Computational and Structural Biotechnology Journal 18 (2020) 3590–3605

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

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

Patrick C.N. Martin [a,b], Nicolae Radu Zabet [a,c,*]

[a] School of Life Sciences, University of Essex, Colchester CO4 3SQ, UK
[b] Biotech Research and Innovation Centre (BRIC), University of Copenhagen, DK-2200 Copenhagen, Denmark
[c] Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK

ARTICLE INFO

Article history:
Received 18 August 2020
Received in revised form 2 November 2020
Accepted 4 November 2020
Available online 12 November 2020

ABSTRACT

Transcription Factors (TFs) bind to DNA and control activity of target genes. Here, we present ChIPanalyser, a user-friendly, versatile and powerful R/Bioconductor package predicting and modelling the binding of TFs to DNA. ChIPanalyser performs similarly to state-of-the-art tools, but is an *explainable model* and provides biological insights into binding mechanisms of TFs. We focused on investigating the binding mechanisms of three TFs that are known architectural proteins CTCF, BEAF-32 and su(Hw) in three Drosophila cell lines (BG3, Kc167 and S2). While CTCF preferentially binds only to a subset of high affinity sites located mainly in open chromatin, BEAF-32 binds to most of its high affinity binding sites available in open chromatin. In contrast, su(Hw) binds to both open chromatin and also partially closed chromatin. Most importantly, differences in TF binding profiles between cell lines for these TFs are mainly driven by differences in DNA accessibility and not by differences in TF concentrations between cell lines. Finally, we investigated binding of Hox TFs in Drosophila and found that Ubx binds only in open chromatin, while Abd-B and Dfd are capable to bind in both open and partially closed chromatin. Overall, our results show that TFs display different binding mechanisms and that our model is able to recapitulate their specific binding behaviour.

1. Background

Decades of research have shown that gene expression plays an essential role in the livelihood of cells and organisms. From development to cellular homoeostasis, the activation or repression of gene expression enables cells, and by extension organisms, to function properly. One of the key components of the regulation of gene expression is Transcription Factors (TFs). The most commonly used experimental method to determine specific regions of DNA where TFs bind is chromatin immunoprecipitation followed by sequencing (ChIP-seq) [1,2]. This technique has become the gold standard to determine the binding profiles of TFs to the genome, but, despite the huge impact on understanding gene regulation, it does not provide a mechanistic model of what drives the binding of TFs to those regions or even how genes are regulated. While we still lack a complete predictive model for gene expression, over the years, many factors have been identified as contributing to context dependant TF binding.

An important aspect to consider concerning TF binding specificity is the DNA sequence itself. While some TFs do not bind in a sequence specific manner, our work focuses on the sequence specific TFs [3–6]. The most common way to describe this motif is in the form of a Position Weight Matrix (PWM); a measure of binding frequency between TFs and DNA weighted by the genomic base pair frequency [3,7]. Nevertheless, TFs can have tens of thousands potential binding sites within each genome, yet only a few hundred will be occupied by TFs [8,9].

Previous studies have shown that some TF binding events are TF concentration dependent [10–13], where varying the concentration of the TF will drive the expression of different sets of genes. However, there are many more spurious sites, rather than functional binding sites where TFs could bind. This still begs the question: how do TFs distinguish between bound and unbound binding sites?.

One way to reduce the number of available sites is to consider DNA accessibility. Are these sites even available for binding in the first place? This assumes that TFs would bind only to sites that are accessible and cannot locate sites within dense chromatin [14,15]. Nevertheless, there is a certain class of TFs known as pioneer TFs are able to bind in closed chromatin. More specifically,

* Corresponding author at: Blizard Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London E1 2AT, UK.
E-mail address: r.zabet@qmul.ac.uk (N.R. Zabet).

https://doi.org/10.1016/j.csbj.2020.11.006
2001-0370/© 2020 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

pioneer TFs can bind sites in closed dense chromatin and subsequently open the chromatin [16–19].

Over the years, many tools and frameworks have aimed to predict transcription factor binding. One of the earliest tools incorporating DNA accessibility was the PIQ algorithm (Protein Interaction Quantification) which implemented a machine learning type algorithm to filter out binding sites located in inaccessible DNA [20]. Later, msCENTIPEDE improved upon CENTIPEDE using multiscale models for inhomogeneous Poisson processes to untangle TF binding with respect to DNA accessibility [21]. Some notable tools that have been developed through DREAM challenges are FactorNet, implementing a deep learning framework [22], Anchor, relying on a XGBoost system [23] and Catchitt making use of supervised machine learning and iterative training [24]. While machine learning methods predict TF binding events with high accuracy, they are often difficult to interpret; i.e., it is not clear what these models actually use to produce their predictions. The interpretability issues have been widely discussed in the last decade [25,26] and machine learning methods are not always best suited to understand the mechanism driving a biological phenomenon.

We previously showed that statistical thermodynamics can be used to model TF binding to DNA with high accuracy [13]. Considering only binding energy between TFs and DNA (estimated by the PWM and a scaling factor modulating the binding energy), the number of bound molecules to the DNA and DNA accessibility, we modelled binding of five TFs in Drosophila embryo. Our results confirmed that, for some TFs, this model is sufficient to explain the majority of observed binding events in ChIP data and we were able to backwards infer number of bound molecules and specificity for five TFs in Drosophila embryo (bcd, cad, gt, hb and Kr).

In this manuscript, we build upon our previous model and developed ChIPanalyser a versatile and user-friendly R/Bioconductor package [27,28]. Furthermore, we used this model to describe the behaviour of several Drosophila TFs (CTCF, BEAF-32, su(Hw), Ubx, Abd-B and Dfd) in different cell lines (BG3, Kc167 and S2).

We show that the performance of ChIPanalyser is at least similar to other TF binding tools available. However, our results provide a mechanistic interpretation of TF binding behaviour and propose a new classification of TFs based on fine details of their binding mechanism. In particular, we found that DNA accessibility is the main driver that explains binding of CTCF, BEAF-32 and su (Hw) in three Drosophila cell lines (BG3, Kc167 and S2) and that relatively medium changes in the concentrations of these TFs lead to only negligible changes in their binding profiles. We also show that TF binding specificity can be achieved by their capacity to bind regions with different levels of DNA accessibility. In particular, we showed that, while Ubx, Abd-B and Dfd have similar binding motifs, the differences in their binding to DNA could be explained by their different capacity to bind dense chromatin, with Ubx binding only in highly accessible chromatin and Dfd and Abd-B binding denser chromatin.

2. Methods

2.1. Model description

ChIPanalyser is an R package available on Bionconductor [28]. The package is an implementation of the statistical thermodynamics model proposed in [13]. Briefly, the model requires *(i)* a PWM (Position Weight Matrix) or PFM (Position Frequency Matrices) of the TF of interest, *(ii)* DNA accessibility data to model binding site accessibility and two additional parameters: *(iii)* λ (a PWM scaling factor) and *(iv)* N (the number of bound molecules) [13]. The probability of a position j on the DNA being occupied is given by [13]:

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

with N the number of molecules bound to the DNA, a_j the accessibility at site j, λ a scaling factor of the PWM score, w_j the PWM score at site j, L the length of the DNA and n is the ploidy level. Given the size of the genome and the range of TF abundances reported in the literature, we can assume that the number of available sites is much larger than the number of bound molecules. Thus, $L \cdot n \cdot \left\langle a_i e^{\left(\frac{1}{\lambda} w_i\right)} \right\rangle_i$ does not consider the number of bound molecules as it describes the rest of the genome that will not be bound by a given TF.

It should be noted that λ represents the relative affinity a TF could have for a binding site [7]. In particular, it represents how well a given TF can discriminate between low/medium and high affinity sites. Also note that number of bound molecules and concentration are not the same since in majority of the cases a large proportion of the molecules are freely diffusing in the nucleoplasm (reviewed in [29]). λ and N are not always readily available in the form of experimental data and, thus, we used ChIP-seq data and select the values of these parameters that maximise (or minimise) the goodness of fit metrics.

2.2. Datasets

To carry out the analysis described in this manuscript, we selected data originating from various sources (see Table S1). We provide the code used in the is manuscript in a GitHub repository (https://github.com/patrickCNMartin/ChIPanalyserSub).

DNA Sequence: Reference Sequences of *Drosophila melanogaster* (dm6) [30,31] and *Homo sapiens* (hg38) [32] were extracted from the Bsgenome R packages [33]. All data sets were either aligned to the dm6 versions of the Drosophila genome or lifted over from dm3 to dm6 using the UCSC genome liftover chain [34].

PWM and PFM: Binding Motif matrices were downloaded from online repositories (JASPAR) [35] or extracted from the MotifDb R package [36], which collects and compiles PFMs and Position Probability Matrices (PPM) from various online repositories (see Figure S1 in *Supplementary Materials*). For the purpose of method comparison requirements (msCENTIPEDE), TF binding sites were extracted using FIMO from the MEME-suit tool kit [37].

ChIP-seq: ChIP enrichment signal and ChIP peaks were downloaded (pre-processed) from modEncode in three Drosophila cell line: Kc167, S2 and BG3. Note that some of these ChIP datasets were generated in RNAi mutant cells. Despite the differences between ChIP-chip and ChIP-seq, they are sufficiently similar to be comparable for the purpose of this analysis [1]. Both describe TF binding events and both are provided in similar formats (.wig,.bed,.bed Graph,.bigWig,.gff,.gff3). Supplementary data sets were downloaded from GEO. GEO datasets were aligned to the genome (dm6) using bowtie-2 (- -non-deterministic). *SAM* files were converted to *BAM* files using samtools [38]. Peaks and pile-up signal were called using macs2 with a 0.01 FDR (-q 0.01) [39] in order to ensure the robustness of the peaks selected. Processed data sets for *Homo sapiens* were directly downloaded from ENCODE were already aligned to hg38. We selected one of the datasets provided and used in the DREAM challenge competition related to TF binding prediction. Peak replicates were combined using the GenomicsRanges package in R. Datasets used for this analysis are described in Table S1 in *Supplementary Materials*.

DNA accessibility: DNase I hypersensitivity data was generated by modEncode for the three cell lines used in this analysis [40]. We aligned fastq files to the dm6 genome build using bowtie-2 (- -non-deterministic). *SAM* files were converted to *BAM* files using samtools [38]. Peaks and read pile-ups were called and produced using macs2 (–broad-call -cutoff 0.05 -q 0.05) [39]. We selected broad peaks and a more relaxed FDR, since DNA accessibility is characterised by broader regions compared to ChIP data. DNase I hypersensitvity data for *Homo sapiens* was directly downloaded from ENCODE and replicates were merged using samtools. DNase peak replicates were combined using the GenomicsRanges package in R. The level of accessibility is consistent with past experiments (see Figure S2). ATAC-seq data for Kc167 cells was used from [41] and ATAC-seq scores were computed using macs2 as described in [41]. We selected a series of ATAC-seq signal thresholds that we would use as a cut-off point to select accessible/inaccessible DNA. These thresholds were based on signal quantiles from 0.05 to 0.95 by 0.05. We also considered 0.99,0.999, 0.9999 quantile thresholds. We will refer to this method as Quantized Density Accessibility (QDA).

RNA-seq In order to rescale TF abundance between cell lines, we used RNA-seq data from [42]. RNA-seq relative abundance was used to rescale the estimated number of bound molecules from one cell line to another.

2.3. Description of ChIPanalyser

The workflow of ChIPanalyser is described in Fig. 1. Briefly, the optimal set of parameters (for λ and N) can be inferred from ChIP data by maximising (or minimising) a goodness of fit metric. Nevertheless, if the user approximates these values by other means, ChIPanalyser does not require any training data at all. Using these values, ChIPanalyser will produce base pair resolution ChIP like profiles for different genomic regions and compare the prediction with the actual ChIP data (if that is provided by the user).

