taking the whole genome into account, the top 10 regions of each chromosome will be within the top 50 or 60 regions genome wide (i.e. of all chromosomes).

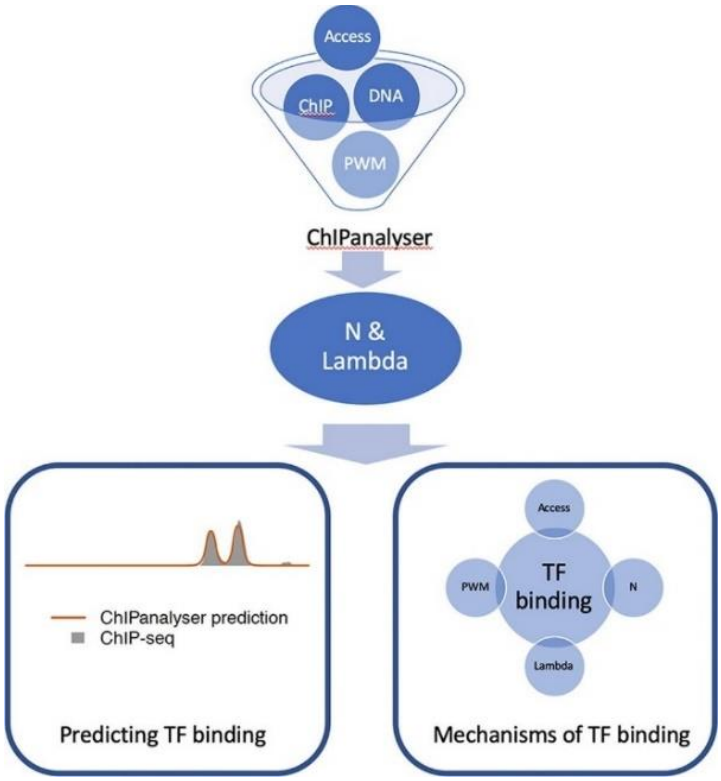

Figure 3: ChIPanalyser workflow in a nutshell. The input data include DNA accessibility and ChIP-seq data; the output data are the optimal parameters of binding from which the TF:DNA predicted binding profiles can be plotted (left), and the TF:DNA binding mechanisms can be better dissected (right) (Martin and Zabet, 2020).

6. Methods

6.1 Data pre-processing

The project's pipeline starts with the data pre-processing. Raw human ChIP-seq *fastq* datasets for all the TFs available in the IMR90 cell line 5 validated AIFs in the literature from IMR90 and HepG2 cell lines were downloaded from ENCODE release 3. Raw human chromatin accessibility *fastq* datasets for IMR90 and HepG2 (in the form of DNase-seq, MNase-seq, ATAC-seq and NOMe-seq) were also downloaded from ENCODE release 3 (Sloan *et al*., 2015). Quality assessment and improvement on these *fastq* datasets was run with *fastqc* version 0.11.7 (Andrews, 2010) and *cutadapt* version 1.18 in order to remove ILLUMINA adapters (Martin, 2011). In order to avoid carrying data contaminated from adapters and low-quality reads into downstream analysis, quality checking must be carried out. For this purpose, FastQC was used as a QC tool (Andrews, 2010), and was run before and after trimming with cutadapt. Then, the data was aligned to the hg38 reference genome downloaded from NCBI (ID: 884148 RefSeq) with *bowtie2* version 2.3.4.1 (Langmead and Salzberg, 2012). Peaks were called with *macs2* version 2.1.2 with a q-value threshold of 0.5 (Zhang *et al*., 2008). Fig. 4, below, summarises the main pre-processing steps.

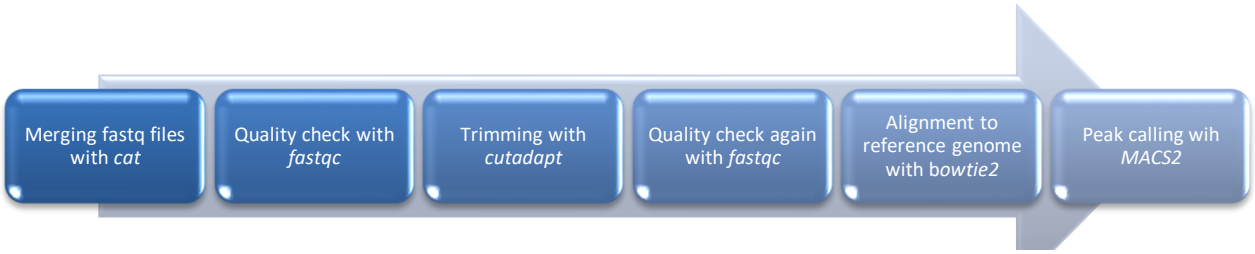

Figure 4: workflow of the data pre-processing of all TFs and all DNA accessibility fastq single-end Illumina FASTQ files. Files were downloaded from ENCODE release 3 and merged using the cat function. FastQC was used before and after cutadapt to check the reads quality. Cutadapt was used to trim and remove adapters. To map the trimmed read sequences to the reference genome – hg38 – Bowtie2 was used. MACS2 was used to call narrow peaks to identify the TF binding sites.

The final number of TFs in the IMR90 and HepG2 cell lines that were analysed is 20. The cell lines and TFs were picked based on data availability on ENCODE. The pool of TFs is small because this analysis has many layers of complexity, including 3 different validation regions and 4 different DNA accessibility assays (see *5.7 Aims of project*). Table 2 includes information on each of the 20 TFs such as whether it does or does not bind specifically to the DNA, and what its primary role is. Fig. 5 shows the DNA sequence logos of the motifs of each of the 20 TFs.

This data pre-processing method applies to both TFs that have a DNA sequence preference as well as other proteins such as the chromatin remodelers (e.g. CHD1) that, instead of binding the DNA specifically, bind nucleosomes, or are recruited by other DNA binding proteins. It is still unclear to what degree chromatin remodelers are swayed by local DNA sequences when shifting a nucleosome to different positions. Chromatin remodelers cluster to form complexes with other proteins: this interaction was reported with other sequence-specific transcriptional regulators, thereby facilitating gene expression (Tyagi, Imam, Verma and Patel, 2016). The PWMs of RCOR1, RAD21, CHD1 and SCM3 (which do not specifically bind the DNA) were constructed *de novo* from ChIP data and, thus, represent a mix of the motifs of the proteins that recruit them, although this does not reflect their binding to the DNA, as it is not specific. In order to check ChIPanalyser's accuracy at recovering the real TF's binding patterns, previously reported AIFs in the literature in the cell lines IMR90 (p53 and MyoD) and HepG2 (FoxA1, GATA4 and CREB1) were used as positive controls for this analysis. The only difference was that the NOMe-seq was not available for HepG2, as it was not possible to download this dataset.

Table 2: The TFs analysed in this project are 20. Most of them specifically bind to the DNA. Raw human TF ChIP-seq datasets were downloaded from ENCODE release 3 for all the TFs available in the IMR90 cell line and some used as controls from IMR90 and HepG2.

TF name	Primary role	Motif repositories	Specific DNA binding
CTCF	Regulating the 3D structure of chromatin	Hsapiens-jaspar2018	Yes
CEBPB	Regulation of genes involved in immune and inflammatory responses	Hsapiens-jaspar2018	Yes
MAFK	Transcriptional repressors when these proteins dimerize among themselves	Hsapiens-jaspar2018	Yes
BHLHE40	Control of circadian rhythm and cell differentiation	Hsapiens-jaspar2018	Yes
MAZ	Dual roles in transcription initiation and termination	Hsapiens-SwissRegulon	Yes
MXI1	Potential tumor suppressor	Hsapiens-jaspar2018	Yes
USF2	Repressors of the human MCT1 (monocarboxylate transporter 1) promoter	Hsapiens-jaspar2018	Yes
NFE2L2	Regulation of metabolism	Hsapiens-jaspar2018	Yes
FOS	Regulating the development of cells destined to form and maintain the skeleton	Hsapiens-jaspar2018	Yes
RFX5	Production of specialized immune proteins called MHC class II proteins	Hsapiens-jaspar2018	Yes
ELK1	Role in neuronal functions	Hsapiens-jaspar2018	Yes
RCOR1	Determining neural cell differentiation	Hsapiens-SwissRegulon	No
RAD21	Sister chromatid cohesion and separation. Subunit of cohesin, therefore does not have sequence preferences	Hsapiens-SwissRegulon	No
CHD1	Alter gene expression possibly by modification of chromatin structure. Chromatin remodeler	Hsapiens-SwissRegulon	No
SMC3	Central component of cohesin, thus does not have sequence preferences	Hsapiens-SwissRegulon	No
P53	Key cell cycle regulator that manages the response to stress signals	Hsapiens-jaspar2018	Yes
MYOD	Regulating muscle differentiation	Hsapiens-jaspar2018	Yes
GATA4	Key role in the development of the heart. It is involved in the induction of cardiac specific gene expression	Hsapiens-HOCOMOCOv10	Yes
CREB1	Plays a role in the regulation of immune responses	Hsapiens-jaspar2018	Yes
FOXA1	Acts as a pioneer factor for ER and AR allowing the nuclear receptors to make targeted contacts with specific genomic regions which induce a transcriptional response	Hsapiens-jaspar2018	Yes

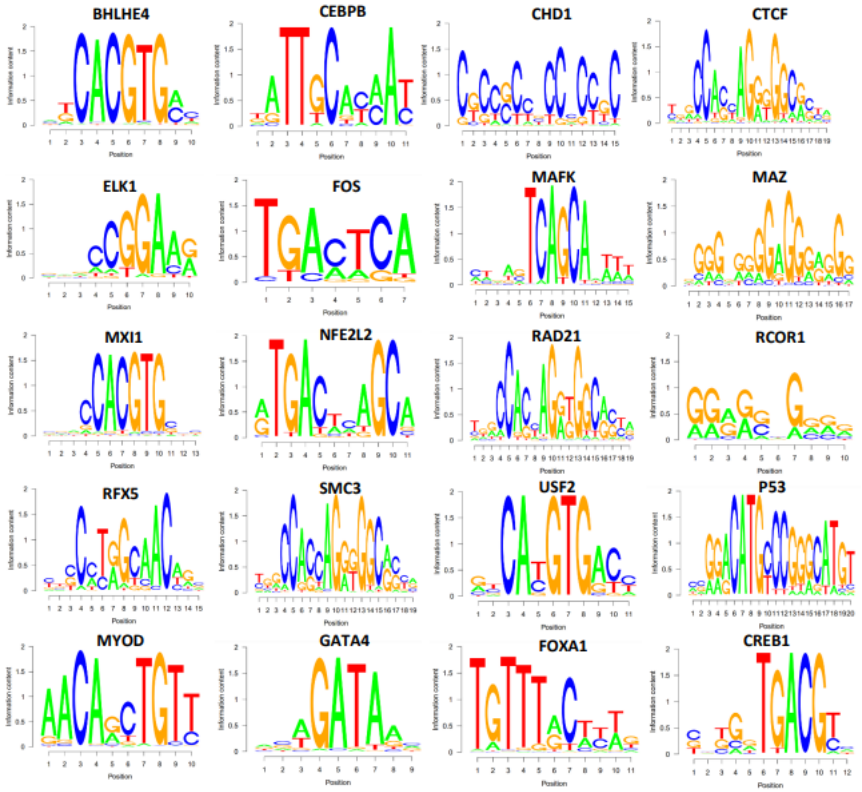

Figure 5: DNA sequence logos of the motifs of the 20 TFs analysed in this project.

Following the pre-processing of the *fastq* datasets of the four DNA accessibility assays in IMR90 and HepG2 which are ATAC-seq, DNase-seq, MNase-seq and NOMe-seq, the DNA accessibility data were subset in R with a vector of quantile density accessibility (QDA) of 12 levels with a window size of 100 base pairs (QDA: from 0 to 0.9, plus 0.95 and 0.99), where level 1 (or QDA=0) corresponds to 0% of chromatin being dense, meaning this QDA file considers all regions of the genome as fully accessible, regardless of their actual accessibility, whilst level 12 (or QDA=0.99) corresponds to 1% of chromatin being accessible (all chromatin is dense, but the top 1% of the genome). MNase-seq datasets are created in a way that the less reads there are, the more accessible the DNA is. Thus, a QDA of 0 correspond to 100% dense DNA instead of 0%. This was considered when creating the QDA files, and the function was adapted in order to be able to directly compare MNase-seq with ATAC-seq, DNase-seq and NOMe-seq.

