Estimating transcription factor binding properties in human cell lines using statistical methods and genomics datasets

Romana T. Pop

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

School of Life Sciences

University of Essex

Submitted October 2020

Acknowledgements

I would like to thank my supervisor Dr. Radu Zabet for his patience and support throughout this project. Thank you also to all members of the Zabet lab, my friends and my family who were there to support me through the challenges of work. Not least of all, thank you to all the teachers who have supported me throughout the years and without whom I would not be here.

CONTENTS

ABSTRACT

Site specific transcription factors recognise and bind DNA motifs to regulate gene expression. Therefore, it is important to understand how and where they interact with the genome. Besides DNA sequence, chromatin accessibility, CpG methylation and cooperative binding with other transcription factors or themselves also impacts transcription factor binding. The era of high throughput sequencing has brought large amounts of genomic data, including chromatin immunoprecipitation and sequencing data for transcription factor binding. As a result, bioinformatics and machine learning tools have become popular for genomic data analysis. When investigating transcription factor activity, it is not enough to understand their function, but understanding the mechanisms behind it is also necessary. Explainable bioinformatics models facilitate the unravelling of mechanistic processes. ChIPanalyser is an R/Bioconductor package that implements a statistical thermodynamics model for transcription factor binding by leveraging binding motifs, chromatin accessibility and transcription factor concentration. This study aimed to use ChIPanalyser on 135 human transcription factors in the K562 cell line and investigate their chromatin accessibility preferences. Quantile density accessibility was used to determine how transcription factor binding changed when considering different levels of chromatin accessibility. In total, 12 quantiles were used and their goodness of fit was determined by AUC. The transcription factors were clustered into four groups based on their AUC trends over all quantiles using two algorithms: k-means and a bespoke algorithm. The four clusters were (i) “pioneer”, containing factors that were indifferent to variations in accessibility, (ii) “partial pioneer”, containing factors with a slight preference for open chromatin, (iii) “traditional”, containing factors with a strong preference for open chromatin, and (iv) “poorly predicted”, containing factors poorly predicted by the model regardless of accessibility. The two methods varied somewhat in their classification, with the “pioneer” and “partial pioneer” groups being larger when using the k-means. This study provided insight into the relationship between transcription factor chromatin accessibility preference and their function, and opened the possibility for further study.

Chapter 1: Introduction and literature review

1.1 Introduction

It is well established that gene expression is the driving force behind most cellular processes, from embryo development (Davidson, 2010; Peter and Davidson, 2011) to cancer development and progression (Yan *et al.*, 2016). Thus, understanding gene expression and the regulatory processes behind it is one of the fundamental problems in molecular biology today. Gene regulation is a complex process that encompasses several layers of expression control. One of the key regulatory elements of the genome are site specific transcription factors (TFs). These are proteins that are capable of binding to the DNA in a site-specific manner, often at specific regulatory sites, known as promoters and enhancers, and as a result of binding to DNA they can regulate gene expression (Spitz and Furlong, 2012; Lambert *et al.*, 2018).TFs vary drastically in function, ranging from “master regulators”, involved in cellular differentiation (Bürglin, 2001), to TFs involved in regulating specific pathways, such as the immune response pathways (Singh, Khan and Dinner, 2014). Expression of TFs also varies greatly, with some TFs being ubiquitously expressed (Adcock and Caramori, 2009), while others are expressed in a cell-type specific manner (Lambert *et al.*, 2018). Furthermore, the same TF can have different functions depending on cell type (Lambert *et al.*, 2018). Due to their heterogeneity as well as their role at the heart of gene expression, there has been a great effort made in recent years to identify and annotate regulatory sites (Libbrecht *et al.*, 2019; Meuleman *et al.*, 2019) and capture their interaction with TFs (Spitz and Furlong, 2012).

In order to control gene expression, TFs must bind to the chromatin at specific short sequences known as “motifs”. TFs recognise and bind their motifs with much higher affinity than any other sequence (Geertz, Shore and Maerkl, 2012). Therefore, TF function is closely linked to DNA sequence and it is necessary to understand how and where they interact with the genome in order to form an understanding of gene regulation. There is a wide array of methods, both in vitro and in vivo, that can be used to determine TF binding sites (Jolma and Taipale, 2011). Chromatin immunoprecipitation (ChIP) based techniques are among the more widely used. Initially, microarrays were used to determine binding of proteins to DNA, in a technique known as ChIP-chip. However, with the advent of next generation sequencing (NGS), came the possibility for higher throughput techniques. Starting in 2007 when they were first published (Johnson *et al.*, 2007), chromatin immunoprecipitation followed by sequencing (ChIP-seq) experiments have become one of the primary techniques for determining TF binding in vitro (Park, 2009). Since then, thousands of ChIP-seq datasets have become available in repositories such as The Encyclopaedia of DNA Elements (ENCODE) (Dunham *et al.*, 2012; Davis *et al.*, 2018).

However, DNA sequence is not the only factor influencing TF binding and the presence of a motif is not sufficient for a TF to bind. Indeed, there is only a partial overlap between motifs present in the genome and actual binding sites, while the location of binding sites themselves

is not necessarily an indicator of what genes are being regulated (Cusanovich *et al.*, 2014). Thus, other factors besides DNA sequence must influence TF binding. Cooperative binding of TFs is a well-established concept (Villar, Flicek and Odom, 2014), with most TFs requiring some co-factors such as epigenetic modifiers, other TFs or even themselves, in homotypic complexes, in order to function correctly (Ravasi *et al.*, 2010). Furthermore, many TFs can regulate different sets of genes (and thus bind different sites) depending on their concentration (Chu *et al.*, 2009; Dangkulwanich *et al.*, 2014; Zabet and Adryan, 2015; Abascal *et al.*, 2020).

In addition to this, TFs in humans and other higher eukaryotes also face the challenge posed by the complex structure of chromatin. Histones are DNA binding proteins that aid the formation of nucleosomes and are strong competitors for DNA binding sites (Bai and Morozov, 2010). Indeed, nucleosomes have long been known to impede TF binding and most TFs bind in nucleosome depleted regions (NDRs), while nucleosome rich regions are associated with transcriptionally inactive chromatin (Lee *et al.*, 2004; Shivaswamy *et al.*, 2008). Thus, nucleosome positioning and DNA accessibility is another significant factor influencing TF binding.

1.2 Pioneer transcription factors and chromatin accessibility

Eukaryotic genomes have a complex structure which poses an obstacle to TF binding. The chromatin forms into nucleosomes by wrapping around a core of 8 histones, thus rendering DNA inaccessible to TFs both through the tightening of chromatin and through competition with histones for binding sites. Nucleosome positioning is not uniform throughout the genome, with transcriptionally active regions being generally nucleosome depleted, while transcriptionally inactive regions being generally nucleosome rich (Tsompana and Buck, 2014). A number of assays exist for investigating chromatin accessibility (thoroughly reviewed in Tsompana and Buck, 2014). The main assays used for this purpose can be split between endonuclease cleavage assays such as DNase-seq and MNase-seq, and assays that do not rely on endonucleases, such as FAIRE-seq (Formaldehyde-Assisted Isolation of Regulatory Elements with sequencing), ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) and NOMe-seq.

The two main endonucleases used for accessibility assays are DNase I and MNase. The former, a double strand, non-specific endonuclease, has been traditionally used in accessibility assays and it preferentially cleaves nucleosome depleted regions. NGS can then be used to sequence the fragments and reconstruct a chromatin accessibility profile (Weintraub and Groudine, 1976; Tsompana and Buck, 2014). DNase-seq has been widely used, especially by the ENCODE consortium to elucidate cell-type specific chromatin accessibility. MNase is a single-strand specific endonuclease that cleaves at inter-nucleosomal sites. It is used in conjunction with NGS to reconstruct a footprint of nucleosomal occupation, and thus indirectly, of chromatin accessibility (Tsompana and Buck, 2014).

However, MNase cleavage is concentration dependent and thus assays can be unreliable unless conditions are tightly regulated.

In addition to endonuclease digestion, several other methods of assessing chromatin accessibility have been developed. FAIRE-seq uses formaldehyde crosslinking of chromatin, followed by DNA fragmentation through sonication. This is followed by phoenol-chloroform extraction during which the nucleosome depleted DNA fragments are relaeased into the aqueous phase of the solution. This is due to histone crosslinking being more efficient than crosslinking of other DNA-binding elements (Tsompana and Buck, 2014). More recently, Nucleosome Occupancy and Methylome Sequencing (NOMe-seq) was developed (Lay, Kelly and Jones, 2018) to measure the relationship between DNA methylation and nucleosome occupancy. The assay uses the methyltransferase M.CviPI which metylates CpG nucleotides unprotected by nucleosomes, thus yielding a chromatin accessibility profile (Lay, Kelly and Jones, 2018).

While the majority of TFs cannot bind nucleosome-rich chromatin, a subset of TFs known as pioneer factors have the ability to interact with nucleosomes and bind their cognate DNA. Not only that, they are also able to displace nucleosomes and make way for other TFs to bind without the use of ATP-dependent chromatin re-modelers (Cirillo *et al.*, 2002). The first pioneer factors to be discovered were FOXA1 and GATA4, two TFs that play an important role in endoderm formation during embryogenesis (Gualdi *et al.*, 1996; Bossard and Zaret, 1998). Later, more members of the FOX and GATA family of TFs were identified as being pioneers (Bossard and Zaret, 1998; Cirillo *et al.*, 2002; Rojas *et al.*, 2005). Several other TFs involved in a variety of processes have been identified as having pioneer function (reviewed in Lai *et al.*, 2018). Of note is the role of pioneer factors in cellular reprogramming. A classic example of this is the pluripotency factors (SOX2, OCT3/4, KLF4 and c-MYC), which are able to de-differentiate cells and essentially revert them to a pluripotent state. These factors have been shown to be pioneers, with the exception of c-MYC which does not appear to be a pioneer itself, but is a co-factor that enhances the activity of SOX2/OCT4 (Soufi, Donahue and Zaret, 2012; Mayran and Drouin, 2018).

In addition to reverting cells to a pluripotent state, pioneer factors also facilitate trans-differentiation of committed cell types into other cell types. For example, PU.1 when expressed in megakaryocyte/erythroid cells, triggers their conversion into myeloid cells, while its absence causes the reverse transformation, often facilitated by GATA1 (Graf, 2002). Furthermore, when collaborating with C/EBPα/β, PU.1 mediates the conversion of fibroblasts, pre-B and T cells into macrophages (Xie *et al.*, 2004; Laiosa *et al.*, 2006; Feng *et al.*, 2008). GATA4 works alongside co-factors MEF2C, and TBX5 in fibroblasts to trigger their transformation into cardiomyocytes (Ieda *et al.*, 2010). Other pioneers of note include tumour suppressor p53, which regulates chromatin state in a variety of tissues (Nili *et al.*, 2010; Sammons *et al.*, 2015; Younger and Rinn, 2017); PAX7, involved in the development of the pituitary (Budry *et al.*, 2012) and ASCL, which is involved in neurogenesis (Wapinski *et al.*, 2013; Park *et al.*, 2017).

The mechanisms by which pioneer factors recognise their motifs in closed chromatin differ between pioneers. For example, FOXA1 was found to travel more slowly through cell nuclei compared to other TFs and it has been postulated that this is due to scanning of heterochromatin for binding sites (Iwafuchi-Doi *et al.*, 2016). Crystal structure analysis of FOXA1 bound to its domain has revealed a triple α-helix structure similar to that of linker histone H5 (Clark *et al.*, 1993). Once it identifies a binding motif, FOXA1 then is able to displace the linker histone H1, thus preventing nucleosomes from aggregating into complex chromatin structures (Iwafuchi-Doi *et al.*, 2016). In contrast, the pluripotency factors SOX2, OCT3/4 and KLF4 are able to recognise and bind partial motifs in nucleosomal DNA and their pioneer activity is boosted by cooperating with each other (Soufi *et al.*, 2015). Notably, when investigating the affinity for nucleosomal DNA of 2-mers of the pluripotency factors, Soufi et al., (2015) found that OCT4 acts as a strong potentiator of binding nucleosomes, while combinations lacking OCT4 exhibited weaker affinity for nucleosomes.

Nucleosome positioning also plays a role in pioneer recognition of their binding motifs. For example, studies have shown that the pioneer tumour suppressor p53 preferentially binds motifs found towards the edges of nucleosomes (Laptenko *et al.*, 2011), unlike FOXA1, which prefers binding sites near the centre of nucleosomes (Iwafuchi-Doi *et al.*, 2016). Furthermore, the rotational positioning of motifs on the nucleosome is also important for binding, with p53 binding preferentially to motifs that are facing the outside of the nucleosome (Sahu *et al.*, 2010; Cui and Zhurkin, 2014; Iwafuchi-Doi, 2019).

“Bookmarking” of binding sites by pioneers (i.e. the continued occupancy of pioneer factors at their binding sites even during transcriptionally inactive phases such as during mitosis) is a mechanism for quick reactivation of transcription sites after cell division. This behaviour has been observed in several TFs such as FOXA1 (Caravaca *et al.*, 2013), GATA1 (Kadauke *et al.*, 2012), as well as the pluripotency factors (Liu *et al.*, 2017) and it serves to maintain cellular differentiation, or lack thereof, in the case of the pluripotency factors.

The mechanisms by which pioneer factors open chromatin are not as well understood. While it has been shown that FOXA1 mediated chromatin relaxation does not require ATP-dependent chromatin re-modelers, it is unclear whether no re-modelers are recruited (Cirillo *et al.*, 2002; Iwafuchi-Doi, 2019). In the case of FOXA1, there is evidence for it directly causing chromatin relaxation through the displacement of linker histones (Iwafuchi-Doi *et al.*, 2016). Other factors, such as OCT4, are known to recruit chromatin re-modelers to their binding site in order to facilitate chromatin opening. For example, several studies (Pardo *et al.*, 2010; van den Berg *et al.*, 2010) identified interactions of OCT4 with the SWI/SNF complex of chromatin re-modelers, particularly Brg1. Later, King and Klose, (2017) showed that OCT4 mediated chromatin accessibility requires the activity of Brg1 which is recruited by OCT4 at its binding sites.

Despite their crucial function at the heart of gene regulation, only a handful of TFs have been identified as pioneer. Some of these have been characterised in depth, however it remains

unclear whether there are other TFs that exhibit similar properties, and to what extent. One interesting example is that of CTCF, which appears to have the ability to displace nucleosomes after cell division and maintain NDRs in some contexts (Owens *et al.*, 2019), while other times its binding is inhibited by the presence of nucleosomes (Teif *et al.*, 2014; Wiehle *et al.*, 2019). This indicates that the relationship between TFs and chromatin is complex and context dependent and more investigation is needed to fully understand it and to determine whether other TFs with similar context dependent binding profiles exist.

1.3 Chromatin methylation and transcription factor binding

In addition to DNA sequence and chromatin accessibility, there are other factors known to influence TF binding. One significant such factor is chromatin methylation. A common type of DNA methylation occurs at the cytosine of CpG dinucleotides. Most CpG sites in the human genome are methylated, however patterns can very between cell types (Bird, 2002), as well as in disease (Robertson, 2005). CpG methylation is not uniform, with CpGs in nucleosome rich regions being generally less methylated than CpGs in the more accessible linker regions (Huff and Zilberman, 2014). Furthermore, methylation is highly abundant in gene bodies (Ball *et al.*, 2009), while regulatory sites remain relatively unmethylated (Hon *et al.*, 2013). The methylome of a cell is inherited through cell division with the aid of the methyl-transferase DNMT1, which essentially copies the methylation patterns of an existing DNA strand onto the newly synthesised one (Schübeler, 2015), and this is thought to contribute to the maintenance of expression patterns in different cell types.

