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

Patrick C.N. Martin and Nicolae Radu Zabet *

School of biological Sciences, University of Essex , Wivenhoe Park Colchester CO4 3SQ

Abstract

At the heart of gene regulation are Transcription Factors (TF), proteins which bind to DNA in a sequence specific manner and drive the activation or repression of genes. Here, we present a statistical thermodynamics framework (ChIPanalyser) which models and predicts binding 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 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 regions displaying intermediate levels of accessibility. 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 prefers open chromatin, while Abd-B and Dfd are capable to bind in partially closed chromatin. Overall, our results show that TFs display different binding mechanisms and that our model is able to recapitulate this diverse repertoire of mechanisms.

Introduction

Decades of research have shown that gene expression is at the heart of many, if not all, cellular processes. 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). TFs are a class of proteins that bind to DNA in a sequence specific manner [1, 2]. The most commonly used experimental method to determine specific regions of DNA where TFs bind is chromatin immunoprecipitation followed by sequencing (ChIP-seq) [3, 4]. 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.

The most fundamental aspect to consider with respect to TF binding specificity is the DNA sequence itself. Most TFs exhibit a preferred binding motif [5, 6, 7, 8]. The most common way to describe this motif is in the form of a Position Weight Matrix (PWM); a measure of binding energy between TFs and DNA weighted by the genomic base pair frequency [5, 9]. Although PWMs have shown their ability to describe TF preferred binding sequences, they have limitations when it comes to predicting TF binding *loci* or sites. In particular, TFs can have tens of thousands of binding sites within each genome, yet they would only bind to a few hundred or thousands of them [10, 11, 12, 13].

Unsurprisingly, there is a large body of early research describing binding of TFs to DNA as a similar problem as ligand to receptor binding [14, 15, 16]. At its core, this approach relies on local concentration of TF (ligand) and receptor availability (binding sites) but also association and dissociation constants between TFs and DNA [17]. Previous studies have shown that some TF binding events are TF concentration dependent [18, 19, 20, 21, 22], where varying the concentration of the TF will drive the expression of different set of genes. A good example is the case of Homeobox genes, more commonly known as Hox genes. In Drosophila, embryo patterning is believed to be a consequence of varying Hox TF concentration along the anterior-posterior axis [23] and recent efforts have shown that Hox TF concentration is sufficient to predict cellular identity along anterior-posterior axis with a high accuracy [24]. However, there are many more spurious sites where TFs could bind than functional binding sites. This still begs the question: how does a TF recognises the right binding motif out of many decoys?

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 in dense chromatin [25, 26, 27, 28, 29]. Nevertheless, there is a certain class of TFs that would ignore accessibility restrictions and these TFs are known as pioneer TFs. More specifically, pioneer TFs can bind sites in closed dense chro-

*To whom correspondence should be addressed. Tel: +44(0)1206872630; Fax: +44(0)1206872592; Email: nzabet@essex.ac.uk

matin and subsequently open the chromatin [30, 31, 32].

We previously showed that statistical thermodynamics can be used to model with high accuracy TF binding to DNA [22]. 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 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 ChIPanaylser a versatile, fast, efficient and user-friendly R/Bioconductor package [33, 34]. We used this model to describe the behaviour of several Drosophila TFs: CTCF, BEAF-32, su(Hw), Ubx, Abd-B and Dfd. Our results provide a mechanistic interpretation of the 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 cells (BG3, Kc167 and S2) and that relative medium changes in the concentrations of these TFs lead to only negligible changes in their binding profiles. Finally, we also show that TF binding specificity can be achieved by their capacity to bind to regions with different levels of DNA accessibility. In particular, we showed that Ubx, Abd-B and Dfd binding to DNA could be explained by their different capacity to bind dense chromatin, with Ubx binding only in highly accessible chromatin and Dfd and Abd-B binding in denser chromatin.

MATERIALS AND METHODS

Model Description

ChIPanalyser is an R package available on Bionconductor [33, 34]. The package is an implementation of the statistical thermodynamics model proposed in [22]. Briefly, the model requires a PWM (Position Weight Matrix) or PFM (Position Frequency Matrices) of the TF of interest, DNA accessibility data to model binding site accessibility and two additional parameters: λ (a PWM scaling factor) and N (the number of bound molecules) [22]. The probability of a position j on the DNA being occupied is given by [22]:

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

λ and N are difficult to estimate from 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.

Datasets

To carry out the analysis described in this manuscript, we selected data originating from various sources.

DNA Sequence

Reference Sequences of *Drosophila melanogaster* were extracted from the Bsgenome R packages [35]. We used both dm3 [36] and dm6 [37] versions of the Drosophila genome. In particular, we only used dm6 for Hox TF analysis and dm3 for the rest of the analysis. It should be noted that the choice between dm3 and dm6 was from a consistency stand point. All modEncode data sets (see section ChIP-seq) use the dm3 build of the genome while ATAC-seq and associated Hox ChIP-seq data were aligned to the dm6 version of genome.

PWM and PFM

Binding Motif matrices were downloaded from online repositories such as JASPAR [38] or extracted from the MotifDb R package [39], which collects and compiles PFMs and Position Probability Matrices (PPM) from various online repositories (see Figure S1 in *Supplementary Material*).

ChIP-seq

Both ChIP-seq enrichment signal and ChIP-seq peaks were downloaded (pre-processed) from modEncode [40] in three Drosophila cell line: Kc167, S2 and BG3. When it was required, supplementary data sets were downloaded from GEO. GEO datasets were aligned to the genome (dm3) using bowtie-2 (- -non-deterministic). Note that for Hox TF analysis we used dm6. Peaks and pile-up signal were called using macs2 with a 0.01 FDR (-q 0.01). As decribed above, choosing between dm3 and dm6 was made based a consitency stand point. modEncode data sets were aligned to the dm3 version of the *Drosophila* genome. Datasets used for this analysis are described in Table S1 in *Supplementary Material*.

DNA accessibility

DNase I hypersensitivity data was generated by modEncode for the three cell lines used in this analysis [40, 41]. We extended the DNase Hypersensitivity Sites (DHS) by 500bp (see Figure S2 in *Supplementary Material*). DNA was either considered accessible (DHS) or non accessible. ATAC-seq data for Kc167 cells was used from [42]. 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 a Quantized Density Accessibility (QDA).

RNA-seq

In order to rescale TF abundance between cell lines we used RNA-seq data from [43], who preprocessed original modEncode datasets [44]. RNA-seq relative abundance was used to rescale the estimated number of bound molecules for one cell line to another.

Package Description

The workflow of ChIPanalyser is described in Figure 1. Briefly, the optimal set of parameters (for λ and N) are inferred from ChIP-seq data by maximising (or minimising) the goodness of fit metric. Using these values, ChIPanalyser will predict ChIP-seq like profiles for different genomic regions and compare the prediction with the actual ChIP-seq data.

ChIPanalyser uses a set of genomic regions to carry out the optimisation of the model. Generally, these regions would be user provided. For our analysis, we investigated regions that contain both ChIP-seq peaks and accessible DNA. To do so, we binned the genome into bins of 20 Kb and used the *processingChIPseq* function provided by ChIPanalyser. This functions returns a normalised ChIP enrichment score for the top n regions with respect to ChIP-seq data (both number of peaks and peak enrichment) and DNA accessibility data. It should be noted that the number of regions selected is user specified.

During this step of the analysis, we also included a noise filtering method. The current model does not consider ChIP depletion therefore all negative score are replaced by 0. With that in mind, ChIPanlyser provides four methods of filtering noise: *Zero*, *Mean*, *Median* and *Sigmoid*. *Zero* removes only depletion score (equivalent to "no noise filtering"). *Mean* and *Median* replace all scores below the mean and the median after filtering out depletion scores. Finally, *Sigmoid* applies logistic weighting to every score. The logistic mid point is set at the 95th quantile of ChIP scores. Lower bound is zero and upper bound is 2. Consequently each score will be multiplied by a weight: if the score is above the 95th quantile the score will be weighted by values between 1 and 2. If the score is below the 95th quantile, score will be weighted by a factor ranging from 0 to 1. All analysis in this manuscript was carried out after using the *Sigmoid* noise filtering method.

Once the *loci* of interest have been selected, we computed the optimal set of parameters by using ChIPanalyser *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 12 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 1). We also developed a novel method that describes the ratio of shared geometric area between curves and difference in area between curves.

