A base-resolution panorama of the *in vivo* impact of cytosine methylation on transcription factor binding

Aldo Hernandez-Corchado[1,2], Hamed S. Najafabadi[1,2,*]

[1] Department of Human Genetics, McGill University, Montreal, QC H3A 1B1, Canada

[2] McGill Genome Centre, Montreal, QC H3A 0G1, Canada

* Corresponding author: H. S. Najafabadi, hamed.najafabadi@mcgill.ca

ABSTRACT

Methylation of the cytosine base at CpG dinucleotides is traditionally considered antagonistic to the DNA-binding activity of the majority of transcription factors (TFs). Recent *in vitro* studies of TF-DNA interactions have revealed a more complex picture, suggesting a heterogeneous cytosine methylation impact that varies across TFs, with over a third of TFs preferring methylated sequences. Expanding these *in vitro* observations to *in vivo* TF binding preferences, however, is challenging, as the effect of methylation of individual CpG sites cannot be easily isolated from the confounding effects of DNA accessibility and regional DNA methylation. As a result, the *in vivo* methylation preferences of most TFs remain uncharacterized.

Here, we introduce Joint Accessibility-Methylation-Sequence (JAMS) models for inferring the effect of CpG methylation on TF binding *in vivo*. JAMS creates quantitative models that connect the strength of the binding signal observed in ChIP-seq to the DNA accessibility of the binding site, regional methylation level, DNA sequence, and base-resolution cytosine methylation. Furthermore, by jointly modeling both the control and pull-down signal in a ChIP-seq experiment, JAMS isolates the TF-specific effects from background effects, revealing how methylation of specific CpGs within the binding site alters the TF binding affinity *in vivo*.

We show that JAMS can quantitatively model the TF binding strength and learn the accessibility-methylation-sequence determinants of TF binding. JAMS models are reproducible and generalizable across cell lines, and can faithfully recapitulate cell type-specific TF binding. Systematic application of JAMS to 2368 ChIP-seq experiments generated high-confidence models for 260 TFs, revealing that 45% of TFs are inhibited by methylation of their potential binding sites *in vivo*. In contrast, only 6% prefer to bind to methylated sites, including 11 novel methyl-binding TFs. Comparison of these *in vivo* models to *in vitro* data confirmed high precision of the methyl-preferences inferred by JAMS. Finally, among the CpG-binding proteins from the ZF-KRAB family of TFs, we observed a disproportionately high preference for methylated sequences (24%), highlighting the role of CpG methylation in determining the genome-wide binding profiles of the TFs from this family.

BACKGROUND

Transcription factors (TFs) are key regulators of gene expression. Each TF usually recognizes a specific sequence motif; however, TF binding is affected by several other variables, among which cytosine methylation is traditionally viewed as having a repressive effect on TF binding [1]. However, this traditional view is gradually changing, as more examples are reported of TFs that bind to methylated sequences. These include studies that have reported increased binding of specific TFs to methylated DNA *in vitro* [2], in addition to reports indicating that, for some TFs, a large fraction of their *in vivo* binding sites is highly methylated [3, 4].

While it is tempting to view these anecdotal cases as exceptions rather than a general trend, a recent systematic analysis of TF CpG methylation preferences revealed that, in fact, a large fraction of TFs may bind to methylated CpGs *in vitro*. Based on this study, the effect of methylation is dependent on its position in the binding site, and is heterogeneous within and across TF families [5]. While this study provides *in vitro* evidence for widespread recognition of methylated CpGs by TFs, a comparable systematic analysis of *in vivo* methylation preferences of TFs is still lacking. This is primarily because observing the specific *in vivo* effect of intra-motif CpG methylation is confounded by binding site-specific factors such as DNA accessibility, regional methylation level, and binding site sequence [6-8]. Experimental approaches to control these confounding factors are complicated and resource-exhaustive [9-11], highlighting the need for computational methods to untangle, from these confounding variables, the base-resolution relationship between TF binding affinity and intra-motif CpG methylation.

A few recent studies have proposed computational methods to identify TFs that are affected by CpG methylation *in vitro*. These include efforts to better distinguish bound from unbound sequences using TF binding models that incorporate CpG methylation status [12, 13], as well as tools that expand the ATGC alphabet by adding symbols for methylated cytosines in order to perform methylation-aware *de novo* motif discovery [14, 15]. These methods, however, only report whether methylation improves TF binding prediction without delineating the direction of the effect [13], lack the resolution to investigate the effect of methylation of individual intra-motif cytosines [13], and/or do not consider the confounding effects of DNA accessibility and regional methylation level [12-15]. As a result, even some of the most classic methyl-binding TFs, such as CEBPB [2] and KAISO [16], are not detected by these methods [12].

To overcome these challenges, we introduce Joint Accessibility-Methylation-Sequence (JAMS) models, a statistical framework for deconvolving the individual contribution of various factors, including intra-motif CpG methylation, on the *in vivo* strength of TF binding as observed by ChIP-seq. We show that JAMS models are reproducible and generalizable, can capture known CpG methyl-preferences of TFs, and can even predict differential TF binding across cell lines based on changes in intra-motif CpG methylation. Finally, we apply JAMS to a large compendium of ChIP-seq experiments to systematically explore the CpG methylation preferences of TFs across different families.

RESULTS

Modeling the joint effect of accessibility, methylation and sequence on TF binding

Several factors work together to determine the TF binding strength, as measured by ChIP-seq, toward a specific binding site. First, the sequence of the binding site determines the TF affinity, given that the majority of TFs are sequence-specific. Secondly, for most TFs, the existing level of DNA accessibility heavily influences TF binding [7, 8]. Finally, regional methylation outside the TFBS may affect the TF binding strength, for example by recruiting Methyl-CpG-binding domain (MBD) proteins, which in turn recruit chromatin remodelers [6]. Therefore, in order to examine the specific effect of methylation of the TFBS on TF binding affinity, we need to jointly model it together with these confounding factors.

For this purpose, we developed Joint Accessibility-Methylation-Sequence (JAMS) models, which quantitatively explain both the pull-down and background signal in ChIP-seq experiments (https://github.com/csglab/JAMS). The JAMS model for each ChIP-seq experiment considers the pull-down read density as a combination of a background signal and a TF-specific signal. On the other hand, the read count profiles obtained from control experiments (e.g. input DNA) purely reflect the background signal (**Fig. 1A**). Each of the background and TF-specific signals, in turn, is modeled as a function of the peak sequence, chromatin accessibility profile along the peak, regional methylation level, and base-resolution intra-motif CpG methylation (**Fig. 1B-C**). JAMS converts these associations into a generalized linear model, whose parameters can be inferred by fitting simultaneously to both pull-down and control read counts. To ensure that JAMS can correctly learn the features associated with both TF-specific and background signals, we fit the model to the read counts across peaks with a wide range of pulldown-to-control signal ratio. These include not only the peaks that have significantly high pull-down signal, but also peaks with low pull-down signal as well as genomic locations with significantly high background signal. For model fitting, an appropriate error model is needed that connects the expected (predicted) signal at each peak to the observed read counts—we use negative binomial with a log-link function in this work (**Fig. 1D**; see **Methods** for details).