Methylated DNA has traditionally been associated with transcriptionally inactive regions and can directly block TF binding (Watt and Molloy, 1988; Iguchi-Ariga and Schaffner, 1989; Gaston and Fried, 1995), or indirectly interfere with TF binding by recruiting other epigenetic factors that trigger chromatin condensation (Schübeler, 2015). However, not all TF binding is inhibited by CpG methylation and some TFs even preferentially bind methylated sites. For example, both YY1 and ETS family TFs are known to regulate the Surf-1 and Surf-2 genes. YY1 binds to the Su1 promoter, which triggers transcription of both Surf-1 and Surf-2 genes, while ETS binds the Su2 promoter regulating the same genes. However, YY1 binding remained unaffected by methylation of CpG sites in the Su1 promoter, while methylation of the Su2 promoter prohibited ETS from binding (Gaston and Fried, 1995). Thus, the same genes can be differentially impacted by CpG methylation. Sp1, which is involved in regulation of housekeeping genes is another TF that is not hindered by CpG methylation (Harrington *et al.*, 1988), while CTCF, a chromatin remodeler and insulator, has been shown to be methylation sensitive (Hnisz, Day and Young, 2016; Wiehle *et al.*, 2019).

Some TFs have been found to have different regulatory activity based on the methylation state of its binding motif (Hu *et al.*, 2013). For example, when testing KLF4 binding to several methylated and unmethylated motifs in vitro, it was discovered that KLF4 was able to only bind the methylated version of the M197 (TCCmCpGCCC) motif, but not its unmethylated

form, while the reverse was true for motif M412 (GCTTTTACG). TFAP2A, ARID3B, and ZMYM3 were also found to behave similarly (Hu *et al.*, 2013).

Interestingly, when investigating the occupancy of several pioneer factors, Donaghey *et al.*, (2018) found that DNA demethylation takes place at a subset of FOXA2 binding sites after being bound by it, albeit in a replication dependent manner, suggesting that one mechanism of pioneer-mediated chromatin accessibility could be through promoting de-methylation. Thus, DNA methylation plays an important and diverse role in TF binding and a better understanding of how each TF is impacted by it is needed.

1.4 Computational models for transcription factor binding

With the advent of high-throughput technologies came a large influx of genomic data and along with it, bioinformatic tools and machine learning has become more and more popular for genomic data analysis. Due to the drastic increase in computational power over the past two decades, there has been a renewed interest in machine learning algorithms such as artificial neural networks (ANN) and deep neural networks (DNNs), which can now be used to predict with high accuracy many types of data. Furthermore, they are able to deal with large datasets, as genomic data often is.

There are a number of tools that have been designed in the past few years that model TF binding in silico. One of the earliest such tools was CENTIPEDE (Pique-Regi *et al.*, 2011), a probabilistic tool that considered DNase I hypersensitivity, histone modifications and DNA sequence to identify TF binding sites. CENTIPEDE works based on two assumptions (i) TF binding leaves DNA more vulnerable to cleavage by DNase I and (ii) the cleavage profiles around their bound sites is different and characteristic to each TF (Pique-Regi *et al.*, 2011). However, this model does not consider variation in the DNase I profile between different sites bound by the same TF, nor variation between DNase I replicates at the same site. msCENTIPEDE (Raj *et al.*, 2015) was therefore later released to improve on the CENTIPEDE framework by using multi-scale models to model this variability.

1.4.1 Linear regression models

Some of the simplest types of machine learning algorithms are linear regression (LR) models. LR models can predict a dependent variable based on an independent variable by fitting data to the equation of the straight line ($y = mx + c$) (Wasserman, 2004). Due to their relative simplicity, regression algorithms can provide explainable biological models. However, regression models have difficulties when it comes to large, multivariate data. Thus, feature selection must be performed on the data before analysis. Several methods for feature selection have been proposed, including the lasso regression (Tibshirani, 1996) and least angle regression selection (LARS), which is a modified version of the lasso (Efron *et al.*, 2004). (Yuan and Lin, (2006) later proposed an extension to both the lasso and LARS models that was more robust to parameter orthonormalization and termed them group lasso and group LARS. Such algorithms can be very efficient at narrowing down the most relevant features. For example, Narlikar *et al.*, (2010) used lasso regression to narrow their data down

from 727 features to 45 when investigating heart enhancers. However, 45 is still a large number of features for regression models and thus the model accuracy suffered. Despite their drawbacks, regression models remain attractive due to their explainability and simplicity and improved algorithms continue to be proposed.

In 2017, Zhou *et al.* developed Big Data Regression for predicting DNase I hypersensitivity (BIRD), a regression model that can use RNA-seq data to predict DNase I hypersensitivity sites. In order to bypass the big data problem that is a steeple of regression models, BIRD clusters similar predictors (i.e. co-expressed genes) together and converts each cluster into a single predictor, thus effectively reducing the dimensionality of the data and optimising computation time. Indeed, when comparing their model to another previously proposed regression model (Yuan and Lin, 2006), they found their model was 105 times faster while maintaining similar predictive power. However, both these algorithms are based on reducing data dimensionality, meaning that information, and therefore predictive power and accuracy, are lost.

Therefore, while regression models might not be able to handle high dimension data and other tools are often better for genomic analysis, such models might be useful in analysing datasets with small cell numbers when it may not be possible to get larger samples. Such might be the case with clinical samples or cell types that are difficult or expensive to obtain in large numbers.

1.4.2 Artificial neural network and deep learning models

When large samples are available, artificial neural networks (ANNs) are much more powerful. Unlike LR models, ANNs can easily handle large, multivariate datasets. ANNs are designed to mimic the way neurons interact in the human brain and are comprised of several interconnected layers of “neurons” that pass information from one to the next and are thus able to learn patterns. Deep neural networks (DNNs) are similar to ANNs, and indeed have evolved from them, but they are more complex and contain many more neuronal layers. Each layer is trained to recognise increasingly complex patterns by compounding on the outputs of the previous layers (Glorot and Bengio, 2010). Because of this, they are able to very accurately predict complex and multidimensional data. Convolutional neural networks (CNNs) are a type of deep learning algorithms that have seen many applications in genomics. CNNs are well suited for analysing multi-dimensional data, such as images (2D), or videos (3D) but are also applicable to one dimensional data such as genomic sequences (Lecun, Bengio and Hinton, 2015; Angermueller *et al.*, 2016; Telenti *et al.*, 2018).

One of the first CNNs to be applied in a genomic context was DeepBind (Alipanahi *et al.*, 2015), an algorithm designed to identify DNA and RNA binding sites. Since then, several other CNNs have been used to investigate TF binding. Phuycharoen *et al.*, (2019), compared the performance of several k-mer methods, a shallow, 1-layer CNN and a deeper CNN when predicting cooperative TF binding by looking at the differential binding in 3 mouse cell lines. To test the different methods, they looked at the MEIS family of TFs and HOXA2, which are

known to cooperatively bind in some mouse branchial tissues. By predicting differential MEIS binding in three different branchial tissues, they were able to predict HOXA2 co-binding sites. Their study indicates that deep networks outperform both shallow networks and k-mer methods and highlights several of the advantages of using deep networks. Namely, deep models are better able to learn the context of motifs, which the commonly used annotation of the genome with a position weight matrix (PWM) (used by k-mer methods) remains insensitive to. Furthermore, deep networks were shown to lower the rate of false positives when compared to k-mer methods. Quang and Xie, (2019) developed FactorNet, a CNN developed through the ENCODE DREAM challenge, that incorporates genomic sequences, genome annotations, gene expression and accessibility data in order to predict TF binding sites in various cell types.

Another type of DNNs that have received some interest in the context of genomics are recurrent neural networks. These are a type of DNN that have feedback loops, thus the input for each iteration of the network is the output of the previous iteration; recurrent neural networks are especially useful for learning relationships in time. In 2016, Quang and Xie, combined CNNs with recurrent neural networks and developed DanQ, an algorithm comprised of a convolutional layer that identifies patterns within the sequence (i.e. motifs), and a recurrent layer which determines long-term dependencies between the motifs. They found that the hybrid algorithm outperformed the previously published DeepSEA model, which was purely CNN. However, when comparing the motifs identified by DanQ to known motifs, only ~50% of them were significantly matched.

As highlighted above, there are many advantages in using DNNs and they are becoming more and more used in the literature. However, DNNs require large training datasets which may not always be possible to obtain. Furthermore, training DNNs can be highly ineffective if the network architecture is not suitable for the intended task and therefore selecting the appropriate architecture is crucial. Overfitting is also a concern with machine learning in general, but especially with DNNs which, due to their great success might be tempting to use in situations where they are unnecessary and simpler models would suffice.

1.4.3 Unsupervised learning models

The models discussed above are “supervised” models, meaning that they must first be trained on a “known” dataset before they are able to predict patterns in “unknown” data. However, genomic data is often unlabeled and heterogeneous and therefore supervised learning methods may not be successful, while deep algorithms may be inefficient to use for identifying distinct subpopulations. This is where unsupervised clustering algorithms are useful and have received some interest.

Single cell sequencing technologies offer the opportunity of studying heterogeneity both between cells and between tissue types. However, such data is difficult to analyse due to sparsity. As a result, it is common for cells to be sorted experimentally, using technologies such as FACS and performing bulk sequencing on the sorted subpopulations to guide the

analysis. In order to bypass such costly and time consuming practices, Zamanighomi *et al.*, (2018) developed a tool called single cell Accessibility Based Clustering (scABC), which uses a weighted K-medioids unsupervised clustering algorithm to computationally separate cell subpopulations based on scATAC-seq data. Using scABC, they were able to distinguish between 6 established cell lines, as well as distinguish between subpopulations of cell lines and primary cells and were able to identify cell-type specific accessibility patterns. Using their accessibility profiles, they were the able to then identify cell-type specific TFs by interrogating their clusters with chromVAR, a tool developed by Schep *et al.*, (2017) that can identify TF motifs in accessible chromatin regions.

In 2018 de Boer and Regev developed BROCKMAN, an unsupervised learning tool similar to chromVAR, that can infer TF binding from the changes in chromatin structure around their binding motifs. BROCKMAN accounts for three groups of TFs: TFs that impact chromatin structure as they bind in a concentration dependent manner, TFs that do not impact chromatin structure as they bind, and TFs that bind cooperatively. Both BROCKMAN and chromVAR use a k-mer approach. Simply, a k-mer is a DNA word or motif of length k that is recognised by a TF. When the TF binds the motif, it causes chromatin changes in the vicinity. Thus by associating chromatin changes to motifs, active sites can be identified. Unlike chromVAR which uses ungapped 7-mers, BROCKMAN uses gapped 8-mers, which potentially makes it more robust to TF motif variability. BROCKMAN was able to distinguish between cell types as well as subpopulations within cell types and to identify TFs that had differential activity between groups. The following year, Jansen *et al*. published another tool (SOMatic) that investigates how changes in chromatin accessibility and changes in gene expression are linked. Unlike de Boer and Regev's k-mer approach, SOMatic uses linked self-organising maps (SOMs) to this end, pointing out that the use of 8-mers might miss TFs with short motifs.

1.4.4 Explainable TF binding models

The main limitation of the types of algorithms described above is that they are essentially “black box” models (Loyola-Gonzalez, 2019), meaning they are not easily interpretable and thus it is difficult to gain mechanistic insights from them.

Thus, more explainable models are needed for a thorough understanding of the mechanisms behind TF binding. There have been several models attempting to mechanistically describe TF binding using statistical thermodynamics (Bintu *et al.*, 2005; Roider *et al.*, 2007; He *et al.*, 2009; Zhao, Granas and Stormo, 2009).

Zabet and Adryan, (2015) developed a model based on the statistical thermodynamics framework which can predict TF binding profiles on a genomic scale based on four parameters: (i) a weighted DNA binding motif referred to as a PWM, (ii) DNA accessibility data, (iii) the number of molecules bound to the DNA (determined experimentally or predicted) and (iv) a factor that modulates TF specificity.

The first two parameters are relatively easy to obtain; PWMs (w) for many TFs are collected in online repositories such as JASPAR (Sandelin *et al.*, 2004), and when not available, can be determined from ChIP-seq experiments with tools such as MEME (Bailey *et al.*, 2006) and Homer (Heinz et al., 2010); DNA accessibility (a) data from DNase I-seq or ATAC-seq experiments is also readily available for many cell types. The latter two parameters, namely the scaling factor λ and the number of bound molecules N, respectively, can be more difficult to measure experimentally; however, the model is able to predict a set of optimal parameters by fitting to already existing ChIP-seq data. Martin and Zabet developed an R/Bioconductor package called ChIPanalyser (Martin and Zabet, 2020), which can accurately recover ChIP-seq profiles based on this model. When comparing ChIPanalyser with several other tools such as msCENTIPEDE, PIQ and FactorNet, they found that ChIPanalyser outperforms all of them when trained on the same data (Abascal *et al.*, 2020).

This study aims to investigate the binding profiles and chromatin accessibility preferences of human TFs using ChIPanalyser, by first training the model on bulk ChIP-seq data in order to estimate TF binding parameters and then using those to determine whether TFs prefer to bind open or nucleosome associated DNA.

Chapter 2: Methods

2.1. Data gathering and pre-processing

Raw human TF ChIP-seq data was downloaded from ENCODE (Dunham *et al.*, 2012; Davis *et al.*, 2018) for 244 TFs available in the K562 cell line at the time (prior to the ENCODE phase 3 release (Abascal *et al.*, 2020)). This cell line was chosen because it had the most available TF ChIP-seq data available on ENCODE at the time. For each TF, all the available replicates at the time were downloaded, however those treated with anything were excluded from further analysis. Furthermore, TFs for which a PWM motif was not available in any online repositories were also excluded from the dataset. All metadata information for the downloaded data can be found in supplementary data Table A1 in the Appendix. The final number of TFs after triage was 135, a list of which is available in Table A2 in the Appendix. Where multiple experiments were available for one TF, the data were concatenated into one file. All scripts used for pre-processing and further analysis can be accessed at https://github.com/rtpop/MSD.

Once the data was downloaded and triaged, fastqc v. 0.11.7 (Andrews, 2010) was then used to assess the quality of the data, followed by removal of the Illumina adapters with trimmomatic v. 0.39 (Bolger, Lohse and Usadel, 2014)(Bolger, Lohse and Usadel, 2014). After trimming, the data was aligned to the human genome (hg38) (NCBI, 2019) downloaded Oct. 2019 from NCBI using bowtie2 v. 2.3.4.1 (Langmead and Salzberg, 2012). The original experimental controls for the ChIP-seq were downloaded from ENCODE (see Table A1 in the Appendix) for each TF and processed in the same way. As before, where there were multiple

control files for one TF, they were concatenated. Initially, SAM files were converted to BAM using samtools, however, for some files this posed an issue with peak calling, so SAM files were used instead. Finally, narrow peaks were called with macs2 (Zhang *et al.*, 2008), with a q-value threshold of 0.05.

In addition to the TF ChIP-seq data, chromatin accessibility data in the form of DNase I hypersensitivity data was also downloaded from ENCODE for the K562 (ECACC 89121407) cell line (experiment accession ENCSR000EOT). This was processed in the same way as the ChIP-seq data until the peak calling stage, where broad peaks were called with a q-value threshold of 0.1 instead.

After peak calling was complete, the accessibility data was subset with a quantile vector (0-0.9, 0.95, 0.99) in R. This resulted in multiple quantile density accessibility (QDA) files, each considering the top 1-n highest scoring regions (based on the DNase peak scores) in the original data as accessible, regardless of their actual accessibility scores; for example QDA 0 considered all the genome accessible (i.e 1-0=1 =>100% of the regions considered accessible), while QDA 0.99 considered only the top 1% of regions as accessible (i.e. 1-0.99 = 0.01 => 1% of regions considered accessible). These QDAs were used for further analysis.

Finally, PWMs were downloaded for all TFs, primarily with the use of the R Bioconductor package MotifDb (Shannon and Richards, 2018). The databases available on MotifDb were queried in the order shown in Table 1.

Where multiple PWMs were available from the same database, the first one was selected; if a PWM was not found in either of the above databases, but there was one from a different database, then the first motif returned by MotifDb was selected. For the TFs that did not have a PWM on MotifDb, the Tf2dna database (Pujato *et al.*, 2014) was downloaded and the PWM was taken from there. If a motif was not available there either, the TF was removed from the analysis. All motif logos are available in the Appendix.