ChIPanalyser uses a set of genomic regions to infer optimal parameters. If the genomic regions are not provided by the user, the top n regions will be selected based on highest ChIP score after binning the genome into bins of 20 Kb (number of regions to be selected and bin width and can be customized). In the context of this analysis, ChIP score refers to the min/max normalised enrichment scores at base pair resolution provided in each data set. For our analysis, we split the entire genome into bins of 20 Kb and selected bins that contained at least one CTCF, BEAF-32 and su (Hw) peaks in at least one biological replicate and at least one cell line. By doing so, we ensure that the regions we will use in this analysis are common between all data sets. This resulted in 3293 bins of 20 Kb that contain at least one peak of any of these architectural proteins and at least one base pair of accessible DNA. In addition, we followed the same process for Hox transcription factors, which resulted in a total of 3838 bins of 20 Kb containing at least one peak for each TF (Ubx, Abd-b, and Dfd). Normalised and ordered bins (based on highest ChIP scores in that bin) were produced by the *processingChIP* function from ChIPanalyser. Following this, we selected the top ten regions in order to train our model (to infer N and λ by maximising or minimising a goodness of fit metric). While the top ten regions contain the strongest peaks (True positive signal), they also contain large segments of DNA that are not bound by a given TF (True Negative signal). The ratio of True positives to True negatives in the top ten regions provides a appropriate set of input data to train our model. Once we had selected the optimal set of parameters based on our training set, we validated our results on the other regions that do not contain the training set. For example, as regions were ordered based on ChIP signal (from strongest signal to lowest signal), we selected the top ten regions with the strongest signal score to train our model (regions 1 to 10) and following twenty regions (11 to 30) for validation.

During this step of the analysis (*processingChIP*), we also included a noise filtering method. The current model does not consider ChIP depletion, therefore all negative scores are replaced by 0. With that in mind, ChIPanlyser provides four methods of filtering noise: *Zero*, *Mean*, *Median* and *Sigmoid*. *Zero* removes only depletion scores (equivalent to “no noise filtering”). *Mean* and *Median* replace all scores below the mean or median after filtering out depletion scores. Finally, *Sigmoid* applies a logistic weighting to every score, modulat-

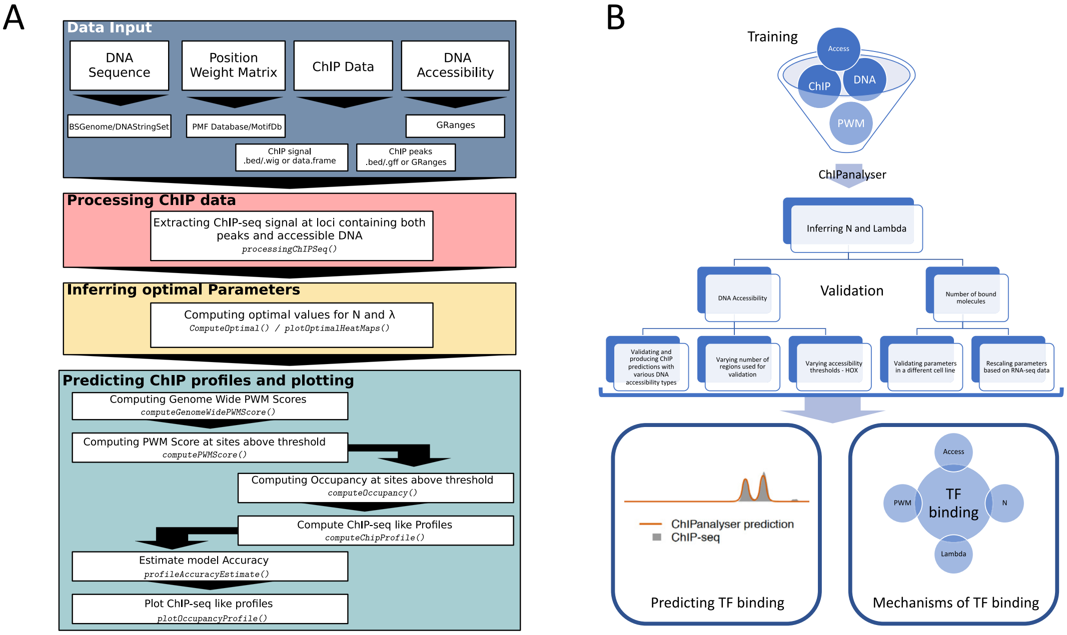

A
Data Input
DNA Sequence
Position Weight Matrix
ChIP Data
DNA Accessibility
BSGenome/DNAStringSet
PMF Database/MotifDb
GRanges
ChIP signal .bed/.wig or data.frame
ChIP peaks .bed/.gff or GRanges
Processing ChIP data
Extracting ChIP-seq signal at loci containing both peaks and accessible DNA
processingChIPSeq()
Inferring optimal Parameters
Computing optimal values for N and λ
ComputeOptimal() / plotOptimalHeatMaps()
Predicting ChIP profiles and plotting
Computing Genome Wide PWM Scores
computeGenomeWidePWMScore()
Computing PWM Score at sites above threshold
computePWMScore()
Computing Occupancy at sites above threshold
computeOccupancy()
Compute ChIP-seq like Profiles
computeChipProfile()
Estimate model Accuracy
profileAccuracyEstimate()
Plot ChIP-seq like profiles
plotOccupancyProfile()
B
Training
Access
ChIP
DNA
PWM
ChIPanalyser
Inferring N and Lambda
Validation
DNA Accessibility
Number of bound molecules
Validating and producing ChIP predictions with various DNA accessibility types
Varying number of regions used for validation
Varying accessibility thresholds - HOX
Validating parameters in a different cell line
Rescaling parameters based on RNA-seq data
ChIPanalyser prediction
ChIP-seq
Predicting TF binding
Access
PWM
TF binding
N
Lambda
Mechanisms of TF binding

Fig. 1. ChIPanalyser workflow. (**A**) ChIPanalyser follows the following work flow. **Data Input:** Data may come in various formats (e.g. bed, wig, gff etc.). **Processing ChIP-seq data:** If ChIP data is used to infer the optimal set of parameters (and/or validate model goodness of fit), ChIP data will be normalised and only regions of interest will be extracted for further analysis. **Inferring optimal parameters:** Inferring optimal parameters will be achieved by maximising (or minimising) a goodness of fit metric. **Predicting ChIP profiles and plotting:** Using the optimal values for number of bound molecules and the PWM scaling factor, ChIPanalyser will produce ChIP like profiles. Both optimal parameter heatmaps and ChIP profiles can be plotted using the package’s plotting functions. (**B**) shows the general workflow and the main computational experiments that were carried out throughout this manuscript.

ing ChIP scores around the 95th quantile point. All analysis in this manuscript was carried out after using the *Sigmoid* noise filtering method.

Once the *loci* of interest have been selected, we inferred the optimal set of parameters by using *computeOptimal* function. The optimal set of parameters are inferred by maximising (or minimising) the average goodness of fit metric over all regions selected. ChIPanalyser offers twelve different 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) (see Table S2 in Supplementary Materials). We also developed a novel method that describes the ratio of shared geometric area between curves and difference in area between curves. ChIPanalyser generates a ChIP like profile at a base pair level resolution, however window size may be adjusted. The goal is to mimic experimental ChIP profiles by smoothing high occupancy binding sites into ChIP like profiles. This approach was described by [13].

For this analysis we used a 100 bp window for validation. Goodness of fit is carried out by comparing our prediction to ChIP score data (as opposed to peak location overlap). The rationale behind using ChIP scores instead of peaks was twofold: *(i)* we consider peaks that are missed by peak calling algorithms and *(ii)* using ChIP scores ensures that we also consider signal enrichment. The latter is particularly relevant when estimating the number of bound molecules.

The evaluation method used by ChIPanalyer is significantly more stringent than methods used in other frameworks. When confusion matrices are required for scoring (AUC, recall, F-score, MCC, Accuracy, precision), ChIPanalyser uses 20 threshold values bound between the lowest occupancy score (predicted or experimental score) and the highest occupancy score (predicted or experimental score). The threshold values are squared in order to ensure a higher density of threshold values close to the lower end of occupancy scores. For every threshold value, ChIPanalyser compares its predicted profile to the experimental profile in 100 bp bins. If they both contain a "signal", we consider that ChIPanalyser has correctly predicted local ChIP enrichment. If ChIPanalyser predicts ChIP enrichment when no experimental signal is present, we consider this bin to be a false positive case. The same approach was used for false negative cases (Experimental enrichment but no predicted enrichment) and true negative cases (No enrichment in either experimental or predicted profiles). This approach ensures that the model is penalised if it fails to predict peak enrichment or conversely over estimates peak enrichment.

The optimal parameters inferred over training can be visualised in the form of a heatmap describing the score associated to each combination of λ and N. Heatmaps are produced using the *plotOptimalHeatMaps* function. Finally, using the optimal set of parameters, ChIPanalyser will produce ChIP like profiles that can be visualised using the *plotOccupancyProfiles* function provided by the package.

3. Results

3.1. Evaluation of ChIPanalyser

Previously, we showed how statistical thermodynamics can be used to mechanistically explain the binding of TFs in *Drosophila* [13]. The optimal set of parameters (see Methods) was inferred by maximising correlation and minimising Mean Squared Error (MSE) between the predicted profile and experimental ChIP data. Nevertheless, we observed that, in some cases, the predicted profiles and ChIP profiles display low correlation coefficient despite the profiles looking similar and vice versa (e.g. see Figure S3A and S3B in *Supplementary Figures*). In some cases, selecting the optimal parameters was hindered by little variation in correlation between parameter combinations and, thus, the selection of these parameters was exclusively driven by MSE (see Figure S3C in *Supplementary Figures*).

To reduce the potential influence of background noise, we tested four noise removal methods: *Zero*, *Mean*, *Median* and *Sigmoid*; see Methods. To test their performance, we used three CTCF datasets (see Table S1 in *Supplementary Tables*): *(i)* a ChIP-chip dataset with very little background noise (modEncode 2639), *(ii)* a ChIP-seq dataset with high background noise (modEncode 3674) and *(iii)* a combination of all ChIP datasets in S2 cells (by adding enrichment signals together at a base pair level). To ensure equal contribution of each data set, we normalised the signal prior to combining data sets. We ran the model on top ten regions (as described in Methods) and searched for the optimal set of parameters (λ and N) that optimised the goodness of fit metric (in this instance – AUC). All four noise filtering methods have little to no effect on ChIP data (see Figure S4 in *Supplementary Figures*). The Sigmoid method showed a slight signal reduction in smaller peaks (especially for noisy datasets), which was then translated into a slight improvement of the mean Area Under Curve Receiver Operator Characteristic (AUC ROC) score between ChIP signal and our predictions (see Figure S4 in *Supplementary Figures*). The distribution describes all AUC ROC scores for the ten regions used for this analysis.

In addition to Pearson correlation and MSE, we tested several goodness of fit metrics to verify the influence of the metrics on our model as described in Methods and Table S2 in *Supplementary Tables*. We used the same three CTCF datasets as described above and observed the emergence of two classes within these metrics: *(i)* similarity metrics that describe how similar the two curves are (correlation coefficients, precision, MCC, Accuracy, F-score and AUC ROC) and *(ii)* dissimilarity metrics that are a measure of how different two curves are (MSE, geometric ratio, recall and Kolmogorov–Smirnov distance). Our results showed that depending on the metric used, the optimal set of parameters varied significantly, but each of the two classes (similarity and dissimilarity metrics) displayed similar yet not identical values for the optimal parameters (see Supplementary Figure S5 A–F).

Goodness of fit metrics influence the way the model selects optimal parameters, but how does this translate to the individual predicted ChIP profile level? We further investigated this behaviour at the individual *loci* using the same three CTCF datasets. Fig. 2A-B shows that similarity metrics (black shades) tend to be less prone to false positive peaks but miss the actual ChIP signal strength within the peak (the height of the peak). On the other hand, dissimilarity metrics (light blue shade) generate far more false positives but accurately recover the height of the peaks.

Overall, the best performing metrics were AUC ROC and MSE. AUC ROC occasionally missed peak enrichment completely however, seemed to recover peak location fairly accurately, while MSE rarely missed peak enrichment but also produced a higher number of false positive peaks. For much of the following analysis, we used AUC ROC and MSE, since they are more widely used estimators and performed best. More specifically, MSE was used as the training metric to select the optimal set of parameters. AUC, recall, Spearman correlation and MSE were used for validating model performance.