These 12 different thresholds of DNA accessibility (namely QDAs) were used to run the ChIPanalyser model to create AUC and MSE heatmaps, which respectively indicate: (i) the estimated number of bound molecules according to the model, and (ii) the model scaling factors, from which the optimal values can be selected. The closer the AUC is to 1, the better the model is estimating the ChIP-seq-like profile. The accessibility percentage per each QDA for all four methods was calculated in R. Finally, motifs for each TFs are obtained in R using the MotifDb package (Shannon and Richards, 2020), under the form of PWMs, by running a function that queried various motif databases. All PWMs were downloaded from the database Jaspar2018 (Ge Tan, 2017). If the motif for a TF was not available, several other databases including SwissRegulon and HOCOMOCO v10 were scanned for the motif, however if the motif was not available, the TF would have been removed from the analysis.

The rationale behind using many QDA values to fit the model is to assess how the various TFs perform under different DNA accessibility levels: an AIF would bind to dense chromatin as its binding is not restricted to accessible regions only, whereas an ADF would not be able to bind to dense chromatin as its binding properties stop it from reaching its target site. To assess the quality of the predictions, two measures are used: the area under the curve (AUC) to highlight how well the peaks are predicted and the mean squared error (MSE) to show how different the predicted profile is from the actual one. The AUC indicates the resemblance of the predicted profile compared with the actual ChIP data (the higher, the more accurate it is). The aim is to minimise the mean squared error (MSE) when comparing the prediction with actual ChIP-seq data and therefore find values that give the lowest error possible. If a TF displays a high AUC value even for the bottom QDAs (0.7 – 0.99), that would indicate that the TF can bind dense chromatin, as the prediction accuracy was high even when most of the regions were considered inaccessible by the model (a QDA of 0.99 corresponds to 99% dense regions).

Instead, if the only AUCs to be high are the top ones (0.0 – 0.4), this would suggest that the TF binds open chromatin only, as a low QDA corresponds to mostly accessible chromatin (Martin and Zabet, 2020). Therefore the question is “how do different DNA accessibility levels affect TF binding?” The hypothesis is that, if no changes are observed in the AUC when the percentage of accessible genome is available, it means that that specific TF is insensitive to DNA accessibility, leading to think it has pioneering functions.

6.2 ChIPanalyser experimental design

ChIPanalyser (Martin and Zabet, 2020) is an R package available on Bioconductor that can reproduce ChIP-seq-like profiles based real ChIP-seq data and therefore is able to model and predict TF:DNA binding, by computing the probability that a binding site is occupied, therefore circumventing costly ChIP-seq experiments. ChIPanalyser version 1.12.0 was run several times on the aforementioned TFs: per each DNA accessibility level (1 to 12), per each DNA accessibility measure (ATAC-seq, DNase-seq, MNase-seq, NOMe-seq), per each TF (total 20, see table 2) and per each validation region (strong, medium and weak-bound). The model was trained on the top 10 regions and validated on different sets of regions – 50 strong, 50 weak and 50 medium-bound regions. The AUCs reported in this thesis are those always belonging to the validation datasets, and never the training datasets. The strong and medium-bound validation region ranges are constant, but the weaker-bound regions varied according to how many real ChIP-seq peaks a TF presented. Therefore in order to decide which regions range to use as *weak*, the number of ChIP-seq peaks was divided by 3 and the first 50 regions of the bottom third were used for validation (see table 3). Note that the number of ChIP-seq peaks of ELK1 was only 145, therefore this TF was only validated on the top regions.

Table 4, below, recapitulates what each parameter mean. Different values for λ and N were tested to evaluate how well the predicted ChIP profiles resembled the actual ChIP-seq data. To assess the quality of the predictions, two measures were used: AUC to highlight how well the peaks are predicted and MSE to show how different the predicted profile is from the actual one. The ChIPanalyser experimental design includes extracting ChIP-scores at the loci of interest, training and validation of the data and finally computing the optimal parameters. Hg38 was downloaded from NCBI and used as the reference genome. One example line is found below (see table 4 for a description of each argument).

Table 3: Number of real ChIP-seq peaks for each TF in IMR90 and HepG2 cell lines and their validation regions. The range for strong and medium-bound regions is fixed. The weaker-bound regions are chosen by dividing the total number of peaks by 3 and multiplying by 2. The weak-bound range is then obtained by taking the top 50 regions. E.g.: total number of peaks=3000. 3000/3 x 2 = 2000. Weak-bound validation range: 2000–2050. Sometimes, the first 50 regions of about half the number of peaks of each TF were used instead. The results were consistent. Note that ELK1 did not present a high enough number of ChIP-seq peaks, therefore it was only validated on its top regions.

TFs	*Total ChIP-seq peaks number*	*Top regions range*	*Middle regions range*	*Bottom regions range*
CEBPB	227924	11-60	500-550	18620 – 18670
MAFK	111711	11-60	500-550	18620- 18670
RAD21	87317	11-60	500-550	14550- 14600
BHLHE40	65406	11-60	500-550	10900-10950
FOS	62082	11-60	500-550	10350-10400
SMC3	58914	11-60	500-550	9820-9870
MAZ	57739	11-60	500-550	9630-9680
CTCF	56686	11-60	500-550	9450-9500
RCOR1	55700	11-60	500-550	9290-9340
MXI1	46379	11-60	500-550	7730-7780
USF2	40012	11-60	500-550	6670-6720
NFE2L2	34719	11-60	500-550	5790-5840
CHD1	27160	11-60	500-550	4530-4580
RFX5	5461	11-60	500-550	910-960
ELK1	145	11-60	NA	NA
p53	9968	11-60	500-550	6645-6695
MYOD	5190	11-60	500-550	700-750
CREB1	205936	11-60	500-550	46940-46990
FOXA1	205372	11-60	500-550	46940-46990
GATA4	36300	11-60	500-550	12610-12660

```
qsub -cwd -j y -q all.q \                          #Standard for qsub jobs in Bash
-o ./CTCF_12.txt \                                 #Name of output file
-b y -N CTCF_12 \                                  #Name of job on the cluster
Rscript ./ChIPanalyser_general.R \                 #Name of the R script to execute
CTCF 12 IMR90 1 10 11 60 topRegions DNase          #The 9 fundamental arguments
```

Once the model had been validated on each region and the optimal binding parameters per each specific TF, with that specific DNA accessibility measure and QDA, had been estimated, the next step was to plot. Fig. 6 summarises ChIPanalyser's main steps.

Figure 6: training and validating ChIPanalyser experimental design. The model was trained on the top 10 regions and validated on other sets of regions. The workflow includes use of the previously pre-processed data, as well as the creation of a few objects, such as PWMs, DNA sequence of the reference genome, DNA-accessibility, as well as GRange objects containing ChIP-seq data. The function *processingChIP* was run to extract ChIP-scores at the loci of interest. Training and validation of the data was completed before moving onto computing the optimal parameters with *computeOptimal* function. Finally, heatmaps were plotted, optimal parameters extrapolated, and ChIP-seq-like profiles were drawn using *plotOccupancyProfile* function, armed with values for number of bound molecules (N) and the PWM scaling factor (λ).

Table 4: **Summary of which parameters and arguments were used to execute ChIPanalyser on the 20 TFs in the cell line IMR90 and HepG2, with a brief description of what they are used for**. If not specified, then the parameters were left as default. ChIPanalyser was run in R v3.6.0 as qsub jobs submitted in Bash.

Parameters/Arguments	Value	Description
TF	See table 2 for a complete list	The TF we are studying.
QDA	From 0 to 0.9, plus 0.95 and 0.99	Used to assess how the TF performs under different DNA accessibility thresholds.
Cell line	IMR90 or HepG2	The cell line of interest.
Regions for model training	1-10	Range of genomic regions used to train the model.
Regions for model validation	See table 3 for a complete list	Range of genomic regions used to validate the model.
Validation regions	Either strong, medium or weak-bound	Used to check if a TF displays pioneering properties at stronger-bound sites only or not.
DNA accessibility assay	Dnase-seq, ATAC-seq, Mnase-seq, NOMe-seq	The DNA accessibility method used.
lociWidth	50000	When no loci are provided, ChIPanalyser will split ChIP data into bins of width equals to lociWidth.
Noise filtering method	Sigmoid	Noise filter applied to ChIP data (could be zero, mean, median or sigmoid). Sigmoid applies a logistic weighting to every score, modulating ChIP scores around the 95th quantile point.
stepSize	100	Used to restrict the size of the ChIP-seq-like profile produced by *computeChIPProfile*. Instead of returning ChIP-seq like score for each base pair, this will return the predicted enrichment score for every "n" base pairs.
chipSd	150	Standard deviation of peak width. SD of peak width is used during the smoothing of ChIP data.
chipMean	150	Average ChIP peak width. Peak width is used during the smoothing of ChIP data.
Range of N	From 1 to 1e+08	Possible number of bound molecules (positive numeric value).
Range of λ	Between 0.25 and 5 with an interval of 0.25	Possible lambda values (positive numeric value).
DNA sequence set	hg38	Reference genome.

6.3 Plotting and estimating optimal parameters

To allow estimation of the optimal combination of parameters, thereby allowing selection of λ and N, heatmaps of 12 different metrics including MSE and AUC (Martin and Zabet, 2020) were plotted with the *plotOptimalHeatmaps* function. ChIP-seq-like profiles for validation were plotted with the *plotOccupancyProfile* function, armed with values for number of bound molecules (N) and the PWM scaling factor (λ) obtained by indexing. Other than plotting the heatmaps for training and profiles for validation, another method of data visualisation, which is TF clustering, was applied. For this, the optimal AUC from the validated datasets was used. Clusters of TFs were made based on their accessibility preference profile using the k-means clustering method. In order to perform it, the data obtained from the ChIPanalyser step was merged with pre-existing data that had been obtained within Dr. Zabet's lab in previous years (Pop R., 2021). This step was necessary in order to have enough data to be able to cluster accurately.

K-means clustering works with an algorithm that allows to cluster data of the same nature in *k* groups, in order to find patterns within the data. An elbow plot (fig. 8) was generated from which the number of optimal k clusters to use was selected (this will be where the curve plateaus and will indicate how many groups there are in the given datasets). In this case, k=5 was selected, and the five groups were labelled as *AIF*, *partial AIF*, *ADF*, *poorly predicted* and *IDF* based on their trends. These plots will be called *QDA-means plots* in the results section and will show the QDA on the x-axis and the optimal AUC on the y-axis. The clustering results were plotted as classification heatmaps, where each colour corresponded to a class of TFs.

To interpret the results, there are 5 possible scenarios. The criteria taken into account are whether there is low (<0.8) or high (>0.8) binding in: areas of open chromatin, areas of dense chromatin, and if the difference between the AUC in dense vs open chromatin was high or low. The 5 classes are AIF, partial AIF, ADF, poorly predicted and IDF. Fig. 7 gives a full explanation of how to interpret the results from the k-means clustering algorithm.

All scripts used for the above analysis are available on:
https://github.com/AlessandraPisante/MSD.git.