Table 1. Order in which motif databases on MotifDb were queried

Jaspar2018
SwissRegulon
HOCOMOCOv10
Cisbp_1.02
Jolma2013

2.2. Estimating binding parameters with ChIPanalyser

ChIPanalyser (Martin and Zabet, 2020) is a bioinformatics tool that implements an approximation of the statistical thermodynamics framework (Zabet and Adryan, 2015) to estimate ChIP-seq like profiles based on four parameters: (i) a weighted DNA binding motif referred to as a position weight matrix (PWM), (ii) DNA accessibility data, (iii) the number of molecules bound to the DNA (determined experimentally or predicted) and (iv) a factor that modulates TF specificity. The model outputs the probability that a TF is bound to a site j, which is given by the statistical weight of site j divided by the total statistical weight, which is given by the sum of the statistical weight when the site İs unoccupied and the statistical weight when the site is occupied. as shown in (Zabet and Adryan, 2015):

$$P_j^{bound}(\lambda, w, N, a) = \frac{N \cdot a_j \cdot e^{\frac{1}{\lambda} w_j}}{N \cdot a_j \cdot e^{\frac{1}{\lambda} w_j} + L \cdot n \cdot \langle a_i \cdot e^{\frac{1}{\lambda} w_j} \rangle_i} \quad (1)$$

Where N is the number of TF molecules bound to the genome, aj is the DNA accessibility at site j, λ is the specificity scaling factor, w is the PWM score and L and n are the length and ploidy of the genome, respectively, while the genome-wide average weight is represented by:

$$L \cdot n \cdot \langle a_i e^{(\frac{1}{\lambda} w_j)} \rangle_i$$

Where i indicates the term is a genome-wide average.

The The PWM along with λ are used to determine the binding energy of the TF at site j, as given by:

$$binding\, energy = \frac{1}{\lambda} w_j \quad (2)$$

Thus, a lower λ value indicates a high affinity of the TF for its motif, while a high λ value indicates low affinity of the TF for the motif.

The PWMs and chromatin accessibility data were downloaded from third parties as described in the previous section; ChIPanalyser was used to estimate the number of bound TF molecules (N) and the specificity factor (λ) for each TF using the ChIP-seq data. The analysis was run 12 times, once for each QDA. For each run of the analysis, all parameters remained constant, with the exception of the QDA; all the parameters can be found in the GitHub repository. If not specified, then they were left as default.

First, the processingChIP function was used to extract ChIP scores at each locus of interest. This requires either a GRanges object containing loci of interest, or the path to a bedGraph (BDG) file from where the top n loci will be extracted after binning the genome into 50kb bins. The number of loci to select and the bin width can be specified by the user. BDG files were used for this analysis with n = 60 and a bin size of 50kb. In addition, the function also takes as arguments a peaks file and DNA accessibility data. For this analysis, this was the QDAs. Finally, a noise-filtering method can also be specified, in this case a “sigmoid” filter was used, which applies a logistic weighting to each score (Martin and Zabet, 2020).

Next, the optimal N and λ were computed with the computeOptimal function. The top 10 regions were used for training the model. The input for this function is the training regions, a PWM, the genome sequence and accessibility data. As before, QDAs were used as accessibility data.

Following the parameter estimation based on the top 10 regions, validation of these parameters was performed on the subsequent 50 loci (as outputted by processingChIP) using 12 goodness of fit metrics: correlation coefficients (Pearson, Spearman and Kendall), Mean Squared Error (MSE), Kolmogorov-Smirnov Distance, precision, recall, accuracy, F-score, Matthew’s correlation coefficient (MCC) and Area Under Curve Receiver Operator

Characteristic (AUC ROC or just AUC) (Martin and Zabet, 2020). However, it was previously shown that the best metric for training is MSE (Martin and Zabet, 2020), while AUC was the best metric for validation.

2.3. Transcription factor clustering

After the parameter training and validation was complete for each QDA, the corresponding AUC was extracted and this was repeated for each TF (see Table A3 in the Appendix). The AUC was calculated as detailed in (Martin and Zabet, 2020). Briefly, the genome was first sectioned into 100 bp bins and the peaks were sliced at different heights. At each level, the number of true positives and the number of false positives were used to construct a sensitivity curve, and the area under the sensitivity curve was calculated. The AUCs for each TF for each QDAs were then combined into a matrix with TFs as rows and QDAs as columns. This was then used to cluster the TFs based on their accessibility preference profile.

Two methods were used for the clustering: k-means and a threshold-based algorithm. The k-means clustering was performed with the native kmeans function in R. This requires as input a matrix or data frame with the data to be clustered, as well as a value k for the number of clusters. The function then generates k random cluster “centres” and assigns each data-point to the cluster whose “centre” is nearest to it. The value of the “centre” is then updated by calculating the mean between the previous value and the newly added data-point. Another parameter of the kmeans function is the “number of starts”, which defines how many random starting points should be attempted. For this analysis, this was set to 20, meaning that the

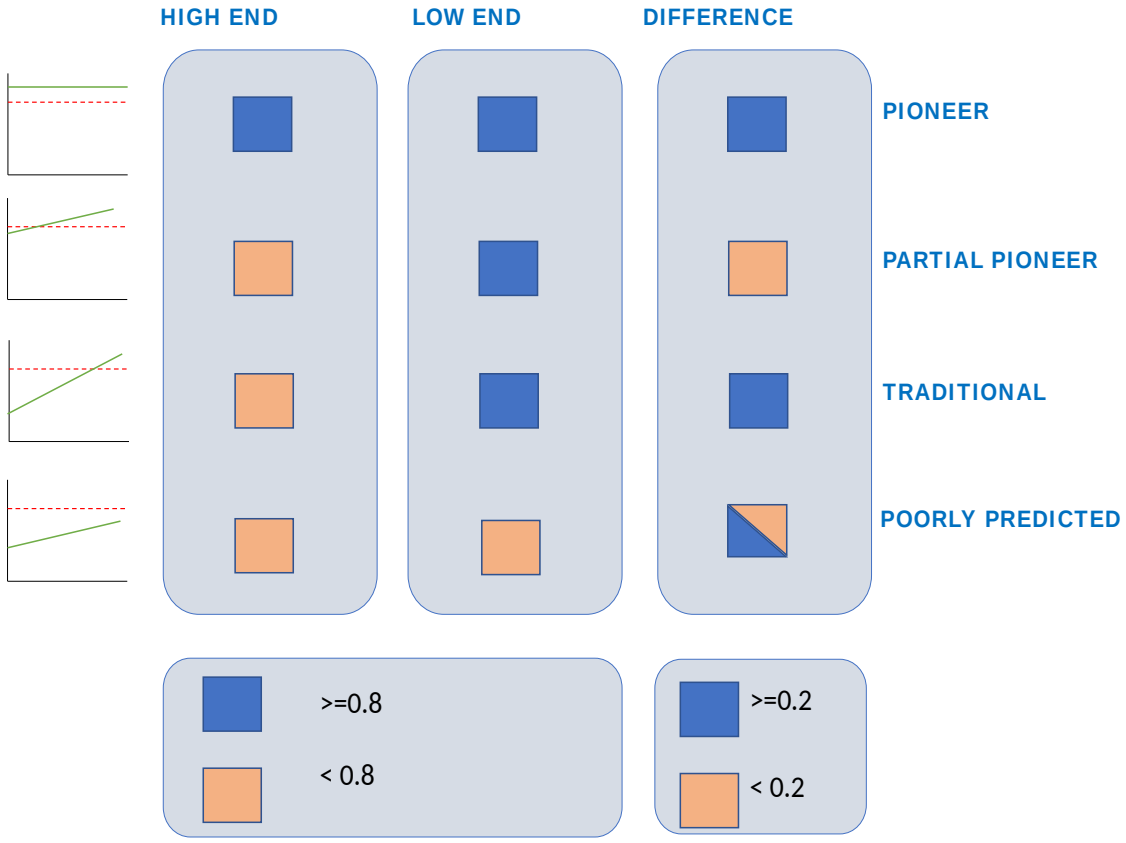

Figure 1. Graphical illustration of the threshold-based clustering algorithm.

algorithm was performed 20 times with different initial cluster centres, after which the best one was selected based on the lowest within cluster variation.

As mentioned, the number of clusters k must be provided by the user. In order to determine the optimal number of clusters for this data, 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 an “elbow plot”, with the “bend” in the plot indicating the optimal number of clusters for a given dataset. In this case, k=4 was selected based on the elbow plot (Figure 5A) and the four clusters were labelled as “pioneer”, “partial pioneer”, “traditional” and “poorly predicted” based on their trends. In order to reproduce the results, the seed used was always 13.

For comparison, a bespoke threshold-based algorithm was also used to cluster the TFs. Briefly, the algorithm considered an AUC of 0.8 to be the threshold for a TF being well predicted by the model. First, an average of the AUC for the low end QDAs (0-0.2) and of the high end QDAs (0.8, 0.9, 0.99) was calculated. Then the difference between the high end and low end averages was calculated and the TFs were sorted into clusters as shown in Figure 1. If both the high end and low end averages were above or equal to the 0.8 threshold, the TF was sorted into the pioneer cluster; if the low end average was below the threshold and the high end average was above it and the difference between them was below or equal to 0.2, the TF was sorted into the “partial pioneer” cluster. If the low end was below the threshold, the high end was above the threshold and the difference between them was greater than 0.2, the TF was sorted into the “traditional” cluster; finally, if both the low and high end averages were below the threshold, the TF was sorted into the “poorly predicted” cluster.

Chapter 3: Parameter estimation with ChIPanalyser

3.1 Data gathering, pre-processing and quality control

In order for ChIPanalyser to estimate ChIP profiles, TF ChIP-seq data, chromatin accessibility data and TF binding motifs were required as input. Bulk transcription factor ChIP-seq and accessibility data (in the form of DNase I-seq) and controls were downloaded from ENCODE for the K562 human cell line and the data was pre-processed as detailed in the Methods and illustrated in Figure 2A. First, FASTQC (Andrews, 2010) was used on all the files to assess read quality before and after trimming (Table A2). The files were then trimmed before being aligned to the genome with bowtie2 (Langmead and Salzberg, 2012). The overall alignment rate was above 80% for the majority of the transcription factors, indicating that overall, the data was of good quality (see Figure 2A). After alignment, macs2 (Zhang *et al.*, 2008) was used to call peaks; the number of peaks varied widely, with the highest number being ~170,000 peaks for MAX, while the lowest after removing TFs with fewer than 60 peaks was

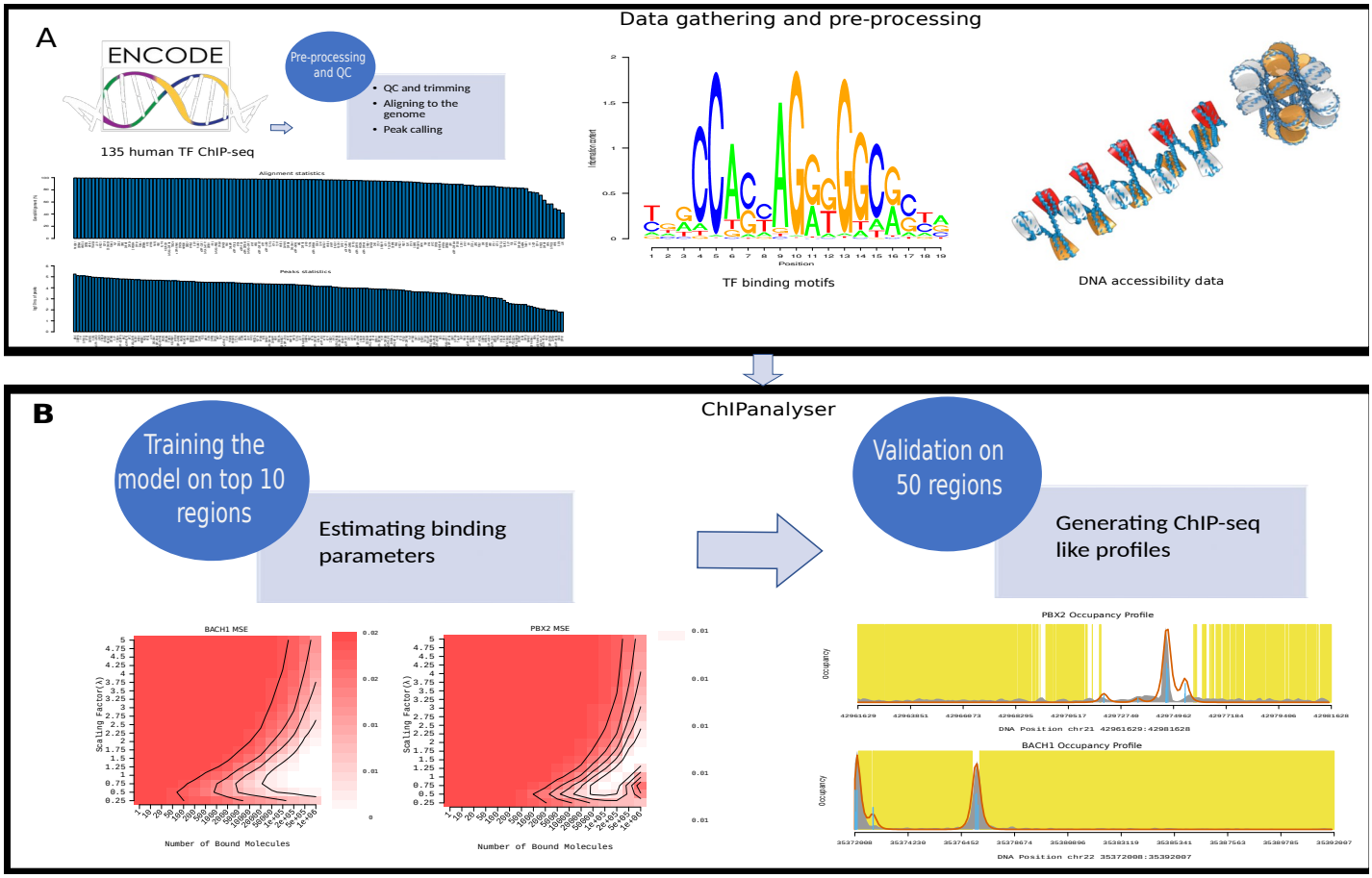

Figure 2. Outline of ChIPanalyser workflow; **A.** TF ChIP-seq, DNA accessibility and TF binding motif data was downloaded from public repositories. The ChIP-seq and accessibility data were QC'd and pre-processed before the ChIPanalyser analysis commenced. **B.** The model was trained on the top 10 ChIP regions for each TF and the optimal binding parameters were estimated. Validation was then done on the subsequent top 50 ChIP regions and ChIP-like profiles were estimated.

ZFP91 with 62 peaks. Any TFs with fewer than 60 peaks were eliminated from the analysis. The wide variation in peak number is likely due to cell-line specific expression.