The optimal parameters can be visualised in the form of a heat map 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-seq like profiles. Profiles can be visualised using the *plotOccupancyProfiles* function provided by the package.

RESULTS

Goodness of fit metrics are context dependent

Previously, we showed how statistical thermodynamics can be used to mechanistically explain the binding of TFs in Drosophila [22]. The optimal set of parameters (see *Materials and Methods*) was inferred by maximising correlation and minimising Mean Squared Error (MSE) between the predicted profile and experimental ChIP-seq data. Nevertheless, we observed that, in some cases, the predicted profiles and ChIP-seq profiles display low correlation coefficient despite the profiles looking similar. Conversely, high correlation coefficients were also associated with poor overlap between predicted and actual ChIP profiles (e.g. see Figure S3A and S3B in *Supplementary Material*). In addition, selecting the optimal parameters was hindered by little variation in correlation between parameter combinations. As a consequence the selection of these parameters was exclusively driven by MSE (see Figure S3C in *Supplementary Material*). We hypothesised that these discrepancies could be due to either background noise in ChIP-seq data or biases in the goodness of fit metrics that we used (Pearson coefficient of correlation and MSE).

To reduce the potential influence of background noise, we tested four noise removal methods: *Zero* (removes only depletion score), *Mean* (replace all score below the mean), *Median* (replace all score below the median) and *Sigmoid* (applies logistic weighting to every score); see *Materials and Methods*. To test the performance of these methods, we used three CTCF datasets: (*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-seq datasets in S2 cells (by adding enrichment signals together at a base pair level); see Table S1 in *Supplementary Material*. We ran the model on the top ten regions and searched for the optimal set of parameters (λ and N) that optimised each goodness of fit metric (see *Materials and Methods*). We normalised the signal in order to ensure equal contribution of each data set (see Table S1 in *Supplementary Material*). All four noise filtering methods have little to no effect on ChIP data. 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 Material*).

In addition to Pearson correlation and MSE, we tested several goodness of fit metrics to verify the influence of the metrics on our model. In particular, we compared correlation (Pearson, Spearman and Kendall), MSE, Kolmogorov-Smirnov Distance, precision, recall, accuracy, F-score, Matthew's correlation coefficient (MCC) and AUC ROC (see Table 1). In addition, we also developed a novel method that describes the ratio of geometric shared area between curves and difference in area between curves (see *Materials and Methods*). We used the same three CTCF

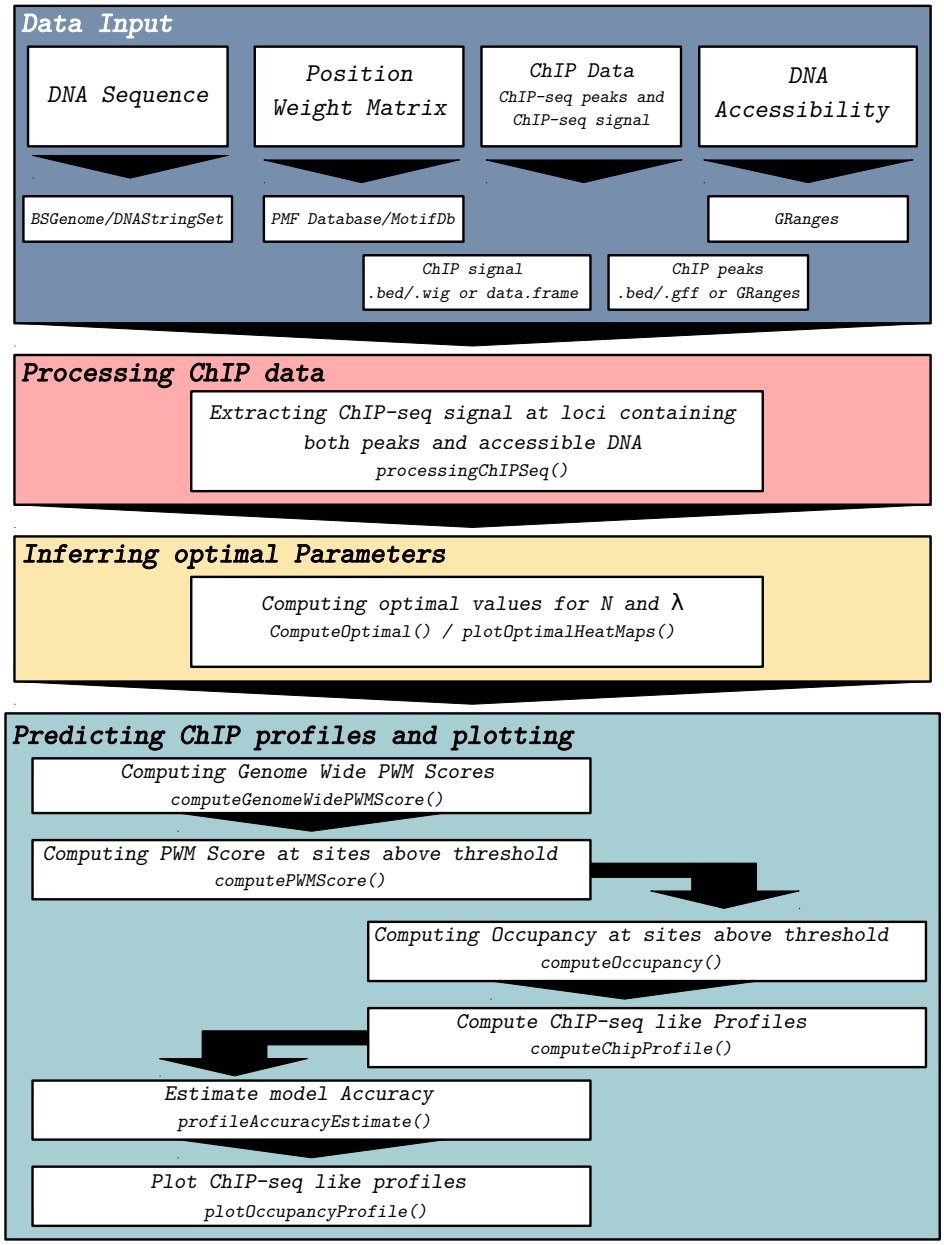

Figure 1: *ChIPanalyser workflow.* ChIPanalyser follows the following work flow. **Data Input:** Data may come in various formats (e.g. bed, wig, gff etc.). **Processing ChIP-seq data:** If ChIP-seq data is used to infer the optimal set of parameters (and/or validate model goodness of fit), ChIP-seq 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:** Armed with values for number of bound molecules and the PWM scaling factor, ChIPanalyser will produce ChIP-seq like profiles. Both optimal parameter heat maps and ChIP profiles can be plotted using the packages plotting functions.

Table 1: **Goodness of Fit metrics:** ChIPanalyser offers 12 goodness of fit metrics grouped into two classes *Dissimilarity* and *Similarity*. Symbolically each metric is either a measure of how different two datasets are (Dissimilarity) or a measure of how similar two datasets are (Similarity).

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

TP - true positives, TN - true negatives, FP - false positives and FN - false negatives. MCC represents Matthews Correlation Coefficient

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 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 would vary significantly, but each of the two classes (similarity and dissimilarity metrics) displayed different values for the optimal parameters (see Figure 2A-C). However in certain instances the optimal set of parameters selected by dissimilarity method would overlap slightly with parameters selected by similarity methods (see Figure 2D-F).

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

Overall, the best performing metrics were AUC ROC, MSE and geometric ratio. AUC ROC occasionally missed peaks completely but seemed to recover peak height fairly accurately, while geometric ratio and MSE rarely missed peaks but also tended to predict 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.

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 of TFs [22]. Here, we selected accessible DNA based on DNase Hypersensitivity Sites (DHS) in three Drosophila cell lines (Kc167, S2 and BG3) and, as a point of comparison, we also considered all DNA to be accessible (No Access). We focussed our analysis on three TFs: CTCF, BEAF-32 and su(Hw) (a break down of each data set can be found in Table S1 in *Supplementary Material*). The influence of accessibility was measured by computing the median AUC score over all regions for the best performing set of parameters. In this instance, we used AUC scores as we were interested in recovering peak location more than peak height. Figure 3 shows that, for CTCF and BEAF-32, the binding predictions were improved when considering DNA accessibility. Nevertheless, su(Hw) displayed a different behaviour, as the mean AUC decreased when DNA accessibility was considered for most ChIP-seq datasets (Figure 3B). Note however, that there are two replicates were this is not the case: Kc167 su(Hw) and BG3 3718 su(Hw).