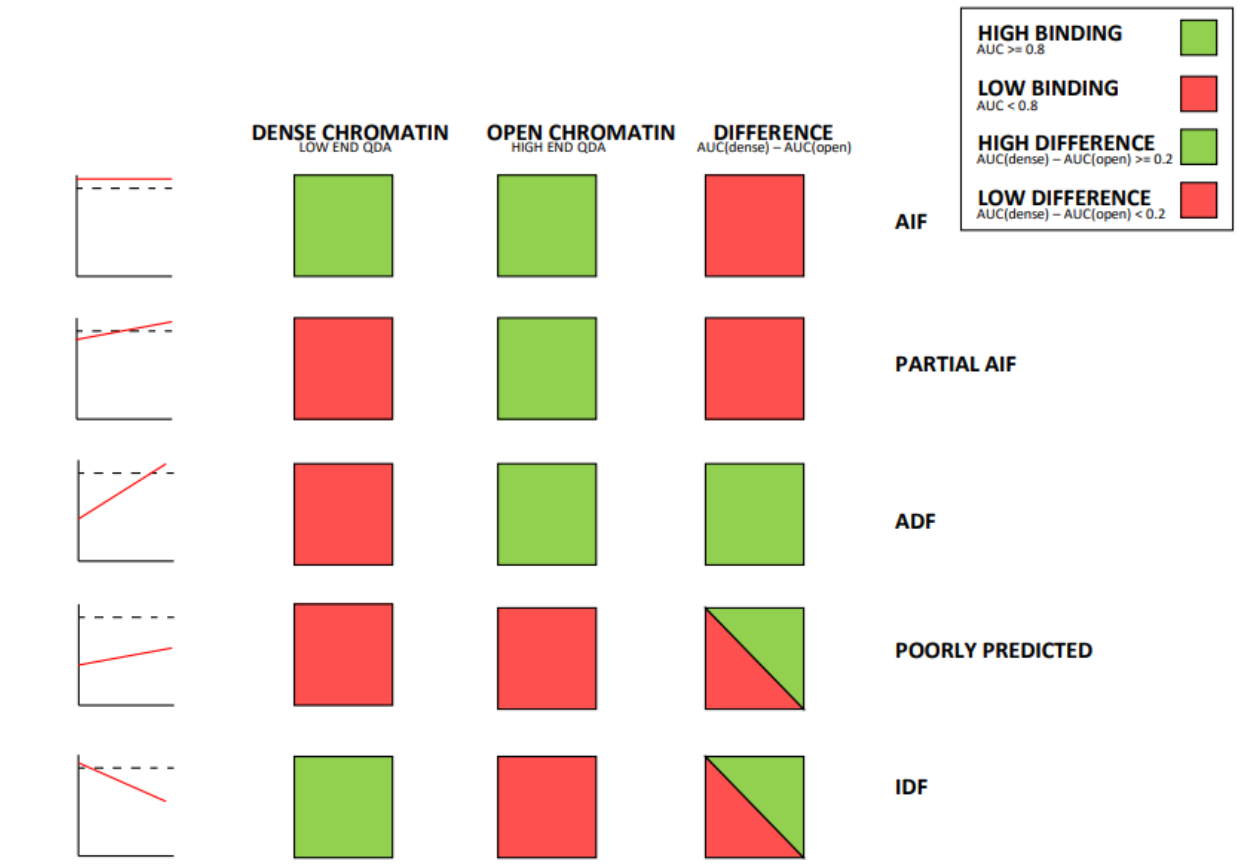

Figure 7: graphical illustration of the results interpretation behind the k-means algorithm. The low-end QDAs are 0 – 0.2 and high-end QDAs are 0.8 – 0.99. The slopes on the left indicate what a QDA plot would look like for each TF class (QDA on x-axis and optimal AUC on y-axis). Depending on the predicted AUC value in dense and open chromatin as well as the difference between these two values, a TF is classified as: i) AIF when AUC >= 0.8 in dense and open chromatin, and therefore AUC(dense) – AUC(open) < 0.2; ii) partial AIF when AUC < 0.8 in dense chromatin and >= 0.8 in open chromatin, but AUC(dense) – AUC(open) is still < 0.2; (iii) ADF when AUC < 0.8 in dense chromatin and >= 0.8 in open chromatin, and therefore AUC(dense) – AUC(open) >= 0.2; (iv) poorly predicted when AUC < 0.8 in dense and open chromatin, and the difference can either be low or high; and (v) IDF when AUC >= 0.8 in dense chromatin and AUC < 0.8 in open chromatin, and the difference can be either low or high. The *difference* column indicates the TF's preference for open compared to dense chromatin. For example, if the difference is low, we would call the TF an AIF or partial AIF.

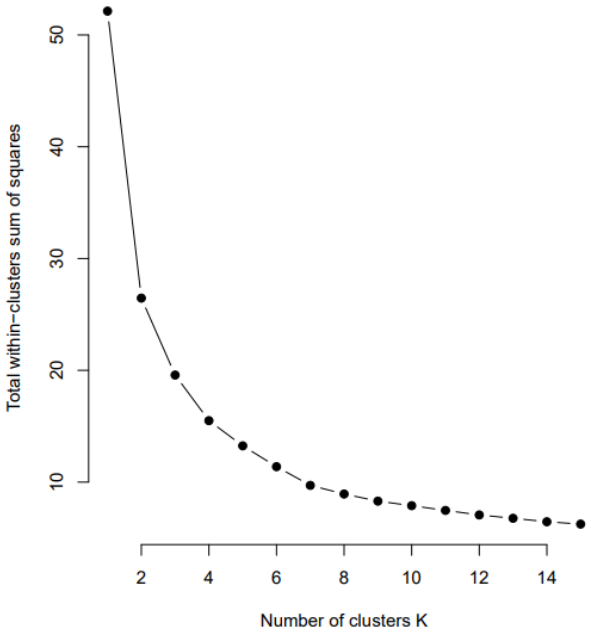

Figure 8: elbow plot showing how many groups there are in a given dataset. The k-means algorithm was run with k values between 1-15. The within-cluster sum of squared errors for each *k* was plotted in order to generate the elbow plot below. The lowest point before the "elbow" of the plot (where the line points stop going down drastically and start to plateau) is indicating the optimal number of clusters for a given dataset. Based on this elbow plot, k=5 was selected.

7 Results

7.1 Pre-processing of TF binding datasets

A significant increase in sequence quality was shown after pre-processing the data with FastQC and cutadapt. *Per Sequence Quality Scores* modules showed that all sequences had an optimal quality score at around 37 Phred or above, which is a measure of the probability of a sequencing error. The *Adapter content* modules showed adapter contamination which, after trimming, was at its lowest. The use of Bowtie2 allowed to align the sequence reads to the reference genome (hg38) and resulted in high alignment rates in all datasets. As a matter of fact, the overall alignment rate was over 90% for all TFs (highest: CEBPB with 97.91%; lowest: MXI1 with 92.55%), as shown in fig. 9, meaning the data was of good quality. Following alignment with *Bowtie2*, peaks were called with *macs2* in order to identify the TF binding sites. The number of peaks was variable, going from the highest value of ~228,000 peaks for CEBPB to 145 peaks for ELK1 (fig. 10); such variation comes from the different expression and binding strength of each TF within the cell line. When a TF presented less than 550 peaks, it was only validated on the top regions (11-60). This was the case of ELK1.

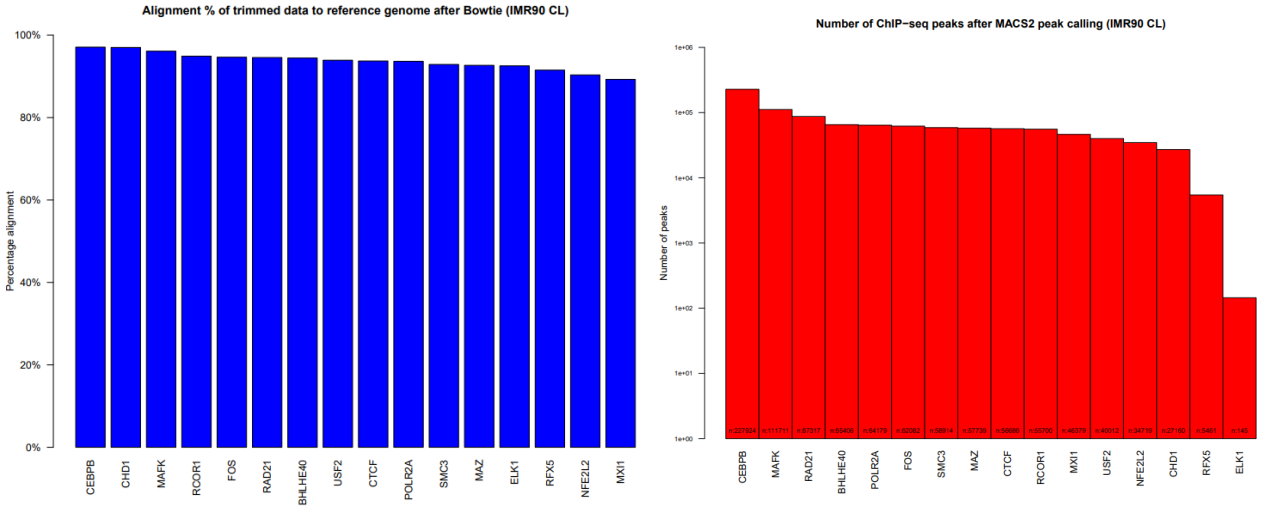

Figure 9 (left): the percentage alignment to the reference genome (hg38) of the pre-processed fastq datasets, following *Bowtie,* is very high. A good alignment of the mapped reads to the reference genome suggests that the data is reliable. When mapped, the reads are assigned to a specific location in the genome (highest: CEBPB with 97.91%; lowest: MXI1 with 92.55%). **Figure 10 (right): logged number of ChIP-seq peaks after *macs2* peak calling function in the IMR90 cell line.** The real (not logged) number of total peaks is found below. Note that ELK1 only presents 145 peaks, meaning it will be only validated on the top regions.

Fig. 11 shows the percentage of accessibility to which each QDA level corresponds for the four DNA accessibility assays used to run ChIPanalyser in this analysis, for both cell lines IMR90 and HepG2. As can be observed from the figure, the trends of accessibility percentages do differ between the four methods. The variability comes from the fact that these are different methods of measuring DNA accessibility. Note that whilst ATAC-seq, DNase-seq and MNase-seq follow the same accessibility pattern (starting at 100% accessible DNA for QDA=0 and then gradually decreasing their percentage of accessible genome), NOMe-seq does not follow this pattern and in this case QDA=0 corresponds to 41% of accessible regions and QDA=0.99 corresponds to 8% accessible regions. MNase-seq does not go down all the way to 0%, which does not agree with the other methods used. Quantifying what percentage of accessible DNA corresponds to each QDA shows how some methods measure different aspects of accessibility and, thus, are affected by different biases.

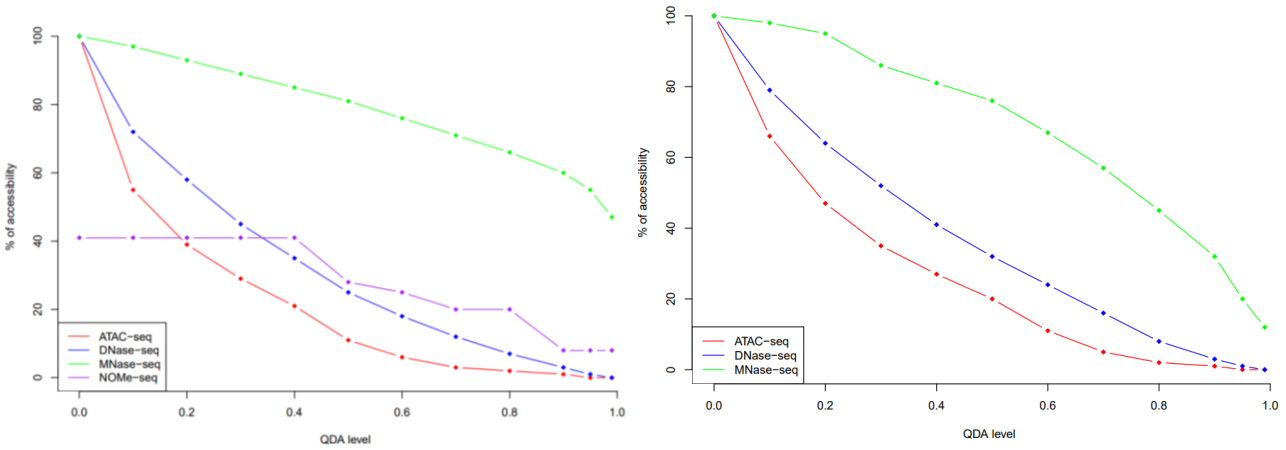

Figure 11: QDA threshold for each DNA accessibility measure in IMR90 (right) and HepG2 (left). It is important to quantify what is the percentage of accessible DNA per each DNA accessibility method in order to know which accessibility percentage corresponds to each QDA. The variability comes from the fact that these are different methods of measuring DNA accessibility.

7.2 Preliminary evaluation of model performance

Several values for N and λ were considered and heatmaps were generated plotting two metrics (AUC and MSE) for all combinations of those. ChIP-seq-like profiles were generated. Some TFs showed to be bound to the DNA regardless of the percentage of accessible chromatin, and it was also shown that their binding was accurately predicted by the model. For example, this was the case of CTCF, which activity is responsible for chromatin architecture and transcription (Luan *et al*., 2021).

Fig. 12 shows AUC and MSE heatmaps for different DNA accessibility levels of CTCF, validated on the stronger-bound regions, i.e. regions with high level of binding. MSE indicates a measure of error and highlights how accurate the model prediction is compared to real ChIP-seq data, whilst AUC highlights how well the peaks are predicted. The closer the AUC is to 1, the better prediction ChIPanalyser is making at modelling real ChIP-seq data. Ideally, the heatmaps would pinpoint optimal parameters that fall into the white area for MSE (meaning the error is minimised) and purple for AUC. A purple colour in the AUC heatmap indicates good resemblance of the predicted profile compared with the actual ChIP data. The lower the λ, the higher the TF affinity for the DNA.