In order to examine the ability of JAMS models to recover the *in vivo* binding preferences of TFs, we first applied it to ChIP-seq data from CTCF, a widely studied TF that is constitutively expressed across cell lines and tissues [17, 18] and has a long residence time on DNA [19]. We initially focused on the cell line HEK293, and generated a JAMS model of CTCF binding in this cell line using previously published ChIP-seq [20], WGBS [21], and chromatin accessibility data [22] (**Methods**). To evaluate the performance of the JAMS model, we used 10-fold cross-validation, and examined the correlation between the predicted TF-specific signal and the observed pulldown-to-control signal ratio across the peak regions. As **Fig. 1E** shows, the JAMSmodel predictions correlate strongly with the pulldown-to-control signal ratio (Pearson r=0.69), suggesting that accessibility-methylation-sequence features can quantitatively predict the CTCF-binding strength.

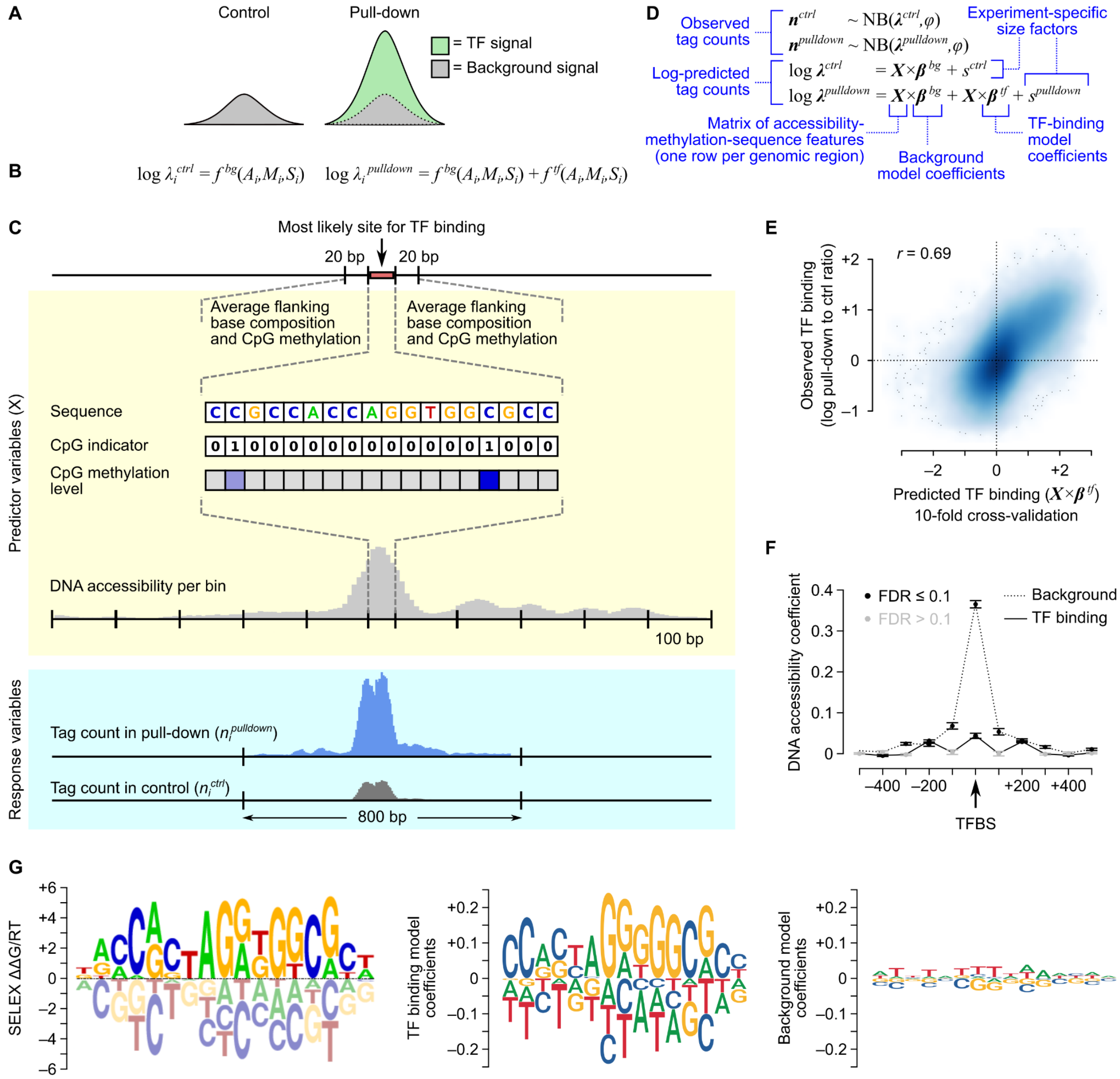
A
Control
Pull-down
= TF signal
= Background signal
B
log λ_i^ctrl = f^bg(A_i,M_i,S_i)
log λ_i^pulldown = f^bg(A_i,M_i,S_i) + f^tf(A_i,M_i,S_i)
C
Most likely site for TF binding
20 bp
20 bp
Average flanking base composition and CpG methylation
Average flanking base composition and CpG methylation
Predictor variables (X)
Sequence
C C G C C A C C A G G T G G C G C C
CpG indicator
0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
CpG methylation level
DNA accessibility per bin
100 bp
Response variables
Tag count in pull-down (n_i^pulldown)
Tag count in control (n_i^ctrl)
800 bp
D
Observed tag counts
n^ctrl ~ NB(λ^ctrl, φ)
n^pulldown ~ NB(λ^pulldown, φ)
Log-predicted tag counts
log λ^ctrl = X×β^bg + s^ctrl
log λ^pulldown = X×β^bg + X×β^tf + s^pulldown
Experiment-specific size factors
Matrix of accessibility-methylation-sequence features (one row per genomic region)
Background model coefficients
TF-binding model coefficients
E
r = 0.69
Observed TF binding (log pull-down to ctrl ratio)
+2
+1
0
−1
−2
0
+2
Predicted TF binding (X×β^tf)
10-fold cross-validation
F
DNA accessibility coefficient
0.4
0.3
0.2
0.1
0
● FDR ≤ 0.1
● FDR > 0.1
Background
TF binding
−400
−200
+200
+400
TFBS
G
SELEX ΔΔG/RT
+6
+4
+2
0
−2
−4
−6
TF binding model coefficients
+0.2
+0.1
0
−0.1
−0.2
Background model coefficients
+0.2
+0.1
0
−0.1
−0.2