To evaluate the performance of our model, we first used a chromosome withholding set up. The model was trained on the top 10 regions (as described in Methods – performed on modEncode 922) on chromosome 3R (Fig. 3A). We then validated our model using two approaches: *(i)* on the top 10 regions found on chr2R (Fig. 3B) and *(ii)* on the top 10 regions on chr3R excluding regions used for training (Fig. 3C). Our results show that ChIPanalyser

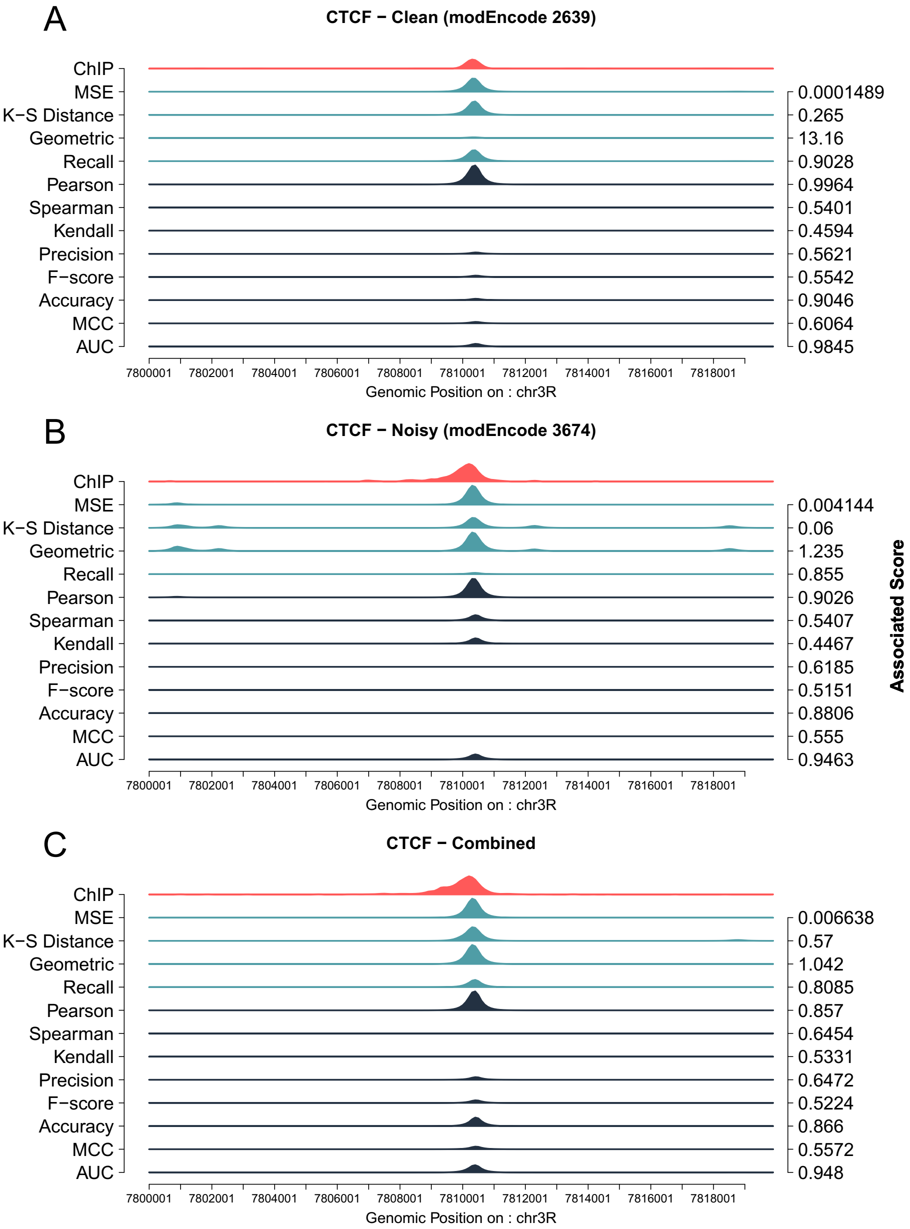
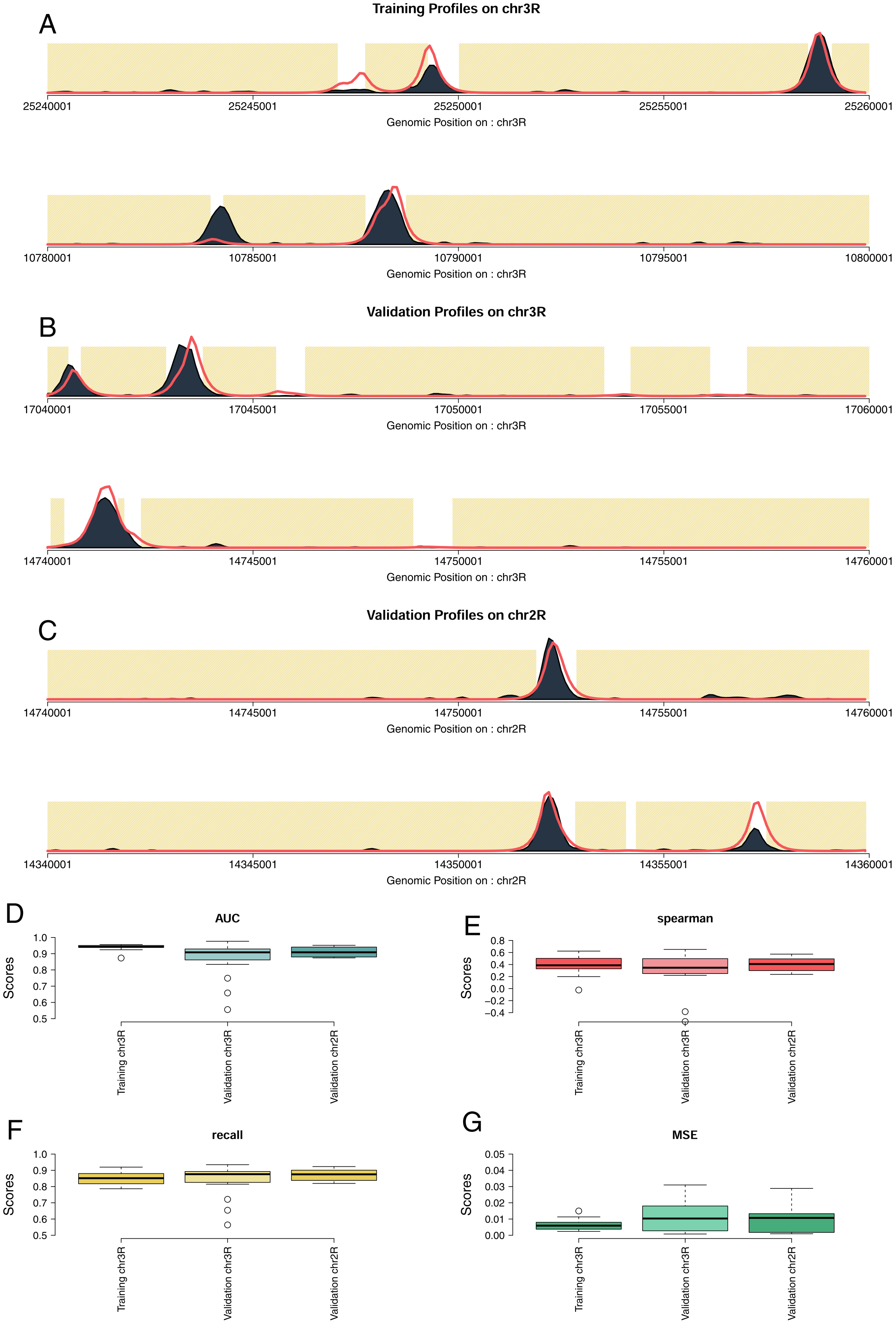


Fig. 2. Goodness of fit Methods are context dependent. (**A**) ChIPanalyser correctly predicts CTCF peaks in a clean ChIP dataset (modEncode 2639) for the majority of metrics used. (**B**) For a noisier dataset (modEncode 3674), dissimilarity metrics capture the height of the peak but also tend to show a high rate of False Positive peaks. In contrast, similarity metrics accurately predict the location of the peak, but tend to underestimate peak height. (**C**) Combining several ChIP replicates (all ChIP-seq datasets in S2 cells; see Table S1 in Supplementary Materials) does not reduce the rate of False Positive peaks for similarity metrics. The red profile shows experimental ChIP peaks, while light blue and dark blue are predicted profiles. Light blue and dark blue as dissimilarity and similarity metrics respectively. Associated scores are the scores for each profile when that metric was used to select optimal parameters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

accurately recovers peak location and enrichment between chromosomes (Fig. 3D–G).

In order to demonstrate that our model accurately recovers TF binding mechanisms, we also produced profiles in chr2R after training in chr3R but with the addition of PWM scores above threshold. In Fig. 4, we show predicted profiles (red lines) compared to experimental ChIP (dark blue) for BEAF-32 (**A**), CTCF (**B**) and su(Hw) (**C**). The vertical lines represent strong PWM scores (top 20%, top 20% and top 30% respectively). We observed that while some regions displayed strong PWM scores, this was not necessarily accompanied by experimental ChIP peaks. By using the statistical thermodynamic model, we were able to recapitulate ChIP peaks more accurately and thus demonstrating that our model predicts TF binding with higher accuracy than PWM scores alone.

Finally, we compared the performance of ChIPanalyser to other TF binding prediction frameworks namely PIQ, msCENTIPEDE and Catchitt (see Table 1). As many available tools and frameworks are restricted to only considering human or mouse data, we selected CTCF ChIP data in astrocyte cells (*Homo sapiens*) as provided by ENCODE (Table S1 in Supplementary Materials). This dataset was also used in the DREAM challenge competition related to TF binding prediction. It should be noted that neither PIQ nor msCENTIPEDE have a validation step and, for this reason, we ran both PIQ and msCENTIPEDE on both chr11 and chr18 of the human genome (full chromosomes). Input BAM files were truncated using samtools to only include these chromosomes. As PIQ and msCENTIPEDE provided discrete TF binding sites, we smoothed scores over 100 bp in order to keep the evaluation window consistent


Training Profiles on chr3R
A
25240001
25245001
25250001
25255001
25260001
Genomic Position on : chr3R
10780001
10785001
10790001
10795001
10800001
Genomic Position on : chr3R
Validation Profiles on chr3R
B
17040001
17045001
17050001
17055001
17060001
Genomic Position on : chr3R
14740001
14745001
14750001
14755001
14760001
Genomic Position on : chr3R
Validation Profiles on chr2R
C
14740001
14745001
14750001
14755001
14760001
Genomic Position on : chr2R
14340001
14345001
14350001
14355001
14360001
Genomic Position on : chr2R
D
AUC
Scores
1.0
0.9
0.8
0.7
0.6
0.5
Training chr3R
Validation chr3R
Validation chr2R
E
spearman
Scores
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Training chr3R
Validation chr3R
Validation chr2R
F
recall
Scores
1.0
0.9
0.8
0.7
0.6
0.5
Training chr3R
Validation chr3R
Validation chr2R
G
MSE
Scores
0.05
0.04
0.03
0.02
0.01
0.00
Training chr3R
Validation chr3R
Validation chr2R

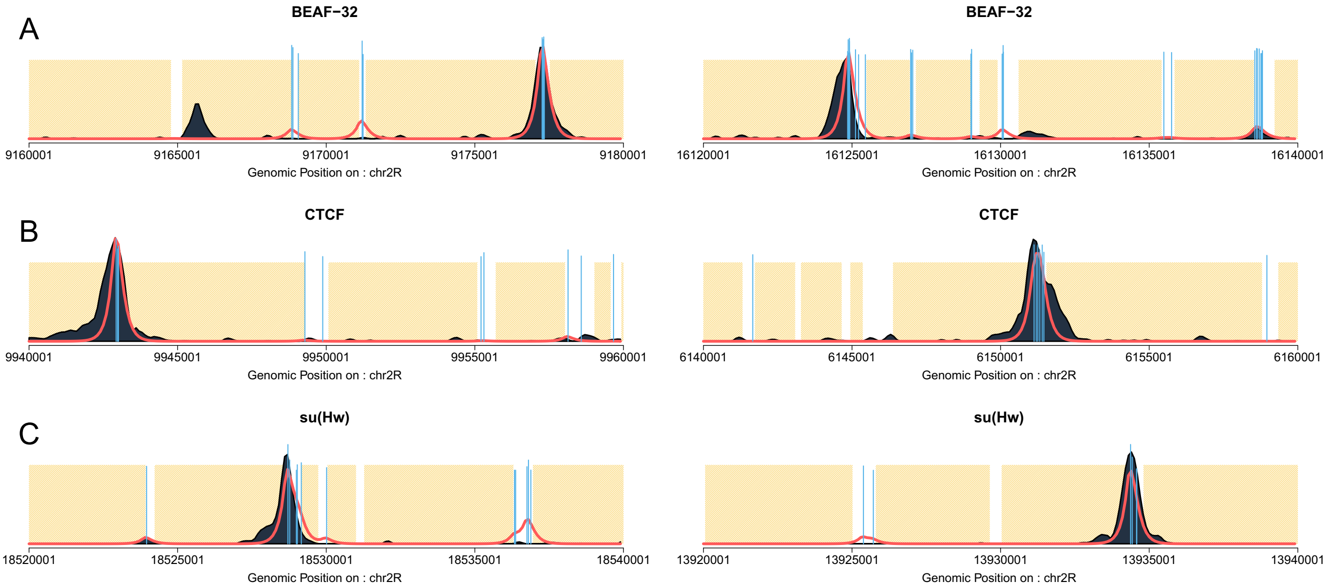


Fig. 4. ChIPanalyser models TF binding with higher accuracy than PWM scores alone. A-C show predicted ChIP-seq profiles for BEAF-32 (modEncode 922), CTCF (modEncode 282) and su(Hw)(modEncode 330). After training the model on chr3R, we validated the model on chr2R. The vertical blue lines represent normalised PWM scores for regions above threshold: 0.8 for BEAF-32 and CTCF and 0.7 for su(Hw). The red line represents our prediction while the dark blue region represent experimental ChIP. Yellow areas are regions on inaccessible DNA. Despite exhibiting strong PWM scores, some regions are not bound by TFs according to ChIP data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Tool and framework comparison We provide a breakdown of a few popular tools and frameworks for TF binding prediction. We assess the ease of use of each tool based on three main factors: ease of installation (package manager), knowledge of underlying code (if it is required to make changes to the underlying code) and finally package support (support and documentation).