From these observations, one can conclude that if no changes are observed in the AUC when varying the percentage of the accessible genome, that specific TF is insensitive to DNA accessibility, indicating that that TF has potential pioneering functions, and so it might be an AIF. This approach can address questions such as: how sensitive is this TF to DNA accessibility? ChIPanalyser predicted for CTCF an estimated number of bound molecules of 100,000 N (x-axis) for both AUC and MSE metrics, and for both accessibility levels of DNase-seq (QDA=0 corresponding to 0% dense regions vs QDA=0.95 corresponding to 95% dense regions). These first results are in line with a 2019 study where around 100,000 molecules of CTCF were found in the nucleus of human cells with some variation given by different cell lines and the highly diverse functions of CTCF, ranging from IDF to AIF (Cattoglio et al., 2019). A different study in 2019 measured that ~95% of CTCF is bound to the DNA (Belaghzal et al., 2019). Moreover, all N values correspond to a low λ of ~1 which indicates a high specificity of the TF to the DNA. Thus, from these results it seems that the percentage of accessible DNA does not drive the binding of CTCF, because our results show that CTCF would bind anyway, regardless of the accessibility restrictions.

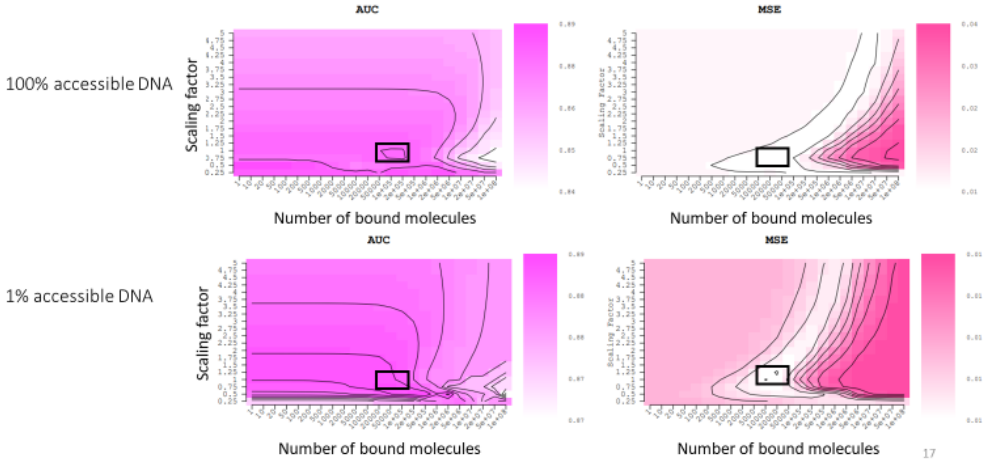

Figure 12: AUC and MSE heatmaps of CTCF (stronger-bound validation regions used) fitted with DNase-seq data, show the same number of DNA-bound molecules (x-axis) to be around 100,000, regardless of the chromatin accessibility restriction used, whether the QDA is 0, corresponding to 100% accessible chromatin, or 0.95, corresponding to 5% of accessible DNA. These heatmaps plot optimal number of bound molecules (N, on the x-axis) vs optimal scaling factor (λ, on the y-axis) and the black boxes represent the parameters (N and λ) pinpointed by the model to be optimal for modelling real binding. The black boxes fall into the white area for MSE (which means the error is minimised) and bright purple for AUC (which indicates good resemblance of the predicted profile compared with the actual ChIP data). The scale on the right-hand side indicates an optimal AUC of ~0.85 (the closer to 1, the better prediction ChIPanalyser is doing at modelling real ChIP-seq data). All values correspond to a low λ of ~1 which is promising of high affinity of the TF to the DNA.

Similar results are observed when looking at different validation regions (fig. 13a: medium-bound validation regions; fig. 13b: weak-bound validation regions). This suggests that CTCF binds to the DNA regardless of the chromatin accessibility restrictions and regardless of the validation regions used, although some variation is observed due to the four different DNA accessibility assay used.

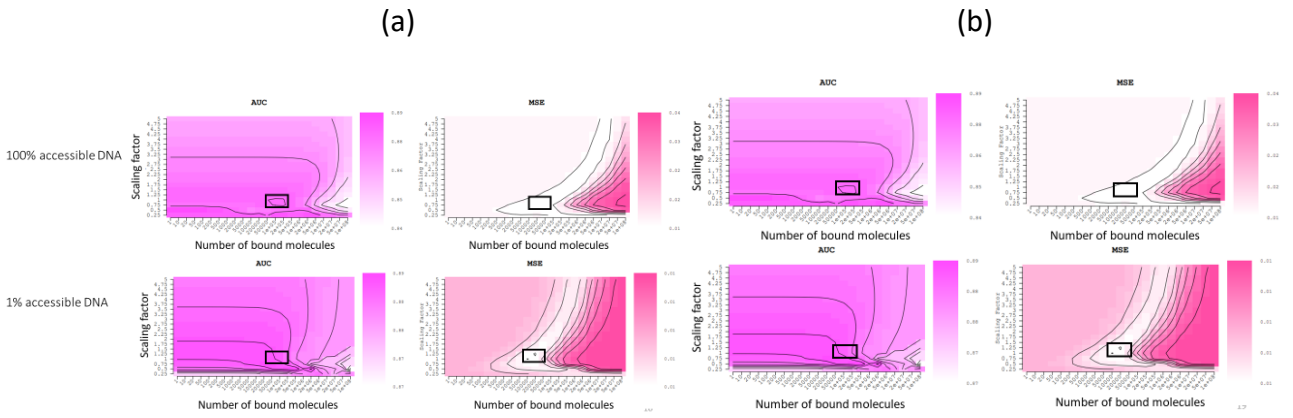

Figure 13: AUC and MSE heatmaps of CTCF (a – medium-bound validation regions; b – weak-bound validation regions) fitted with DNase-seq data, show the same number of bound molecules (N) to be around 100,000, regardless of the chromatin accessibility restriction used, whether the QDA is 0, corresponding to 100% accessible chromatin, or 0.95, corresponding to 5% accessible DNA. These heatmaps plot optimal number of bound molecules (N, on the x-axis) vs optimal scaling factor (λ, on the y-axis) and the black boxes represent the parameters (N and λ) pinpointed by the model to be optimal for modelling real binding. The black boxes fall into the white area for MSE (which means the error is minimised) and bright purple for AUC (which indicates good resemblance of the predicted profile compared with the actual ChIP data). The scale on the right-hand side indicates an optimal AUC of ~0.85 (the closer to 1, the better prediction ChIPanalyser is doing at modelling real ChIP-seq data). All values correspond to a low λ of ~1 which is promising of high affinity of the TF to the DNA. These results are identical to those obtained by validating the model on the stronger-bound regions, suggesting that CTCF is indifferent to chromatin accessibility restrictions.

Fig. 14 shows a different TF called MAZ. MAZ (Myc-associated zinc-finger protein) is a TF with dual roles in transcription initiation and termination; deregulation of its expression is associated with the progression of pancreatic cancer (Maity *et al.*, 2018). As fig. 14 shows, the optimal parameters for the MSE of this TF are actually variable, based on whether the model assumes that chromatin is dense, with an optimal N of between ~20,000-50,000 in accessible chromatin and of ~1,000,000 in dense chromatin. This suggests that this TF in particular **is** indeed dependent upon DNA accessibility levels, otherwise the predicted optimal parameters to model its binding would be equal whether chromatin was open or not. This is an example of a TF which does not display pioneering functions and thus is classed as an ADF.

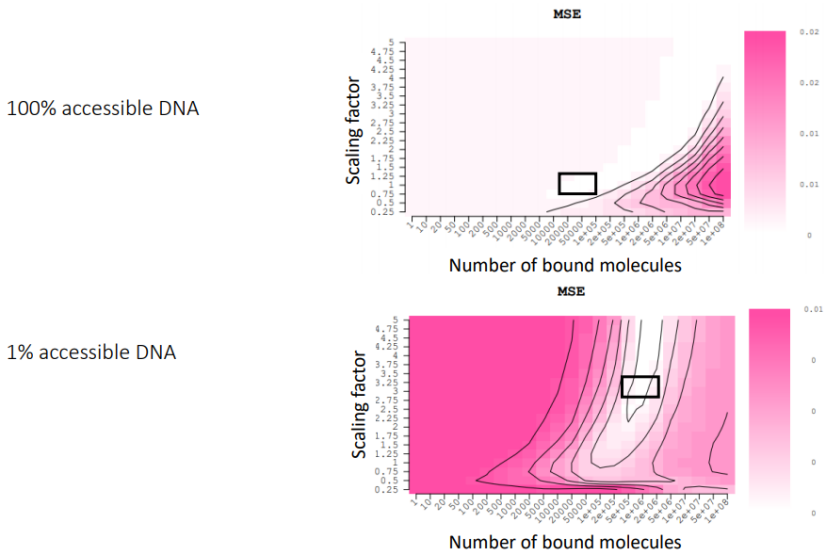

Figure 14: MAZ is a TF that does not display pioneering properties, as the optimal number of DNA-bound molecules (N) when chromatin is dense does not match the optimal N of when chromatin is accessible (optimal N of ~3,000 in accessible chromatin and of ~1,000,000 in dense chromatin). These heatmaps plot optimal number of bound molecules (N, on the x-axis) vs optimal scaling factor (λ, on the y-axis) and the black boxes represent the parameters (N and λ) pinpointed by the model to be optimal for modelling real binding. The black boxes fall into the white area for MSE (which means the error is minimised) and bright purple for AUC (which indicates good resemblance of the predicted profile compared with the actual ChIP data). The scale on the right-hand side indicates a poor AUC of 0 (the closer to 1, the better prediction ChIPanalyser is doing at modelling real ChIP-seq data). λ values are variable ~1 - ~3.5 which indicates high to low affinity of the TF to the DNA, depending on the percentage of accessible DNA.

The AUC and MSE heatmaps indicated good agreement between the predicted profiles and the actual ChIP-seq dataset. Thus, upon analysis of the heatmaps, one can infer optimal parameters to understand how well the model predicts TF binding by plotting occupancy profiles. To confirm that the model was predicting accurately ChIP profiles, we plotted the predicted ChIP-seq-like profiles using the optimal values for number of bound molecules – N – and the PWM scaling factor – λ.

Fig. 15 shows an example of a predicted ChIP-seq-like profile from the analysis of CTCF, using DNase-seq data and assuming all DNA is accessible. This profile shows that the model prediction is very accurate in both areas of open and dense chromatin, meaning that ChIPanalyser can recover the real ChIP-seq data with high accuracy, regardless of whether we restrict its binding to accessible DNA or not. Moreover, the fact that most of the background is yellow, but the peaks overlap with a white background, indicates that the peaks are found in open chromatin (white background = euchromatin) even when we make the model assume that chromatin is dense (yellow background = heterochromatin), which suggests that CTCF is a potential AIF: this is the steady state of the cell – therefore wherever CTCF has bound to, the chromatin has relaxed.

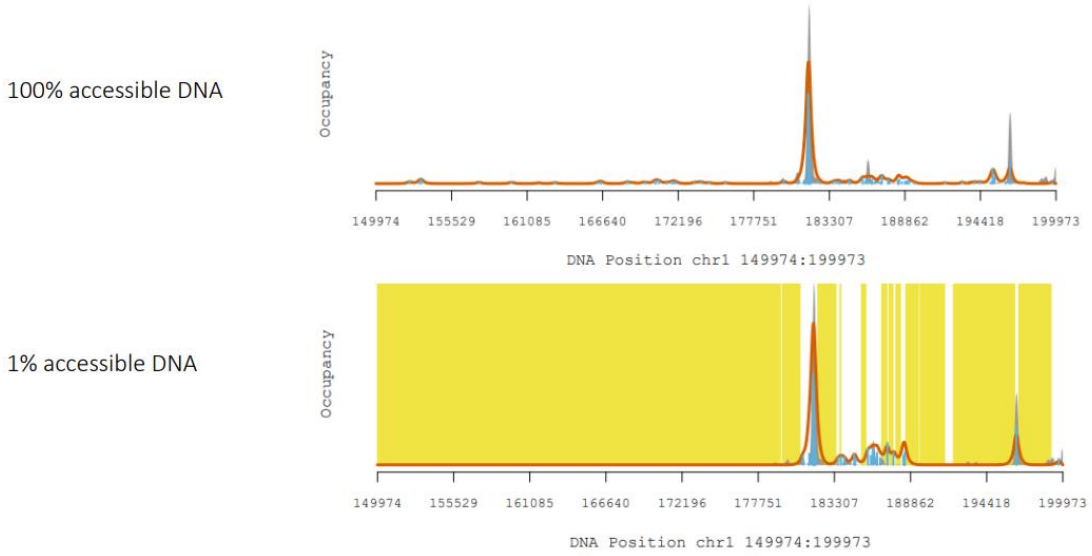

Figure 15: ChIP-seq-like (occupancy) profile of ChIPanalyser prediction of CTCF binding in areas of 100% accessible DNA (white background) vs 5% accessible DNA. Yellow and white background shadings of these occupancy profiles respectively represent inaccessible and accessible DNA. The grey areas are real peaks, from real ChIP data. The red lines are ChIPanalyser's predicted peaks. The blue lines indicate the percentage of binding site occupancy. These profiles show that the model prediction is very accurate in both areas of open and dense chromatin, as the red line perfectly outlines the grey peaks. If the peaks are on a white background, even when chromatin is assumed to be dense (profile below) that means those peaks are found in open chromatin. These are the same genomic regions; the only difference is that to plot the profile below ChIPanalyser was run assuming 99% of chromatin was dense.