While DNA accessibility seems to improve the predictions, we also observed that the number of bound molecules (N) and scaling factor (λ) show a reduced influence when DNA accessibility is considered (Figure 3). In particular, we observed less variation in AUC for different set of parameters, when DNA accessibility is 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. The trend is true for CTCF and BEAF-32, but less strong for su(Hw).

As described, the genome was split into tiles of 20 Kb

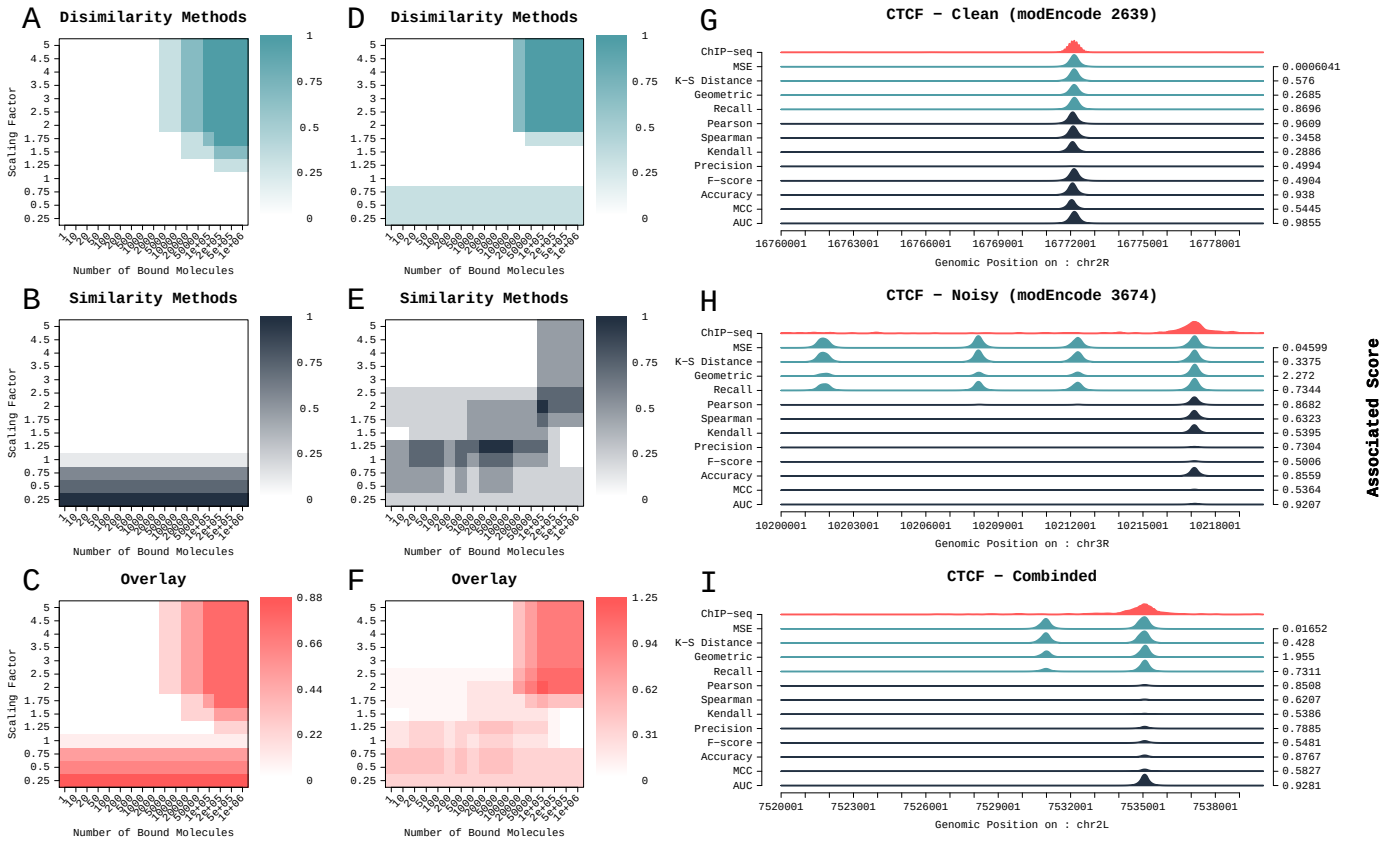

Figure 2: **Goodness of fit Methods are context dependant**. (**A-F**) Heat maps show the overlap of best performing (top 10 %) combination of parameters for similarity and dissimilarity method as well as an overlay of all methods. We produced these heatmaps using the noisy and clean data sets. (**G**) ChIPanalyser correctly predicts CTCF peaks in a clean ChIP dataset (modEncode 2639) for the majority of metrics used. (**H**) For a noisier dataset (modEncode 3674) , dissimilarity metrics capture the height of the peak but also tend to show a high rate of False Positive peaks. In contrast, similarity metrics accurately predicted the location of the peak, but tend to fall short in terms on peak height. (**I**) Combining several ChIP replicates (all ChIP-seq datasets in S2 cells; see Table S1 in *Supplementary Material*) does not reduce the rate of False Positive peaks for similarity metrics.

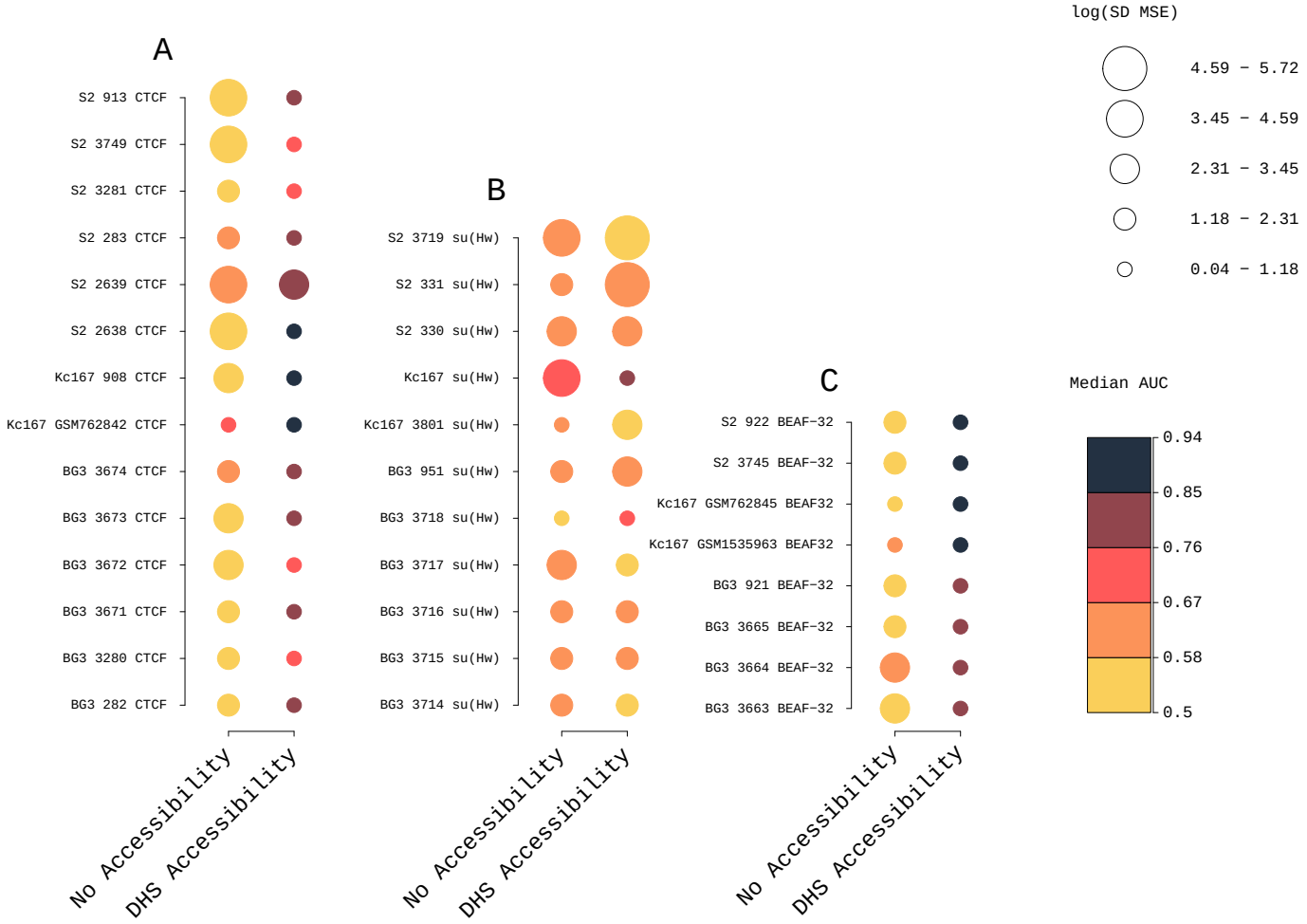

Figure 3: **DNA accessibility, number of molecules and binding energy have different roles in TF binding**. We selected the AUC of best performing combination of parameters (see Table S2 in *Supplementary Material*) and then computed the median values of these AUC score over all selected regions. We 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 AUC scores for all combination of parameters. We then log transformed the final result for data compaction and visualisation purposes. Smaller circles indicate less variability in AUC when different parameters are used and larger circles more variability.