Figure 1. Overview of JAMS model. (**A**) At each genomic region *i*, the JAMS model considers the control tag count (left) or the pull-down tag count (right) as a combination of background and/or TF-binding signals at that position. (**B**) Each of these signals are then modeled as a function of accessibility (A_i), methylation (M_i), and sequence (S_i) at each region *i*. (**C**) Schematic summary of the predictor features extracted for each genomic location and the outcome variables. (**D**) The specifications of the generalized linear model used by JAMS. (**E**) Comparison between the observed and predicted CTCF binding signal in HEK293 cells [20]. (**F**) DNA accessibility coefficients learned by the CTCF JAMS model; each dot corresponds to the effect of accessibility at a 100bp-bin. (**G**) Sequence motif logos representing the known CTCF binding preference (based on SELEX [54] (left), the TF binding specificity learned by JAMS (middle), and the effect of sequence on the background signal (right). JAMS motif logos are plotted using ggseqLogo [55], with letter heights representing model coefficients; SELEX motif logo was obtained from the CIS-BP database [45].

Examining the coefficients of the fitted JAMS model, we observed that DNA accessibility, especially at the peak center, has a strong effect on the TF-specific signal (which only affects the pull-down read count), but limited effect on the background ChIP-seq signal (which affects both the control and pull-down read counts; **Fig. 1F**). Nonetheless, the effect on background signal was still statistically significant, consistent with previously observed bias of DNA sonication toward accessible chromatin regions [23]. Importantly, sequence features at the TF binding site are strongly predictive of the CTCF binding strength, while they have limited and diffuse effect on the background signal (**Fig. 1G**). Importantly, the sequence model learned by JAMS is highly correlated with the known motif for CTCF (r=0.86, **Fig. 1G**), suggesting that JAMS models can recapitulate the underlying biology of TF binding.

JAMS models reveal the contribution of CpG methylation to TF binding

By jointly considering the contribution of accessibility, methylation and sequence to TF binding, JAMS models should be able to deconvolve the specific effect of methylation from the confounding effect of other variables. To begin to explore this possibility, we examined the JAMS model of CTCF. For this purpose, in addition to the widely used sequence motif logos, we developed “dot plot logos” to enable easier visual inspection of JAMS coefficients that correspond to sequence and methylation effects. As **Fig. 2A** shows, the JAMS model of CTCF binding in HEK293 cells suggests that CpG methylation in the 2nd and 12th positions of the binding site has a significantly negative effect on CTCF binding (but not on the background signal; **Supplementary Fig. 1**). In other words, while a large fraction of CTCF binding sites have CpGs at those two positions, CTCF preferentially binds when these CpGs are not methylated.

To ensure that this observation is not confounded by other variables such as accessibility and the average local methylation level, we also trained JAMS models with all the variables except the CpG methylation level at each binding site position; we then compared these reduced models to the full model using a likelihood ratio test. This analysis revealed that removing the CpG methylation levels at positions 2 or 12 of the binding site significantly reduces the fit of the model to the observed data (**Supplementary Fig. 2**). Therefore, the CpG methylation level in these positions is informative about CTCF binding signal even after considering the effect of other confounding variables such as sequence, accessibility, and the average methylation of flanking regions. The independent effect of CpG methylation on CTCF binding can also be observed after stratification of CTCF peaks based on the confounding variables. Specifically, we repeated the JAMS modeling after removing the variables that represent the TF-specific contribution of methylation at positions 2 and 12, and sorted the peaks by the residual of this model (i.e. by the ChIP-seq signal that could not be explained by the reduced model). As **Fig. 2B** shows, even if we focus on the peaks with similar DNA sequence and accessibility, the residual of the reduced model still correlates negatively with CpG methylation at positions 2 and 12. In other words, peaks whose signal is smaller than what the reduced model predicts have higher CpG methylation, supporting the negative effect of CpG methylation on CTCF binding. Importantly, our observation that CpG methylation negatively affects CTCF binding is consistent with previous reports on CTCF methylation preferences *in vivo* [24] and *in vitro* [25]. Our results are also reproducible across different cell lines, as we obtained similar JAMS

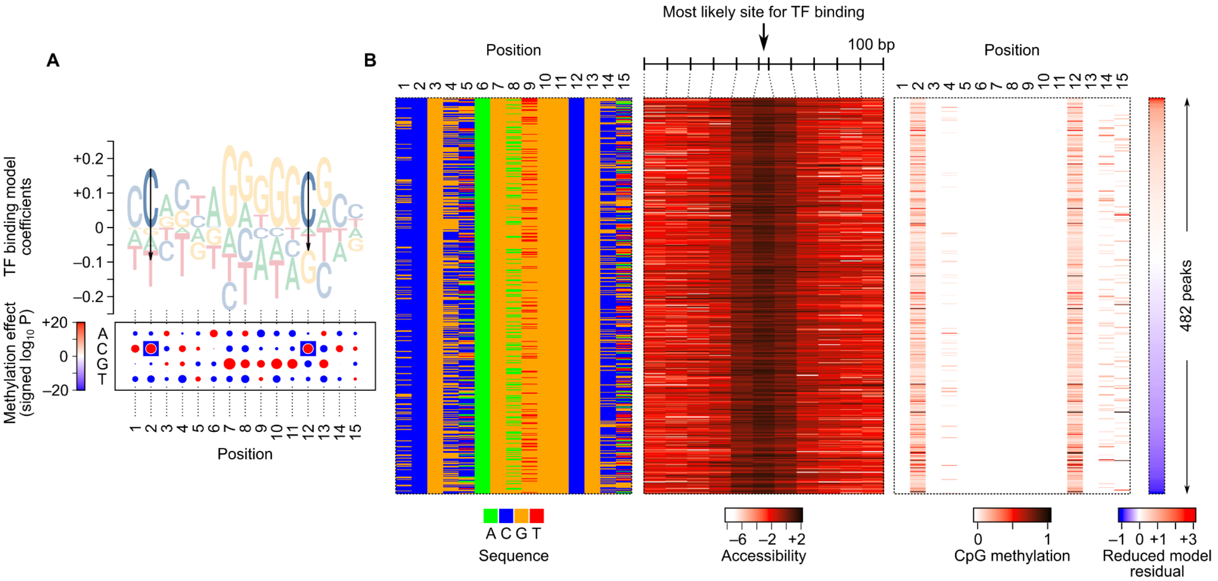

Figure 2. CpG methylation preference of CTCF in HEK293 cells. (**A**) Motif logo and dot plot representations of the sequence/methylation preference of CTCF. The logo (top) shows methylation coefficients as arrows, with the arrow length proportional to the mean estimate of methylation effect. The heatmap (bottom) shows the magnitude of the preference for each nucleotide at each position using the size of the dots, with red and blue representing positive and negative coefficients, respectively. The signed logarithm of P-value of the methylation coefficient is shown using the color of the squares around the dots, with red and blue corresponding to increased or decreased binding to methylated C, respectively (only significant methylation coefficients at FDR$<1\times10^{-5}$ are shown). (**B**) Heatmap representation of the sequence, accessibility, and CpG methylation, for a subset of CTCF peaks that have high DNA accessibility, a close sequence match to the initial CTCF motif, and CpGs at positions 2 and 12. Peaks are sorted by the residual of a reduced JAMS model that does not use the methylation level of C2 and C12 for predicting the CTCF binding signal.

models using CTCF ChIP-seq, WGBS, and accessibility data from several other cell lines (**Supplementary Fig. 3**). These results overall suggest that JAMS models have the potential to faithfully recapitulate the methylation preferences of TFs using ChIP-seq data.