7.3 Analysing the general binding behaviour for the TFs

For each accessibility quantile, the parameters that optimise the MSE on the training data are selected. The AUC was computed and plotted for each validation set (strong, medium and weak-bound regions) against the quantile vector, to create what will be referred to as a *QDA plot* (see *6. Methods*). QDA plots show a threshold to call DNA accessibility on the x-axis, where 0 means everything is accessible, and 1 means chromatin is completely dense, and the AUC on the y-axis, so they plot DNA accessibility against the measure of quality of the prediction. If a straight line is plotted, it means that the TF has no preference for open or dense chromatin, independently of the accessibility of DNA (which is how an AIF would behave).

Fig. 16 shows that ChIPanalyser has predicted several TFs to be AIFs using DNase-seq as the DNA accessibility measure and the strongest binding regions as validation regions. Those TFs – proposedly AIFs – bound with good accuracy to dense chromatin as well, as shown by the high AUC values averaging around 0.88, highlighting that, for them, it does not matter how stringent the DNA accessibility value is, the model will still predict their binding with high accuracy, as depicted by the slope in fig. 16, which is high regardless of the QDA value.

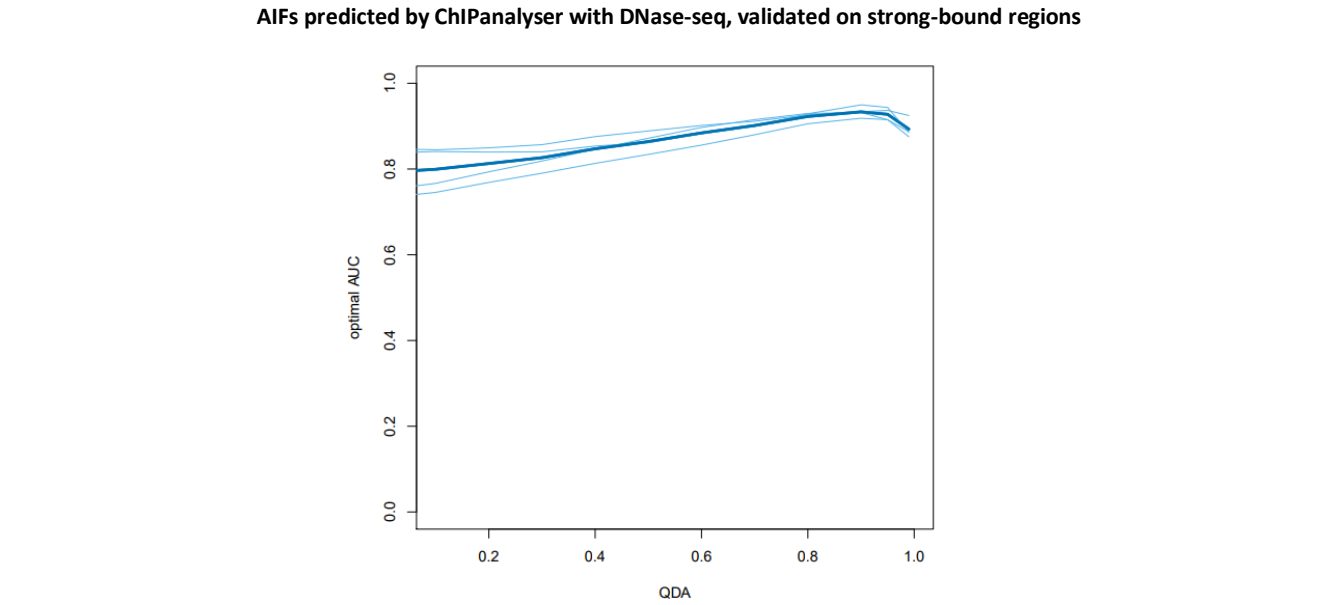

Figure 16: from the pool of 20 TFs analysed in IMR90 and HepG2, ChIPanalyser has predicted 5 of them to be AIFs. This QDA plot plots QDA (x-axis) vs optimal AUC (y-axis) of the 20 TFs analysed, when ChIPanalyser was fitted with DNase-seq and was validated on the strong-bound regions. From the pool of 20 TFs, 5 of them showed pioneering-like properties and were therefore classed as AIFs, as shown here: when the AUC is high (equal to or above 0.8), regardless of the QDA value (0 to 1), it means that that specific TF is insensitive to chromatin accessibility restrictions. This is because the model had outputted an optimal AUC regardless of how dense chromatin was. This graph shows all of the predicted AIFs in one plot, and those are: USF2, MXI1, BHLHE40, RFX5 and CREB1. The thick line is the average AUC vs QDA. This is based on the k-means clustering algorithm (see *Methods*). To see what other kinds of QDA plots there are and to learn how to interpret them see fig. 7.

It follows that ChIPanalyser does not predict everything as AIFs, hence why showing negative controls is important. Fig. 17 portrays QDA-plots where data from this analysis was bound to previous data in Dr. Zabet's lab (data had to be bound due to the small sample pool in this analysis), where TFs were analysed in the same way but in different cell lines (K562 and HepG2) with DNase-seq as accessibility data (Pop R., 2021). These plots show how ChIPanalyser had also predicted a number of partial AIFs (fig. 17a), some ADFs (fig. 17b) and some poorly predicted TFs (fig. 17c). ADFs have a preference for open chromatin, as shown by the abrupt slope in fig. 17b. Partial AIFS (fig. 17a) do also have a preference for open chromatin, but not as strong as ADFs.

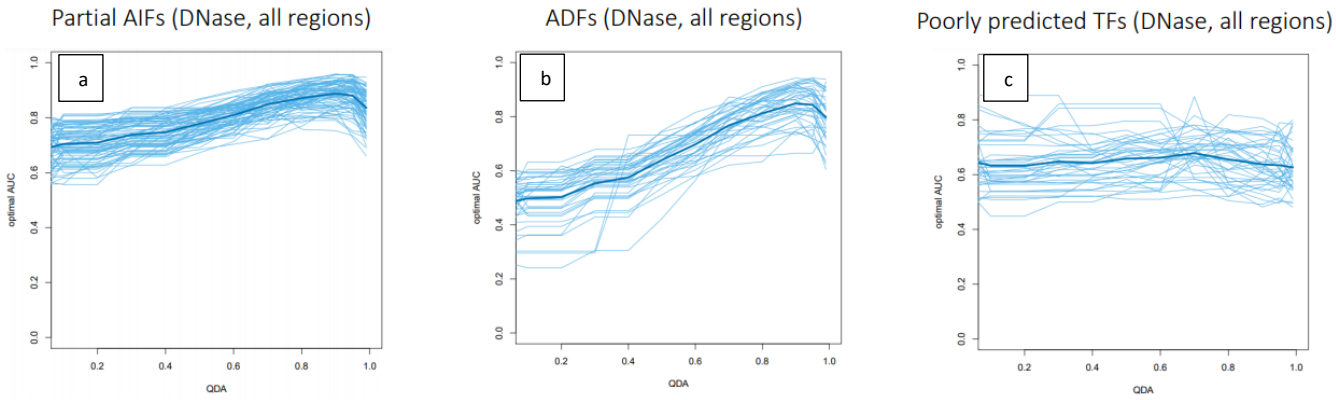

Figure 17: ChIPanalyser can predict several classes of TFs based on their binding properties. QDA-plots of IMR90 and HepG2 TFs bound to data from previous TFs in K562. These QDA plots are showing the change in AUC vs QDA for partial AIFs, ADFs and poorly predicted TFs. From left to right: (a) partial AIFs (straight line corresponding to high AUC but not as straight and high as AIFs (fig. 16)); (b) ADFs (slope from low to medium AUC); (c) poorly predicted TFs (straight line but low AUC). This graphical classification is based on fig. 7 – how to interpret the k-means clustering results.

In order to visually summarise how ChIPanalyser classed each TF in this study, run with each DNA accessibility assay, per each validation region, a classification heatmap was plotted as per fig. 18, which sums up all the results that can be observed from individual plots into one heatmap. This is because different DNA accessibility measures, as well as different model validation regions, will influence the outcome of how a TF is classified based on its DNA binding properties. Some of these results are consistent: for instance, USF2 seems to behave like an AIF under all circumstances. On the other hand, other TFs display a different behaviour based on the parameters ChIPanalyser was run with. An example would be CTCF, which is predicted as an AIF throughout the whole of the DNase-seq, MNase-seq and NOMe-seq analysis, but it actually behaves like an IDF when ChIPanalyser is run with ATAC-seq. This would make sense, given its role in chromatin loops and TADs (Wutz *et al*., 2017).

In every case, a decreasing pattern is observed: when the model was validated on the strong regions, a number of TFs were predicted as being able to bind in areas of dense chromatin, whether they were classified as AIFs (USF2, MXI1, BHLHE40 and more), partial AIFs (CTCF, CEBPB, FOS and more), or IDFs (RAD21, SMC3, MAFK and more). As the validation regions became weaker-bound, the prediction accuracy decreased so that, eventually, most of the TFs were classed as *poorly predicted,* except for CTCF and SMC3. This is an expected behaviour which will be explained in the discussion chapter of this thesis. Note that ELK1 had missing values for every medium and weak-bound validation region because of the low number of ChIP-seq peaks (145), which made it impossible to run the model on any other validation range other than the top one (11-60).

The only TF that is classified by ChIPanalyser as an ADF is MAZ (ATAC-seq and DNase-seq only). Whilst the ADF classification is consistent in the top and medium-bound validation regions of ATAC-seq, with DNase-seq MAZ is classified as an AIF in the top regions and as an ADF in the medium regions. MNase-seq and NOMe-seq, instead, classed MAZ as poorly predicted throughout the whole analysis.

CHD1 and RCOR1 were the only two TFs which binding was classed as poorly predicted on each validation region and with each chromatin accessibility assay used. When ChIPanalyser was run with MNase-seq as chromatin accessibility assay, the prediction accuracy was at its lowest compared to the other three assays, and in fact there were four TFs that were poorly predicted in the top regions (FOS, CHD1, RCOR1 and MAZ) compared to the other methods which only had two (ATAC-seq and DNase-seq: CHD1, RCOR1) or three (NOMe-seq: CHD1, RCOR1 and MAZ).

From the literature, p53 and MyoD are validated AIFs in IMR90. However, it seems that ChIPanalyser did not recover their binding properties, and in fact fig. 18 shows that ChIPanalyser classified those two TFs as poorly predicted. This result could be explained by other factors that are not accounted for by ChIPanalyser. For instance, these TFs might need the presence of co-factors for cooperative binding, or they might be sensitive to DNA methylation. Neither of these possibilities are considered by the model. Instead GATA4, FOXA1 and CREB1, which are previously reported AIFs in the HepG2 cell line, gave different results. GATA4 was predicted as an ADF in the strong and medium-bound regions of ATAC-seq, and in the strong-bound regions of DNase-seq, whilst FOXA1 was predicted as a partial AIF in those same validation regions. Both GATA1 and FOXA1 were poorly predicted with the MNase-seq dataset, across all validation regions. CREB1 was predicted as an AIF in the strong and middle-bound regions of ATAC-seq and DNase-seq, whilst it was poorly predicted using MNase-seq.