and then parsed to ChIPanlyser. To factor in for potentially differences in the capacity of the model to predict binding in regions with strong or weak ChIP signal, we selected the top 20, 50, 100, 150, 200, 300 and 500 regions in terms of ChIP signal that also contained accessible DNA. We then looked at how the AUC score changes when regions with weaker binding are included in the analysis or when DNA accessibility is considered. For each number of regions selected and for each data set, we subtracted the AUC score when no accessibility was considered from the AUC score with DHS accessibility. Our results indicate that CTCF, BEAF-32 and su(Hw) are all influenced by DNA accessibility albeit in a different manner. First, we observed that both CTCF and BEAF-32 displayed significantly higher AUC scores when DNA accessibility was included, supporting the previous findings (Figure 4A-B and Figure S5 in *Supplementary Material*). In addition, AUC scores for CTCF decreased as the number of regions selected for analysis increased (Figure 4A and Figure S5 in *Supplementary Material*), while BEAF-32 AUC scores were not affected by the increase in the number of regions (Figures 4B and E Figure S5 in *Supplementary Material*). This means that BEAF-32 performs much better when DNA accessibility is considered, but its binding does not seem to be influenced by the number of regions selected. BEAF-32 would bind anywhere along the genome as long as it has an accessible site. CTCF also displays better AUC scores when accessibility is considered, but, in contrast to BEAF-32, analysing more regions (also with weaker binding) negatively affect the performance of the predictions for CTCF. This implies that CTCF binds in accessible DNA but preferentially binds to genome hotspots. We call the BEAF-32 a global binder and CTCF a hotspot TF.

Furthermore, Figure 4C and F shows that there is a small but 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. Our su(Hw) predictions worsened as the number of regions increased, but only when DNA accessibility was considered (see Figure S5 in *Supplementary Material*). The opposite became true when DNA accessibility was not considered (see Figure S5 in *Supplementary Material*). While, su(Hw) did not generally 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 preferred binding sites of su(Hw) are found in inaccessible DNA. Increasing the number of regions only increases the probability of including these high affinity sites in the analysis.

This analysis was performed by optimising AUC scores, but we also run a similar analysis for optimising MSE and our findings are also supported in that case (Figure S6 in *Supplementary Material*).

Number of bound molecules and TF specificity influences TF binding.

The number of bound molecules and the scaling factor of the PWM scores have an impact on the binding profiles of TFs [22]. Building upon this idea, we sought to identify the optimal set of parameters by minimising the MSE between our predicted ChIP profile and experimentally produced ChIP-seq profiles. In this part of the analysis, we used MSE as the question at hand requires the predicted curve to follow experimental curve not only in location but for relative enrichment as well.

The first step was to show that the optimal parameters selected were consistent between different biological replicates. If this was not case, we would not be able to ascertain the validity of our inferred parameters. The optimal set of parameters can be visualized as a heat map showing goodness of fit score for a set of bound molecules and scaling factors. Despite strong variations between experimental data, we show that the predicted optimal set of parameters remained similar between biological replicates and differences tend to arise from differences between cell lines (see Figure 5). This suggests that despite biological and technical variation between replicates performed by different labs using different protocols, our model robustly infers similar number of bound molecules and scaling factor for a given TF. Interestingly the same robustness carries over to other goodness of fit metrics (see Figures S7 and S8 in *Supplementary Material*).

Furthermore, we found that CTCF seems to be more abundant in BG3 cells than Kc167 cells, with S2 cells displaying intermediate levels (see Figure 5). In contrast, BEAF-32 and su(Hw) seem to have similar number of bound molecules in the three cell lines. The optimal parameter estimates for the best performing number of regions can be found in Table S2 and Table S3 in *Supplementary Material*for both AUC and MSE.

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 combination, we can then concluded that they do not strongly influence our predicted profiles. We analysed the standard deviation of MSE between different set 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 Figure 3). CTCF and BEAF-32 showed a decrease in sensitivity to number of bound molecules and the scaling factor when accessibility is considered (Figure 3A and C). This means that DNA accessibility would be the strongest driver towards predicting TF binding. Restricting the amount of available binding motifs would be more influential than TF copy number and the ability of a TF to discriminate between high and low affinity sites and we can see this in the decrease of parameter sensitivity with DHS. When all DNA is considered accessible, these parameters have a stronger influence on modulating the predicted curves. It should be noted that this behaviour was also observed when using other metrics than MSE e.g. AUC in Figure 3).