Differential TF binding across cell lines can be explained using JAMS models

A model that encodes the intrinsic binding preference of a TF should be able to predict the ChIP-seq signal of that TF in new contexts, such as in previously unseen cell types that were not used in model training. We began to examine this possibility by investigating the transferability of the CTCF model that was learned in HEK293 cells to other cell types. We used DNase-seq and WGBS data (**Methods** and **Supplementary Table 1**) from six cell lines (H1, GM12878, HeLa-S3, HepG2, and K562) to predict the CTCF binding signal (using the HEK293-trained JAMS model), and compared the predictions to experimental CTCF ChIP-seq data obtained for each cell type. We observed that the CTCF JAMS model that was trained on HEK293 data could successfully predict the ChIP-seq pulldown-to-control ratio in other cell types, with a performance comparable to JAMS models that were specifically trained on the data from each type (**Table 1**). These results support the transferability of JAMS models across cell types.

The above analysis shows that the JAMS models learned from one cell type can be transferred to another cell type. However, the majority of CTCF binding sites are shared across different cell types; therefore, it is not immediately clear to what extent this transferability corresponds to cell-invariant features of the JAMS model (sequence) as opposed to potentially cell type-specific features (methylation and accessibility). In fact, one of the most challenging aspects of modeling TF binding is the ability to identify TF binding sites that are differentially occupied across cell types [26]. To understand the extent to which differential accessibility and methylation of DNA drives differential CTCF binding, and the extent to which these effects can be captured by JAMS, we decided to use the JAMS model learned from HEK293 cells to predict differential binding of CTCF in other cell lines. We started by identification of differentially bound CTCF peaks in pairwise comparisons of cell lines listed in **Table 1**. For any given two cell lines, we used the log-fold change (logFC) in the pulldown-to-control ratio as the measure of differential binding (**Fig. 3A**). The mean and standard error of mean (SEM) of this metric was calculated using a statistical model that assumes a negative binomial distribution for the tag counts, which also allows us to calculate a P-value for the null hypothesis that logFC is equal to zero (see **Methods**). Application of this method to all pairwise cell comparisons revealed the largest number of differentially bound CTCF peaks between GM12878 and HeLa-S3 cells (**Fig. 3B**); therefore, we focused on prediction of the differential peaks between these two cell lines using the HEK293 JAMS model of CTCF. Specifically, we used the JAMS model to predict the CTCF binding signal in each of the GM12878 and HeLa-S3 cell lines (based on the accessibility and methylation data of each cell line), and then calculated the difference of the JAMS predictions (in log-scale) between the two cells. As shown in **Fig. 3C**, the JAMS-predicted changes in CTCF binding are strongly correlated with the experimental logFC values (*r*=0.40 across peaks with logFC standard error of mean <1.28; see **Supplementary Fig. 4** for details on the choice of cutoff). These results suggest that the CTCF JAMS model can quantitatively predict the change in CTCF binding strength based on differential accessibility and methylation. Importantly, for the set of peaks that pass the statistical significance threshold for differential binding between the two cell lines (FDR<0.1), the correlation between JAMS predictions and experimental logFC reaches as high as 0.84 (**Fig. 3C**), with JAMS being able to distinguish GM12878-specific from HeLa-S3-specific binding events with 95% accuracy.

Table 1: Pearson correlation (*r*) between observed and predicted CTCF-binding across cell types. The third column shows *r* between observed and cross-validated JAMS predictions for models that were trained on each individual cell type. The fourth column shows the *r* between the predictions of the JAMS model that was trained on HEK293 and the observed ChIP-seq data in other cell lines.

Cell line	ChIP-seq peaks (*n*)	10-fold CV	HEK293-trained *r*
HEK293	135,717	0.69	-
H1	128,123	0.72	0.62
GM12878	39,535	0.69	0.54
HeLa-S3	65,865	0.72	0.60
HepG2	81,188	0.73	0.64
K562	85,122	0.74	0.68

We note that many of the CTCF binding sites are differentially accessible between GM12878 and HeLa-S3 (**Fig. 3D**), which may drive the differential binding. To specifically examine the role of differential methylation