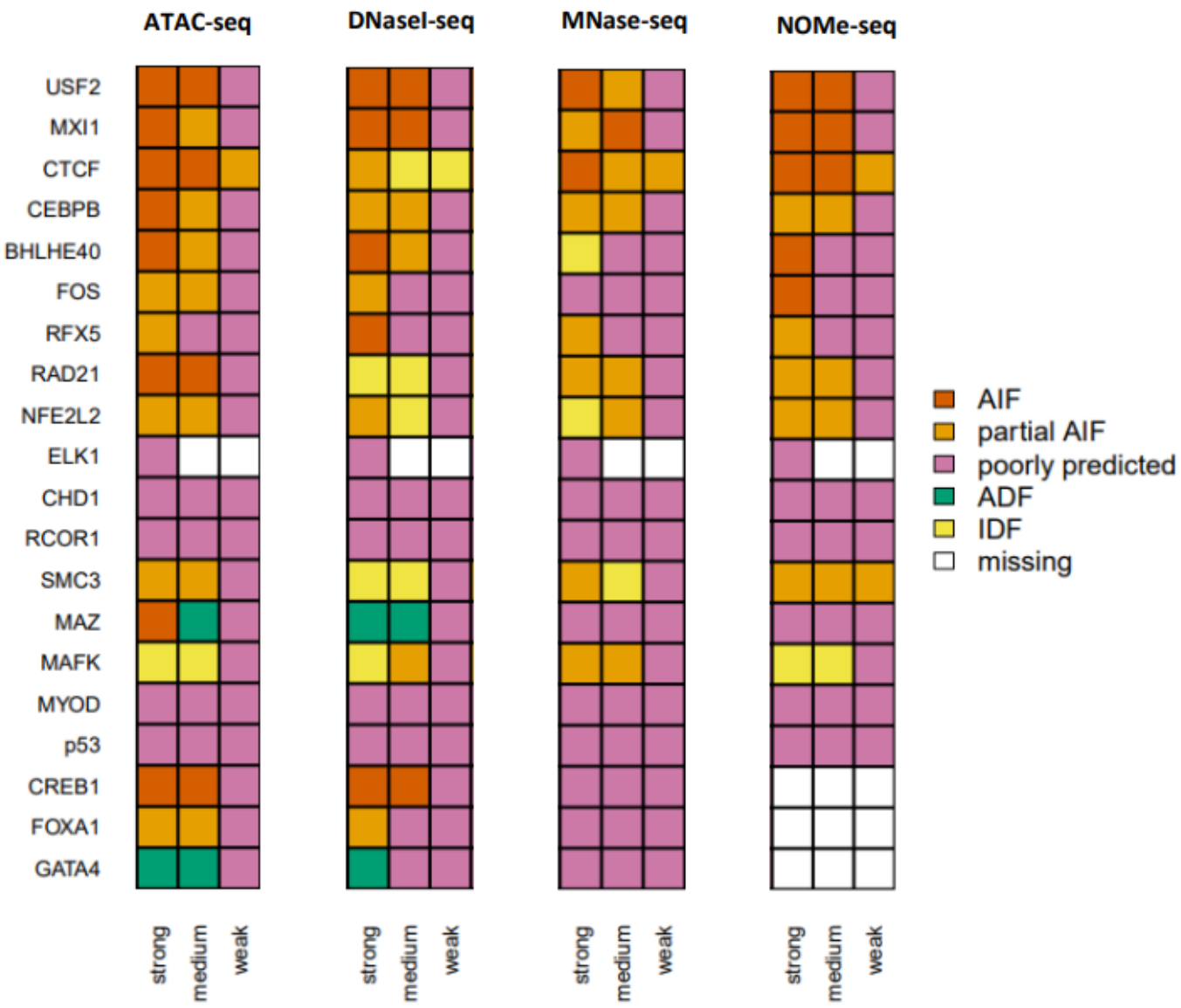

Figure 18: ChIPanalyser's prediction of classification of 20 TFs in the IMR90 and HepG2 cell lines, according to their DNA binding properties. Different DNA accessibility measures as well as different model validation regions will influence the outcome of how a TF is classified based on its DNA binding properties. Some of these results are consistent, such as USF2 which seems to behave like an AIF under all circumstances, whereas other TFs display a different behaviour based on the parameters ChIPanalyser was run with, such as CTCF which is predicted as an AIF throughout the whole of the DNase-seq, MNase-seq and NOMe-seq analysis, but it actually behaves like an IDF when the analysis is run with ATAC-seq. P53 and MYOD are the validated AIFs used as controls in IMR90 and CREB1 , FOXA1 and GATA4 are the ones in HepG2.

8 Discussion, conclusions, and further work

8.1 Research purpose and scope of the analysis

Binding sites (whether they are proximal, distant or there are clusters of them) are genomic regions where TFs bind in order to regulate the process of transcription or, in other words, promote or repress gene expression (Inukai *et al*., 2019). ChIP-seq techniques are widely used to study protein:DNA interactions in order to observe molecular regions and pinpoint where are TFs bound in the genome (Nakato and Sakata, 2020). ChIPanalyser is an R/Bioconductor package that models and predicts the binding of TFs to the DNA, and outputs ChIP-seq-like profiles based on (i) DNA-accessibility threshold to model binding site accessibility, (ii) a Position Weight Matrix (PWM – a representation of patterns in a biological sequence which TFs will preferentially bind to), (iii) the number of bound TFs to the DNA (represented by the letter N) and (iv) a modulator for the binding specificity (represented by the letter λ) (Martin and Zabet, 2020 and Zabet and Adryan, 2015). In short, starting from a DNA sequence binding motif, DNA accessibility (e.g. ATAC-seq or DNase-seq etc.) and ChIP data, one can infer the specificity of the DNA-bound TFs.

The research scope of this dissertation was to classify TFs based on their DNA binding properties using ChIPanalyser. The three different classes of TFs were described as follows: **AIFs** have pioneering-like properties and can bind chromatin regardless of its accessibility constrictions, to then subsequently relax it and make it accessible for **ADFs** which, instead, can only bind already accessible chromatin. The third and last class includes **IDFs**, which will preferentially bind dense chromatin to then maintain and/or reinforce the compacted status. Classifying TFs aims to tackle a tough question in the Genomics field as TFs are at the very heart of every single cellular process. Learning about their properties can benefit the Genomic world beyond the simple application of expanding our state of the art Biological know-how, but will also offer a solid and analytical base to tackle more complex questions and gain a better understanding into the chain of events that lead to certain diseases, to then be able to target and inhibit specific cellular pathways for therapeutic approaches.

8.2 Our main findings

According to our study, the shortlisted AIFs (and partial AIFs) in the strong-bound regions of ATAC-seq, DNase-seq, MNase-seq and NOMe-seq are USF2, MXI1, CEBPB, BHLHE40, FOS and RFX5. Those TFs could be responsible for opening the chromatin at their strongest sites. Two cohesion subunits (RAD21 and SMC3) and MAFK have been classified as inaccessibility-dependent factors, suggesting they prefer partially closed chromatin. CTCF was classified as either an accessibility-independent factor or inaccessibility-dependent factors. This indicates that CTCF could be either a pioneer or an insulator, and both roles have previously been proposed for this TF in the literature.

Since ChIPanalyser could not accurately predict previously reported TFs that can bind dense chromatin such as p53 and MyoD, possibly due to the absence of their co-factors, it means that the AIFs predicted by our model do not need co-factors for binding – but might still need co-factors to open the chromatin – otherwise ChIPanalyser, which does not include cooperativity as one of the modelling criterion, could have not properly predicted their binding patterns.

GATA4, FOXA1 and CREB1 were respectively predicted as and ADF, partial AIF and AIF in the strongest sites of ATAC-seq and DNase-seq. The predictions for FOXA1 and CREB1 were expected, which is a positive sign of model reliability. Instead, the pioneering properties of GATA4 were not

recovered. This could potentially be because GATA4 is recruited by an AIF rather than being and AIF itself. Moreover, a recent study from 2019, however, proposed a model to explain GATA4's DNA-binding properties as a cooperative complex with TBX5. This GATA4:TBX5 complex has a role in normal heart development (Rosado-Rodríguez, Rodríguez-Ríos and Rodríguez-Martínez, 2019). Influenced by the need for cooperativity, ChIPanalyser's results in predicting GATA4 binding properties might have also been affected, like with p53 and MYOD. It is hypothesised that co-factor expression affects the binding mechanisms of some TFs and perhaps leads to increased pioneering activity at target sites (Donaghey et al., 2018).

8.3 Interpreting ChIPanalyser's results

We used ChIPanalyser to computationally assume that 100% of the chromatin was accessible and TFs could bind everywhere (QDA=0, i.e. 0% of the chromatin is dense, 100% is accessible), to then compare the predicted profile with the real, experimentally obtained one. By assuming that the TF can bind everywhere according to the model, one could see if there are more peaks predicted in inaccessible DNA compared to another TF which binding was not restricted but predicted fewer peaks in dense chromatin. In this case the first TF would be a traditional TF which binding is restricted by DNA accessibility levels, whilst the second would bind regardless of the chromatin accessibility restrictions. This means that an AIF does not have predicted peaks in dense chromatin, whilst ADFs do. In other words, the fact that there is no predicted binding in dense chromatin is a suggestion of a potential AIF.

We also assumed that TF binding was restricted only to small areas of accessible DNA (e.g. QDA=0.9, i.e. 90% of the chromatin is dense) and gained mechanistical and biological insights about the binding of that TF. By running ChIPanalyser assuming that 90% of the chromatin is dense, we distinguished between an AIF and an IDF because an insulator has ChIP peaks which are present only in dense chromatin, while an AIF will display peaks only in open chromatin.

8.4 Limitations of this study

8.4.1 Limitations with accessibility datasets

This dissertation aimed to characterise 20 TFs in the IMR90 and HepG2 cell lines. Moreover, the analysis was run multiple times with different QDAs (1 to 12) which were used in order to investigate the preference of each TF for chromatin accessibility. Each QDA considers accessible the top 1-n regions for every DNA accessibility assay used (ATAC-seq, DNase-seq, MNase-seq and NOMe-seq). A QDA of 0, theoretically, should correspond to 1-0 regions being accessible (i.e. all regions will be accessible). On the contrary, a QDA of 0.99 should mean that 1-0.99 regions are open (i.e. 1% is accessible, 99% is dense).

Having said that, it must be noted that the levels of accessible DNA varied amongst the four DNA accessibility assay used, as shown in fig. 10. A QDA of 0.9 did not always correspond to 90% of dense chromatin – this was especially true for the MNase-seq and NOMe-seq datasets. Such variability comes from the fact that these are different methods of measuring DNA accessibility. Specifically, the MNase-seq technology is based on the fact that the DNA that is not wrapped around nucleosomes will be digested, hence MNase-seq signals correspond to regions where the nucleosomes are bound. These correspond to the non-accessible regions. Then, in order to obtain the accessible ones, the signal is normalised to be comparable with the other datasets. Having done

that, the problem persisted in that a QDA≥0.9 did not correspond to 1-0.9 accessible region (0.1 i.e. 10%), but it was 60-47%, meaning some quantiles did not match the accessibility percentages. One explanation could be that many regions in the MNase-seq datasets had missing signals, meaning that they were not covered, probably because reads did not align. It is hypothesised that the MNase-seq we used had missing data and although there was binding, there was no accessibility data available. Since we want to measure accessible regions, if the heterochromatin repetitive elements are in dense chromatin they will not show as accessible. This means that certain regions might be inaccessible even though in certain regions the reads do not map. This could justify why most of the TFs in medium and weak-bound validation regions are poorly predicted. In order to correct for this condition, we could have selected specific regions where signals are present instead of running the analysis genome-wide. Moreover, another difficult relationship to unravel was the one between DNA methylation and nucleosome occupancy, which is what the NOMe-seq dataset measures. In this analysis, the DNA methylation patterns were difficult to align, as a QDA of 0 corresponded to 41% accessible regions rather than 100%. The MNase-seq and NOMe-seq results, therefore, needed to be interpreted taking these limitations into account. Overall, these patterns suggest that these two methods do not have the dynamic range to distinguish well between different levels of accessibility.

8.4.2 Explaining why some TFs are poorly predicted by ChIPanalyser

In addition to noticing that the MNase-seq and NOMe-seq datasets followed a different pattern of accessibility levels compared the expected one (which was instead what we observed in ATAC-seq and DNase-seq datasets), it is also important to notice how much better ChIPanalyser did in predicting the TFs' binding properties in the top-bound regions compared to the medium and weak-bound. Better prediction in strong bound regions is a feature of almost all tools, both biophysical (Kaplan *et al*., 2011) and machine learning (Martin and Zabet, 2020). The fact that the model performs well in strong-bound region but that we observe less and less TF binding events as the validation regions become weaker-bound suggests that the binding event is not driven by any of the four parameters that ChIPanalyser considers, which are DNA accessibility, binding energy, binding motif or the concentration of the TF bound to the DNA. Simply put, it must be something outside of those four criteria, something that is not taken into account by any of these models. For example, some of these TFs might work cooperatively, and as a matter of fact cooperative binding is not one of ChIPanalyser's modelled features. This hypothesis was supported by Nagy *et al*. in 2016 where they showed that RAD21 and SMC3, which are both part of the pool of 20 TFs analysed in IMR90 and HepG2, are cohesin subunits and as such they do not bind specifically to the DNA as they do not have a sequence preference. The DNA is extruded by the cohesin subunits RAD21 and SMC3 and this extrusion is blocked at CTCF sites (Schwarzer *et al*., 2017). This is just one instance that could explain the drop in model performance at the lower affinity binding sites. Looking into this limitation is important because it might mean that an AIF shows pioneering properties only at the strong-bound sites. Another hypothesis regarding the behaviour of RAD21 and SCM3 is that they bind to the DNA without help from other TFs and then get stopped at CTCF sites. RCOR1 and ELK1 were also predicted poorly; again, perhaps they need co-factors for cooperativity or are sensitive to DNA methylation, neither of which possibilities were explored in this thesis. According to a 2016 paper by Lu, Mucaki and Rogan, 2016, RCOR1 is also a non-DNA binding protein and, as such, will not have specific sequence preferences (Lu, Mucaki and Rogan, 2016).