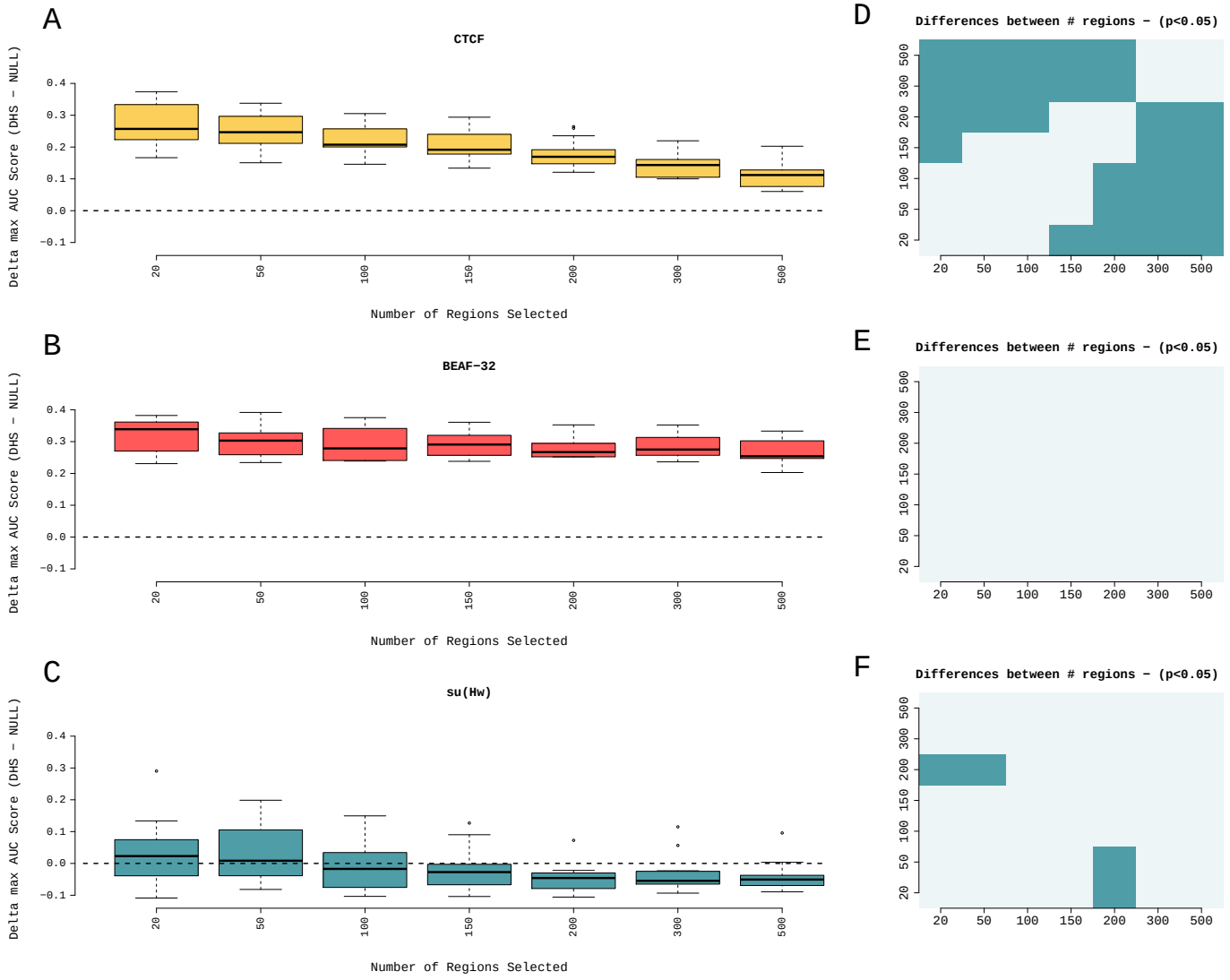

Figure 4: **Number of selected regions sheds light on TF behaviour.** (**A-C**) Boxplot representing the difference in AUC between the model with and without DNA accessibility for several biological replicates and different number of selected bins. (**D-F**) ANOVA test to asses whether the differences are statistically significant (blue indicates statistically significant differences). (**A** and **D**) Predictions of CTCF binding are improved by DNA accessibility, but this improvement is present only at top bound regions. (**B** and **E**) Predictions of BEAF-32 binding are improved by DNA accessibility and are not affected by number of regions selected. (**C** and **D**) su(Hw) performs better when all DNA is considered accessible.

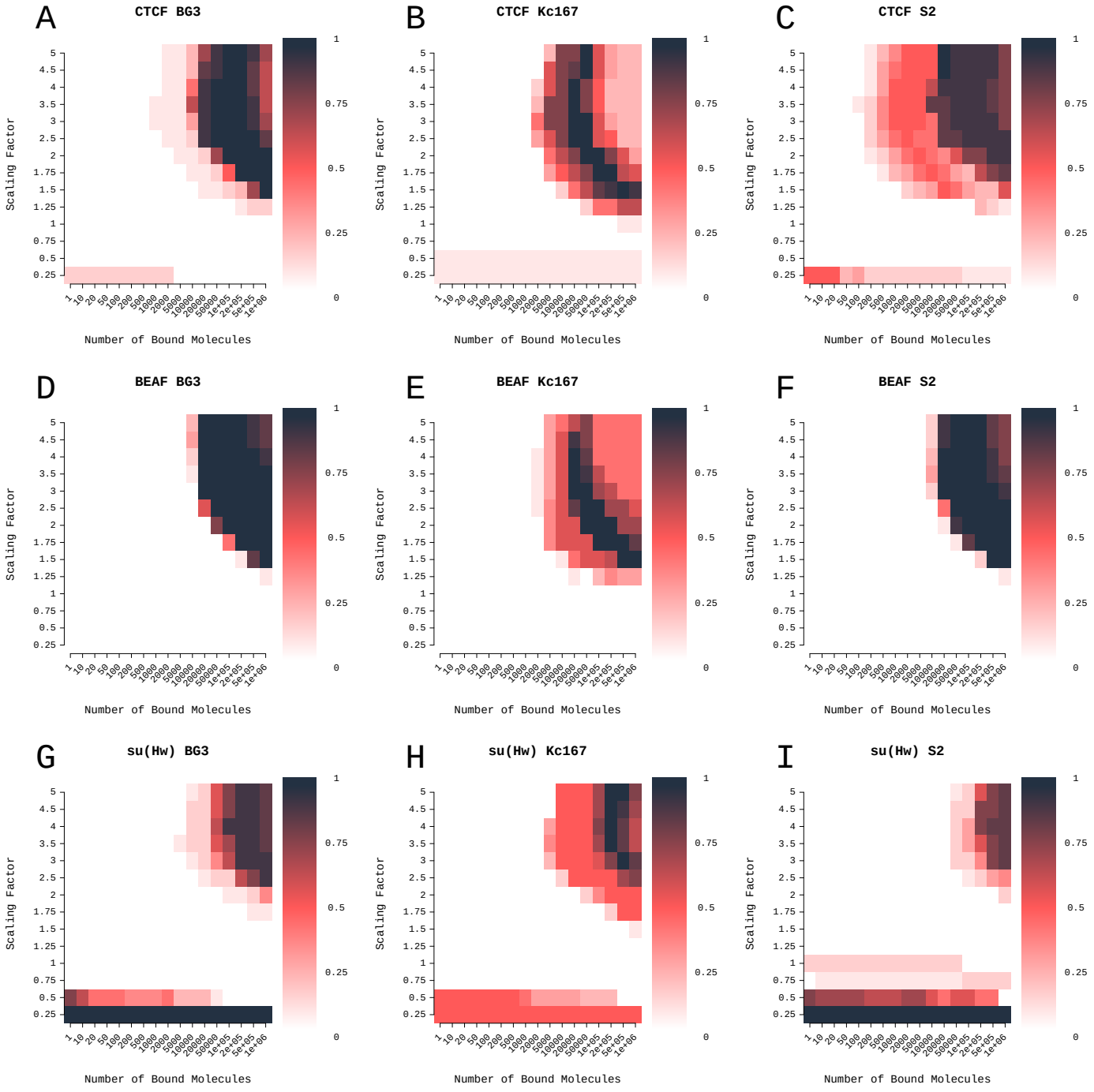

Figure 5: **Optimal parameters consistency among biological replicates.** Heat maps show an overlay of the top 10 % combination of parameters when minimising MSE for: (**A-C**) CTCF, (**D-F**) BEAF-32 and (**G-I**) su(Hw). We plot the following cell lines: (**A**, **D** and **G**) BG3, (**B**, **E** and **H**) Kc167 and (**C**, **F** and **I**) BG3.

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

We wanted to further investigate the predictive capabilities our model and also demonstrate its mechanistic soundness for CTCF, BEAF-32 and su(Hw) in the three 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 that would minimise MSE as $\lambda = 2.5$ and $N = 2 \times 10^5$ over the top 20 regions (see *Materials and Methods*). By rescaling N based on relative RNA-seq levels of CTCF in the two cells lines, we could approximate the number of CTCF molecules bound to DNA in BG3 cells ($\approx 3.3 \times 10^5$). This together with BG3-specific DNA accessibility data is capable to predict with high accuracy the ChIP-seq profile in BG3 cells (see Figure 6A and B). RNA rescaling seems to recover both the number of peaks and their location with high accuracy. Moreover, the rescaling of number of bound molecules did not lead to any difference in terms of MSE variation between estimated and rescaled (Figure 6G). The same analysis was performed for BEAF-32 (Figure 6C, D and H), where we estimated parameters in BG3 cells ($\lambda = 2$ and $N = 2\times10^5$) and rescaled the number of molecules in S2 cells ($\approx N = 1.2\times10^5$). Once again, the model correctly predicts ChIP profiles in both location and relative enrichment. Finally, for su(Hw) (Figure 6E, F and I) we estimated parameters in Kc167 cells ($\lambda = 3$ and $N = 5\times10^4$) and rescaled the number of molecules in S2 cells ($\approx N = 3\times10^4$). Again, the predictions of the model are accurate.

Our results show that changes between cells in DNA accessibility and number of molecules are enough to explain the changes in TF binding profiles. 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 there is no change or a one thousand time reduction in the number of bound molecules and repeated the analysis. Figure 6 shows that using the same TF abundance as in the original cell line did not change the predictions quality at all. In fact, we observed a significant reduction in the predicted profile only when reducing the number of bound molecules by 1000. These results show that cell differences in binding profiles of TFs 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 downregulated. The fact that number of bound molecules variations between cell lines have small effects on the binding profiles is not so surprising. Figures 5 and 3 show that there are wide range of values for the number of bound molecules that lead to optimising the goodness of fit metrics (see also Figures S7 and S8 in *Supplementary Material*).

Hox genes show differentially 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 [42]. 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 accesibility. DNA accesibility level were selected based on quantile distribution of ATAC-seq scores (see *Materials and Methods*). Briefly, this means that higher QDA scores lead to the fewer regions being marked as accessible.