3.2. Binding parameter and ChIP profile estimation

Once the pre-processing was finalised, ChIPanalyser was used to estimate the optimal number of bound molecules (N) and specificity factor (λ) for each TF by fitting the ChIPanalyser model to the existing ChIP-seq data (see Figure 2B). This was necessary because these two parameters are not easily determined experimentally and, thus, not widely available for most TFs. The genome was tiled into 50kb bins and the model was then trained on the top 10 regions with the highest ChIP score from the ChIP-seq data and MSE was

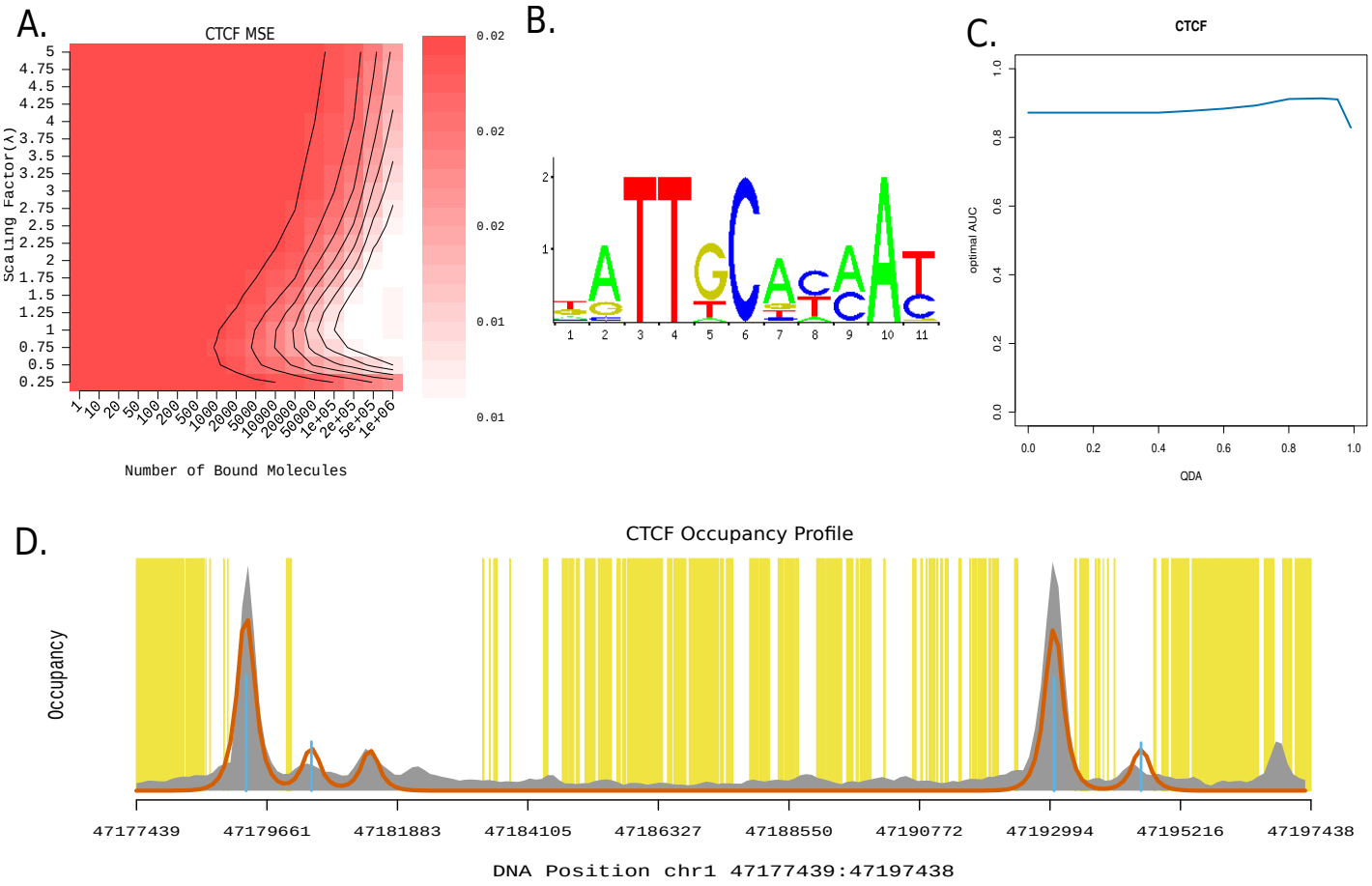

Figure 3. Optimal parameter and ChIP-like profile estimation for CTCF. **A.** Heatmap showing the optimal range for the optimal QDA for CTCF; **B.** Logo of the CTCF motif used in the analysis; **C.** AUC of the MSE for the optimal parameters estimated for CTCF for all QDAs; **D.** ChIP profile estimated with ChIPanalyser based on the optimal parameters. The grey shaded area represents the ChIP signal, the orange line represents the ChIPanalyser estimation of the ChIP signal, the blue lines represent occupancy at each locus, the yellow shaded areas represent closed chromatin and the white shaded areas represent open chromatin.

minimised as a goodness of fit measure. Following training, the estimated parameters were validated on the subsequent top 50 highest scoring regions of the ChIP-seq data.

The accessibility data was subset with a quantile vector, resulting in QDAs between 0-0.99. This means that for each QDA, the model considers a % of the top regions as accessible, regardless of their actual accessibility scores; for example, for QDA 0, 100% of regions were considered accessible, while for QDA 0.9, only the top 10% of regions were considered accessible (see section 3.1). In order to observe the chromatin accessibility preferences of the various TFs, the analysis was run for each QDA. The AUC of the MSE was used as a goodness of fit metric for the model accuracy. If the AUC for the bottom QDAs is high, it indicates that the TF can bind inaccessible chromatin, as the prediction accuracy remains high even though all or most of the genome is considered accessible by the model (including dense chromatin), while if the AUC is high for the high end QDAs, it indicates that the TFs bind open chromatin, as only the most open regions are considered accessible by the model. Here, CTCF is used as an example (Figure 3), however a complete list of the estimated optimal parameters for all TFs is available in Table A2 in the Appendix.

For visualisation purposes, heatmaps of the optimal parameters were generated using the plotOptimalHeatMaps function native to ChIPanalyser. Figure 3A shows the optimal binding parameter heat map for the optimal QDA for CTCF (QDA 0.9). The estimated λ for CTCF was 0.75, indicating that CTCF has a high affinity for its motif (shown in Figure 3B). The binding affinity of a TF for its motif is given by its binding energy when interacting with its motif (see equation 2); thus, the lower the λ, the higher the TF affinity. ChIPanalyser estimated 500,000 molecules of CTCF to be the optimal number of molecules bound to the genome in order to replicate the ChIP-seq data. Previously, around ~200,000 CTCF molecules were reported in the nucleus of HeLa cells, although the numbers varied throughout the cell cycle (Holzmann *et al.*, 2019). Another study, found ~100,000 molecules of CTCF in human U2OS cells and ~200,000 in mouse embryonic stem cells (Cattoglio *et al.*, 2019). Therefore, variations between cell lines may also be a factor. Belaghzal *et al.*, (2019) found that ~95% of nuclear CTCF is bound to chromatin. As can be seen from the heatmaps, varying the number of bound molecules between ~200,000 – ~500,000 does not seem to affect the peak prediction accuracy much. Indeed, even 100,000 bound molecules (BM) is only slightly outside the optimal range, with an MSE of 0.01163792. The MSE for 200,000 BM was 0.009522740, while the optimal MSE for 500,000 BM was 0.007814508. As can be seen in Figure 3C, CTCF was well predicted regardless of the QDA, indicating that it can bind chromatin regardless of its level of accessibility. This is in line with the highly diverse function of CTCF, ranging from insulator to gene regulation and maintaining chromatin open (Kitchen and Schoenherr, 2010).

Based on the estimated optimal parameters, as well as the PWM, a ChIP-like profile was estimated to validate these parameters. The model was able to recover peak location with high accuracy, as indicated by the blue lines in Figure 3D. The peak hight was also recovered, however not always completely.

When looking at the estimated optimal parameters for all TFs, theTFs (85) that were predicted to have 1,000,000 bound molecules (Figure 4A). This is likely a result of the default maximum number of bound molecules set by ChIPanalyser of 10^6 being too low and some of the TFs in this group having an even higher optimal number of bound molecules. It is also possible that ChIPanalyser overestimates the number of bound molecules to compensate for factors that may influence TF binding that are not considered by the model, such as DNA methylation, cofactors or chromatin architecture. There is little available data about the number of TF molecules in cell nuclei and virtually none for the K562 cell line, so it is difficult to validate these findings without further experimental investigation. The data that is available indicates that in humans, TF abundance ranges from a few thousand to a few million, depending on TF, cell type and cell cycle stage (Biggin, 2011; Cattoglio *et al.*, 2019; Holzmann *et al.*, 2019). Thus, while the predictions made by ChIPanalyser may not be accurate for use outside of the model, and may not reflect absolute values, it is nevertheless sufficiently accurate for the model to accurately estimate the range of these values and binding profiles for the majority of these TFs.

Table 2. TFs with a predicted λ above 3

TF name	Predicted λ
ZEB2	5.00
SOX6	4.25
PKNOX1	4.00
ZNF407	3.75
E2F8	3.50
MGA	3.25
YBX1	3.25
ZMYM3	3.25
NR3C1	3.00
ZFP91	3.00

The predicted λ for the majority of TFs was below 1.5 (Figure 4B), indicating a good affinity for their motifs. There were 10 TFs with a predicted λ of 3 or higher, listed in Table. Of these, all but one were poorly predicted by the ChIPanalyser model (see Table A2 in the Appendix and Figure 5 F). This is likely due to the model not accounting for cooperative binding. Indeed, some of these TFs are known to be cooperative, such as the E2F family TFs which form a complex with DP and have low affinity for DNA as monomers (Zheng *et al.*, 1999; Morgunova

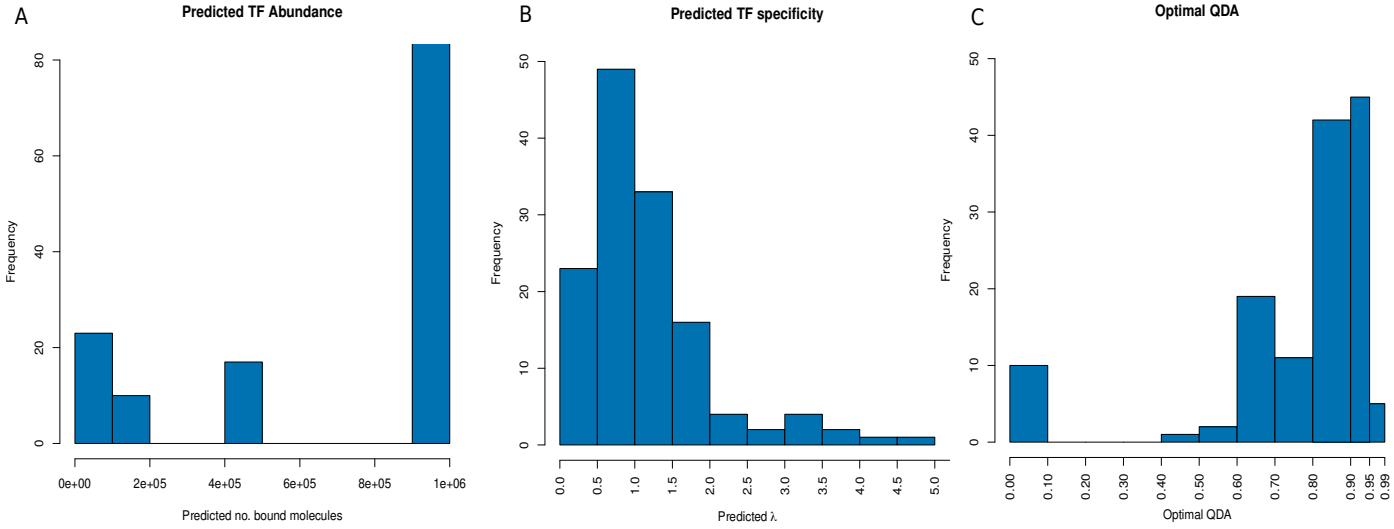

Figure 4. A. Bar plot of the predicted N for all TFs; **B.** Bar plot of the predicted λ for each TF; **C.** Bar plot of the optimal QDA based on the ChIPanalyser model.

et al., 2015; Morgunova and Taipale, 2017). Furthermore, E2F family TFs are also known to be sensitive to DNA methylation (Gaston and Fried, 1995), another factor not taken into account by the model; SOX6, which was found to bind interactively with BCL11A to regulate Ɣ-globin (Xu *et al.*, 2010), as well as interact with other members of the SOX family to regulate cell fate (Lefebvre, Li and De Crombrugghe, 1998; Lefebvre, 2010).

As seen in Figure 4C, the majority of TFs had their optimal QDA among the high end QDAs (0.7-0.99), suggesting that they bind best in open chromatin, as would be expected. However, there were 9 TFs that had their optimal QDA as 0, suggesting that their preferred chromatin type is closed chromatin. Three of these (ZNF146, REST and ZNF274), were poorly predicted by ChIPanalyser, as mentioned previously, likely due to cooperative binding, methylation or other such factors not taken into account by the current model. For the remaining 6, this preference might be an indicator of function.

Chapter 4: Clustering TFs based on their chromatin accessibility preference

In order to further investigate the chromatin state preference of the different TFs and how that might relate to function, the AUC of each TF for each QDA was calculated as detailed in Martin and Zabet, 2020 and summarised in the Methods section. The TFs were then clustered based on how the AUC changed with QDA (Figures 5-6). Two methods were used for clustering: k-means and a bespoke threshold-based algorithm and the two were compared.

4.1. K-means clustering

The k-means algorithm native to R was used to cluster the TFs (Figure 5). The value of k was determined with the use of an elbow plot (Figure 5A). Because the “bend” in the elbow was not clear, the clustering was performed with k = 3, 4 and 5, however the best fit for the data was when k=4 was used. Three main trends were observed that were labelled “pioneer”, “partial pioneer” and “traditional”, with an additional fourth cluster of TFs that were poorly predicted by the model. In addition to these, a fifth cluster with a downward trend (high AUC in the low QDAs and low AUC in the high QDAs) was expected based on similar work in other organisms (unpublished work), however this does not appear to be present in this dataset. The largest group was the “partial pioneer” group, with 54 TFs, followed by the “pioneer” group with 42 TFs, the “traditional” group with 20 TFs and finally, the poorly predicted group with 19 TFs (Figure 5B). Table A2 in the Appendix contains a full list of TFs and which clusters they were assigned to.

As can be seen in Figure 5C, the “pioneer” group contains TFs that were well predicted by the model regardless of QDA. As mentioned previously, the lower the QDA, the more of the

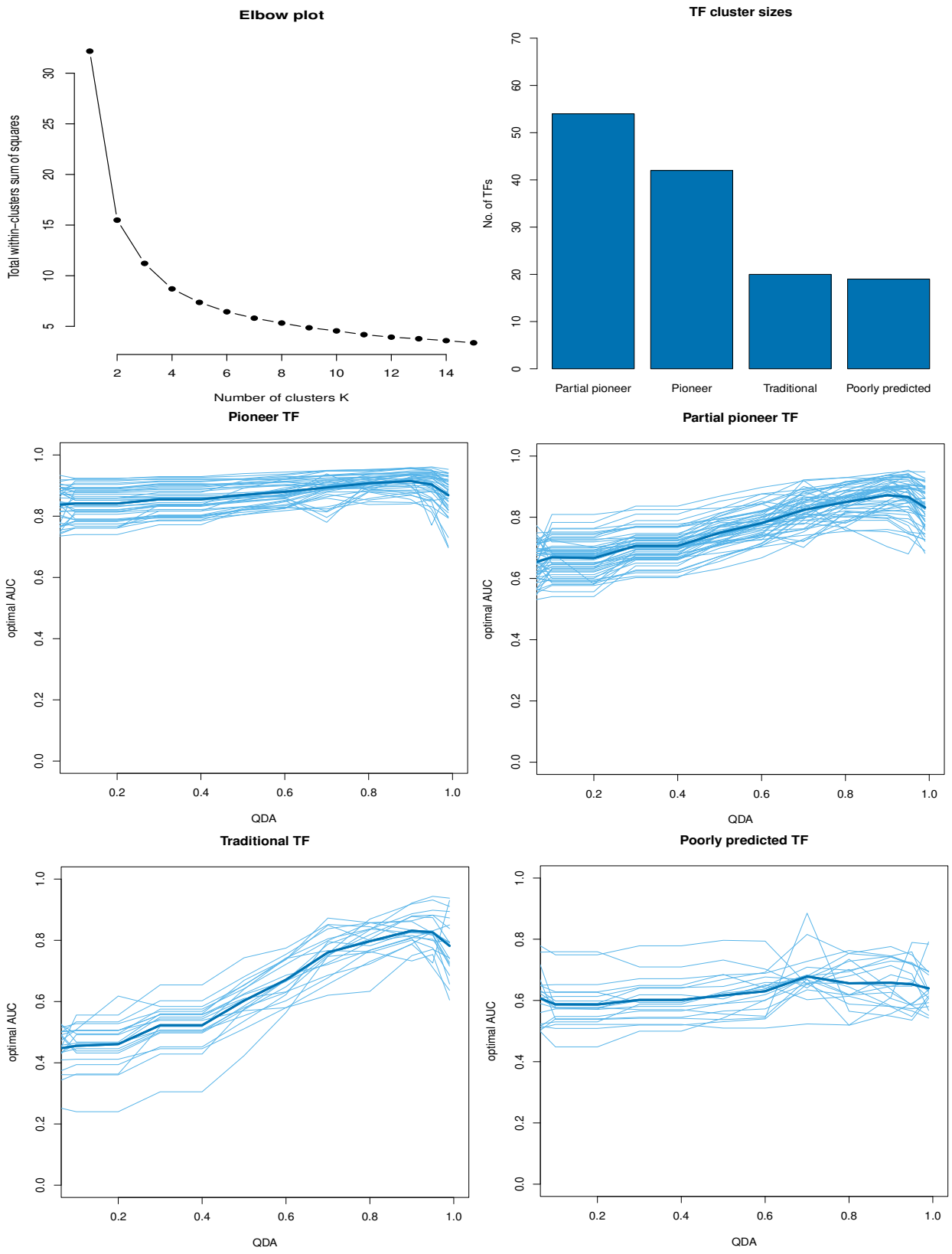

Figure 5. TF clustering with the k-means method. A. Elbow plot showing the optimal number of clusters for the dataset. **B.** Bar plot of TF cluster sizes; **C-F.** Line plots showing the change in AUC with QDA for the "pioneer", "partial pioneer", "traditional" and "poorly predicted" clusters, respectively with the k-means method.

genome is considered “accessible” by the ChIPanalyser model regardless of the actual chromatin accessibility in the region. Thus, if the model recovers the real ChIP-seq data with high accuracy under those conditions, the TF in question must be able to bind anywhere in the genome where its motif is present, regardless of accessibility. These were labelled as “pioneer” on the premise that a pioneer TF binds closed chromatin in order to open it, but the chromatin surrounding it then becomes accessible and it can therefore bind in both open and dense chromatin.