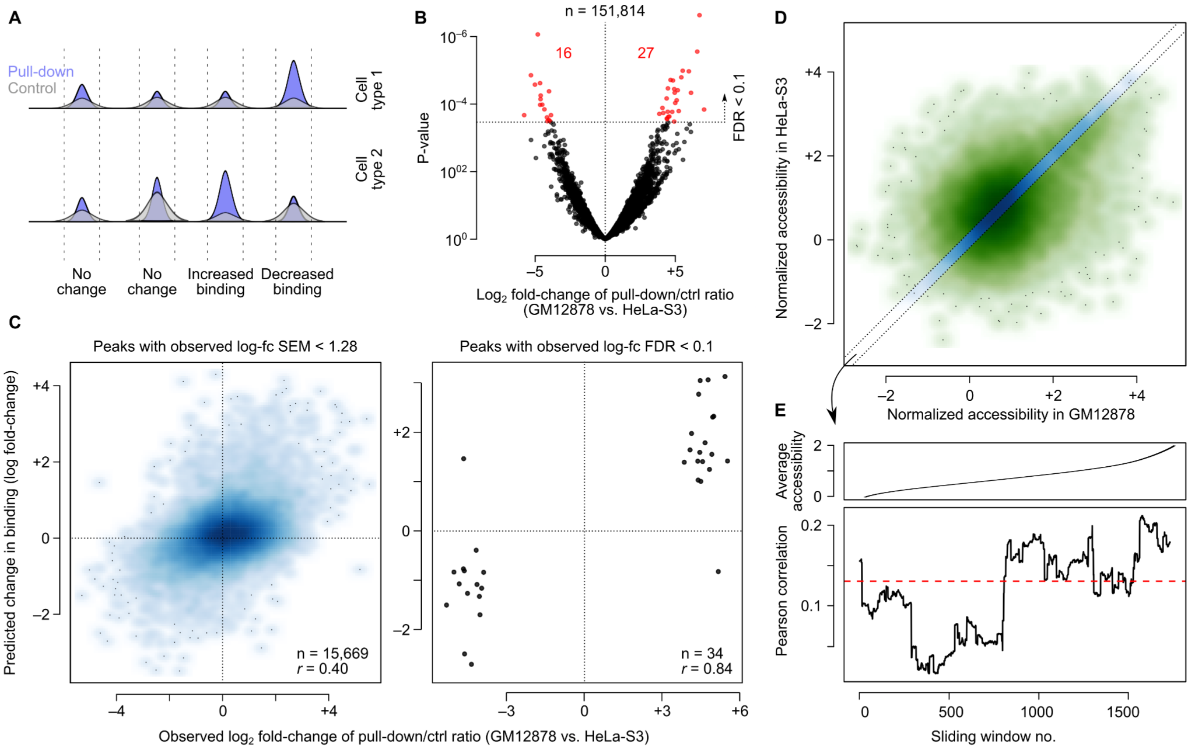

Figure 3. Prediction of differentially bound CTCF peaks using JAMS. (**A**) Schematic representation of identifying differentially bound peaks based on the combination of pulldown and control signal in two cell lines. See **Methods** for details. (**B**) Volcano plot showing differential binding of ChIP-seq peaks between GM12878 and HeLa-S3. Significant peaks at FDR < 0.1 are shown in red. (**C**) Left: Scatter plot of JAMS-predicted changes in CTCF binding and observed differential binding between GM12878 and HeLa-S3 cells. Peaks with observed logFC SEM <1.3 are included. Right: Limited to peaks that pass FDR<0.1 for differential binding of CTCF. (**D**) Comparison of the accessibility of putative CTCF peaks between two cell lines. The diagonal band in the middle (blue) shows the region that was selected as no-change in accessibility (difference in accessibility < 0.2). (**E**) Predicting differential CTCF binding for peaks with no change in accessibility. Peaks were ranked by accessibility, and the correlation between predicted and observed logFC of CTCF binding was calculated for sliding windows of 500 peaks (bottom). The average accessibility for each sliding window is shown on top.

in driving cell type-specific CTCF binding, we further limited our analysis to the set of peaks that had similar accessibility in both cell lines (**Fig. 3D**), and also removed all the JAMS predictor variables corresponding to accessibility. We observed that this reduced JAMS model can still predict differential CTCF binding among the peaks that are not differentially accessible (r=0.14 between predicted and observed logFC across n=2232 peaks; **Fig. 3E**). This correlation increases to 0.22 for the set of peaks that have high accessibility in both cell lines (**Fig. 3E**), suggesting that the effect of CpG methylation is most noticeable when the putative CTCF binding site is accessible.

Overall, these analyses suggest that JAMS models can predict differential TF binding across cell types, including differential TF binding events that are driven by changes in the methylation of the putative binding sites. The ability of JAMS models to predict cell type-specific TF binding events further highlight their reliability in capturing the determinants of TF binding using ChIP-seq data.

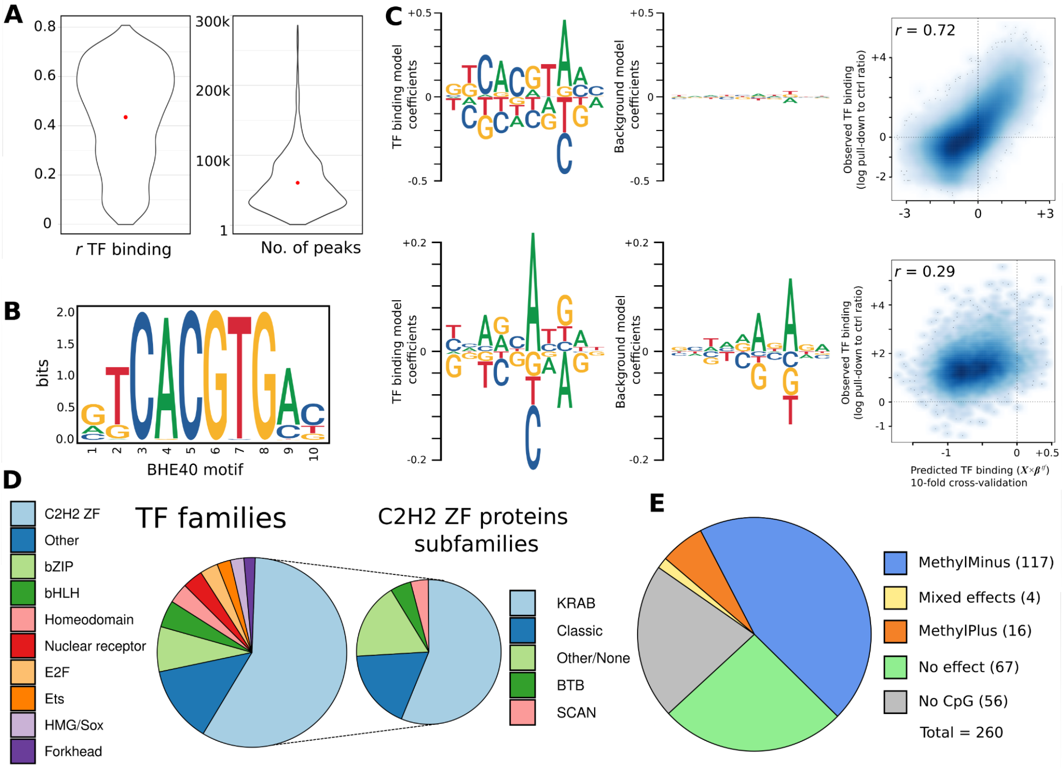

Figure 4. Systematic application of JAMS. (**A**) Left: Violin plot showing the distribution of Pearson correlation between the observed and predicted TF binding signal. Right: Distribution of the number of peaks used to create JAMS models. The violin plots represent a total of 2368 ChIP-seq experiments that were analyzed by JAMS. (**B**) Known BHE40 motif obtained from the CIS-BP database, shown as an example [45]. (**C**) Results from a high-quality (top) and a low-quality (bottom) JAMS model for BHE40. Inferred sequence coefficients for TF binding (left) and background (middle), as well as the predicted vs. observed TF binding signal (right) are shown. (**D**) Pie charts of the main TF families (left) and C2H2 ZF proteins subfamilies (right) for TFs with at least one high-quality JAMS model. (**E**) Pie chart of the methyl-binding preferences of TFs with at least one high quality JAMS model. We obtained high-quality models for a total of 260 TFs.