We selected regions containing both peaks and accessible regions and then, for each QDA accessibility, selected the optimal set of parameters by maximising the AUC score between our prediction and ChIP-seq data. Our results show that Ubx exhibits a preference towards open chromatin. In Figure 7A, the maximum AUC score for Ubx increases with increasing the QDA score. This can be explained by the fact that increasing the threshold of ATAC-seq score ensures that open chromatin regions are truly opened and not an intermediate state. 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 (99th quantile of ATAC-seq scores - AUC 0.862), while Abd-B and Dfd with 0.95 QDA and 0.8 QDA respectively (see Figure 7B).

The model recovers the position of peaks accurately especially for Ubx (see Figure 7C-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 [42]. Our model does not include cooperative interactions and this could explain the reduced performance for Dfd and Abd-B.

DISCUSSION

Our analysis shows that ChIPanalyser and its underlying model predicts binding profiles of TFs (ChIP) with high accuracy and 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. The optimal set of parameters seem to be grounded into the biology itself. Here we untangled the interplay between DNA accessibility and the number of bound molecules and how that impacts on the binding of TFs to the DNA.

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

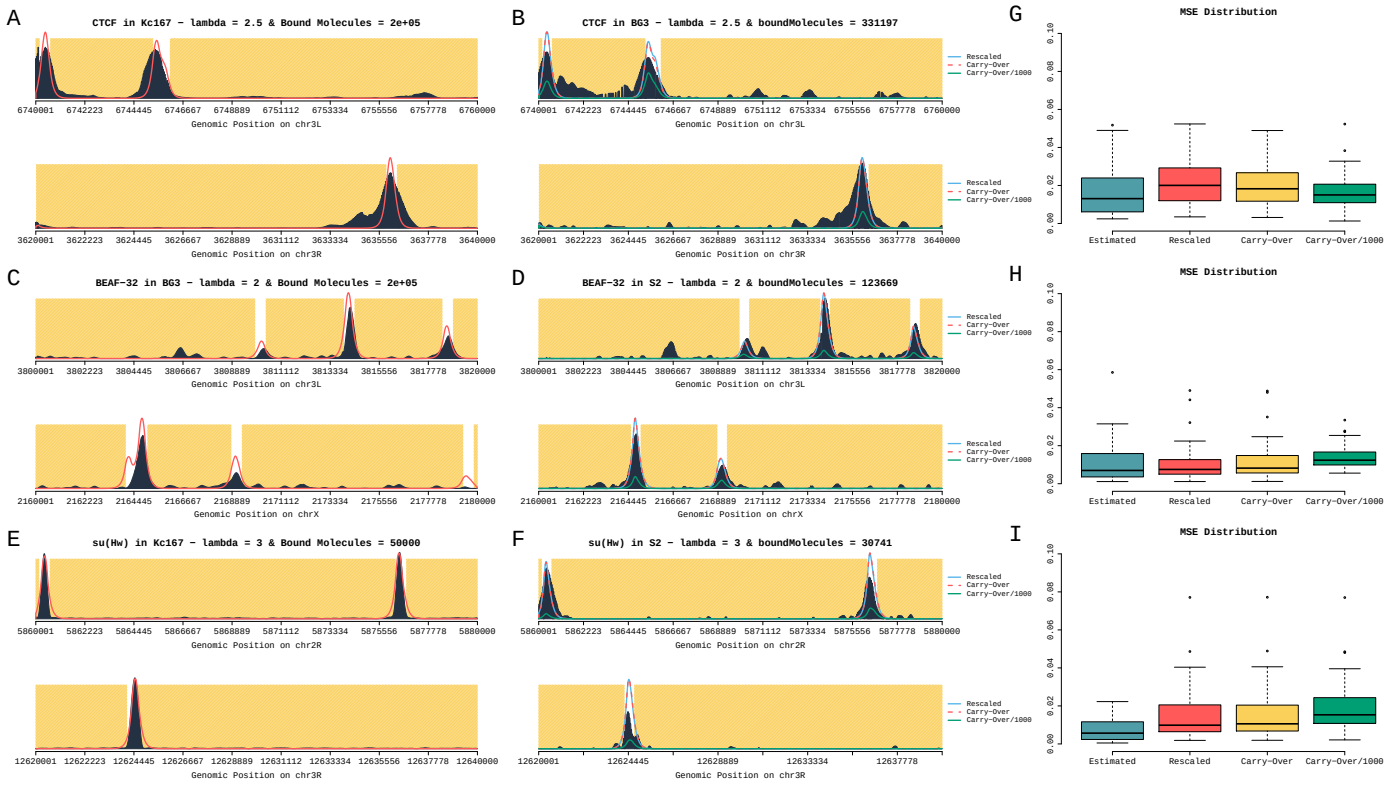

Figure 6: **Predicting TF binding in different cell lines by considering relative mRNA abundance. A-F** show predicted ChIP-seq profiles with our TF abundance estimated based on RNA-seq. The yellow area represents inaccessible DNA, the dark grey area represents experimental ChIP signal and the lines are our predicted profiles. We estimated the number of bound molecules in one cell line (**A**, **C** and **E**) and rescaled our estimate using relative mRNA abundance in an other cell line (**B**, **D** and **F**). (**B**, **D** and **F**) The blue line represents the rescaled value of number of bound molecules based on RNA-seq, the dashed red line the original value estimated in (**A**, **C** and **E**) and the green line the original estimated value reduced 1000 times. (**G** , **H** and **I**) Boxplots with MSE for all cases in the estimated and predicted profiles at top 50 regions.

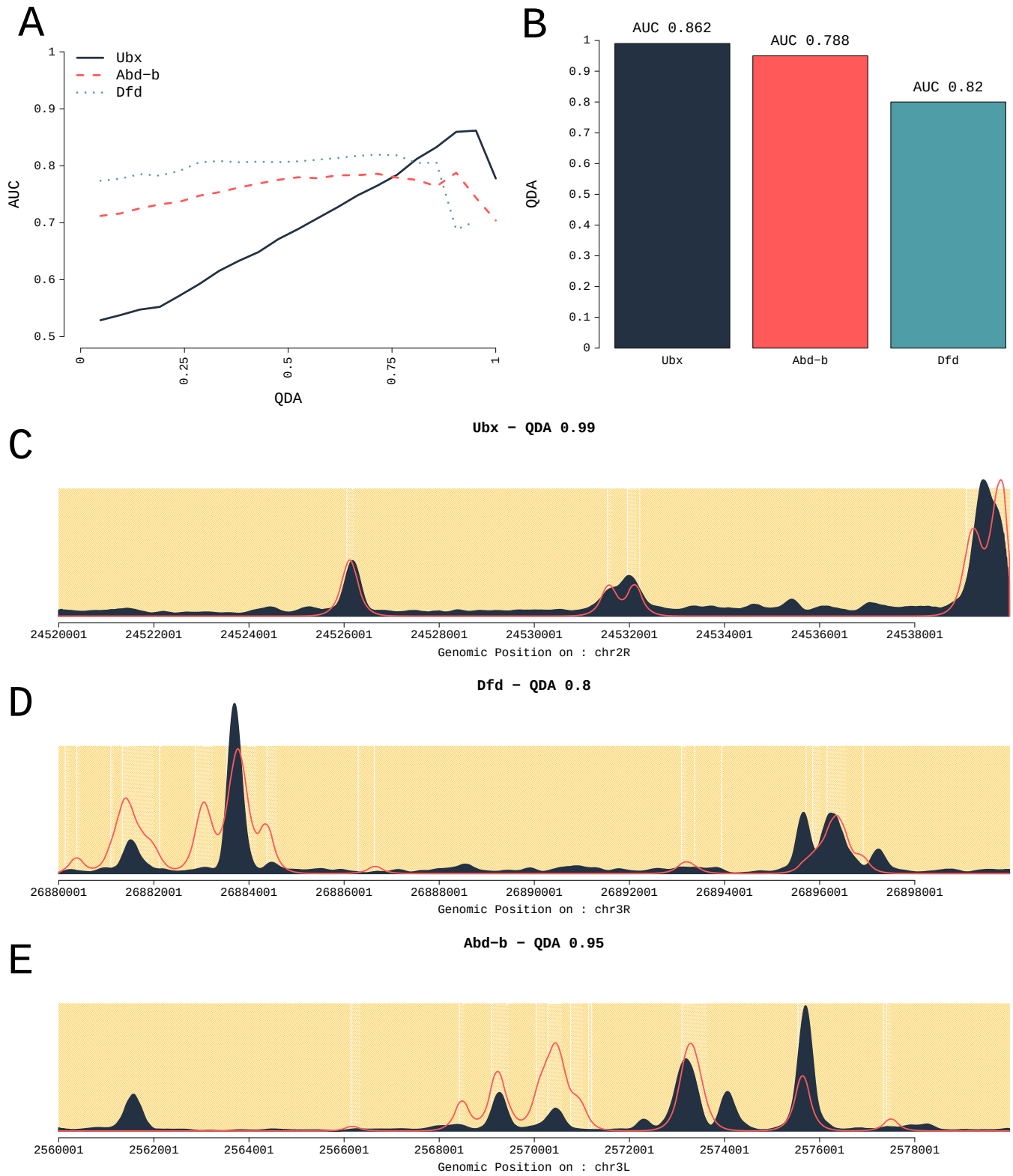

Figure 7: **Hox genes show binding preferences towards DNA accessibility.** We tested our model using different DNA accessibility stringencies. (**A**) Maximum AUC score as a function of stringency of DNA accessibility (the higher the QDA value the less DNA is called accessible) for three Hox TFs: Ubx, 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-seq data at individual *loci* for the three TFs.