The TFs in this cluster included known pioneer TFs such as GATA1, which belongs to the GATA family of pioneer factors that are involved in cell differentiation among other things (Bossard and Zaret, 1998; Lai *et al.*, 2018), NRF1 (Lai *et al.*, 2018; Mayran and Drouin, 2018), POU5F1 (Oct4) which although it has weak pioneer activity when binding on its own, it is a strong pioneer when working with Sox2 (Soufi *et al.*, 2015; Lai *et al.*, 2018; Michael *et al.*, 2020), NFYA/NFYB which was shown to contribute to cell differentiation of murine embryonic stem cells (Oldfield *et al.*, 2014), FOS (Biddie *et al.*, 2011; Alvarez-Dominguez *et al.*, 2019). FOS was notable for being present in the dataset twice, one version tagged with a GFP marker, and one version untagged. Interestingly, the tagged version was classed as a pioneer, however, the untagged version was classed as traditional. The likely cause of this is data quality, as indeed the tagged version had a better overall alignment rate (97.78% vs. 83.34 % for the untagged version) and a lower λ (0.75 vs. 1.25 for the untagged version), indicating better affinity for its motif. This is an interesting observation, as 18 out of the 135 TFs within the dataset were tagged with either a GFP marker or a 3xFLAG marker. In addition to FOS, ATF1, CUX1, GATA2, JUND and USF2 were also present in both tagged and untagged version, however, there were no notable differences between the two versions (Table A2).

The remaining TFs were only present in either tagged or untagged form in the dataset. In addition to these, this cluster also contained some pioneer co-factors such as MYC which works alongside Oct4/Sox2 (Soufi, Donahue and Zaret, 2012). This is because the model is unable to differentiate between TFs that have pioneer function and their cofactors but can only determine whether a TF is able to bind dense chromatin.

Other TFs in this cluster include IRF2 and while there is no evidence yet of it having pioneer function, three other members of the IRF family, IRF1, IRF4 and IRF8, have been previously shown to have some pioneer activity (Decker *et al.*, 2009; Alvarez-Dominguez *et al.*, 2019), although none of them were present in the dataset analysed. However, when considering domain architecture (available through the SMART database (Letunic, Doerks and Bork, 2015), IRF2 is similar to IRF1 and thus may share its function. Similarly, ELF1, CEBPB and RFX5 also have other members of their family that have been found to have some pioneer function. ELF3, RFX2-4 were found to have some chromatin opening function (Alvarez-Dominguez *et al.*, 2019) and CEBPA is also known to be a pioneer factor (Lai *et al.*, 2018).

Interestingly, CTCF was also a part of the pioneer group. There are several conflicting results within the literature regarding CTCF. Some studies have found that CTCF binding is impaired by nucleosomes (Teif *et al.*, 2014; Wiehle *et al.*, 2019), while others show that CTCF can displace nucleosomes and maintain chromatin accessibility throughout the cell cycle (Owens *et al.*, 2019). In addition, CTCF is also known to bind at the boundary between heterochromatin and euchromatin in order to stop heterochromatin spreading, and acts as an insulator (Lee, 2003). Thus, CTCF is a highly versatile TF whose function seems to vary widely depending on context. CTCF is known to have clusters of multiple adjacent binding sites (Kentepozidou *et al.*, 2020; Nanni, Ceri and Logie, 2020). Since the training and validation for the model were performed on the top 10 and subsequent 50 highest scoring ChIP-seq regions, respectively, it is possible that these regions contained such clusters which are heavily populated by binding sites and likely inherently accessible, thus explaining why CTCF appears to be a pioneer. Another possibility is that the presence of multiple CTCF molecules in the same region can displace nucleosomes and thus CTCF acts as a pioneer through homo-cooperativity, however further investigation is needed to elucidate this.

Figure 5D shows the “partial pioneer” cluster, which contains TFs that showed a slight preference for open chromatin (i.e. AUC for the lower end QDAs was slightly lower) and were labelled as “partial” pioneers because they do not appear to have as strong affinity for closed chromatin as the “pioneer” cluster TFs. This cluster contained some factors that are known to have pioneer function, such as FOXA1 and FOXK2, two members of the FOX family which are some of the best characterised pioneer TFs (Cirillo *et al.*, 2002; Zaret and Carroll, 2011; Iwafuchi-Doi *et al.*, 2016; Zaret and Mango, 2016; Lai *et al.*, 2018). RUNX1 was also part of this cluster and along with PU.1 is known to prime transcription of the c-fms/csf1R gene involved in hematopoietic differentiation. However, RUNX1 was shown to only be needed transiently (Krysinska *et al.*, 2007; Hoogenkamp *et al.*, 2009), which might explain why it is picked up as being a “weaker” pioneer.

In addition to pioneer factors, this group also contained pioneer co-factors, such as Ep300. While not itself a TF, Ep300 is a histone acetyltransferase that has been shown to work with a number of pioneer factors such as FOXO1 (Perrot and Rechler, 2005). Ep300 is not a site specific transcription factor and does not bind DNA on its own (Perrot and Rechler, 2005), however ChIPanalyser can model the binding of any protein for which a motif is provided. Often de novo motifs are generated from ChIP-seq experiments, which are problematic due to the fact that they originate from the ChIP-seq data in order to explain the ChIP-seq data. As Ep300 is recruited by various pioneers at their binding sites, the “motif” discovered for it is likely a consensus between all the TFs that it is being recruited by, thus giving it the appearance of being a pioneer factor itself. However, while the “binding” of Ep300 in itself has no biological meaning, its classification as a “partial pioneer” supports its role as a co-factor for pioneers. Indeed, histone acetylation has long been associated with the relaxation of chromatin.

The pioneer and partial pioneer groups are very similar in the kinds of TFs that they contain, both having a mixture of known pioneers and pioneer cofactors, along with a number of factors for which no pioneer function is known. Together, they seem to account for most of the TFs in the dataset, with 96 out of 135 TFs between them, which seems contrary to the widely accepted idea that pioneer factors are a small subset of TFs. While it is possible that the model has picked up some previously unknown TFs, there is also some bias that might explain the unusually high number of “pioneer” TFs. One such bias is the use of DNase I data as the accessibility data, as this assay takes into account large regions of the genome, and some small accessible regions might be considered inaccessible by the assay if they are surrounded by dense regions, thus introducing a bias in the model. Using accessibility data with higher resolution, such as micrococcal nuclease digestion with sequencing (MNase-seq) (Chereji, Bryson and Henikoff, 2019) could be a way of eliminating this bias and improving the model accuracy. Another possible bias is that the training and validation of the model were done on the top 10 and subsequent 50 highest scoring regions, respectively. These could potentially all be regions highly populated with binding sites and thus accessible DNA. In that case, modifying how much of the genome the model considers accessible (i.e. using different QDAs) would not make a difference if none of the 60 regions were found in dense chromatin. A possible way of overcoming this bias would be to perform the validation on lower scoring regions and see if the same trends persist.

The “traditional” cluster (Figure 5E) contains TFs that behave in the “traditional” way of binding in open chromatin and not binding in dense chromatin. The TFs in this group had a strong preference for open chromatin and included ZBTB40, a TF primarily involved in bone formation and density (Doolittle *et al.*, 2020), but is also associated with stress response (Bae and Lee, 2018); PKNOX1, a tumour suppressor (Longobardi *et al.*, 2010) that is also involved in hematopoiesis (Di Rosa *et al.*, 2007); and TEAD4, a developmental TF involved in trophectoderm lineage formation during mammalian embryogenesis (Yagi *et al.*, 2007). An outlier in this cluster is NEUROD1, which was found to have pioneer function in mouse neurons (Lai *et al.*, 2018). However, this is a specific TF only expressed in the brain and GI tissues and is not present in lymphatic tissue or bone marrow. Indeed, there was only one replicate on for the K562 cell like on ENCODE. Thus, data on this TF is not conclusive in the K562 cell line, as this is a lymphoma cell line.

Finally, the “poorly predicted” cluster (Figure 5F) contains TFs that were poorly predicted for all QDAs. This could be due to some of the following reasons. Some of them might need to bind cooperatively, such as FOS has long been known to bind the AP-1 site cooperatively with JUN, but is unable to bind it on its own (Rauscher *et al.*, 1988). Others might be methylation sensitive, for example ZMYM3 was found to preferentially bind the M203 motif when it is methylated, while it only bound the unmethylated version of the M85 motif (Hu *et al.*, 2013), while Oct4 binds preferentially to CpG methylated motifs (Yin *et al.*, 2017). And some might be due to poor data, such as MGA and ZNF24 which only had 165 peaks and 89 peaks, respectively.

4.2. Threshold-based clustering

Since the “pioneer” and “partial pioneer” clusters were unexpectedly large and accounted for more than half the dataset, we wanted to see if the clusters would be maintained if we set a hard threshold for what was “well predicted” and what was not. In order to do this, we used the k-means clustering to inform the creation of a threshold-based algorithm to cluster the TFs with AUC >= 0.8 being considered as “well predicted” and AUC < 0.8 as “poorly predicted” with the stipulation that if the low end QDAs were below 0.8, but the high end QDAs were above 0.8 and the difference between the two was less than 0.2, they would still be considered “well predicted” and sorted into the “partial pioneer” factors (Figure 1). As can be seen in Figure 6, the clusters were largely maintained, however the “pioneer” and “partial pioneer” groups are slightly smaller, reflecting the stricter criteria. Thus, the “pioneer” cluster was reduced to 30 TFs from 42 in the k-means method and the “partial pioneer” group was reduced to 32 from 54 in the k-means method. This number of “pioneer” TFs is more in line with literature, however it is still an unusually large number of “pioneer” factors, totalling 62 TFs between the two clusters. The “traditional” and “poorly predicted” groups both gained TFs, with the “traditional” cluster now containing 36 TFs up from 20 in the k-means method and the poorly predicted cluster with 37 TFs up from 19 in the k-means method.

While the number of “pioneer” factors is more reasonable with this method, the number of “poorly predicted” factors has almost doubled. However, as can be seen in Figure 6E, the “poorly predicted” AUC trend closely resembled the “partial pioneer” AUC trend, only with an overall lower AUC. Indeed, many of the TFs in this cluster look as if they would fit the “partial pioneer” or “traditional” groups much better. Figure 6F shows the TFs that were sorted as “poorly predicted” by the threshold algorithm, but not the k-means one. The majority of these have an AUC only slightly below 0.8 in the high end QDAs and some of them are even above 0.8. However, due to the algorithm using averages of the top and bottom QDAs, they were classed as “poorly predicted”. Thus, this algorithm could be improved upon by using dynamic thresholds, instead of a hard threshold.

As can be seen in the Sankey diagram (Figure 7), the largest migration was from the “partial pioneer” group to the “traditional” group, with 24 TFs shifting. Among these were EP300, which as mentioned previously is a histone acetyltransferase which might explain the shift, and FOXA1, which is a pioneer factor. However, its average AUC for the high end QDAs was above 0.9, while the average AUC for the lower end QDAs was ~0.69 and thus, the difference between the two was greater than 0.2 and as such the algorithm classed it as traditional.

A total of 11 TFs moved from “pioneer” to “partial pioneer”, including CEBPB and CEBPZ. While CEBPA is known to be a pioneer (Lai *et al.*, 2018), it is unclear whether all members of the family share this function. 11 factors moved from “partial pioneer” to “poorly predicted”, 9 from “traditional” to “poorly predicted” and 1 from “pioneer” to “traditional”.

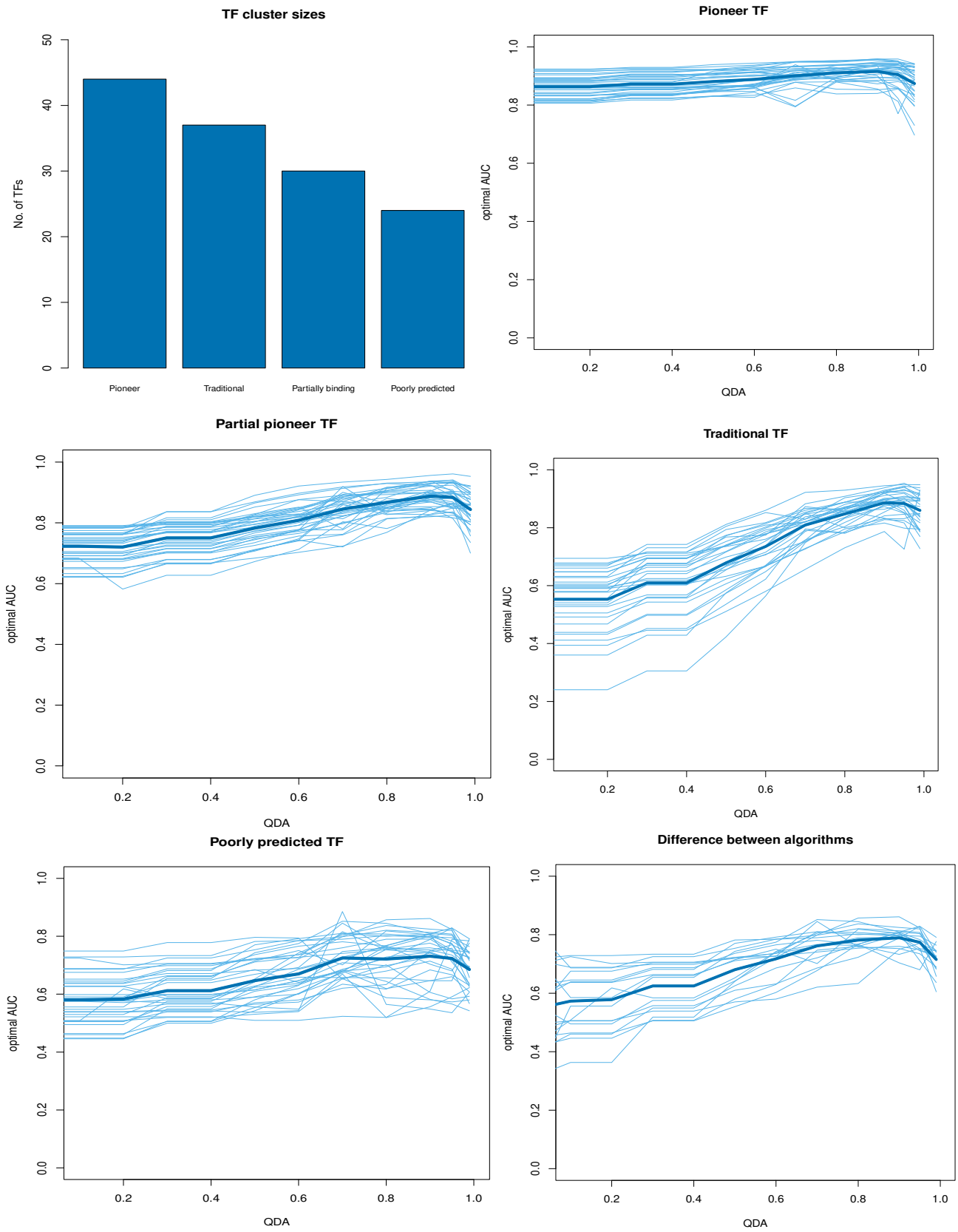

Figure 6. Clustering of TFs with the threshld-based algorithm. A. Bar plot showing the number of TFs in each cluster; **B-E** Line plots showing the change in AUC with QDA for the "pioneer", "partial Pioneer", "traditional" and "poorly predicted" clusters, respectively with the threshold method. **F.** TFs classed as "poorly predicted" by the threshold method, but not by the k-means method.