Moreover, ChIPanalyser also did not perform well in predicting the properties of some previously reported AIFs in the literature such as p53 and MyoD that indeed were poorly predicted throughout

the whole of the analysis. One plausible explanation could be that AIFs need more co-factors than the average TF, hence why the issue with recovering some of their binding properties. Recent work confirmed binding of p53 to *cis*-regulatory elements, which are non-coding segments of DNA that control transcription through the combinatorial binding of local TFs (Catizone *et al*., 2020). These co-regulatory factors and local sequences acting at p53-bound *cis*-regulatory elements are comparatively understudied. The data suggest that p53 has the flexibility to cooperate with a multitude of TFs in order to regulate *cis*-regulatory elements' activity. These findings support the hypothesis that p53 would be poorly predicted without its co-factors that are indeed essential for its binding. They hypothesised that global p53 activity is guarded against loss of any one regulatory partner, allowing for dynamic and redundant control of p53-mediated transcription. Martin and Zabet in 2020 showed that cooperative TFs such as Hox are poorly predicted by ChIPanalyser, therefore these results are expected. Another recent study revealed new insights into how p53 and p63 (previously reported AIFs regulating chromatin structure) can both positively and negatively influence each other to direct cell fate (Woodstock, Sammons and Fischer, 2021). Both AIFs appear to regulate a mostly unique set of target genes and have non-overlapping cellular roles, however evidence suggests that p53 and p63 cooperate to regulate DNA damage-induced apoptosis in mouse embryonic fibroblasts (Flores *et al*., 2002). Therefore, both studies offer strong support to claim that ChIPanalyser could not accurately predict p53 due to the absence of its co-factors.

The second known AIF that we analysed in IMR90 as positive control is MyoD. The MyoD gene family encode transcription factors that are essential regulators of skeletal muscle lineage determination and differentiation in vertebrates (L'honore *et al*., 2003). MyoD is a well-established AIF in myogenic cell lineage specification during development and trans-differentiation. Its role is to regulate the expression of myogenic specific genes through its binding to regions containing cis-regulatory elements. MyoD also constitutively binds at tens of thousands of additional sites throughout the genome in proliferating muscle stem cells (Chen et al., 2020). Nonetheless, ChIPanalyser was not able to recover its binding properties. In 2003, L'Honore *et al*. showed that the distal regulatory region of the MyoD gene contains a conserved SRF (serum response factor) binding CArG-like element (CArG boxes are present in the promoters of smooth muscle cell genes), which is involved in the control of the gene expression in skeletal myoblasts and in mature muscle satellite cell activation during muscle regeneration. This MyoD-CArG sequence is active in modulating endogeneous MyoD gene expression because microinjection of oligonucleotides corresponding to the MyoD-CArG sequence specifically and rapidly suppressed MyoD expression in myoblasts (L'honore *et al*., 2003). These findings could explain once again why the model predictions are poor.

8.5 Further work and conclusions

Cross-referencing with known pioneer TFs in the same cell line is one approach for proof of concept. The question is then to distinguish between a pioneer and a co-factor of a pioneer, which would require further analysis (Jiang et al., 2019). For example, one could computationally hypothesise a TF's pioneering properties with ChIPanalyser. Then in order to validate them, the analysis needs to be complemented by experimental set up where there is an overexpression of the TF, to then check if the predicted regions become open. By running ChIP-seq and ATAC-seq in the wild type and then in the overexpression mutant, one can compare whether in the mutant there are more open sites, in which case we know that that TF has accessibility independent properties. On the opposite, if there are no novel peaks, then that TF was probably one with accessibility dependent properties. This would be strong experimental evidence to the question "is this TF a pioneer?". A TF with pioneering properties would never overlap with dense chromatin, because it would have already

opened it. However, unless every region where the supposedly AIF bound was open, it could not be excluded that the TF was accessibility dependent. Overall, it can be concluded that one cannot get the definite answer to such questions using solely a computational approach, but ChIPanalyser can work as a filter to classify TFs and shortlist the possible ADFs, AIFs and IDFs without running multiple wet-lab experiments. Also, ChIPanalyser allows to look at the bigger picture compared to individual patterns without the need for previous experimental data, if used for prediction. However, it is important to mention that the model is restricted to whatever is included in it: for example, ChIPanalyser does not include cooperative binding as one of the parameters to explain a TF's binding pattern. Therefore, predicting the binding properties of certain TFs will probably be inaccurate if they rely on factors which the model does not consider, just like the aforementioned TFs SMC3 and RAD21.

DNA methylation could be responsible, together with cooperative binding, for the drop in model performance observed in the weaker-bound regions. It is known that, in some cases, DNA methylation impacts some TFs' binding: a recent study showed that while many TFs do not affect the methylation status of their binding sites, a group of AIFs called protective pioneer transcription factors (PPFs) prevent acquisition of DNA methylation, while another group called super pioneer transcription factors (SPFs) induce DNA demethylation at their methylated binding sites (Vanzan *et al*., 2021). Importantly, the authors of that study also showed that SOX2 inhibits DNMT-1 dependent maintenance of methylation during replication and that this inhibition is amplified by the co-binding of OCT4. However, enhancing chromatin accessibility is not achieved by the mechanism of this binding and the resulting demethylation alone, but interaction with multiple TFs is likely to be required too.

As we did not include DNA methylation in our analysis, it might be that some of the differences in the model performance come from this. The relationship between TFs and DNA methylation does not appear to be a straightforward one: although it has been long thought that DNA methylation represses TF binding, a more recent study has proposed that actually TF binding can inhibit DNA methylation (Héberlé and Bardet, 2019). DNA methylation and TFs do, in fact, affect each other, in a number of different scenarios: for example, the shape of DNA – i.e. its 3D structure – has been thought to affect TF binding. Moreover, epigenetic regulation such as nucleosome positioning and histone modifications can also modulate the physical access of TFs to DNA (Slattery *et al*., 2014). The views regarding the sensitivity of TFs to DNA methylation are changing with emerging studies: it is now hypothesised that DNA methylation could be a consequence of TF binding, rather than a cause (Héberlé and Bardet, 2019 and Blattler and Farnham, 2013). In fact, established models suggest that DNA methylation represses TF binding, whilst emerging models suppose that TF binding to methylated regions induces demethylation. According to a recent 2019 study by Héberlé and Bardet, there are four possible scenarios by which TF binding could impact DNA methylation: firstly, TF binding to DNA **promotes DNA methylation** (e.g. MYC): when TFs bind to unmethylated DNA, recruitment of enzymes called DNA methyltransferases (DNMTs), responsible for the transfer of a methyl group to DNA, promotes *de novo* DNA methylation. Secondly, TF binding to DNA **triggers DNA demethylation** (e.g. CTCF): TFs recruit TET proteins, enzymes that catalyse the demethylation of 5-methylcytosine (Yang *et al*., 2020 and Wiehle *et al.*, 2019), thereby triggering DNA demethylation. Thirdly, TF binding to DNA does not affect **methylated regions**, therefore they **are maintained** (e.g. ZFP57): in this case, transcriptional repression is maintained due to the binding of methyl-CpG-binding domains (MBDs) to methylated genomic islands (called mCpGs), which inhibits TF binding to CpG-rich regions. Lastly, TF binding to DNA does not affect **unmethylated regions**, therefore they **are maintained** (e.g. CFP1): a transcriptionally active state is recognised and

maintained thanks to TF binding which protects from *de novo* methylation by DNMTs. To conclude, the link between DNA methylation and TF binding appears to be variable and dependent upon factors such as the presence of mCpGs at specific positions within TF motifs. This means that there is not one specific model that fits all cases with regards to TF binding to methylated/unmethylated DNA. Instead, TF sensitivity is variable: the same TF could be responsive to a methylated genomic area and indifferent when CpGs are unmethylated (Héberlé and Bardet, 2019).

In order to conduct the TF:methylated DNA parallel study, QDMs (quantile density methylation) could be used instead of QDAs, where the top methylated parts of the DNA as well as the least methylated ones are leveraged to understand whether a specific TF (i) prefers binding to methylated DNA, (ii) avoids methylated DNA or (iii) is indifferent to DNA methylation conditions. The hypothesis is that AIFs are indifferent to DNA methylation levels and can therefore bind with no preference to methylated or unmethylated DNA. There are also numerous TFs that show poor model prediction when 100% of DNA is considered accessible, whilst such prediction precision increases towards the top accessible regions. Considering this as further work following this project, instead of varying the top accessible regions, it could be interesting looking at the top methylated vs the top un-methylated regions to observe if any sensitivity in TF binding is observed there. TFs could be classified based on whether they have this sensibility or not. Investigating the interplay between TF binding and DNA methylation is therefore one plausible application for ChIPanalyser.

In conclusion, then, ChIPanalyser can be used to model and predict ChIP-seq data and learn new biological insights; it can be used to predict TF binding events between cell lines with partial data in a cell line, or for a screening process in order to understand what type of behaviour is associated to a TF. This model has the potential to be applied to any cell line. It is concluded that, regardless of the limitations discussed above, most of the predictions made by ChIPanalyser so far are in line with published data and, therefore, ChIPanalyser has been proved to be a valid predictive tool. These results will hopefully allow to gain new insights into the mechanisms of TF binding and open up new doors for further investigation.

9 Bibliography

In alphabetical order:

Activemotif.com. 2021. NOMe-Seq (Nucleosome Occupancy and Methylome Sequencing). At: https://www.activemotif.com/catalog/864/nome-seq#:~:text=NOMe%2DSeq%20works%20by%20treating,nucleosome%20or%20protein%20binding%20sites. [Accessed 29 March 2021].

Ahn, J., Lee, J., 2008. X chromosome: X inactivation. *Nature Education* 1(1):24

Allison, Lizabeth A. (2012). Fundamental Molecular Biology. *New Jersey: John Wiley & Sons*, Inc. pp.300–301.

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. At: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Accessed 25 October2020].

Barski, A. and Zhao, K., 2009. Genomic location analysis by ChIP-Seq. *Journal of Cellular Biochemistry,* 107(1), pp.11-18.

Belaghzal, H., Borrman, T., Stephens, A., Lafontaine, D., Venev, S., Weng, Z., Marko, J. and Dekker, J., 2019. Compartment-dependent chromatin interaction dynamics revealed by liquid chromatin Hi-C.

Blattler, A. and Farnham, P., 2013. Cross-talk between Site-specific Transcription Factors and DNA Methylation States. *Journal of Biological Chemistry*, 288(48), pp.34287-34294.

Catizone, A., Uzunbas, G., Celadova, P., Kuang, S., Bose, D. and Sammons, M., 2020. Locally acting transcription factors regulate p53-dependent *cis*-regulatory element activity. *Nucleic Acids Research*, 48(8), pp.4195-4213.

Cattoglio, C., Pustova, I., Walther, N., Ho, J., Hantsche-Grininger, M., Inouye, C., Hossain, M., Dailey, G., Ellenberg, J., Darzacq, X., Tjian, R. and Hansen, A., 2019. Determining cellular CTCF and cohesin abundances to constrain 3D genome models. *eLife*, 8.

Chen, Q., Chen, F., Wang, R., Shi, M., Chen, A., Ma, Z., Li, G., Wang, M., Li, H., Zhang, X., Ma, J., Zhong, J., Chen, M., Zhang, M., Zhang, Y., Chen, Y. and Zhu, D., 2020. MyoD is a structure organizer of 3D genome architecture in muscle cells.

Cirillo, L., Lin, F., Cuesta, I., Friedman, D., Jarnik, M. and Zaret, K., 2002. Opening of Compacted Chromatin by Early Developmental Transcription Factors HNF3 (FoxA) and GATA-4. *Molecular Cell*, 9(2), pp.279-289.

Cuatrecasas, P., Edelhoch, H. and Anfinsen, C., 1967. Fluorescence studies of the interaction of nucleotides with the active site of the nuclease of Staphylococcus aureus. *Proceedings of the National Academy of Sciences,* 58(5), pp.2043-2050.