also play roles in transcription regulation and insulation [45, 46, 47, 48, 49, 50]. Moreover, it was shown that these three TFs showed distinct binding behaviours and were classified into three subclasses with respect to chromatin architecture [51, 52]. In our analysis we show that CTCF, BEAF-32 and su(Hw) all exhibit different behaviours with respect to DNA binding.

CTCF has been shown to play a role in loop formation and participating in Topologically Associated Domains (TADs) boundary maintenance [50]. However, only a subset of CTCF sites are involved in these structures and that many CTCF sites do not conform to this rule [53, 54]. In our analysis, CTCF displayed strong sensitivity to DNA accessibility but reduced sensitivity to the number of bound molecules and scaling factor when DNA accessibility was considered (see Figures 3 and 4). Our findings suggest that CTCF binds to hotspots along the genome and this could be explained by the observation that the strongest peaks (based on our selection method - see *Materials and Methods*) are in fact highly conserved binding sites. As the number of sites increase, the conservation of binding sites decreases, as does the goodness fit. Thus, CTCF binding to highly conserved sites can be explained by our model, but something else is is responsible for the reduce binding at less conserved sites (i.e. cell specific CTCF binding) [55].

BEAF-32 is a Drosophila specific genetic insulator [56] that shows preferential binding towards TAD boundaries, but also is involved in transcription itself. More specifically, BEAF-32 was identified as a cis-regulatory element separating close head-to-head genes with different transcription regulation modes [57]. In Drosophila, there is a high density of these genes through out the genome and BEAF-32 tends to bind closely to the TSS [58]. This is further confirmed by studies showing that BEAF-32 has uniform binding along the entire genome [51]. TSSs are generally considered open chromatin and, if BEAF-32 binds in close proximity of the TSSs, it comes to no surprise that BEAF-32 would show a high sensitivity towards DNA accessibility. Our results confirm that BEAF-32 shows a strong preference towards DNA accessibility and, to a lower extent to local abundance.

Furthermore, we show that su(Hw) binds in both open and closed chromatin and also displays a high sensitivity towards number of bound molecules and scaling factor. There is a significant body of work showing the role su(Hw) plays in chromatin insulation and remodelling [59, 60, 61, 62]. It had been suggested that the role of insulator is only possible when paired with other DNA binding proteins such as Cp190 and mdg4. su(Hw) is also a primary actor in the interaction between the genome and nuclear lamina (also know as Lamina Associated Domains) [63, 64]. Both chromatin insulation and LADs would induce closed chromatin in order to maintain chromosomal structure and this would explain why su(Hw) can bind in both open and closed chromatin. In this context, ChIP-seq peaks might not overlap well with DNase hyper sensitivity data, which would be even more the case for our highly stringent method of selecting DNA accessibility sites from DHS sites (see *Materials and Methods*).

It has been shown that su(Hw) binding sites tend to cluster together (with varying number of sites) and that these sites are constitutively bound by su(Hw) [65, 66]. 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. If cluster binding sites are constitutively bound and the density of these cluster vary along the genome, this would suggest that the number of bound molecules and how well they discriminate between low and high affinity sites, is a strong driver towards su(Hw) binding. We show that if DNA accessibility is not considered su(Hw) was sensitive to these two factors (see Figure 3 and Figures S5 and S6 in the *Supplementary Material*) thus suggesting a mechanistic explanation of this behaviour.

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

Our results show that DNA accessibility and number of bound molecules control the binding profiles of TFs (Figures 3 and 4). 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 RNA-seq), we found a good agreement between our predictions and the actual ChIP-seq dataset (see Figure 6). 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 Figure 6 B, D and F). This indicates that biological relevant fluctuations in TF numbers between different cell lines would have little effect on the differences in binding profiles of TFs and those differences are mainly driven by changes in DNA accessibility. Furthermore, only when reducing the TF concentration by 1000, we observed a noticeable decrease in the predicted ChIP profile, which suggests that only strong knockdowns or overexpression would affect binding of TFs and, consequently, lead to changes in the expression of target genes. It should be noted that the TFs we analysed here (CTCF, BEAF-32 and su(Hw)) are highly expressed architectural and insulator proteins and, thus, they would be saturating their binding sites.

Why would relative medium changes in concentration of the TF have such a limited effect on the binding of the TF? One potential explanation is that TFs control the expression of essential genes that should be tightly regulated despite fluctuations in number of molecules that affect the cell [67, 68]. This would be a buffering mechanism for the fluctuations in protein numbers in the cell.

Finally, we also investigate the capacity of our model to differentiate between TFs that can bind only in open chromatin and TFs that can also bind in less opened chromatin. For that, we looked at three Hox TFs displaying different preferences for DNA accessibility. Our results showed that while Ubx displays a strong sensitivity to open chromatin and binds in the top 1% accessible sites, the binding of Dfd and Abd-B is less influenced by DNA accessibility (with Abd-B and Dfd binding in top 5% and 20% respectively accessible regions); see Figure 7. Interestingly, our statisti-

cal thermodynamics model is better predicting the binding profile of Ubx (AUC of 0.862) compared to Abd-B and Dfd (with AUC 0.79 and 0.82 respectively).

Hox TFs are known for having a similar motif, but displaying differences in their binding profiles [69, 70, 71]. It was hypothesised that binding cooperativity could explain the difference in binding profiles coupled with protein sequence changes [72, 73, 74]. Here, we showed that DNA accessibility could also be responsible for the difference in binding profiles of Hox TFs (see Figure 7). Interestingly, our results support a model where Hox TFs would be able to bind to regions of DNA showing different level of accessibility and the DNA accessibility would be sufficient to explain these differences in the binding profiles of Hox TFs. Nevertheless, we also observed a poorer quality in modelling the binding profiles of TFs that can bind in dense chromatin (e.g., Abd-B or Dfd), which suggests that cooperative binding would be required to explain their binding. Due to the fact that our model does not include cooperativity, the predictions for these TFs would not be as accurate as in the case of TFs that preferentially bind to open chromatin.

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 impede our ability to accurately assess the goodness of fit of the model. This ability is a corner stone in our understanding of the biological implication arising from our findings. Despite our approaches to reduce background noise, it seems that ChIP-seq data will always suffer from unspecific DNA pull-down [75]. More complex method of signal filtering are available, and applying these methods could potentially lead to significant reduction in the noise of ChIP-seq signal.

Another possibility is that the noise in ChIP signal could be the results of unspecific binding of TFs to DNA followed by one-dimensional random walk along the genome [76, 77, 78, 79, 80]. For the purpose of our analysis, we selected only sites of DNase hypersensitivity and considered these regions as strictly open. However, regions that were marked as closed chromatin between clusters of open regions might in fact either be partially open or dynamically opened, thus, leaving time and space for 1D molecular walks along the genome [81]. Discerning real TF binding and experimental artefacts remains extremely challenging.