Of the 42 TFs in the k-means “pioneer” cluster, 30 maintained their cluster with the threshold-based method, including CTCF, GATA1 and FOS. 21 of the 54 TFs in the k-means “partial pioneer” group maintained their group, including SP1, Jun and KLF16. 11 “traditional” TFs maintained their cluster including SMAD1, PKNOX1 and NR2C1. Finally, all the “poorly predicted” TFs maintained their cluster when using the threshold method. The Sankey diagram is also available as a html widget in Figure A1 in the Appendix and a complete comparison between the classification given to each TF by the two methods can be found in Table A2 in the Appendix.

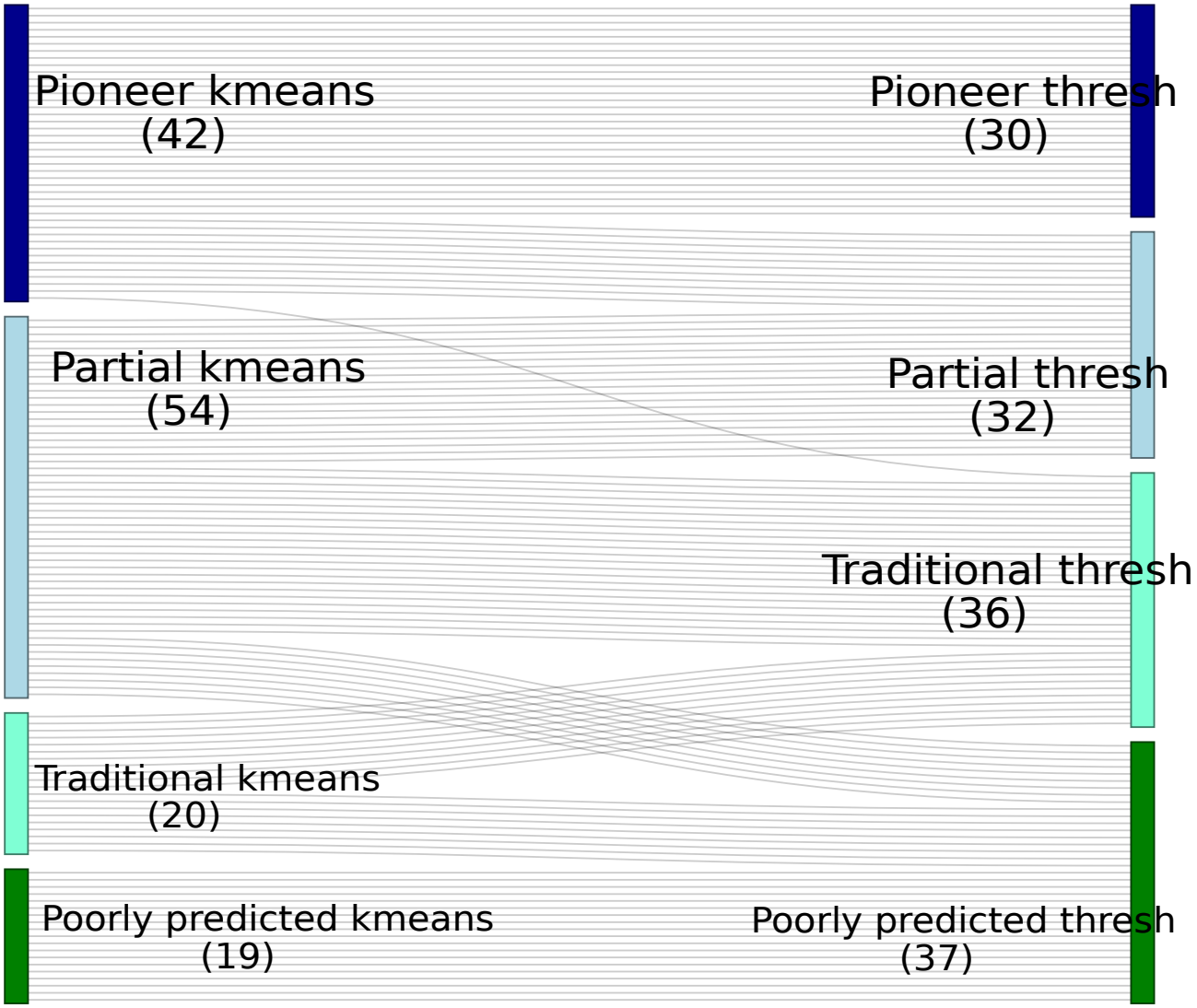

Figure 7. Sankey diagram showing the shift between clusters when using the k-means and threshld-based algorithms to classify TFs based on their accessibility preference.

Chapter 5: Summary, conclusions and future work

5.1. Summary and conclusions

This dissertation set out to investigate the chromatin accessibility preferences of 135 human TFs in the K562 cell line. First, bulk ChIP-seq data and DNase I data from ENCODE was used to estimate the number of bound molecules (N) and the TF binding specificity factor (λ) for each TF with the R Bioconductor package ChIPanalyser. N proved to be generally overestimated compared to experimental values, however this did not impact the prediction accuracy of the model. Indeed, the overestimation was likely due to the model compensating for other factors that influence TF binding, such as cooperativity or DNA methylation, but which were not considered by the model. Furthermore, there was a large number of TFs that had a N of 106, which is the default upper threshold for ChIPanalyser. However, it is possible that this was not high enough for all TFs and had the threshold been set higher, the prediction would have been higher as well. The λ for the majority of the TFs was 0.5-1, suggesting good affinity of the TFs for their motifs. Some of the TFs had high λ values (see Table 2), indicating low affinity for their motifs. This is likely due to them binding cooperatively, which was something not considered by the current model.

In order to investigate the preference of each TF for chromatin accessibility, the analysis was run multiple times with different QDAs. Simply put, this means that each iteration of the model considered the top 1-n regions of the DNase I data as accessible, regardless of the scores. Thus, a QDA of 0 meant the top 1-0 (i.e. all) regions were considered as accessible, while a QDA of 0.99 meant the top 1-0.99 (i.e. 1%) of the regions were considered as accessible. Essentially, this meant that any TFs that were well predicted with the lower QDAs were not affected by chromatin accessibility, since there was no distinction made between accessible and inaccessible regions; the reverse is true for higher QDAs. ChIPanalyser has 12 goodness of fit metrics incorporated (Martin and Zabet, 2020), however the best ones to use for this type of data are MSE and AUC (Martin and Zabet, 2020) and therefore, they were used to determine the accuracy of the model.

In order to gain a better overview of TF accessibility preferences, the TFs were clustered based on the trend in their AUC over all QDAs. Two different methods were used and compared, k-means and a bespoke threshold-based algorithm (Figure 1). Four clusters were observed with both methods. The first cluster appeared to be unaffected by DNA accessibility and remained well predicted regardless of QDA (Figures 5C, 6B) and was dubbed “pioneer TFs”. The second cluster showed a slight preference for more open chromatin, but was still relatively well predicted regardless of QDA (Figures 5D, 6C) and was dubbed “partial pioneer TFs”. The third cluster showed a strong preference for open chromatin and was not well predicted when lower QDAs were used (Figures 5E, 6D) and was dubbed “traditional TFs”. Finally, the fourth group was poorly predicted regardless of QDA (Figures 5F, 6E). The “pioneer” and “partial” pioneer groups accounted for a much larger percentage of the data than anticipated, containing several known pioneer factors such as GATA1 (Bossard and

Zaret, 1998), Oct4 (Soufi *et al.*, 2015; Yu and Buck, 2020), NFYA/B (Oldfield *et al.*, 2014) and forkhead TFs (Cirillo *et al.*, 2002; Zaret and Carroll, 2011; Zaret and Mango, 2016). In addition to known pioneers, these groups also contained some known pioneer co-factors such as c-MYC (Soufi, Donahue and Zaret, 2012). This is due to the model not being able to distinguish between pioneers and their co-factors. Thus, some of the other TFs in these clusters are potentially pioneer co-factors, explaining the unusually large size of the clusters.

When using the threshold-based algorithm to cluster the TFs, the “pioneer” and “partial pioneer” clusters were significantly reduced, while the “traditional” cluster gained some TFs bringing them more in line with the consensus of there only being a small number of pioneer factors. 30 TFs were classed as “pioneer” by both including CTCF, GATA1 and FOS, while 11 “traditional” factors maintained their cluster across the two methods, including SMAD1, PKNOX1 and NR2C1. However, this came at the cost of many TFs being sorted in the “poorly predicted” group despite being better suited to the “partial pioneer” or “traditional” groups, based on their AUC trends (Figure 6F). This could potentially be rectified by calculating the threshold for a TF being well predicted dynamically, instead of using a hard threshold. 30 TFs were classed as “pioneer” by both including CTCF, GATA1 and FOS, while 11 “traditional” factors maintained their cluster across the two methods, including SMAD1, PKNOX1 and NR2C1.

5.2. Future work

This work prompts many potential future investigations, both to improve upon this work and to expand its scope. One limitation of the current implementation of ChIPanalyser is that is does not consider cooperative binding (Martin and Zabet, 2020) and some TFs cannot bind well on their own (Ravasi *et al.*, 2010). For example, E2F8 was poorly predicted by ChIpanalyser and some members of the E2F family are known to bind cooperatively and have low affinity for DNA when on their own (Zheng *et al.*, 1999).

5.2.1. Optimising the current analysis

A few parameter optimisations could be made to improve on both the ChIPanalyser and the clustering analyses, which due to time constraints, were not included in this thesis. As shown in figures 5F & 6E, there were a number of TFs that were poorly predicted by ChIPanalyser. There are a number of potential factors influencing this, and some could be mitigated by tweaking some of the parameters of the analysis. One example would be to rerun the analysis for the poorly predicted TFs with another PWM motif. Due to the size of the data, it was not possible to select the optimal PWM for each TF and this might have negatively impacted the predictions in some cases. Another would be to rerun the analysis with a higher upper threshold for the number of bound molecules, which may result in better prediction of some TFs. Moreover, multiple replicates for the same TFs were merged where available. This was done for simplicity and speed, considering the large quantity of date that was analysed. An alternative would be running the analysis on each individual replicate.

As discussed, the “pioneer” and “partial” pioneer groups were unexpectedly large, in contradiction of the general consensus of only a small subset of TFs having pioneer function. This indicates that there might be some bias in the analysis. One such bias could be the use of DNase I data as a determinant of accessibility, as the assay accounts for relatively large genomic regions and thus some accessible sites might be “hidden” within larger surrounding regions of inaccessible DNA, thus giving the appearance of TF binding in dense chromatin. One way to mitigate this would be to use another type of accessibility data, such as MNase-seq, to rerun the analysis, and compare it to the current results.

Furthermore, another potential bias could be the use of the top 60 highest scoring peak regions when training and validating the model. TFs usually bind at highly accessible promoter and enhancer regions that contain many binding sites for many TFs. This might result in the top 60 regions for some of these TFs not containing any dense chromatin, and thus they appear to be unaffected by chromatin accessibility. Furthermore, one study found that at low affinity sites, the binding of CTCF was impaired by nucleosomes (Teif *et al.*, 2014; Wiehle *et al.*, 2019) This bias could be overcome by performing the validation on lower scoring regions which are more likely to contain some dense chromatin. If the AUC trends persist, then it is likely that the TF is indeed able to bind dense chromatin.

5.2.2. Expanding the scope of the analysis

There are many further avenues of investigation which were opened up by this research. First, there are several more human cell lines for which ChIP-seq data is available on ENCODE, and with the release of ENCODE phase 3 (Abascal *et al.*, 2020), more data for the K562 cell line has become available as well. Incorporating them into the analysis would provide a more complete picture of TF binding in humans, how it is affected by chromatin accessibility and how it varies between different tissues. Furthermore, cooperativity between TFs is a well-established phenomenon (Villar, Flicek and Odom, 2014) and a full understanding of TF binding cannot be gleaned without taking it into account. The underlying framework for modelling cooperativity into ChIPanalyser already exists (unpublished work), however it has yet to be implemented into the package. Incorporating it into the analysis could both improve the predictive power of the model, as well as provide new insights into the interactions between TF when binding to the genome. Furthermore, this could also improve the parameter estimation and bring the estimates for N more in line with experimental evidence, by removing the need to compensate for lower binding affinity by predicting increased TF concentrations.

Secondly, it has been established that CpG methylation plays a role in TF binding (Yin *et al.*, 2017). However, much like is the case with DNA-accessibility, CpG methylation impacts TFs differently. Some TFs, such as ETS (Gaston and Fried, 1995) and CTCF (Hnisz, Day and Young, 2016) are inhibited by CpG methylation at their motifs, while others, such as YY1 (Gaston and Fried, 1995) and SP.1 (Harrington *et al.*, 1988) remain unaffected. Moreover, some TFs such as KLF4 preferentially bind the methylated versions of some motifs, while

preferring the unmethylated form of others (Hu *et al.*, 2013). Thus, the methylation preference of TFs merits further investigation. Furthermore, some pioneer factors such as FOXA2 were found to promote demethylation around their binding site (Donaghey *et al.*, 2018), suggesting that a possible mechanism for pioneer activity is promoting demethylation. It would be interesting to compare the methylation preference profiles of TFs to their accessibility preference profiles and investigate how they correlate. Bisulfite sequencing data is widely used for methylome studies (Grunau, Clark and Rosenthal, 2001) and similarly to DNase I, there are many publicly available datasets. Thus, ChIPanalyser could be used to perform a similar analysis to what has been described in this thesis, using bisulfite data.

This project has considered the relationship between TF binding and DNA accessibility and has revealed the chromatin accessibility preferences of 135 human TFs in the K562 cell line. Of these, 96 were identified as known or putative pioneer factors and co-factors, and thus raised further questions about the mechanisms of TF-DNA interactions.

References

Abascal, F. *et al.* (2020) ‘Expanded encyclopaedias of DNA elements in the human and mouse genomes’, *Nature*, 583(7818), pp. 699–710. doi: 10.1038/s41586-020-2493-4.

Adcock, I. M. and Caramori, G. (2009) ‘Transcription factors’, in *Asthma and COPD*. Elsevier Ltd, pp. 373–380. doi: 10.1016/B978-0-12-374001-4.00031-6.