A high-confidence compendium of JAMS models for 260 TFs

To identify TFs whose *in vivo* binding is positively or negatively affected by methylation of intra-motif CpGs, we decided to apply JAMS to a comprehensive compendium of ChIP-seq data for a wide range of TFs. We collected and uniformly processed data from 2368 ChIP-seq and ChIP-exo experiments [20, 22, 27], covering the *in vivo* binding profiles of 421 TFs in six cell lines, along with the WGBS and DNase-seq assays in those cell lines. On average, we identified ~60k peaks per ChIP-seq experiment using the permissive P-value threshold of 0.01 (**Fig. 4A**). We then used the peak tag counts to fit a JAMS model to each ChIP-seq experiment. We noticed that the quality of the JAMS models, measured by the Pearson correlation between the predicted and observed TF-specific signal, varied substantially across the experiments, with correlations ranging from 0 to 0.8 (median 0.48, **Fig. 4A** and **Supplementary Fig. 5**). This variation may reflect a multitude of factors, including the ChIP-seq data quality as well as the extent to which the TF signal can be explained by our model specifications. We therefore decided to keep only a subset of high-confidence models. Specifically, we selected at most one representative model per TF based on the following criteria: (i) the model should have used at least 10,000 peaks for training, (ii) Pearson correlation >0.2 between the predicted and observed TF-specific signal after cross-validation, (iii) Pearson correlation >0.3 between the known and JAMS-inferred sequence motif, (iv) and low contribution of the sequence to the background signal compared to the TF-specific signal (control-to-pulldown ratio of the sequence coefficients mean < 0.4). As an example, in **Fig. 4C** we show two JAMS models for BHE40, obtained from two different ChIP-seq experiments, only one of which passes all the criteria mentioned above. Overall, we obtained high-confidence JAMS models for 260 TFs, spanning a range of TF families (**Fig. 4D**).

Systematic inference of the *in vivo* TF methyl-binding preferences

After selecting one JAMS model per TF, we used the JAMS-inferred effects of methylation to classify the TFs according to their inferred methyl-binding preferences. We use a notation similar to Yin et al. [5]. Specifically, we classified a TF as (a) methyl-minus if its JAMS model included at least one significantly negative mCpG effect ($FDR<1\times10^{-5}$), (b) methyl-plus if the model included at least one significantly positive mCpG effect, (c) mixed-effect if the model included both significantly positive and negative mCpG effects, (d) and no-effect if the motif included a CpG but its methylation level did not have a significant effect. Overall, we found 117 methyl-minus TFs, 16 methyl-plus TFs, four mixed-effect TFs, and 67 TFs with no significant mCpG effects; we also identified a set of 56 TFs without a CpG site in their binding site (**Fig. 4E**).

To understand whether our JAMS-based classification captures known methyl-binding preferences of TFs, we started by examining a few TFs whose methyl-binding preferences have been extensively studied *in vitro* and *in vivo*, including CEBPB and NRF1. Using protein-binding microarrays (PBMs), Mann et al. have previously reported enhanced binding of CEBPB to its CpG-containing target sequence when the array probes were methylated [2], consistent with the observation that a large fraction of the genomic binding sites of CEBPB is highly methylated *in vivo* [3]. The JAMS model for CEBPB (**Fig. 5A-D**) is concordant with these previous reports, showing that CpG methylation at the 6th position of CEBPB target sequence has a positive effect on its binding strength. This effect is in fact highly reproducible, and is present in three out of four JAMS models that we obtained using different CEBPB ChIP-seq experiments.

Another well studied TF is NRF1, which has been found to be sensitive to CpG methylation of DNaseI-hypersensitive sites in murine stem cells [10]. Moreover, Cusack et al. found that NRF1 preferentially binds to unmethylated DNA even after accounting for changes in DNA accessibility caused by the recruitment of HDACs to methylated CpGs through MBD proteins [9]. Consistent with these reports, we found that CpG methylation of the 3rd and 9th positions of the NFR1 target sequence has a negative effect on its binding (**Fig 5E-H**); these effects were consistent across all the cell lines we analyzed.