We showed that choosing a goodness of fit method is context dependant. Interestingly, similarity methods (such as correlation, F-score or AUC) had the tendency of correctly calling peaks location but greatly underestimated the enrichement on the peak (see Figure 2). This behaviour results from the fact that these method are highly penalised by false positive hits. They show a wide range of optimal values for the number of bound molecules, but they tended to prefer low values for the scaling factor. This scaling factor can be described as how well a TF discriminates between binding sites; i.e., how much a TF will prefer a strong binding site over a weaker one. High values for the scaling factor translates 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. Smaller peaks could be caused by lower affinity binding or suboptimal binding sites along the genome as described by [82], but these binding sites would not be picked up by the similarity methods. The number of bound molecules on the other hand tend to affect the height of the peak (relative local enrichment). Similarity method would avoid inflating these sites as this would penalised their goodness of fit score more severely than dissimilarity methods. Dissimilarity methods (such as MSE or geometric ratio) showed a much higher number of bound molecules and a high value for the scaling factor (see Figure 2).

It is interesting to see that each method is penalised by different aspects of the model. For these reasons, we believe that choosing the right method will depend on the question at hand. Similarity methods could be used to determine peak location, but, if the TF local enrichment is of interest, a dissimilarity metric would be more appropriate.

FUNDING

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

ACKNOWLEDGEMENTS

We thank Dr Rob White for sharing the Hox ChIP-seq and ATAC-seq data and for comments on this manuscript. We also thank Professor Sarah Bray and Zabet lab 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 on using the cluster.

Conflict of interest statement.

None declared.

References

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

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

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

[4] Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B. E.,

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

[5] Berg, O. G. and von Hippel, P. H. (1987) Selection of DNA Binding Sites by Regulatory Proteins Statistical-mechanical Theory and Application to Operators and Promoters. *Journal of Molecular Biology,* **193**(4), 723–750.

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

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

[8] Stormo, G. D. and Zhao, Y. (2010) Determining the specificity of protein-DNA interactions. *Nature Reviews,* **11**, 751–760.

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

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

[11] Li, X.-y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D. A., Iyer, V. N., Hechmer, A., Simirenko, L., Stapleton, M., Hendriks, C. L. L., Chu, H. C., Ogawa, N., Inwood, W., Sementchenko, V., Beaton, A., Weiszmann, R., Celniker, S. E., Knowles, D. W., Gingeras, T., Speed, T. P., Eisen, M. B., and Biggin, M. D. (feb, 2008) Transcription Factors Bind Thousands of Active and Inactive Regions in the Drosophila Blastoderm. *PLoS Biol.,* **6**(2), e27.

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

[13] Skalska, L., Stojnic, R., Li, J., Fischer, B., Cerda-Moya, G., Sakai, H., Tajbakhsh, S., Russell, S., Adryan, B., and Bray, S. J. (2015) Chromatin signatures at Notch-regulated enhancers reveal large-scale changes in H3K56ac upon activation. *The EMBO Journal,* **34**(14), 1889–1904.

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

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

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

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

[18] Small, S., Blair, A., and Levine, M. (nov, 1992) Regulation of even-skipped stripe 2 in the Drosophila embryo.. *EMBO J.,* **11**(11), 4047–57.

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

[20] Cheng, Q., Kazemian, M., Pham, H., Blatti, C., Celniker, S. E., Wolfe, S. A., Brodsky, M. H., and Sinha, S. (2013) Computational Identification of Diverse Mechanisms Underlying Transcription Factor-DNA Occupancy. *PLoS Genet,* **9**(8), e1003571.

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

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

[23] Pearson, J. C., Lemons, D., and McGinnis, W. (dec, 2005) Modulating Hox gene functions during animal body patterning. *Nat. Rev. Genet.,* **6**(12), 893–904.

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

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

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

[27] Peng, P.-C., Khoueiry, P., Girardot, C., Reddington, J. P., Garfield, D. A., Furlong, E. E., and Sinha, S. (oct, 2018) The role of chromatin accessibility in cis-regulatory evolution. *bioRxiv,* p. 319046.

[28] Thomas, S., Li, X.-Y., Sabo, P. J., Sandstrom, R., Thurman, R. E., Canfield, T. K., Giste, E., Fisher, W., Hammonds, A., Celniker, S. E., Biggin, M. D., and Stamatoyannopoulos, J. A. (2011) Dynamic reprogramming of chromatin accessibility during Drosophila embryo development.. *Genome Biol.,* **12**(5), R43.

[29] Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos, G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E., and Stamatoyannopoulos, J. A. (sep, 2012) The accessible chromatin landscape of the human genome. *Nature,* **489**(7414), 75–82.

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

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

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

[33] R Development Core Team (2014) R: A language and environment for statistical computing.. *R Foundation for Statistical Computing,*.

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

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

[36] Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y.-H. C., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Miklos, Abril, J. F., Agbayani, A., An, H.-J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., Pablos, B. d., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M.-H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D. C., Scheeler, F., Shen, H., Shue, B. C., Sidén-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z.-Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., Woodage, T., Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R.-F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., and Venter, J. C. (2000) The Genome Sequence of Drosophila melanogaster. *Science,* **287**(5461), 2185–2195.

[37] dos Santos, G., Schroeder, A. J., Goodman, J. L., Strelets, V. B., Crosby, M. A., Thurmond, J., Emmert, D. B., Gelbart, W. M., and the FlyBase Consortium

(11, 2014) FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Research,* **43**(D1), D690–D697.

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

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

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

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

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

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

[44] Cherbas, L., Willingham, A., Zhang, D., Yang, L., Zou, Y., Eads, B. D., Carlson, J. W., Landolin, J. M., Kapranov, P., Dumais, J., Samsonova, A., Choi, J.-H., Roberts, J., Davis, C. A., Tang, H., van Baren, M. J., Ghosh, S., Dobin, A., Bell, K., Lin, W., Langton, L., Duff, M. O., Tenney, A. E., Zaleski, C., Brent, M. R., Hoskins, R. A., Kaufman, T. C., Andrews, J., Graveley, B. R., Perrimon, N., Celniker, S. E., Gingeras, T. R., and Cherbas, P. (2011) The transcriptional diversity of 25 Drosophila cell lines. *Genome Research,* **21**(2), 301–314.

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

[46] Li, L., Lyu, X., Hou, C., Takenaka, N., Nguyen, H. Q., Ong, C.-T., Cubeñas-Potts, C., Hu, M., Lei, E. P., Bosco, G., Qin, Z. S., and Corces, V. G. (2015) Widespread Rearrangement of 3D Chromatin Organization Underlies Polycomb-Mediated Stress-Induced Silencing. *Molecular Cell,* **58**(2), 216–231.

[47] Hug, C. B., Grimaldi, A. G., Kruse, K., and Vaquerizas, J. M. (2017) Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription.. *Cell,* **169**(2), 216 – 228.e19.

[48] Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K. C., Grüning, B. A., Villaveces, J., Habermann, B., Akhtar, A., and Manke, T. (dec, 2018) High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications,* **9**(1), 189.

[49] Wang, Q., Sun, Q., Czajkowsky, D. M., and Shao, Z. (dec, 2018) Sub-kb Hi-C in D. melanogaster reveals conserved characteristics of TADs between insect and mammalian cells. *Nature Communications,* **9**(1), 188.

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

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

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

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

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

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

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

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

[58] Rennie, S., Dalby, M., Lloret-Llinares, M., Bakoulis, S., Dalager Vaagensø, C., Heick Jensen, T., and Andersson, R. (2018) Transcription start site analysis reveals widespread divergent transcription in D. melanogaster and core promoter-encoded enhancer activities. *Nucleic Acids Research,* **46**(11), 5455–5469.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[75] Teytelman, L., Thurtle, D. M., Rine, J., and van Oudenaarden, A. (2013) Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *PNAS,* **110**(46), 18602,Äì18607.

[76] Berg, O. G., Winter, R. B., and von Hippel, P. H. (1981) Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory.. *Biochemistry,* **20**(24), 6929–6948.

[77] Elf, J., Li, G.-W., and Xie, X. S. (2007) Probing Transcription Factor Dynamics at the Single-Molecule Level in a Living Cell. *Science,* **316**, 1191–1194.

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

[79] Hammar, P., Leroy, P., Mahmutovic, A., Marklund, E. G., Berg, O. G., and Elf, J. (2012) The lac Repressor Displays Facilitated Diffusion in Living Cells. *Science,* **336**(6088), 1595–1598.

[80] Zabet, N. R. and Adryan, B. (2012) Computational models for large-scale simulations of facilitated diffusion.. *Molecular BioSystems,* **8**(11), 2815–2827.

[81] Ezer, D., Zabet, N. R., and Adryan, B. (2014) Physical constraints determine the logic of bacterial promoter architectures. *Nucleic Acids Research,* **42**(7), 4196–4207.

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