Alipanahi, B. *et al.* (2015) ‘Predicting the sequence specificities of DnA-and RnA-binding proteins by deep learning’, *Nature Biotechnology*, 33. doi: 10.1038/nbt.3300.

Alvarez-Dominguez, J. R. *et al.* (2019) ‘Dissecting mechanisms of human islet differentiation and maturation through epigenome profiling’, *bioRxiv*, p. 613026.

Andrews, S. (2010) ‘FastQC: a quality control tool for high throughput sequence data’. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Angermueller, C. *et al.* (2016) ‘Deep learning for computational biology’, *Molecular Systems Biology*, 12(7), p. 878. doi: 10.15252/msb.20156651.

Bae, S. C. and Lee, Y. H. (2018) ‘Meta-analysis of gene expression profiles of peripheral blood cells in systemic lupus erythematosus’, *Cellular and Molecular Biology*, 64(10), pp. 40–49. doi: 10.14715/cmb/2018.64.10.7.

Bai, L. and Morozov, A. V. (2010) ‘Gene regulation by nucleosome positioning’, *Trends in Genetics*. Trends Genet, pp. 476–483. doi: 10.1016/j.tig.2010.08.003.

Bailey, T. L. *et al.* (2006) ‘MEME: Discovering and analyzing DNA and protein sequence motifs’, *Nucleic Acids Research*, 34(WEB. SERV. ISS.), pp. W369–W373. doi: 10.1093/nar/gkl198.

Ball, M. P. *et al.* (2009) ‘Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells’, *Nature Biotechnology*, 27(4), pp. 361–368. doi: 10.1038/nbt.1533.

Belaghzal, H. *et al.* (2019) ‘Compartment-dependent chromatin interaction dynamics revealed by liquid chromatin Hi-C’, *bioRxiv*, p. 704957. doi: 10.1101/704957.

van den Berg, D. L. C. *et al.* (2010) ‘An Oct4-Centered Protein Interaction Network in Embryonic Stem Cells’, *Cell Stem Cell*, 6(4), pp. 369–381. doi: 10.1016/j.stem.2010.02.014.

Biddie, S. C. *et al.* (2011) ‘Transcription Factor AP1 Potentiates Chromatin Accessibility and Glucocorticoid Receptor Binding’, *Molecular Cell*, 43(1), pp. 145–155. doi: 10.1016/j.molcel.2011.06.016.

Biggin, M. D. (2011) ‘Animal Transcription Networks as Highly Connected, Quantitative Continua’, *Developmental Cell*. Elsevier, pp. 611–626. doi: 10.1016/j.devcel.2011.09.008.

Bintu, L. *et al.* (2005) ‘Transcriptional regulation by the numbers: Applications’, *Current Opinion in Genetics and Development*. Elsevier Ltd, pp. 125–135. doi: 10.1016/j.gde.2005.02.006.

Bird, A. (2002) ‘DNA methylation patterns and epigenetic memory’, *Genes and Development*. Cold Spring Harbor Laboratory Press, pp. 6–21. doi: 10.1101/gad.947102.

de Boer, C. G. and Regev, A. (2018) ‘BROCKMAN: deciphering variance in epigenomic regulators by k-mer factorization’, *BMC Bioinformatics*, 19(1), p. 253. doi: 10.1186/s12859-018-2255-6.

Bolger, A. M., Lohse, M. and Usadel, B. (2014) ‘Trimmomatic: A flexible trimmer for Illumina sequence data’, *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

Bossard, P. and Zaret, K. S. (1998) ‘GATA transcription factors as potentiators of gut endoderm differentiation’, *Development*, 125(24), pp. 4909 LP – 4917. Available at: http://dev.biologists.org/content/125/24/4909.abstract.

Budry, L. *et al.* (2012) ‘The selector gene Pax7 dictates alternate pituitary cell fates through its pioneer action on chromatin remodeling’, *Genes and Development*, 26(20), pp. 2299–2310. doi: 10.1101/gad.200436.112.

Bürglin, T. (2001) ‘Homeobox’, in *Encyclopedia of Genetics*. Elsevier, pp. 958–962. doi: 10.1006/rwgn.2001.0625.

Caravaca, J. M. *et al.* (2013) ‘Bookmarking by specific and nonspecific binding of FoxA1 pioneer factor to mitotic chromosomes’, *Genes and Development*, 27(3), pp. 251–260. doi: 10.1101/gad.206458.112.

Cattoglio, C. *et al.* (2019) ‘Determining cellular CTCF and cohesin abundances to constrain 3D genome models’, *eLife*, 8. doi: 10.7554/eLife.40164.001.

Chereji, R. V., Bryson, T. D. and Henikoff, S. (2019) ‘Quantitative MNase-seq accurately maps nucleosome occupancy levels’, *Genome biology*, 20(1), p. 198. doi: 10.1186/s13059-019-1815-z.

Chu, D. *et al.* (2009) ‘Models of transcription factor binding: Sensitivity of activation functions to model assumptions Author’s Accepted Manuscript Models of transcription factor binding: Sensitivity of activation functions to model assumptions Models of Transcription Factor Binding: Sensitivity of Activation Functions to Model Assumptions’, *Journal of Theoretical Biology*, 257(3), p. 419. doi: 10.1016/j.jtbi.2008.11.026ï.

Cirillo, L. A. *et al.* (2002) ‘Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4’, *Molecular Cell*, 9(2), pp. 279–289. doi: 10.1016/S1097-2765(02)00459-8.

Clark, K. L. *et al.* (1993) ‘Co-crystal structure of the HNF-3/fork head DNA-recognition motif resembles histone H5’, *Nature*, 364(6436), pp. 412–420. doi: 10.1038/364412a0.

Cui, F. and Zhurkin, V. B. (2014) ‘Rotational positioning of nucleosomes facilitates selective binding of p53 to response elements associated with cell cycle arrest’, *Nucleic Acids Research*, 42(2), pp. 836–847. doi: 10.1093/nar/gkt943.

Cusanovich, D. A. *et al.* (2014) ‘The Functional Consequences of Variation in Transcription Factor Binding’, *PLoS Genetics*. Edited by Y. Pilpel, 10(3), p. e1004226. doi: 10.1371/journal.pgen.1004226.

Dangkulwanich, M. *et al.* (2014) ‘Molecular mechanisms of transcription through single-molecule experiments’, *Chemical Reviews*. American Chemical Society, pp. 3203–3223. doi: 10.1021/cr400730x.

Davidson, E. H. (2010) ‘Emerging properties of animal gene regulatory networks’, *Nature*, pp. 911–920. doi: 10.1038/nature09645.

Davis, C. A. *et al.* (2018) ‘The Encyclopedia of DNA elements (ENCODE): Data portal update’, *Nucleic Acids Research*, 46(D1), pp. D794–D801. doi: 10.1093/nar/gkx1081.

Decker, T. *et al.* (2009) ‘Stepwise Activation of Enhancer and Promoter Regions of the B Cell Commitment Gene Pax5 in Early Lymphopoiesis’, *Immunity*, 30(4), pp. 508–520. doi: 10.1016/j.immuni.2009.01.012.

Donaghey, J. *et al.* (2018) ‘Genetic determinants and epigenetic effects of pioneer-factor occupancy’, *Nature Genetics*, 50(2), pp. 250–258. doi: 10.1038/s41588-017-0034-3.

Doolittle, M. L. *et al.* (2020) ‘Genetic analysis of osteoblast activity identifies Zbtb40 as a regulator of osteoblast activity and bone mass’, *PLOS Genetics*. Edited by G. A. Cox, 16(6), p. e1008805. doi: 10.1371/journal.pgen.1008805.

Dunham, I. *et al.* (2012) ‘An integrated encyclopedia of DNA elements in the human genome’, *Nature*, 489(7414), pp. 57–74. doi: 10.1038/nature11247.

Efron, B. *et al.* (2004) ‘Least angle regression’, *Annals of Statistics*, 32(2), pp. 407–499. doi: 10.1214/009053604000000067.

Feng, R. *et al.* (2008) ‘PU.1 and C/EBPα/β convert fibroblasts into macrophage-like cells’, *Proceedings of the National Academy of Sciences of the United States of America*, 105(16), pp. 6057–6062. doi: 10.1073/pnas.0711961105.

Gaston, K. and Fried, M. (1995) ‘CpG methylation has differential effects on the binding of YY1 and ETS proteins to the bi-directional promoter of the Surf-1 and surf-2 genes’, *Nucleic Acids Research*, 23(6), pp. 901–909. doi: 10.1093/nar/23.6.901.

Geertz, M., Shore, D. and Maerkl, S. J. (2012) ‘Massively parallel measurements of molecular interaction kinetics on a microfluidic platform’, *Proceedings of the National Academy of Sciences of the United States of America*, 109(41), pp. 16540–16545. doi: 10.1073/pnas.1206011109.

Glorot, X. and Bengio, Y. (2010) ‘Understanding the difficulty of training deep feedforward neural networks’, in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256.

Graf, T. (2002) ‘Differentiation plasticity of hematopoietic cells’, *Blood*. American Society of Hematology, pp. 3089–3101. doi: 10.1182/blood.V99.9.3089.

Grunau, C., Clark, S. J. and Rosenthal, A. (2001) ‘Bisulfite genomic sequencing: systematic investigation of critical experimental parameters’, *Nucleic acids research*, 29(13), pp. 65–65. doi: 10.1093/nar/29.13.e65.

Gualdi, R. *et al.* (1996) ‘Hepatic specification of the gut endoderm in vitro: Cell signaling and transcriptional control’, *Genes and Development*, 10(13), pp. 1670–1682. doi: 10.1101/gad.10.13.1670.

Harrington, M. A. *et al.* (1988) ‘Cytosine methylation does not affect binding of transcription factor Sp1’, *Proceedings of the National Academy of Sciences of the United States of America*, 85(7), pp. 2066–2070. doi: 10.1073/pnas.85.7.2066.

He, X. *et al.* (2009) ‘A biophysical model for analysis of transcription factor interaction and binding site arrangement from genome-wide binding data’, *PLoS ONE*, 4(12). doi: 10.1371/journal.pone.0008155.

Hnisz, D., Day, D. S. and Young, R. A. (2016) ‘Leading Edge Review Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control; Phil-lips-Cremins and Corces, 2013); here, we focus on the insulated neighborhood as a model for further exploration of the principles that underpin gene control in mammalian systems’. doi: 10.1016/j.cell.2016.10.024.

Holzmann, J. *et al.* (2019) ‘Absolute quantification of cohesin, CTCF and their regulators in human cells’, *eLife*, 8, p. e46269. doi: 10.7554/eLife.46269.

Hon, G. C. *et al.* (2013) ‘Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues’, *Nature Genetics*, 45(10), pp. 1198–1206. doi: 10.1038/ng.2746.

Hoogenkamp, M. *et al.* (2009) ‘Early chromatin unfolding by RUNX1: A molecular explanation for differential requirements during specification versus maintenance of the hematopoietic gene expression program’, *Blood*, 114(2), pp. 299–309. doi: 10.1182/blood-2008-11-191890.

Hu, S. *et al.* (2013) ‘DNA methylation presents distinct binding sites for human transcription factors’, *eLife*, 2013(2). doi: 10.7554/eLife.00726.

Huff, J. T. and Zilberman, D. (2014) ‘Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes’, *Cell*, 156(6), pp. 1286–1297. doi: 10.1016/j.cell.2014.01.029.

Ieda, M. *et al.* (2010) ‘Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors’, *Cell*, 142(3), pp. 375–386. doi: 10.1016/j.cell.2010.07.002.

Iguchi-Ariga, S. M. and Schaffner, W. (1989) ‘CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation.’, *Genes & development*, 3(5), pp. 612–619. doi: 10.1101/gad.3.5.612.

Iwafuchi-Doi, M. *et al.* (2016) ‘The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation’, *Molecular Cell*, 62(1), pp. 79–91. doi: 10.1016/j.molcel.2016.03.001.

Iwafuchi-Doi, M. (2019) ‘The mechanistic basis for chromatin regulation by pioneer transcription factors’, *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 11(1), p. e1427. doi: 10.1002/wsbm.1427.

Jansen, C. *et al.* (2019) ‘Building gene regulatory networks from scATAC-seq and scRNA-seq using Linked Self Organizing Maps’, *PLOS Computational Biology*. Edited by C. S. Leslie, 15(11), p. e1006555. doi: 10.1371/journal.pcbi.1006555.

Johnson, D. S. *et al.* (2007) ‘Genome-wide mapping of in vivo protein-DNA interactions’, *Science*, 316(5830), pp. 1497–1502. doi: 10.1126/science.1141319.

Jolma, A. and Taipale, J. (2011) ‘Methods for analysis of transcription factor DNA-binding specificity in vitro’, *Sub-Cellular Biochemistry*, 52, pp. 155–173. doi: 10.1007/978-90-481-9069-0_7.

Kadauke, S. *et al.* (2012) ‘Tissue-specific mitotic bookmarking by hematopoietic transcription factor GATA1’, *Cell*, 150(4), pp. 725–737. doi: 10.1016/j.cell.2012.06.038.

Kentepozidou, E. *et al.* (2020) ‘Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains’, *Genome Biology*, 21(1). doi: 10.1186/s13059-019-1894-x.

King, H. W. and Klose, R. J. (2017) ‘The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells’, *eLife*, 6. doi: 10.7554/eLife.22631.

Kitchen, N. S. and Schoenherr, C. J. (2010) ‘Sumoylation modulates a domain in CTCF that activates transcription and decondenses chromatin’, *Journal of Cellular Biochemistry*, 111(3), pp. 665–675. doi: 10.1002/jcb.22751.

Krysinska, H. *et al.* (2007) ‘A Two-Step, PU.1-Dependent Mechanism for Developmentally Regulated Chromatin Remodeling and Transcription of the c-fms Gene’, *Molecular and Cellular Biology*, 27(3), pp. 878–887. doi: 10.1128/mcb.01915-06.

Lai, X. *et al.* (2018) ‘Pioneer factors in animals and plants—colonizing chromatin for gene regulation’, *Molecules*. MDPI AG. doi: 10.3390/molecules23081914.

Laiosa, C. V. *et al.* (2006) ‘Reprogramming of Committed T Cell Progenitors to Macrophages and Dendritic Cells by C/EBPα and PU.1 Transcription Factors’, *Immunity*, 25(5), pp. 731–744. doi: 10.1016/j.immuni.2006.09.011.

Lambert, S. A. *et al.* (2018) ‘The Human Transcription Factors’, *Cell*. Cell Press, pp. 650–665. doi: 10.1016/j.cell.2018.01.029.

Langmead, B. and Salzberg, S. L. (2012) ‘Fast gapped-read alignment with Bowtie 2’, *Nature Methods*, 9(4), pp. 357–359. doi: 10.1038/nmeth.1923.

Laptenko, O. *et al.* (2011) ‘p53 binding to nucleosomes within the p21 promoter in vivo leads to nucleosome loss and transcriptional activation’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(26), pp. 10385–10390. doi: 10.1073/pnas.1105680108.

Lay, F. D., Kelly, T. K. and Jones, P. A. (2018) ‘Nucleosome occupancy and methylome sequencing (NOMe-seq)’, in *Methods in Molecular Biology*. Humana Press Inc., pp. 267–284. doi: 10.1007/978-1-4939-7481-8_14.

Lecun, Y., Bengio, Y. and Hinton, G. (2015) ‘Deep learning’, *Nature*. Nature Publishing Group, pp. 436–444. doi: 10.1038/nature14539.

Lee, C. K. *et al.* (2004) ‘Evidence for nucleosome depletion at active regulatory regions genome-wide’, *Nature Genetics*, 36(8), pp. 900–905. doi: 10.1038/ng1400.

Lee, J. T. (2003) ‘Molecular links between X-inactivation and autosomal imprinting: X-inactivation as a driving force for the evolution of imprinting?’, *Current Biology*. Cell Press, pp. R242–R254. doi: 10.1016/S0960-9822(03)00162-3.