	ChIPanalyser	Catchitt	FactorNet	Anchor	PIQ	msCENTIPEDE
Language	R	java	python 2.7	python3.6/perl 5.1	R	python 3.6
Organisms	All*	All	Human	Human	All*	Human
Training & Validation	Yes	Yes	Yes	Yes	No	No
Plotting	Yes	No	No	No	No	Limited
Support & Documentation	Yes	Yes	No	Incomplete	Yes	Yes
Knowledge of Underlying code	No	No	Yes	Yes	No	No
Package Manager	Yes	No	No	No	No	No
Availability	Bioconductor	GitHub	GitHub	GitHub	bitbucket	GitHub

between all tools. Catchitt was trained on chr18 while ChIPanalyser was trained on the top ten regions of chr18. We then validated each tool on varying number of regions on chr11 (20, 50, 100, 200, 500, 1000 and 6755 bin of 20 Kb). ChIPanalyser outperforms all other tools when the number of regions used for validation is below 500 (see Fig. 5A–C). When using more than 500 regions for validation, we observed that all tools performed similarly poorly. This trend holds true when using AUC, recall and MSE as goodness of fit metrics (although less clear with MSE). Furthermore, we trained all tools (when possible) on whole chr18 and validated on whole chr11. This ensures that all tools were trained and validated using the same data. We show that all tools perform similarly poorly when using the ChIP enrichment method to estimate goodness of fit (see Fig. 5D–F). While ChIPanalyser outperforms other tools, this could also be due to the method used to assess model performance (see Methods, Discussion and Figure S6).

3.2. DNA accessibility plays a key role in the binding of TFs

Steric hindrance can influence the binding of some TFs to DNA, meaning that a TF molecule would only bind stretches of DNA if they are accessible. Any given genomic region can be considered either accessible or inaccessible and that is sufficient to explain the binding profiles of most TFs [13]. Here, we selected accessible DNA based on DNase Hypersensitivity Sites (DHS) in three *Drosophila* cell lines (Kc167, S2 and BG3). In these circumstances, DNA was either considered accessible (score of 1) or inaccessible (score of 0). As a point of comparison, we also considered all DNA to be accessible (No Access – all regions are assigned a score of 1) and also used a min–max normalised DNase score as continuous DNA accessibility level (values between 0 and 1). We focused our analysis on three TFs: CTCF, BEAF-32 and su(Hw). We trained our model on the top 10 regions for each data set. Then, we validated

Fig. 3. Chromosome withholding setup for model validation. We analysed BEAF-32 ChIP in S2 cells (modEncode 922) and we trained ChIPanlayser on the top 10 regions on chromosome 3R. Top regions were selected from the 3293 regions described in *Materials and Methods*. We then validated our model on the top 20 regions on chromosome 2R and, for comparison, on top 10 regions on chromosome 3R that did not contain the training set. (**A**) shows example profiles obtained during training. (**B**) shows validation profiles obtained on chromosome 3R. (**C**) are profiles obtained during validation on chromosome 2R. Finally, (**D–G**) are the associated metrics for training and validation: AUC, Spearman correlation, recall and MSE respectively.

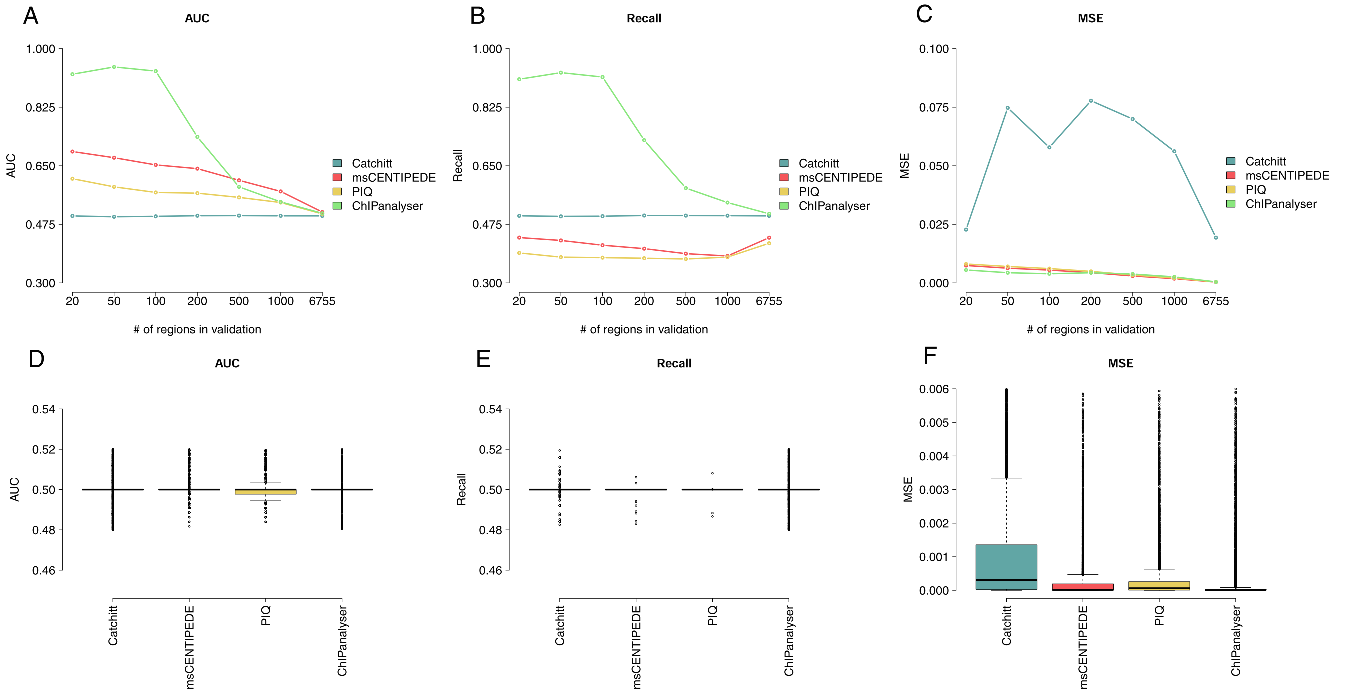


Fig. 5. Performance comparison to other TF binding predictions tools After training each model in their respective training set, we validated each tool using varying number of validations regions. ChIPanalyser outperforms other tools when number of validation regions remains below 500. This demonstrates ChIPanalyser's ability to describe TF binding behaviour with respect to peak strength. (**A**) shows AUC scores between Catchitt, msCENTIPEDE, PIQ,and ChIPanalyser over the selected validation regions in chr11 on *Homo sapiens*. (**B** and **C**) are respectively recall and MSE over validation regions for each tool. Finally, (**D**), (**E**), and (**F**) show the performance of all tools when trained on whole chr18 and validated on whole chr11. It should be noted that these results were performed using the ChIP enrichment method (see Methods) and that this approach considers both ChIP peak location as well as local peak enrichment.

our results using the optimal parameters selected during training. The optimal parameters were selected by minimising MSE between experimental ChIP profiles and predicted ChIP profiles. Validation was carried out on the top 100 regions for each dataset (excluding the ones used for training). Fig. 6 shows that, for BEAF-32, the binding predictions were improved when considering DNA accessibility. Nevertheless, su(Hw) and CTCF displayed a different behaviour, as the mean AUC decreased when DNA accessibility was considered for most ChIP-seq datasets (Fig. 6A–B). This difference is especially striking in the case of su(Hw). The performance of the model improves drastically when all DNA was considered accessible or when we used continuous values for DNA accessibility. CTCF showed a similar trend although improvement was not as striking as in the case of su(Hw). This would indicate that only a small number of CTCF peaks are located in closed chromatin regions that display intermediary levels of accessibility.

While DNA accessibility seems to play a role in the quality of our predictions, we also observed that the number of bound molecules (N) and scaling factor (λ) show a reduced influence when DNA accessibility is considered for CTCF (Fig. 6). In particular, we observed less variation in MSE for different sets of parameters, when DNA accessibility was included, i.e., larger circles indicate that number of bound molecules and λ have a more important role in TF binding, while smaller circles indicate that they have a less important role. This opposite trend is seen in the case of su(Hw) where N and λ show an increased influence when DNA accessibility is considered. BEAF-32 on the other hand is negligibly influenced by N and λ independently of whether or not we consider DNA accessibility. The rational behind this approach was that if different combinations of parameters produce a strong difference in goodness of fit, then N and λ play an important role in producing the predicted profiles. On the other hand, if we observed low variation in MSE, we could conclude that regardless of the values assigned to these parameters, the predicted profiles would remain similar.

To factor in for potential differences in the capacity of the model to predict binding in regions with strong or weak ChIP signal, we trained ChIPanalyser on the top 10 regions (see Methods) for each data set and then selected the top 20, 50, 100, 150, 200, 500, 1000 and 3283 regions for validation (excluding regions used for training). We looked at how the median AUC scores (over all data sets) changes when regions with weaker binding are included in the analysis or when DNA accessibility is considered. For each number of regions selected for validation and for each data set, we subtracted the mean AUC score when no accessibility was considered from the AUC score with DHS accessibility (Delta mean AUC). First, we observed that CTCF exhibited a slightly lower AUC score when DNA accessibility was considered (Figure S7A and D; see also Figures S8A–S11A in *Supplementary Figures*). The decrease in AUC scores observed upon considering more regions (see Figure S8A in *Supplementary Figures*) implies that CTCF binds preferentially to genome hotspots. CTCF shows strong binding at only a subset of binding sites. Interestingly, the same results for CTCF were found in human data sets as described in Fig. 5A–C. In contrast to CTCF, BEAF-32 displayed higher AUC scores when DNA accessibility was included, supporting the previous findings (Figure S7B and E; see also Figures S8B–S11B in *Supplementary Figures*). BEAF-32 AUC scores were not affected by the increase in the number of regions (Figures S7B and E and Figures S8B–S11B in *Supplementary Figures*), which means that BEAF-32 binding is not influenced by the number of regions selected. In other words, BEAF-32 would bind anywhere along the genome as long as it has an accessible site. In this context, we call BEAF-32 a global binder and CTCF a hotspot TF.