Cussat-Blanc, S., Harrington, K. and Banzhaf, W., 2019. Artificial Gene Regulatory Networks—A Review. *Artificial Life*, 24(4), pp.296-328.

Ding, X., Pan, X., Xu, C. and Shen, H., 2010. Computational Prediction of DNA-Protein Interactions: A Review. *Current Computer Aided-Drug Design*, 6(3), pp.197-206.

Donaghey, J., Thakurela, S., Charlton, J., Chen, J., Smith, Z., Gu, H., Pop, R., Clement, K., Stamenova, E., Karnik, R., Kelley, D., Gifford, C., Cacchiarelli, D., Rinn, J., Gnirke, A., Ziller, M. and Meissner, A., 2018. Genetic determinants and epigenetic effects of pioneer-factor occupancy. *Nature Genetics*, 50(2), pp.250-258.

Ehsani, R., Bahrami, S. and Drabløs, F., 2016. Feature-based classification of human transcription factors into hypothetical sub-classes related to regulatory function. *BMC Bioinformatics*, 17(1).

Flores, E., Tsai, K., Crowley, D., Sengupta, S., Yang, A., McKeon, F. and Jacks, T., 2002. p63 and p73 are required for p53-dependent apoptosis in response to DNA damage. *Nature*, 416(6880), pp.560-564.

Ge Tan (2017). JASPAR2018: Data package for JASPAR 2018. R package version 1.1.1. http://jaspar.genereg.net/

Gentleman, R., Carey, VJ., Bates, DM., 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, AJ., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, JY., Zhang, J., 2015. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10).

Héberlé, É. and Bardet, A., 2019. Sensitivity of transcription factors to DNA methylation. *Essays in Biochemistry*, 63(6), pp.727-741.

Inukai, S., Kock, K. H., & Bulyk, M. L. (2017). Transcription factor-DNA binding: beyond binding site motifs. *Current opinion in genetics & development*, 43, 110–119.

Iwafuchi-Doi, M. and Zaret, K., 2014. Pioneer transcription factors in cell reprogramming. *Genes & Development*, 28(24), pp.2679-2692.

Jiang, G., Wang, X., Sheng, D., Zhou, L., Liu, Y., Xu, C., Liu, S. and Zhang, J., 2019. Cooperativity of co-factor NR2F2 with Pioneer Factors GATA3, FOXA1 in promoting ERα function. *Theranostics*, 9(22), pp.6501-6516.

Jiramongkol, Y. and Lam, E., 2020. FOXO transcription factor family in cancer and metastasis. *Cancer and Metastasis Reviews*, 39(3), pp.681-709.

Kaplan, T., Li, X., Sabo, P., Thomas, S., Stamatoyannopoulos, J., Biggin, M. and Eisen, M., 2011. Quantitative Models of the Mechanisms That Control Genome-Wide Patterns of Transcription Factor Binding during Early Drosophila Development. *PLoS Genetics*, 7(2), p.e1001290.

Keene, M., 1981. Micrococcal nuclease as a probe of DNA sequence organization and chromatin structure. *Cell*, 27(1), pp.57-64.

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 Computational Biology*, 13(1), p.e1005311.

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

Langmead, B., Salzberg, S., 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*. 9:357-359.

Latchman, D., 1997. Transcription factors: An overview. *The International Journal of Biochemistry & Cell Biology*, 29(12), pp.1305-1312.

Latchman DS, 1996. Inhibitory transcription factors. *The International Journal of Biochemistry & Cell Biology*. 28 (9): 965–74

Lay, F., Kelly, T. and Jones, P., 2017. Nucleosome Occupancy and Methylome Sequencing (NOMe-seq). *Methods in Molecular Biology*, pp.267-284.

L'honore, A., Lamb, N., Vandromme, M., Turowski, P., Carnac, G. and Fernandez, A., 2003. MyoD Distal Regulatory Region Contains an SRF Binding CArG Element Required for MyoD Expression in Skeletal Myoblasts and during Muscle Regeneration. *Molecular Biology of the Cell*, 14(5), pp.2151-2162.

Lu, R., Mucaki, E. and Rogan, P., 2016. Discovery and validation of information theory-based transcription factor and cofactor binding site motifs. *Nucleic Acids Research*, 45(5), pp.e27-e27.

Luan, J., Xiang, G., Gómez-García, P., Tome, J., Zhang, Z., Vermunt, M., Zhang, H., Huang, A., Keller, C., Giardine, B., Zhang, Y., Lan, Y., Lis, J., Lakadamyali, M., Hardison, R. and Blobel, G., 2021. Distinct properties and functions of CTCF revealed by a rapidly inducible degron system. *Cell Reports*, 34(8), p.108783.

Magnani L, Eeckhoute J, Lupien M. Pioneer factors: directing transcriptional regulators within the chromatin environment. *Trends Genet*. 2011;27(11):465–74.

Maity, G., Haque, I., Ghosh, A., Dhar, G., Gupta, V., Sarkar, S., Azeem, I., McGregor, D., Choudhary, A., Campbell, D., Kambhampati, S., Banerjee, S. and Banerjee, S., 2018. The MAZ transcription factor is a downstream target of the oncoprotein Cyr61/CCN1 and promotes pancreatic cancer cell invasion via CRAF–ERK signaling. *Journal of Biological Chemistry*, 293(12), pp.4334-4349.

Marmorstein, R. and Trievel, R., 2009. Histone modifying enzymes: Structures, mechanisms, and specificities. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1789(1), pp.58-68.

Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1), p.10.

Martin, P. and Zabet, N., 2020. Dissecting the binding mechanisms of transcription factors to DNA using a statistical thermodynamics framework. *Computational and Structural Biotechnology Journal,* 18, pp.3590-3605.

Mayran, A., Sochodolsky, K., Khetchoumian, K., Harris, J., Gauthier, Y., Bemmo, A., Balsalobre, A. and Drouin, J., 2019. Pioneer and nonpioneer factor cooperation drives lineage specific chromatin opening. *Nature Communications*, 10(1).

Meyer, C. and Liu, X., 2014. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics*, 15(11), pp.709-721.

Nagy, G., Czipa, E., Steiner, L., Nagy, T., Pongor, S., Nagy, L. and Barta, E., 2016. Motif oriented high-resolution analysis of ChIP-seq data reveals the topological order of CTCF and cohesin proteins on DNA. *BMC Genomics*, 17(1).

Nakato, R. and Sakata, T., 2020. Methods for ChIP-seq analysis: A practical workflow and advanced applications. *Methods*.

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

Pennacchio, L., Bickmore, W., Dean, A., Nobrega, M. and Bejerano, G., 2013. Enhancers: five essential questions. *Nature Reviews Genetics*, 14(4), pp.288-295.

Phillips JE, V, Corces G. CTCF: master weaver of the genome. *Cell*. 2009;137(7):1194–211.

Phillips-Cremins, J. and Corces, V., 2013. Chromatin Insulators: Linking Genome Organization to Cellular Function. *Molecular Cell*, 50(4), pp.461-474.

Rojano, E., Seoane, P., Ranea, J. and Perkins, J., 2018. Regulatory variants: from detection to predicting impact. *Briefings in Bioinformatics*, 20(5), pp.1639-1654.

Pop, R., 2021, and Nagy, D., 2020. Unpublished Master's theses, University of Essex.

Rosado-Rodríguez, E., Rodríguez-Ríos, J. and Rodríguez-Martínez, J., 2019. Uncovering DNA binding properties of the GATA4 and TBX5 transcription factor complex. *The FASEB Journal*, 33(S1).

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N., Huber, W., Haering, C., Mirny, L. and Spitz, F., 2017. Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, 551(7678), pp.51-56.

Shannon P, Richards M (2020). *MotifDb: An Annotated Collection of Protein-DNA Binding Sequence Motifs*. R package version 1.32.0.

Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford DK. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol*. 2014;32(2):171–8.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A., Gordân, R. and Rohs, R., 2014. Absence of a simple code: how transcription factors read the genome. *Trends in Biochemical Sciences*, 39(9), pp.381-399.

Sloan, C., Chan, E., Davidson, J., Malladi, V., Strattan, J., Hitz, B., Gabdank, I., Narayanan, A., Ho, M., Lee, B., Rowe, L., Dreszer, T., Roe, G., Podduturi, N., Tanaka, F., Hong, E. and Cherry, J., 2015. ENCODE data at the ENCODE portal. *Nucleic Acids Research*, 44(D1), pp.D726-D732.

Song, L. and Crawford, G., 2010. DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. *Cold Spring Harbor Protocols*, 2010(2).

Song, L., Zhang, Z., Grasfeder, L., Boyle, A., Giresi, P., Lee, B., Sheffield, N., Graf, S., Huss, M., Keefe, D., Liu, Z., London, D., McDaniell, R., Shibata, Y., Showers, K., Simon, J., Vales, T., Wang, T., Winter, D., Zhang, Z., Clarke, N., Birney, E., Iyer, V., Crawford, G., Lieb, J. and Furey, T., 2011. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Research*, 21(10), pp.1757-1767.

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), pp.555-568.

Sun, Y., Miao, N. and Sun, T., 2019. Detect accessible chromatin using ATAC-sequencing, from principle to applications. *Hereditas*, 156(1).

Taleb, A., Aisha, H., Al-Mosaib, S., Alnoud, M. and Souksavanh, V., 2019. Peak Calling Algorithms and Their Applications for Next-Generation Sequencing Technologies. *Indian Journal of Natural Sciences*, 9(52).

Teif, V., Vainshtein, Y., Caudron-Herger, M., Mallm, J., Marth, C., Höfer, T. and Rippe, K., 2012. Genome-wide nucleosome positioning during embryonic stem cell development. *Nature Structural & Molecular Biology*, 19(11), pp.1185-1192.

Tyagi, M., Imam, N., Verma, K. and Patel, A., 2016. Chromatin remodelers: We are the drivers!!. *Nucleus*, 7(4), pp.388-404.

Vandel, J., Cassan, O., Lèbre, S., Lecellier, C. and Bréhélin, L., 2019. Probing transcription factor combinatorics in different promoter classes and in enhancers. *BMC Genomics*, 20(1).

Vanzan, L., Soldati, H., Ythier, V., Anand, S., Braun, S., Francis, N. and Murr, R., 2021. High throughput screening identifies SOX2 as a super pioneer factor that inhibits DNA methylation maintenance at its binding sites. *Nature Communications*, 12(1).

Wang, J., Lawry, S., Cohen, A. and Jia, S., 2014. Chromosome boundary elements and regulation of heterochromatin spreading. *Cellular and Molecular Life Sciences*, 71(24), pp.4841-4852.

West, A., 2002. Insulators: many functions, many mechanisms. *Genes & Development*, 16(3), pp.271-288.

Wieczór, M. and Czub, J., 2017. How proteins bind to DNA: target discrimination and dynamic sequence search by the telomeric protein TRF1. *Nucleic Acids Research*, 45(13), pp.7643-7654.

Wiehle, L., Thorn, G., Raddatz, G., Clarkson, C., Rippe, K., Lyko, F., Breiling, A. and Teif, V., 2019. DNA (de)methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries. *Genome Research*, 29(5), pp.750-761.

Wingender, E., Schoeps, T., Haubrock, M. and Dönitz, J., 2014. TFClass: a classification of human transcription factors and their rodent orthologs. *Nucleic Acids Research*, 43(D1), pp.D97-D102.

Woodstock, D., Sammons, M. and Fischer, M., 2021. p63 and p53: Collaborative Partners or Dueling Rivals?. *Frontiers in Cell and Developmental Biology*, 9.

Wutz, G., Várnai, C., Nagasaka, K., Cisneros, D., Stocsits, R., Tang, W., Schoenfelder, S., Jessberger, G., Muhar, M., Hossain, M., Walther, N., Koch, B., Kueblbeck, M., Ellenberg, J., Zuber, J., Fraser, P. and Peters, J., 2017. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *The EMBO Journal*, 36(24), pp.3573-3599.

Yang, J., Bashkenova, N., Zang, R., Huang, X. and Wang, J., 2020. The roles of TET family proteins in development and stem cells*. Development*, 147(2), p.dev183129.

Yang, J. and Corces, V., 2011. Chromatin Insulators: A Role in Nuclear Organization and Gene Expression*. Advances in Cancer Research*, pp.43-76.

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

Zaret, K. and Carroll, J., 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes & Development*, 25(21), pp.2227-2241.

Zhang, Y., Liu, T., Meyer, C., Eeckhoute, J., Johnson, D., Bernstein, B., Nussbaum, C., Myers, R., Brown, M., Li, W. and Liu, X., 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9), p.R137.