Lefebvre, V. (2010) ‘The SoxD transcription factors - Sox5, Sox6, and Sox13 - are key cell fate modulators’, *International Journal of Biochemistry and Cell Biology*. NIH Public Access, pp. 429–432. doi: 10.1016/j.biocel.2009.07.016.

Lefebvre, V., Li, P. and De Crombrugghe, B. (1998) ‘A new long form of Sox5 (L-Sox5), Sox6 and Sox9 are coexpressed in chondrogenesis and cooperatively activate the type II collagen gene’, *EMBO Journal*, 17(19), pp. 5718–5733. doi: 10.1093/emboj/17.19.5718.

Letunic, I., Doerks, T. and Bork, P. (2015) ‘SMART: Recent updates, new developments and status in 2015’, *Nucleic Acids Research*, 43(D1), pp. D257–D260. doi: 10.1093/nar/gku949.

Libbrecht, M. W. *et al.* (2019) ‘A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types’, *Genome Biology*, 20(1), p. 180. doi: 10.1186/s13059-019-1784-2.

Liu, Y. *et al.* (2017) ‘Widespread Mitotic Bookmarking by Histone Marks and Transcription Factors in Pluripotent Stem Cells’, *Cell Reports*, 19(7), pp. 1283–1293. doi: 10.1016/j.celrep.2017.04.067.

Longobardi, E. *et al.* (2010) ‘Prep1 (pKnox1)-deficiency leads to spontaneous tumor development in mice and accelerates EμMyc lymphomagenesis: A tumor suppressor role for Prep1’, *Molecular Oncology*, 4(2), pp. 126–134. doi: 10.1016/j.molonc.2010.01.001.

Loyola-Gonzalez, O. (2019) ‘Black-Box vs. White-Box: Understanding Their Advantages and Weaknesses From a Practical Point of View’, *IEEE Access*, 7, pp. 154096–154113. doi: 10.1109/ACCESS.2019.2949286.

Martin, P. C. N. and Zabet, N. R. (2020) ‘Dissecting the binding mechanisms of transcription factors to DNA using a statistical thermodynamics framework’, *Computational and Structural Biotechnology Journal*, 18, pp. 3590–3605. doi: 10.1016/j.csbj.2020.11.006.

Mayran, A. and Drouin, J. (2018) ‘Pioneer transcription factors shape the epigenetic landscape’, *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology Inc., pp. 13795–13804. doi: 10.1074/jbc.R117.001232.

Meuleman, W. *et al.* (2019) ‘Index and biological spectrum of accessible DNA elements in the human genome’, *bioRxiv*, p. 822510. doi: 10.1101/822510.

Michael, A. K. *et al.* (2020) ‘Mechanisms of OCT4-SOX2 motif readout on nucleosomes’, *Science (New York, N.Y.)*, 368(6498), pp. 1460–1465. doi: 10.1126/science.abb0074.

Morgunova, E. *et al.* (2015) ‘Structural insights into the DNA-binding specificity of E2F family transcription factors’, *Nature Communications*, 6. doi: 10.1038/ncomms10050.

Morgunova, E. and Taipale, J. (2017) ‘Structural perspective of cooperative transcription factor binding’, *Current Opinion in Structural Biology*. Elsevier Ltd, pp. 1–8. doi: 10.1016/j.sbi.2017.03.006.

Nanni, L., Ceri, S. and Logie, C. (2020) ‘Spatial patterns of CTCF sites define the anatomy of TADs and their boundaries’, *Genome Biology*, 21(1), p. 197. doi: 10.1186/s13059-020-02108-x.

Narlikar, L. *et al.* (2010) ‘Genome-wide discovery of human heart enhancers’, *Genome Research*, 20(3), pp. 381–392. doi: 10.1101/gr.098657.109.

NCBI (2019) *GRCh38 - hg38 - Genome - Assembly - NCBI*. Available at: https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.26/ (Accessed: 7 September 2020).

Nili, E. L. *et al.* (2010) ‘p53 binds preferentially to genomic regions with high DNA-encoded nucleosome occupancy’, *Genome Research*, 20(10), pp. 1361–1368. doi: 10.1101/gr.103945.109.

Oldfield, A. J. *et al.* (2014) ‘Histone-Fold Domain Protein NF-Y Promotes Chromatin Accessibility for Cell Type-Specific Master Transcription Factors’, *Molecular Cell*, 55(5), pp. 708–722. doi: 10.1016/j.molcel.2014.07.005.

Owens, N. *et al.* (2019) ‘CTCF confers local nucleosome resiliency after dna replication and during mitosis’, *eLife*, 8. doi: 10.7554/eLife.47898.

Pardo, M. *et al.* (2010) ‘An Expanded Oct4 Interaction Network: Implications for Stem Cell Biology, Development, and Disease’, *Cell Stem Cell*, 6(4), pp. 382–395. doi: 10.1016/j.stem.2010.03.004.

Park, N. I. *et al.* (2017) ‘ASCL1 Reorganizes Chromatin to Direct Neuronal Fate and Suppress Tumorigenicity of Glioblastoma Stem Cells’, *Cell Stem Cell*, 21(2), pp. 209-224.e7. doi: 10.1016/j.stem.2017.06.004.

Park, P. J. (2009) ‘ChIP-seq: Advantages and challenges of a maturing technology’, *Nature Reviews Genetics*, pp. 669–680. doi: 10.1038/nrg2641.

Perrot, V. and Rechler, M. M. (2005) ‘The coactivator p300 directly acetylates the forkhead transcription factor Foxo1 and stimulates Foxo1-induced transcription’, *Molecular Endocrinology*, 19(9), pp. 2283–2298. doi: 10.1210/me.2004-0292.

Peter, I. S. and Davidson, E. H. (2011) ‘Evolution of gene regulatory networks controlling body plan development’, *Cell*, pp. 970–985. doi: 10.1016/j.cell.2011.02.017.

Phuycharoen, M. *et al.* (2019) ‘Uncovering Tissue-Specific Binding Features from Differential Deep Learning’, *bioRxiv*, p. 606269. doi: 10.1101/606269.

Pique-Regi, R. *et al.* (2011) ‘Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data’, *Genome Research*, 21(3), pp. 447–455. doi: 10.1101/gr.112623.110.

Pujato, M. *et al.* (2014) ‘Prediction of DNA binding motifs from 3D models of transcription factors; identifying TLX3 regulated genes’, *Nucleic Acids Research*, 42(22), pp. 13500–13512. doi: 10.1093/nar/gku1228.

Quang, D. and Xie, X. (2016) ‘DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences’, *Nucleic Acids Research*, 44(11). doi: 10.1093/nar/gkw226.

Quang, D. and Xie, X. (2019) ‘FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data’, *Methods*, 166, pp. 40–47. doi: 10.1016/j.ymeth.2019.03.020.

Raj, A. *et al.* (2015) ‘msCentipede: Modeling Heterogeneity across Genomic Sites and Replicates Improves Accuracy in the Inference of Transcription Factor Binding’, *PLOS ONE*. Edited by D. Zheng, 10(9), p. e0138030. doi: 10.1371/journal.pone.0138030.

Rauscher, F. J. *et al.* (1988) ‘Fos and Jun bind cooperatively to the AP-1 site: reconstitution in vitro.’, *Genes & development*, 2(12 B), pp. 1687–1699. doi: 10.1101/gad.2.12b.1687.

Ravasi, T. *et al.* (2010) ‘An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man’, *Cell*, 140(5), pp. 744–752. doi: 10.1016/j.cell.2010.01.044.

Robertson, K. D. (2005) ‘DNA methylation and human disease’, *Nature Reviews Genetics*. Nat Rev Genet, pp. 597–610. doi: 10.1038/nrg1655.

Roider, H. G. *et al.* (2007) ‘Predicting transcription factor affinities to DNA from a biophysical model’, *Bioinformatics*, 23(2), pp. 134–141. doi: 10.1093/bioinformatics/btl565.

Rojas, A. *et al.* (2005) ‘Gata4 expression in lateral mesoderm is downstream of BMP4 and is activated directly by Forkhead and GATA transcription factors through a distal enhancer element’, *Development*, 132(15), pp. 3405–3417. doi: 10.1242/dev.01913.

Di Rosa, P. *et al.* (2007) ‘The homeodomain transcription factor Prep1 (pKnox1) is required for hematopoietic stem and progenitor cell activity’, *Developmental Biology*, 311(2), pp. 324–334. doi: 10.1016/j.ydbio.2007.08.031.

Sahu, G. *et al.* (2010) ‘p53 binding to nucleosomal DNA depends on the rotational positioning of DNA response element’, *Journal of Biological Chemistry*, 285(2), pp. 1321–1332. doi: 10.1074/jbc.M109.081182.

Sammons, M. A. *et al.* (2015) ‘TP53 engagement with the genome occurs in distinct local chromatin environments via pioneer factor activity’, *Genome Research*, 25(2), pp. 179–188. doi: 10.1101/gr.181883.114.

Sandelin, A. *et al.* (2004) ‘JASPAR: an open-access database for eukaryotic transcription factor binding profiles’, *Nucleic acids research*, 32, pp. D91-4. doi: 10.1093/nar/gkh012.

Schep, A. N. *et al.* (2017) ‘ChromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data’, *Nature Methods*, 14(10), pp. 975–978. doi: 10.1038/nmeth.4401.

Schübeler, D. (2015) ‘Function and information content of DNA methylation’, *Nature*. Nature Publishing Group, pp. 321–326. doi: 10.1038/nature14192.

Shannon, P. and Richards, M. (2018) ‘An Annotated Collection of Protein-DNA Binding Sequence Motifs’, *Bioconductor-MotifDb*. doi: 10.18129/B9.bioc.MotifDb.

Shivaswamy, S. *et al.* (2008) ‘Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation’, *PLoS Biology*, 6(3), pp. 0618–0630. doi: 10.1371/journal.pbio.0060065.

Singh, H., Khan, A. A. and Dinner, A. R. (2014) ‘Gene regulatory networks in the immune system’, *Trends in Immunology*. Elsevier Ltd, pp. 211–218. doi: 10.1016/j.it.2014.03.006.

Soufi, A. *et al.* (2015) ‘Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming’, *Cell*, 161(3), pp. 555–568. doi: 10.1016/j.cell.2015.03.017.

Soufi, A., Donahue, G. and Zaret, K. S. (2012) ‘Facilitators and impediments of the pluripotency reprogramming factors’ initial engagement with the genome’, *Cell*, 151(5), pp. 994–1004. doi: 10.1016/j.cell.2012.09.045.

Spitz, F. and Furlong, E. E. M. (2012) ‘Core promoter Transcription factors: from enhancer binding to developmental control’, *NATURE REVIEWS | GENETICS*, 13, p. 613. doi: 10.1038/nrg3207.

Teif, V. B. *et al.* (2014) ‘Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development’, *Genome Research*, 24(8), pp. 1285–1295. doi: 10.1101/gr.164418.113.

Telenti, A. *et al.* (2018) ‘Deep learning of genomic variation and regulatory network data’, *Human Molecular Genetics*. Oxford University Press, pp. R63–R71. doi: 10.1093/hmg/ddy115.

Tibshirani, R. (1996) ‘Regression Shrinkage and Selection Via the Lasso’, *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), pp. 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Tsompana, M. and Buck, M. J. (2014) ‘Chromatin accessibility: A window into the genome’, *Epigenetics and Chromatin*. BioMed Central Ltd., pp. 1–16. doi: 10.1186/1756-8935-7-33.

Villar, D., Flicek, P. and Odom, D. T. (2014) ‘Evolution of transcription factor binding in metazoans-mechanisms and functional implications’, *Nature Reviews Genetics*. Nature Publishing Group, pp. 221–233. doi: 10.1038/nrg3481.

Wapinski, O. L. *et al.* (2013) ‘XHierarchical mechanisms for direct reprogramming of fibroblasts to neurons’, *Cell*, 155(3), p. 621. doi: 10.1016/j.cell.2013.09.028.

Wasserman, L. (2004) ‘Linear and Logistic Regression’, in, pp. 209–229. doi: 10.1007/978-0-387-21736-9_13.

Watt, F. and Molloy, P. L. (1988) ‘Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter.’, *Genes & development*, 2(9), pp. 1136–1143. doi: 10.1101/gad.2.9.1136.

Weintraub, H. and Groudine, M. (1976) ‘Chromosomal subunits in active genes have an altered conformation’, *Science*, 193(4256), pp. 848–856. doi: 10.1126/science.948749.

Wiehle, L. *et al.* (2019) ‘DNA (de)methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries’. doi: 10.1101/gr.239707.118.

Xie, H. *et al.* (2004) ‘Stepwise reprogramming of B cells into macrophages’, *Cell*, 117(5), pp. 663–676. doi: 10.1016/S0092-8674(04)00419-2.

Xu, J. *et al.* (2010) ‘Transcriptional silencing of γ-globin by BCL11A involves long-range interactions and cooperation with SOX6’, *Genes and Development*, 24(8), pp. 783–789. doi: 10.1101/gad.1897310.

Yagi, R. *et al.* (2007) ‘Transcription factor TEAD4 specifies the trophectoderm lineage at the beginning of mammalian development’, *Development*, 134(21), pp. 3827–3836. doi: 10.1242/dev.010223.

Yan, W. *et al.* (2016) ‘Biological networks for cancer candidate biomarkers discovery’, *Cancer Informatics*. Libertas Academica Ltd., pp. 1–7. doi: 10.4137/CIN.S39458.

Yin, Y. *et al.* (2017) ‘Impact of cytosine methylation on DNA binding specificities of human transcription factors’, *Science*, 356(6337). doi: 10.1126/science.aaj2239.

Younger, S. T. and Rinn, J. L. (2017) ‘P53 regulates enhancer accessibility and activity in response to DNA damage’, *Nucleic Acids Research*, 45(17), pp. 9889–9900. doi: 10.1093/nar/gkx577.

Yu, X. and Buck, M. J. (2020) ‘Pioneer factors and their in vitro identification methods’, *Molecular Genetics and Genomics*. Springer, pp. 825–835. doi: 10.1007/s00438-020-01675-9.

Yuan, M. and Lin, Y. (2006) *Model selection and estimation in regression with grouped variables*, *J. R. Statist. Soc. B*.

Zabet, N. R. and Adryan, B. (2015) ‘Estimating binding properties of transcription factors from genome-wide binding profiles’, *Nucleic Acids Research*, 43(1), pp. 84–94. doi: 10.1093/nar/gku1269.

Zamanighomi, M. *et al.* (2018) ‘Unsupervised clustering and epigenetic classification of single cells’, *Nature Communications*, 9(1). doi: 10.1038/s41467-018-04629-3.

Zaret, K. S. and Carroll, J. S. (2011) ‘Pioneer transcription factors: Establishing competence for gene expression’, *Genes and Development*. Cold Spring Harbor Laboratory Press, pp. 2227–2241. doi: 10.1101/gad.176826.111.

Zaret, K. S. and Mango, S. E. (2016) ‘Pioneer transcription factors, chromatin dynamics, and cell fate control’, *Current Opinion in Genetics and Development*. Elsevier Ltd, pp. 76–81. doi: 10.1016/j.gde.2015.12.003.

Zhang, Y. *et al.* (2008) ‘Model-based analysis of ChIP-Seq (MACS)’, *Genome Biology*, 9(9). doi: 10.1186/gb-2008-9-9-r137.

Zhao, Y., Granas, D. and Stormo, G. D. (2009) ‘Inferring binding energies from selected binding sites’, *PLoS Computational Biology*, 5(12), p. 1000590. doi: 10.1371/journal.pcbi.1000590.

Zheng, N. *et al.* (1999) ‘Structural basis of DNA recognition by the heterodimeric cell cycle transcription factor E2F-DP’, *Genes and Development*, 13(6), pp. 666–674. doi: 10.1101/gad.13.6.666.

Zhou, W. *et al.* (2017) ‘Genome-wide prediction of DNase i hypersensitivity using gene expression’, *Nature Communications*, 8(1). doi: 10.1038/s41467-017-01188-x.