Furthermore, Supplementary Figure S7C and S7F shows that there is a strong and statistically significant ($p < 0.05$) reduction in AUC score for su(Hw) when DNA accessibility is included, which indicates that su(Hw) would bind in less accessible DNA (also Supplementary Figures S8C–S11C). While, su(Hw) did not generally

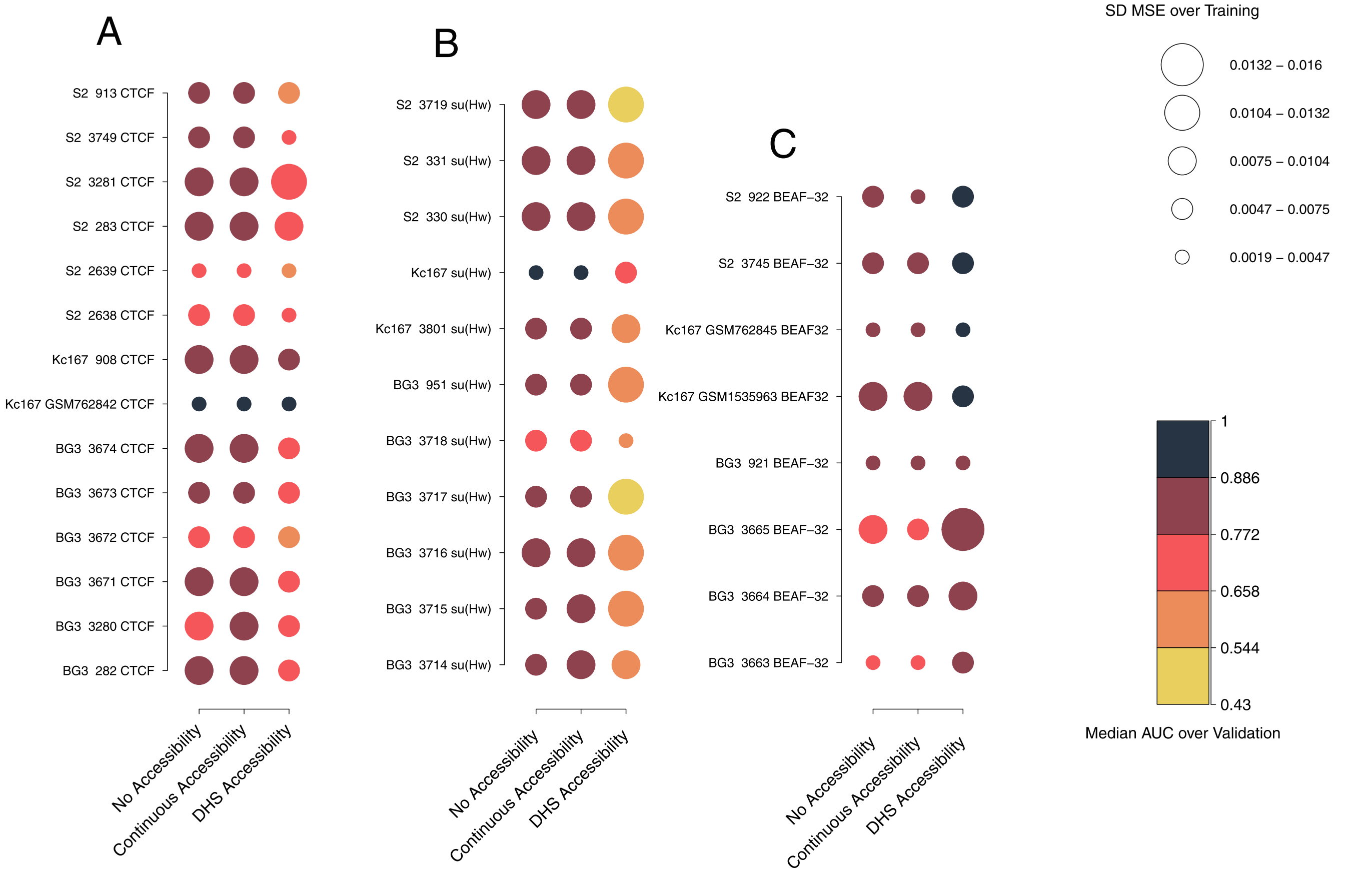


Fig. 6. DNA accessibility, number of molecules and binding energy have different roles in TF binding. We selected optimal parameters by minimising MSE over the training set (see Table S3 in Supplementary Materials) and then computed the median AUC scores over the top 100 regions in the validation set. We considered different ChIP replicates in S2, Kc167 and BG3 cells for: (**A**) CTCF, (**B**) su(Hw) and (**C**) BEAF-32. Darker colours indicate higher AUC scores, while lighter colours lower AUC scores. We also investigated the influence of number of bound molecules and scaling factor on TF binding by computing the standard deviation of MSE scores for all combination of parameters over the training set. Smaller circles indicate less variability in MSE when different parameters are used and larger circles more variability. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

perform well when DNA accessibility is considered, the performance of our model to predict su(Hw) binding is also tied to the number of regions selected and our results show that the strongest su(Hw) binding sites are found within inaccessible DNA. As the model uses experimental ChIP data for training, these results suggest that many su(Hw) peaks are located in inaccessible DNA.

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

To investigate the robustness of our estimated parameters, we computed the optimal parameters for different biological replicates. Despite strong variations between experimental data, we show that the predicted optimal set of parameters when using MSE remained similar between biological replicates (see Fig. 7). This suggests that despite biological and technical variation between replicates performed by different labs using different protocols, our model robustly infers a similar number of bound molecules and scaling factor for a given TF. Note the consistency between optimal parameters of different ChIP dataset despite some of the ChIP being performed in RNAi mutants (not more than one sample per TF and cell), which could be explained by the fact that TFs are not strongly depleted.

The importance of method selection is clearly shown when considering other metrics, where MSE produces most clear heatmaps compared to AUC, Recall or Spearman correlation (see Figures S12–S14 in *Supplementary Figures*). The optimal parameters estimated over the training set can be found in Table S3, Table S4, Table S5 and Table S6 in *Supplementary Tables* for MSE, AUC, recall and Spearman correlation coefficient.

To investigate the influence of these parameters, we assumed that a high variation of goodness of fit score for each combination of parameters would suggest a strong influence of these parameters on TF binding. If goodness of fit scores varied little between parameter combinations, we can then conclude that they do not strongly influence our predicted profiles. We then analysed the standard deviation of MSE over training between different sets of parameters and we found that some TFs are not strongly influenced by the number of bound molecules or the scaling factor (described by circle size in Fig. 6).

CTCF showed a slight decrease in sensitivity to number of bound molecules and the scaling factor when accessibility was considered (Fig. 6A), while, for BEAF-32, N and λ showed reduced influence on the binding profile (Fig. 6C). In contrast to CTCF, su(Hw) displayed an increased sensitivity to N and λ only when DNA accessibility was considered (Fig. 6B). This means that DNA accessibility would be the strongest driver towards predicting TF binding of these architectural proteins. Restricting the amount of available binding motifs would be more influential than the number of TFs and the ability of a TF to discriminate between high and low affinity sites. Interestingly, this still holds in the case of su(Hw); we show that su(Hw) binding sites are most likely found in less accessible DNA. Our results suggest that relative TF abundance only play a role on binding sites found in accessible DNA.

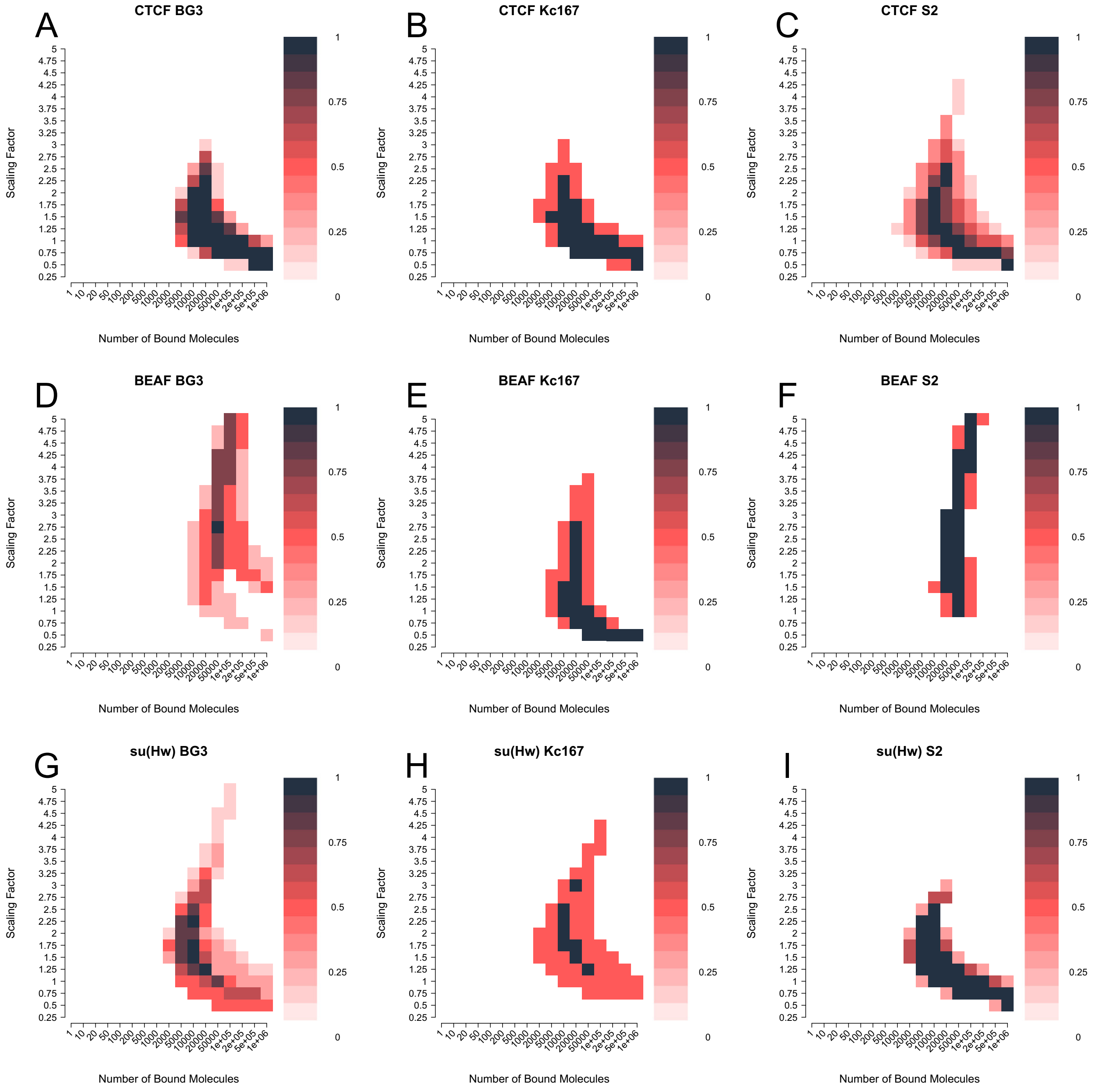


Fig. 7. Optimal parameters consistency among biological replicates for MSE using DHS accessibility. Heatmaps show an overlay of the top 10 % combinations of parameters when minimising MSE for: (**A–C**) CTCF, (**D–F**) BEAF-32 and (**G–I**) su(Hw). We plot the following cell lines: (**A**, **D** and **G**) BG3, (**B**, **E** and **H**) Kc167 and (**C**, **F** and **I**) BG3. The colour legend represents the proportion of data sets inferring each parameter combinations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4. ChIPanalyser recapitulates TF binding profiles in different cell lines by considering relative mRNA abundance.

We wanted to further investigate the predictive capabilities of our model and also demonstrate its mechanistic soundness for CTCF, BEAF-32 and su(Hw) in the three selected cell lines. For that, we estimated the optimal set of parameters in one cell line and aimed to predict TF binding in a different cell line taking into account changes in DNA accessibility using DHS data and changes in number of bound molecules using relative changes in RNA abundance. For example, we estimated the optimal set of parameters for CTCF in Kc167 cells (GSM762842) that would minimise MSE as $\lambda = 1.5$ and $N = 10^4$ over the top 10 regions (see Methods). By rescaling N based on relative RNA-seq levels of CTCF in the two cell lines, we could approximate the number of CTCF molecules bound to DNA in BG3 cells ($N \approx 1.6 \times 10^4$). This together with BG3-specific DNA accessibility data is capable of predicting the ChIP-seq profile in BG3 cells (see Fig. 8A and B – modEncode 282). RNA rescaling of the number of bound molecules seems to recover both the number of peaks and their location with high accuracy. The rescaling of number of bound molecules lead to differences in terms of MSE variation between estimated and rescaled (Fig. 8G).