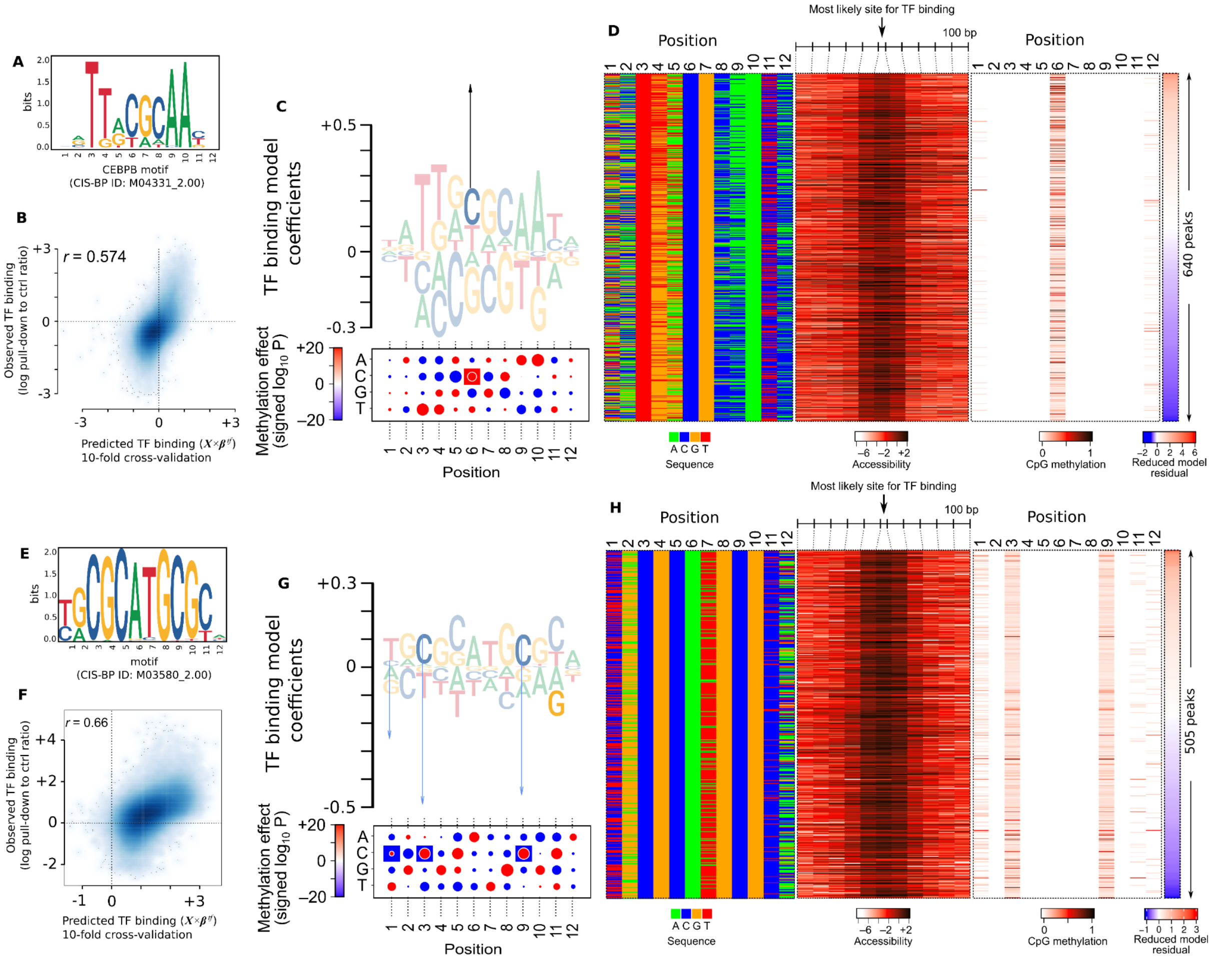

Figure 5. Examples of known TF methyl-binding preferences that were also captured by JAMS. Panels **A-D** correspond to CEBPB, a known methyl-plus TF. Panels **E-H** correspond to NRF1, a known TF whose binding is inhibited by methylation. (**A**) Known motif for CEBPB. (**B**) Scatter plot of JAMS-predicted vs. observed TF binding signal for CEBPB. (**C**) Motif logo and dot plot representations of the sequence/methylation preference of CEBPB as inferred by JAMS (see **Figure 5** for how these representations should be interpreted). (**D**) Heatmap representation of the sequence, accessibility, and CpG methylation, for a subset of CEBPB peaks that have high DNA accessibility. Peaks are sorted by the residual of a reduced JAMS model that does not use the methylation level for predicting the TF binding signal. (**E-H**) Similar to panels **A-D**, but for NFR1.

The above examples suggest that JAMS models are consistent with previously reported methylation preferences of TFs. However, there are only a handful of TFs whose methylation preferences have been validated *in vivo*. Therefore, to systematically evaluate our JAMS-based classification of TFs, we compared our inferred methyl-binding preferences with *in vitro* preferences obtained using methyl-SELEX and/or bisulfite-SELEX [5]. Overall, 76 out of the 260 TFs that we studied here have methyl/bisulfite-SELEX data (**Table 2**). These included 44 TFs that we classified as methyl-minus based on *in vivo* data; 29 of these TFs (~66%) were also identified as methyl-minus by SELEX, and another 7 TFs (16%) were identified as mixed-effect. This suggests that our approach has ~82% precision for identification of TFs that are negatively affected by CpG methylation in at least one position in their target sequence. On the other hand, out of 39 methyl-minus TFs found by SELEX, 31 were also classified as either methyl-minus or mixed-effect by JAMS, suggesting that ~79% of *in vitro*-observed methyl-minus effects can be captured using *in vivo* data.

Table 2: Contingency table of TF classifications by JAMS (rows) and methyl/bisulfite-SELEX [5] (columns).

		SELEX call				
		Methyl-minus	Methyl-plus	Mixed-effect	Little effect	Novel
JAMS calls	Methyl-minus	29	4	7	4	73
	Methyl-plus	1	4	0	0	11
	Mixed-effect	2	1	0	0	1
	No Effect	7	11	4	2	43

Similarly, out of five JAMS-based methyl-plus TFs that have bisulfite-SELEX data [5], four were classified as methyl-plus based on SELEX (**Table 2**), suggesting a precision of ~80%. However, despite this high precision, only 5 out of 20 SELEX-based methyl-plus TFs are identified as either methyl-plus or mixed-effect by JAMS—this suggests that a relatively small fraction of *in vitro* methyl-plus effects can also be observed *in vivo*. Nonetheless, we found 11 methyl-plus TFs that were previously unclassified—this is in addition to 73 previously unclassified methyl-minus and one novel mixed-effect TF, highlighting the ability of JAMS models in revealing novel TF methyl preferences.

Fig. 6 shows the distribution of different methyl-preferences across main TF families. We noticed that a disproportionately large number of methyl-plus TFs belong to the KRAB domain-containing members of the C2H2-ZF family (also shown in **Table 3** and **Supplementary Fig. 6**). Specifically, among KRAB-ZF TFs whose binding is significantly affected by methylation, ~24% preferentially bind to methylated CpGs, compared to only ~12% of non-KRAB TFs (Fisher's exact test P<0.009, **Supplementary Table 2**). This is an intriguing observation, given that a majority of KRAB-ZF proteins evolved to specifically bind and repress transposable elements, which largely reside in highly methylated genomic regions[28]. It is notable that we observed this methyl-plus effect even though we removed all repetitive genomic regions from our analysis (see **Methods**). Our observation suggests that many of the KRAB-ZF proteins preferentially bind to methylated instances of their target sequence, potentially allowing them to distinguish the transposable elements from other genomic regions that contain their preferred binding sequence. In fact, ~56% of all methyl-plus TFs that we identified are KRAB-ZF proteins, suggesting that recognition of methylated transposable elements might have been a primary force in the evolution of methyl-binding TFs.

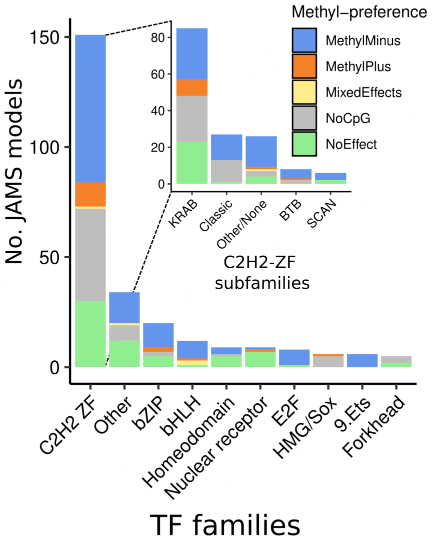

Figure 6. Methylation preferences per TF family. Stacked bar plots showing the distribution of TF methylation preferences inferred with JAMS, grouped by TF families. The inset shows the distribution of methylation preferences for C2H2-ZFP subfamilies.

Overall, our results demonstrate that the methylation preferences of TFs can be reliably inferred from their *in vivo* binding profiles, and provide a comprehensive resource for classification of TF methyl-preferences.

Table 3. TFs with methyl-plus and mixed-effect methyl-binding preferences, as inferred by JAMS using in vivo data. For mixed-effect TFs, both the position at which a positive methylation effect was observed as well as the position with a negative methylation effect are indicated.