The estimated MSE (MSE over the training set) in one cell line is lower than its counter parts in the other cell line. However, we

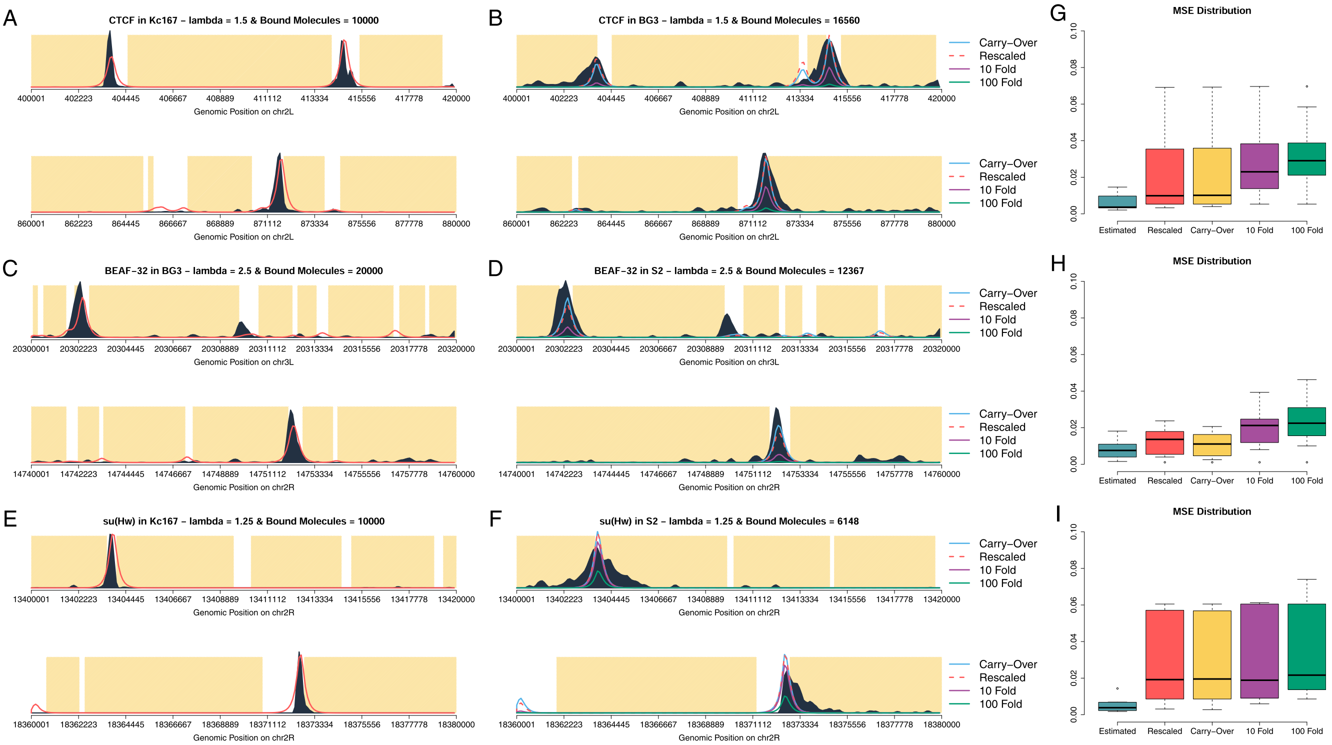


Fig. 8. TF abundance remains stable between different cell lines when considering relative mRNA abundance. A–F show predicted ChIP-seq profiles with TF abundance estimated based on RNA-seq. The yellow area represents inaccessible DNA, the dark area represents experimental ChIP signal and the red lines are our predicted profiles. We estimated the number of bound molecules in one cell line (**A**, **C** and **E**) (GSM762842, modEncode 921,GSM762839 respectively) and rescaled our estimate using relative mRNA abundance in an other cell line (**B**, **D** and **F**) (modEncode 282, modEncode 922, modEncode 331 respectively). (**B**, **D** and **F**) The dashed red line represents the rescaled value of number of bound molecules based on relative RNA-seq abundance, the light blue the original value estimated in (**A**, **C** and **E**).The purple line and the green line represent the original estimated value reduced 10 and 100 times respectively. (**G**, **H** and **I**) Boxplots with MSE for all cases in the estimated and predicted profiles at top 10 regions for both training and validation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

attribute this change to differences quality and nature of peaks/signal between cell lines. Here we assume that more narrow peaks and less spurious binding events represent higher quality data. Nevertheless, depending on the biological question, the data quality assessment can vary. This is especially striking in Fig. 8 A, B, E, and F. ChIP peaks in the training cell line (see Fig. 8A and E) display sharper peaks and much less background signal then ChIP peaks in the validation cell line (see Fig. 8B and F). As described in Methods, ChIPanalyser estimates goodness of fit using ChIP enrichment scores and therefore is sensitive to background signal and/or wider than expected peaks. The same analysis was performed for BEAF-32 (Fig. 8C, D and H - modEncode 921 & modEncode 922), where we estimated parameters in BG3 cells ($\lambda = 2.5$ and $N = 2 \times 10^4$) and rescaled the number of molecules in S2 cells ($N \approx 1.2 \times 10^4$). Once again, the model correctly predicts ChIP profiles in both location and relative enrichment. Finally, for su(Hw) (Fig. 8E, F and I – GSM762839 & modEncode 331) we estimated parameters in Kc167 cells ($\lambda = 1.25$ and $N = 10^4$) and rescaled the number of molecules in S2 cells ($N \approx 6 \times 10^3$). Again, the predictions of the model are accurate.

Our results show that ChIPanalyser can accurately recapitulate ChIP profiles between cell lines using cell specific DNA accessibility data and number of bound molecules. Nevertheless, we still do not know which of the two is the more important factor or whether both have similar contributions. To address this, we also assumed that in the predicted profile that there is *(i)* no change (same number of bound molecules is used in both cells), *(ii)* a 10-fold reduction and *(iii)* one 100 fold reduction in the number of bound molecules and repeated the analysis. Fig. 8 shows that using the same TF abundance as in the original cell line produces extremely similar ChIP like profiles. In fact, we observed a significant reduction in the predicted profile only when reducing the number of bound molecules by 100 (for su(Hw)) or 10 (for CTCF and BEAF-32) fold. These results show that cell differences in binding profiles of TFs, at their *strong binding regions*, would mainly come from differences in DNA accessibility and not relatively small changes in TF abundance. The only way that TF abundance could impact the binding profile (and, consequently, lead to changes in gene regulation) is when the expression of the TF is strongly down-regulated.

3.5. Hox genes show differential binding preferences with respect to DNA accessibility.

Hox proteins are key players during development. Recently it has been suggested that Hox proteins show different binding preferences with respect to DNA accessibility [41]. Most notably, Ubx and Abd-A would bind predominately in open chromatin, while other Hox TF (Lab, Pg, Dfd, Scr and Abd-B) would prefer closed chromatin. We selected three Hox TFs (Ubx, Dfd and Abd-B) and ran our model using different levels of DNA accessibility. DNA accessibility levels were selected based on quantile distribution of ATAC-seq scores (see Methods). This means that higher QDA scores lead to fewer regions being marked as accessible.

We trained our model on the top ten regions selected from the 3838 selected for the Hox analysis (see Methods) for each QDA accessibility. Our results show that Ubx exhibits a preference towards open chromatin. In Fig. 9A, the maximum AUC score for Ubx increases with the increase of the QDA score. Dfd and Abd-B on the other hand were not strongly influenced by QDA accessibility. This means that these TFs can bind in inaccessible DNA. According to our model, Ubx performed best with 0.99 QDA (top

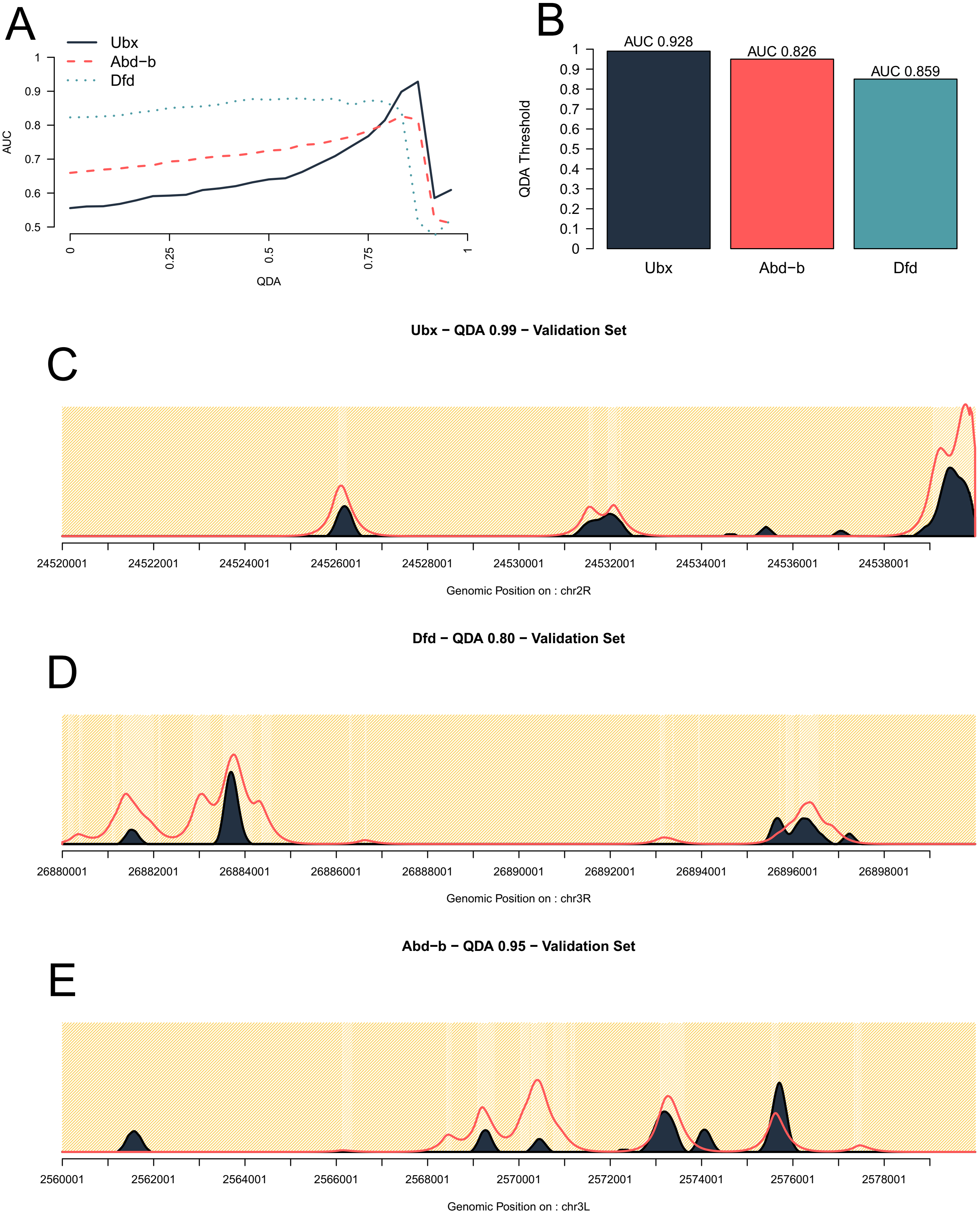


Fig. 9. Hox genes show binding preferences towards DNA accessibility. We tested our model using different DNA accessibility stringencies. (**A**) Maximum AUC score as a function of stringency of DNA accessibility (the higher the QDA value the less DNA is called accessible) for three Hox TFs: Ubx, Dfd and Abd-B. (**B**) The best performing QDA accessibility in terms of AUC. (**C**, **D** and **E**) Binding profiles and prediction of the ChIP data at individual *loci* taken from the validation set for the three TFs.

1% ATAC-seq scores – AUC 0.928), while Abd-B and Dfd with 0.95 QDA (top 5% ATAC-seq scores) and 0.8 QDA (top 20% ATAC-seq scores) respectively (see Fig. 9B). It should be noted that these scores are on the training set as the goal was to understand how QDA would effect the training of our model. We then validated our model on the top 100 regions (excluding the ones used for training) using the optimal set of parameters inferred during training and plotted the predicted profiles for Hox TF (see Fig. 9).

The model recovers the position of peaks accurately especially for Ubx (see Fig. 9C–E). While for Dfd and Abd-B most of the peaks are detected, their height is not always an accurate representation of the strength of the ChIP-seq signal. Hox TFs are known to display cooperative interactions and there are reports that both Dfd and Abd-B have a higher number of sites in the bound peaks, suggesting they bind cooperatively to open the chromatin [41]. Our model does not include cooperative interactions and this could explain the reduced performance for Dfd and Abd-B. Furthermore, TF binding event can also be mediate by protein–protein interactions and post-translational modifications [43], which are not consider in our model.

4. Discussion

Our analysis shows that ChIPanalyser and its underlying model predicts binding profiles of TFs (ChIP) with high accuracy and most importantly it can also shed light on the binding mechanism of TFs. We show how ChIPanalyser not only predicts location of peaks, but can correctly predict the enrichment of a TF at a given location.

4.1. TFs used different binding mechanisms

In this analysis, we focused our attention on three DNA binding proteins: CTCF, BEAF-32 and su(Hw). All three TFs are known architectural proteins in *Drosophila* but also play roles in transcription regulation and insulation [44,45]. Moreover, it was shown that these three TFs have distinct binding behaviours and were classified into three subclasses with respect to chromatin architecture [46,47]. In our analysis we show that they all exhibit different behaviours with respect to DNA binding.

Our findings suggest that CTCF binds to hotspots along the genome and this could be explained by the observation that the strongest peaks are in fact highly conserved binding sites. CTCF binding to highly conserved sites can be explained by our model, but something else is responsible for the reduced binding at less conserved sites (i.e. cell specific CTCF binding) as seen by the decay in performance with increased number of regions used for validation [48].

BEAF-32 is a *Drosophila* specific insulator [49] that shows preferential binding towards TAD boundaries, but also is involved in transcription itself [50]. Previous studies showed that BEAF-32 has uniform binding along the entire genome [46]. Our results confirm that BEAF-32 shows a strong preference towards accessible DNA and that the majority of accessible sites would be bound. We notice a drop in model performance when all regions are used to validated the model, but this is likely due to an increase in false positive peaks as those many regions will not contain any peaks at all.

Furthermore, we show that su(Hw) binds in both open and closed chromatin. su(Hw) plays a role in chromatin insulation and remodelling [51] and is also a primary actor in the interaction between the genome and nuclear lamina [52]. This would explain why su(Hw) can bind in both open and closed chromatin and why ChIP peaks might not overlap well with DNase hypersensitivity data. It has also been shown that su(Hw) binding sites tend to cluster together (with varying number of sites) and that these sites are constitutively bound by su(Hw) [53]. Interestingly, it seems that only isolated high affinity sites had a role in transcriptional regulation and the clustered sites were more involved in chromatin architecture.

4.2. DNA accessibility is the main driver of binding to DNA for architectural TFs and Hox TFs

Our results show that DNA accessibility and number of bound molecules control the binding profiles of TFs (Fig. 6). When we estimated the binding parameters (λ and N) in one cell line and then predicted TF binding profiles in a different cell line based on changes in DNA accessibility and number of TF molecules (using changes in mRNA), we found a good agreement between our predictions and the actual ChIP-seq dataset (see Fig. 8). Nevertheless, the changes in number of TF molecules between the two cell lines did not seem to make any difference to the predicted profiles (compare blue and dashed red line in Fig. 8B, D and F). This means that biologically relevant fluctuations in TF numbers between different cell lines would have little effect on the differences in binding profiles of TFs, which would be mainly driven by changes in DNA accessibility. Furthermore, only very strong knock-downs would decrease or deplete ChIP peaks. It should be noted that CTCF, BEAF-32 and su(Hw) are highly expressed architectural and insulator proteins and, thus, they would be expected to saturate their binding sites. Interestingly, only strong depletion (undetectable by western blot) of CTCF in mammalian cells (using Auxin Inducible Degradation) was able to lead to noticeable changes in 3D chromatin loops controlled by CTCF [54]. It should be noted that our analysis focused on the regions displaying strongest binding, which means that strong depletion is required for the binding of these TFs at their strongest sites to be affected.

While our model and current results do not demonstrate a strong role of TF abundance (as in TF concentration) in the binding of these TFs, this does not mean that concentration as well as binding site affinity does not a play a role in TF binding. Indeed, concentration fluctuation have been shown to play a role in gene expression during embryonic development in *Drosophila* [55,56]. More recently, studies have demonstrated the role of TF concentration by direct measurement of TF concentration instead of relying on mRNA abundance as a TF concentration proxy [57]. Modelling of TF binding by considering TF concentration and binding kinetics have been described by [58,7]. Changes in concentration is expected to change the binding affinity of TFs only if the concentration is within a certain range. Indeed, the K_d (dissociation constant) of TF buffers binding events if the concentration is below a certain threshold and sites will be oversaturated above a certain TF concentration (reviewed in [59]). Furthermore, it has been suggested that DNase I footprinting can display sequence specificity depending on experimental conditions. This would influence the location of open chromatin and in turn influence the performance of our model [60].

Why would changes in concentration of the TF have such a limited effect on their binding? One potential explanation is that these TFs control the expression of essential genes that should be tightly regulated to buffer fluctuations in number of molecules that affect the cell [61].

Finally, we also investigate the capacity of our model to differentiate between TFs that can bind only in open chromatin or also partially opened chromatin. Our results showed that while Ubx displays a strong sensitivity to open chromatin and binds in the top 1% accessible sites, the binding of Abd-B and Dfd is less influenced by DNA accessibility (with Abd-B and Dfd binding in top 5% and 20% respectively accessible regions); see Fig. 9. Hox TFs are known for having a similar motif, but display differences in their binding profiles [62]. It was hypothesised that binding cooperativity could explain the difference in binding profiles coupled with protein sequence changes [63]. Here, we showed that differential capacity to bind in dense chromatin could also be responsible for the difference in binding profiles of Hox TFs (see Fig. 9).

4.3. Background noise and experimental artefacts remain a challenge in TF binding predictions

We found that many ChIP datasets suffer from significant background noise that would reduce our ability to accurately assess the goodness of fit of the model. Despite our approaches to reduce background noise, it seems that ChIP data will always suffer from unspecific DNA pull-down [64]. It should be noted that more recent methods such as ChIP-exo or Cut'n'Run demonstrate sharper peaks and reduced background noise. Using these newer methods could reduce the influence of background noise on the performance of the model [65,66]. Finally, differences in data sets could be the consequence of these data sets being produced by different laboratories with potentially different protocols. However, we demonstrate that ChIPanalyser produces similar results between data sets for a given TF in a given cell line.

Another possibility is that the noise in ChIP signal could be the result of unspecific binding of TFs to DNA followed by one-dimensional random walk along the genome [67,68]. Nevertheless, the washing steps in the ChIP protocol would remove this non-specific binding from the final ChIP signal [2].

We showed that choosing a goodness of fit method is context dependent. Interestingly, similarity methods (correlation, F-score or AUC) had the tendency to correctly call peak location but greatly underestimate the enrichment on the peak (see Fig. 2). This behaviour results from the fact that these methods are highly penalised by false positive hits. The scaling factor can be described as how well a TF discriminates between a strong binding site over a weaker one. High values for the scaling factor translate to poorer ability for the TFs to discriminate between high and low affinity sites, which leads both to a higher number of false positive peaks and the model picking up smaller peaks. The number of bound molecules on the other hand, tend to affect the height of the peak (relative local enrichment). Similarity methods would avoid high values for N and λ as this would penalise their goodness of fit score more severely as opposed to dissimilarity methods (see Fig. 2).

Choosing the right method will depend on the question at hand and similarity methods could be used to determine peak location, while dissimilarity metrics would be more appropriate to investigate the TF local enrichment.

4.4. ChIP enrichment scores provide a highly stringent method to assess model performance

ChIPanalyser evaluates goodness of fit using ChIP enrichment scores (see Methods). This ensures that the model considers peak enrichment during the optimisation step (both location and height of the peak). Competing tools generally assess model performance by overlapping predicted TF binding sites with ChIP peaks and do not explicitly account for peak enrichment (they assume that there is very little difference between a strong and a weak ChIP peak). While we recognise that our scoring method is best suited for TF binding events described both by peak location and peak enrichment, we selected this approach as ChIPanalyser describes a mechanistic interpretation of TF binding.

When comparing ChIPanalyser to other frameworks, we observed that all tools performed poorly when trained and validated on a full chromosomes. When ChIPanalyser was trained on the top 10 regions of chr18 and validated on varying number of regions in chr11, it outperforms other tools and frameworks if the number of validation regions did not exceed 500. The rational to train ChIPanalyser on top 10 regions is to ensure a balance between True Positive and True Negative signals, which results in a more effective parameter inference. Many regions along the genome might not contain any ChIP signal and this lack of signal will affect the profiles produced by ChIPanalyser and result in a drop in performance for our model (increase in False positive).

Finally, the goal of ChIPanalyser is not only to predict TF binding events but also shed light on the mechanisms driving TF binding. In the case of CTCF in human astrocytes (as used in the comparison with other tools), ChIPanalyser showed a decay in performance after 500 regions used for validation (see Fig. 5A–C). PIQ, msCentipede and Catchitt did not display such a clear behaviour. Interestingly, we observed a similar effect for CTCF in *Drosophila*. Our results suggest that CTCF binds to highly conserved sites [48] and this holds true in different organisms. Most importantly, ChIPanalyser was able to recapitulate this behaviour.

5. Conclusion

ChIPanalyser is a user-friendly R package available on Bioconductor for predicting the binding of Transcription Factors to DNA. The package performs similarly if not better than competing tools and frameworks. More importantly, the model also provides an insight into the binding mechanisms of various DNA binding proteins. We show the nuanced role of DNA accessibility in the binding of three architectural proteins CTCF, BEAF-32 and su(Hw) in *Drosophila*. Furthermore, we demonstrate that architectural proteins are robust to relative changes in protein abundance. Finally, we recover the binding preferences of Hox TFs with respect to chromatin compaction. ChIPanlyser provides both predictive and biological modelling capabilities.

6. Funding

This work was supported by University of Essex and by the Wellcome Trust grant [202012/Z/16/Z].

CRediT authorship contribution statement

Patrick C.N. Martin: Conceptualization, Methodology, Software, Investigation, Writing - original draft. **Nicolae Radu Zabet:** Conceptualization, Methodology, Writing - review & editing, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Dr. Rob White, Prof Sarah Bray and Zabet lab (Keerthi Chathoth, Jareth Wolfe, Romana Pop and Olivia Grant) for useful discussion and comments on the project and the manuscript. We would also like to thank Dr. Gorrie-Stone for his comments and suggestions during the development of ChIPanalyser. The analysis was performed on the HPC at University of Essex and we would like to thank Stuart Newman for his support

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.csbj.2020.11.006.

References

[1] Park PJ. Chip-seq: advantages and challenges of a maturing technology. Nat Rev Genet 2009;10(10):669–80. https://doi.org/10.1038/nrg2641.

[2] Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, Chen Y, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 2012;22 (9):1813–31. https://doi.org/10.1101/gr.136184.111.

[3] Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins statistical-mechanical theory and application to operators and promoters. J Mol Biol 1987;193(4):723–50. https://doi.org/10.1016/0022-2836(87)90354-8.

[4] Ptashne M, Gann A. Transcriptional activation by recruitment. Nature 1997;386(6625):569–77. https://doi.org/10.1038/386569a0.

[5] Stormo GD, Zhao Y. Determining the specificity of protein-DNA interactions. Nat Rev 2010;11:751–60. https://doi.org/10.1038/nrg2845.

[6] Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet 2012;13(9):613–26. https://doi.org/10.1038/nrg3207.

[7] Roider HG, Kanhere A, Manke T, Vingron M. Predicting transcription factor affinities to DNA from a biophysical model. Bioinformatics 2007;23(2):134–41. https://doi.org/10.1093/bioinformatics/btl565.

[8] Farnham PJ. Insights from genomic profiling of transcription factors. Nat Rev Genet 2009;10(9):605–16. https://doi.org/10.1038/nrg2636.

[9] Skalska L, Stojnic R, Li J, Fischer B, Cerda-Moya G, Sakai H, Tajbakhsh S, Russell S, Adryan B, Bray SJ. Chromatin signatures at notch-regulated enhancers reveal large-scale changes in h3k56ac upon activation. EMBO J 2015;34 (14):1889–904.

[10] Chu D, Zabet NR, Mitavskiy B. Models of transcription factor binding: Sensitivity of activation functions to model assumptions. J Theor Biol 2009;257(3). https://doi.org/10.1016/j.jtbi.2008.11.026.

[11] Kaplan T, Li X-Y, Sabo PJ, Thomas S, Stamatoyannopoulos JA, Biggin MD, Eisen MB. Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. PLoS Genet 2011;7(2):1001290. https://doi.org/10.1371/journal.pgen.1001290.

[12] Simicevic J, Schmid AW, Gilardoni PA, Zoller B, Raghav SK, Krier I, Gubelmann C, Lisacek F, Naef F, Moniatte M, Deplancke B. Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics. Nat Methods 2013;10:570–6. https://doi.org/10.1038/nmeth.2441.

[13] Zabet NR, Adryan B. Estimating binding properties of transcription factors from genome-wide binding profiles. Nucl Acids Res 2015;43(1):84–94. https://doi.org/10.1093/nar/gku1269.

[14] Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet 2019;1. https://doi.org/10.1038/s41576-018-0089-8.

[15] Lamparter D, Marbach D, Rueedi R, Bergmann S, Kutalik Z. Genome-wide association between transcription factor expression and chromatin accessibility reveals regulators of chromatin accessibility. PLOS Comput Biol 2017;13(1):1005311. https://doi.org/10.1371/journal.pcbi.1005311.

[16] Soufi A, Garcia M, Jaroszewicz A, Osman N, Pellegrini M, Zaret K. Pioneer transcription factors target partial DNA motifs on nucleosomes to initiate reprogramming. Cell 2015;161(3):555–68. https://doi.org/10.1016/j.cell.2015.03.017.

[17] Zhu F, Farnung L, Kaasinen E, Sahu B, Yin Y, Wei B, Dodonova SO, Nitta KR, Morgunova E, Taipale M, Cramer P, Taipale J. The interaction landscape between transcription factors and the nucleosome. Nature 2018;562 (7725):76–81. https://doi.org/10.1038/s41586-018-0549-5.

[18] Michael AK, Grand RS, Isbel L, Cavadini S, Kozicka Z, Kempf G, Bunker RD, Schenk AD, Graff-Meyer A, Pathare GR, Weiss J, Matsumoto S, Burger L, Schübeler D, Thomä NH. Mechanisms of OCT4-SOX2 motif readout on nucleosomes. Science (New York, N.Y.) 2020;368(6498):1460–5. https://doi.org/10.1126/science.abb0074.

[19] Lerner J, Gomez-Garcia PA, McCarthy RL, Liu Z, Lakadamyali M, Zaret KS. Two-parameter mobility assessments discriminate diverse regulatory factor behaviors in chromatin. J Clean Prod 2020. https://doi.org/10.1016/j.molcel.2020.05.036.

[20] Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, van Hoff JP, Karun V, Jaakkola T, Gifford DK. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat Biotechnol 2014;32(2):171–8. https://doi.org/10.1038/nbt.2798.

[21] Raj A, Shim H, Gilad Y, Pritchard JK, Stephens M. msCentipede: modeling heterogeneity across genomic sites and replicates improves accuracy in the inference of transcription factor binding. PLoS One 2015;10(9):0138030. https://doi.org/10.1371/journal.pone.0138030.

[22] Quang D, Xie X. FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods 2019;166:40–7. https://doi.org/10.1016/J.YMETH.2019.03.020.

[23] Li H, Quang D, Guan Y. Anchor: trans-cell type prediction of transcription factor binding sites. Genome Res 2019;29(2):281–92. https://doi.org/10.1101/gr.237156.118.

[24] Keilwagen J, Posch S, Grau J. Accurate prediction of cell type-specific transcription factor binding. Genome Biol. 2019;20(1):9. https://doi.org/10.1186/s13059-018-1614-y.

[25] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206–15. https://doi.org/10.1038/s42256-019-0048-x.

[26] Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Nat Acad Sci 2019;116:22071–80. https://doi.org/10.1073/PNAS.1900654116.

[27] R Development Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing; 2014. .

[28] Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini A, Sawitzki G, Smith C, Smyth G, Tierney L, Yang J, Zhang J. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5(10):80. https://doi.org/10.1186/gb-2004-5-10-r80.

[29] Mueller F, Stasevich TJ, Mazza D, McNally JG. Quantifying transcription factor kinetics: At work or at play? Crit Rev Biochem Mol Biol 2013;48(5):492–514. https://doi.org/10.3109/10409238.2013.833891.

[30] Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al., The genome sequence of drosophila melanogaster. Science 2000;287 (546);2185–95. doi: 10.1126/science.287.5461.2185. https://science.sciencemag.org/content/287/5461/2185.full.pdf.

[31] dos-Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, etal., The FlyBase Consortium: FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. Nucl Acids Res 2014:43(D1);690–697. doi: 10.1093/nar/gku1099. http://oup.prod.sis.lan/nar/article-pdf/43/D1/D690/7317662/gku1099.pdf.

[32] Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, Murphy TD, Pruitt KD, Thibaud-Nissen F, Albracht D, Fulton RS, Kremitzki M, Magrini V, Markovic C, McGrath S, Steinberg KM, Auger K, Chow W, Collins J, Harden G, Hubbard T, Pelan S, Simpson JT, Threadgold G, Torrance J, Wood JM, Clarke L, Koren S, Boitano M, Peluso P, Li H, Chin C-S, Phillippy AM, Durbin R, Wilson RK, Flicek P, Eichler EE, Church DM. Evaluation of grch38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res 2017;27(5):849–64. https://doi.org/10.1101/gr.213611.116. http://genome.cshlp.org/content/27/5/849.full.pdf+html.

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

[34] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, etal., The human genome browser at ucsc. Genome Res 2002:12;996–1006. doi: 10.1101/gr.229102.

[35] Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen C-Y, Chou A, Ienasescu H, Lim J, Shyr C, Tan G, Zhou M, Lenhard B, Sandelin A, Wasserman WW. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucl Acids Res 2014;42(D1):142–7. https://doi.org/10.1093/nar/gkt997.

[36] Shannon P, Richards M. MotifDb: An annotated collection of protein-dna binding sequence motifs. R package version 1.24.1. .

[37] Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics 2011;27(7):1017. https://doi.org/10.1093/BIOINFORMATICS/BTR064.

[38] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. The sequence alignment/map format and SAMtools. Bioinformatics 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352.

[39] Zhang Y, Liu T, Meyer C, Eeckhoute J, Johnson D, Bernstein B, Nusbaum C, Myers R, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). Genome Biol 2008;9(9):137. https://doi.org/10.1186/gb-2008-9-9-r137.

[40] Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, Linder-Basso D, Plachetka A, Shanower G, Tolstorukov MY, Luquette LJ, Xi R, Jung YL, Park RW, Bishop EP, Canfield TP, Sandstrom R, Thurman RE, MacAlpine DM, Stamatoyannopoulos JA, Kellis M, Elgin SCR, Kuroda MI, Pirrotta V. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature 2010. https://doi.org/10.1038/nature09725.

[41] Porcelli D, Fischer B, Russell S, White R. Chromatin accessibility plays a key role in selective targeting of Hox proteins. Genome Biol 2019;20(1):115. https://doi.org/10.1186/s13059-019-1721-4.

[42] Lee H, McManus CJ, Cho D-Y, Eaton M, Renda F, Somma MP, Cherbas L, May G, Powell S, Zhang D, Zhan L, Resch A, Andrews J, Celniker SE, Cherbas P, Przytycka TM, Gatti M, Oliver B, Graveley B, MacAlpine D. Dna copy number evolution in drosophila cell lines. Genome Biol 2014;15(8):70. https://doi.org/10.1186/gb-2014-15-8-r70.

[43] Filtz TM, Vogel WK, Leid M. Regulation of transcription factor activity by interconnected post-translational modifications. Trends Pharmacol Sci 2014;35:76–85. https://doi.org/10.1016/j.tips.2013.11.005.

[44] Van Bortle K, Nichols MH, Li L, Ong C-T, Takenaka N, Qin ZS, Corces VG. Insulator function and topological domain border strength scale with

architectural protein occupancy. Genome Biol 2014;15(5):82. https://doi.org/10.1186/gb-2014-15-5-r82.
[45] Chathoth KT, Zabet NR. Chromatin architecture reorganisation during neuronal cell differentiation in drosophila genome. Genome Res 2019;29:613–25. https://doi.org/10.1101/gr.246710.118.
[46] Bushey AM, Ramos E, Corces VG. Three subclasses of a Drosophila insulator show distinct and cell type-specific genomic distributions. Genes Dev 2009;23 (11):1338–50. https://doi.org/10.1101/gad.1798209.
[47] Vogelmann J, Le Gall A, Dejardin S, Allemand F, Gamot A, Labesse G, Cuvier O, Nègre N, Cohen-Gonsaud M, Margeat E, Nöllmann M. Chromatin insulator factors involved in long-range DNA interactions and their role in the folding of the drosophila genome. PLoS Genet 2014;10(8). https://doi.org/10.1371/journal.pgen.1004544.
[48] Vietri-Rudan M, Barrington C, Henderson S, Ernst C, Odom D, Tanay A, Hadjur S. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep 2015;10(8):1297–309. https://doi.org/10.1016/j.celrep.2015.02.004.
[49] Schoborg TA, Labrador M. The phylogenetic distribution of non-CTCF insulator proteins is limited to insects and reveals that BEAF-32 is drosophila lineage specific. J Mol Evol 2010;70(1):74–84. https://doi.org/10.1007/s00239-009-9310-x.
[50] Jiang N, Emberly E, Cuvier O, Hart CM. Genome-wide mapping of boundary element-associated factor (BEAF) binding sites in Drosophila melanogaster links BEAF to transcription. Mol Cell Biol 2009;29(13):3556–68. https://doi.org/10.1128/MCB.01748-08.
[51] Kurshakova M, Maksimenko O, Golovnin A, Pulina M, Georgieva S, Georgiev P, Krasnov A. Evolutionarily conserved E(y)2/Sus1 protein is essential for the barrier activity of Su(Hw)-dependent insulators in drosophila. Mol Cell 2007;27(2):332–8. https://doi.org/10.1016/J.MOLCEL.2007.05.035.
[52] van Bemmel JG, Pagie L, Braunschweig U, Brugman W, Meuleman W, Kerkhoven RM, van Steensel B. The insulator protein SU(HW) fine-tunes nuclear lamina interactions of the Drosophila genome. PLoS One 2010;5 (11):15013. https://doi.org/10.1371/journal.pone.0015013.
[53] Adryan B, Woerfel G, Birch-Machin I, Gao S, Quick M, Meadows L, Russell S, White R. Genomic mapping of Suppressor of Hairy-wing binding sites in Drosophila. Genome Biol 2007;8(8):167. https://doi.org/10.1186/gb-2007-8-8-r167.
[54] Nora EP, Goloborodko A, Valton A-L, Gibcus JH, Uebersohn A, Abdennur N, Dekker J, Mirny LA, Bruneau BG. Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. Cell 2017;169(5):930–94422. https://doi.org/10.1016/J.CELL.2017.05.004.
[55] Moens CB, Selleri L. Hox cofactors in vertebrate development. Dev Biol 2006;291(2):193–206. https://doi.org/10.1016/j.ydbio.2005.10.032.
[56] Petkova MD, Tkačik G, Bialek W, Wieschaus EF, Gregor T. Optimal decoding of cellular identities in a genetic network. Cell 2019;176(4):844–85515. https://doi.org/10.1016/j.cell.2019.01.007.
[57] Papadopoulos DK, Skouloudaki K, Engström Y, Terenius L, Rigler R, Zechner C, Vukojević V, Tomancak P. Control of hox transcription factor concentration and cell-to-cell variability by an auto-regulatory switch. Dev 2019;146(12). https://doi.org/10.1242/dev.168179.
[58] Wang Y, Guo L, Golding I, Cox EC, Ong NP. Quantitative transcription factor binding kinetics at the single-molecule level. Biophys J 2009;96(2):609–20. https://doi.org/10.1016/j.bpj.2008.09.040.
[59] Bintu L, Buchler NE, Garcia HG, Gerland U, Hwa T, Kondev J, Kuhlman T, Phillips R. Transcriptional regulation by the numbers: Applications. Curr Opin Genet Devel 2005;15:125–35. https://doi.org/10.1016/j.gde.2005.02.006.
[60] Koohy H, Down TA, Hubbard TJ. Chromatin accessibility data sets show bias due to sequence specificity of the dnase i enzyme. PLoS ONE 2013;8:69853. https://doi.org/10.1371/journal.pone.0069853.
[61] Schoech AP, Zabet NR. Facilitated diffusion buffers noise in gene expression. Phys Rev E 2014;90(3):32701. https://doi.org/10.1103/PhysRevE.90.032701.
[62] Gehring WJ, Qian YQ, Billeter M, Furukubo-Tokunaga K, Schier AF, Resendez-Perez D, Affolter M, Otting G, Wüthrich K. Homeodomain-DNA recognition. Cell 1994;78(2):211–23. https://doi.org/10.1016/0092-8674(94)90292-5.
[63] Hayashi S, Scott MP. What determines the specificity of action of Drosophila homeodomain proteins? Cell 1990;63(5):883–94. https://doi.org/10.1016/0092-8674(90)90492-W.
[64] Teytelman L, Thurtle DM, Rine J, van Oudenaarden A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. PNAS 2013;110(46):18602–7. https://doi.org/10.1073/pnas.1316064110.
[65] Skene PJ, Henikoff S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. eLife 2017:6. doi: 10.7554/eLife.21856. .
[66] Serandour AA, Brown GD, Cohen JD, Carroll JS. Development of an Illumina-based ChIP-exonuclease method provides insight into FoxA1-DNA binding properties. Genome Biol 2013;14(12):147. https://doi.org/10.1186/gb-2013-14-12-r147.
[67] Zabet NR, Adryan B. A comprehensive computational model of facilitated diffusion in prokaryotes. Bioinformatics 2012;28(11):1517–24. https://doi.org/10.1093/bioinformatics/bts178.
[68] Hammar P, Leroy P, Mahmutovic A, Marklund EG, Berg OG, Elf J. The lac repressor displays facilitated diffusion in living cells. Science 2012;336 (6088):1595–8. https://doi.org/10.1126/science.1221648.