			Effect of methylation by position		
Protein	**Family**	**JAMS call**	**Positive**	**Negative**	**SELEX call** [5]
ZNF793	C2H2 ZF (KRAB)	Methyl-plus	7		
ZKSCAN1	C2H2 ZF (KRAB+SCAN)	Methyl-plus	2		
CEBPB	bZIP	Methyl-plus	6		Methyl-plus
ZNF141	C2H2 ZF (KRAB)	Methyl-plus	17		
ZNF320	C2H2 ZF (KRAB)	Methyl-plus	17		
ZNF605	C2H2 ZF (KRAB)	Methyl-plus	15		
NR2F2	Nuclear receptor	Methyl-plus	5, 8		
ZNF479	C2H2 ZF (KRAB)	Methyl-plus	11		
SP1	C2H2 ZF	Mixed-effect	5	8	Methyl-plus
ZNF490	C2H2 ZF (KRAB)	Methyl-plus	7		
ZNF506	C2H2 ZF (KRAB)	Methyl-plus	5		
ZNF417	C2H2 ZF (KRAB)	Methyl-plus	16		
USF1	bHLH	Mixed-effect	7	5	Methyl-minus
USF2	bHLH	Mixed-effect	7	5	Methyl-minus
TCF7	HMG/Sox	Methyl-plus	2		Methyl-minus
ZBTB33 (KAISO)	C2H2 ZF (BTB)	Methyl-plus	5, 7		Methyl-plus
TFAP4	bHLH	Methyl-plus	7		
NFYB	NFYB/HAP3	Mixed-effect	9	13	
SCRT1	C2H2 ZF	Methyl-plus	3		Methyl-plus
CEBPG	bZIP	Methyl-plus	6		Methyl-plus

DISCUSSION

In this study, we built Joint Accessibility-Methylation-Sequence (JAMS) models to capture the relationship between TF binding and DNA methylation *in vivo*. Our approach models the TF binding as a function of DNA accessibility, sequence and methylation at and around TF binding sites, while separating the background from TF-specific signals.

We started by applying this method to CTCF, which revealed that CpG methylation at the 2nd and 12th positions of the CTCF motif is associated with decreased TF binding. This methylation sensitivity is reproduced in multiple cell lines, can be observed even among highly accessible genomic regions, and can explain differential CTCF binding between different cell lines. As mentioned in the Results section, methylation-sensitivity of CTCF has been previously reported [24]. An intriguing observation in this regard was made by Zuo et al., who used a high-throughput *in vitro* method to quantify the effect of CpG methylation on CTCF binding: they found a substantial negative effect of the CpG methylation at the 2nd position of the motif [25], which is also one of the CpG sites we identified. However, we also identified a second CpG site at the 12th position whose methylation reduces CTCF binding, which was not reported by Zuo et al. [25]. Using a likelihood ratio test we showed that the observed effect of methylation at this position cannot be simply explained by its correlation with the first CpG site (**Supplementary Fig. 2**), suggesting that we may have identified a novel CpG methylation effect.

One possible explanation as to why the methylation effect at position 12 could not be observed *in vitro* is that it may reflect the direct competition between CTCF and MBD proteins, which are not included in the *in vitro* assay. While JAMS is able to capture the effect of changes in DNA accessibility that result from chromatin remodelling factors recruited by MBD proteins, it currently does not model the direct competition of TFs and MBD proteins. This undetected direct competition between MBD proteins and TFs for the binding sites could affect the interpretation of our model parameters: methylation coefficients obtained by JAMS models should be more accurately interpreted as the affinity of a TF toward mCpG sites relative to the affinity of MDB proteins.

Accordingly, a positive methylation coefficient means that the TF binds more strongly to the mCpG than MDB proteins do, therefore outcompeting them. This interpretation may in fact explain why a large number of *in vitro*-detected methyl-plus TFs [5] could not be identified by JAMS: even though these TFs can bind to mCpGs *in vitro*, competition with MDB proteins might attenuate this effect *in vivo*. On the other hand, a negative JAMS methylation coefficient could mean that the MDB proteins outcompete the TF *in vivo*, or that the TF simply does not bind to mCpGs even without considering the effect of MBDs. Since the majority of methyl-mins TFs that we identified match *in vitro* observations [5], the latter scenario is likely more prevalent, with most negative coefficients reflecting the intrinsic preference of the TF for unmethylated CpGs even without considering competition with MBDs. We would like to note the possibility of directly deconvolving these scenarios (i.e. intrinsic preference for unmethylated CpGs vs. competition with MBDs) by including the MBD protein occupancy profiles as additional variables in JAMS models.

One potential limitation of inference of methyl-preferences of TFs from *in vivo* data is that it is difficult to establish the direction of causality: while it is likely that the observed associations reflect the effect of methylation on TF binding, it could also be that they reflect the effect of TF binding on the DNA methylation level [29]**.** However, TFs that influence DNA methylation most often have an effect on the local neighborhood of their binding sites, which can span tens of nucleotides [30-32]. JAMS takes into account the neighboring methylation levels, and tries to identify the site-specific methylation effects within the motif that cannot be explained by (or are independent of) the flanking methylation levels. We note that, when available, the majority of our methyl-minus and methyl-plus findings (>80%) match the results of *in vitro* experiments, in which binding site methylation levels are established before introducing the TF into the system [5, 25, 33]. Therefore, it is more likely that we are observing the effect that CpG methylation has on TF binding, rather than the effect of TF on CpG methylation.

Our results represent, to our knowledge, the largest resource for exploring the *in vivo* effect of methylation on TF binding: only a handful of studies have previously investigated methylation preferences of a limited number of TFs *in vivo* while accounting for changes in DNA accessibility. Our results match what has been reported in these studies [2, 4, 9, 10, 16], but also reveals a substantial number of novel TFs that are affected by CpG methylation. Of particular interest, our study revealed a significant number of methyl-plus TFs consistent with *in vitro* studies [5], in stark contrast to previous methods that have attempted to infer the effect of DNA methylation on TF binding [12-15]. Notably, a large proportion of methyl-plus TFs belonged to the C2H2-ZF family. C2H2-ZF proteins recognize DNA with an array of zinc fingers (ZFs) [34], with each ZF recognizing three or four nucleotides using a specific set of base-contacting residues [35]. Identifying the methylation preferences of C2H2-ZF proteins opens the possibility of associating the identity of base-contacting residues to mCpG binding: with a sufficiently large number of methyl-binding ZFs, we could potentially identify an "mCpG recognition code" for these TFs.

METHODS

Methods overview

To understand the relationship between DNA methylation and TF binding, we began by retrieving and analyzing WGBS, ChIP-seq, and DNase-seq data from different TFs in several cell lines. We developed a method to jointly model these data sets to predict TF-specific binding, and benchmarked it on CTCF ChIP-seq data in HEK293 cells. We expanded our CTCF studies by obtaining differential binding sites of CTCF between different cell lines, and examined whether, using our method, we can predict differential binding that was caused by DNA methylation changes. Finally, we applied our method to a comprehensive collection of ChIP-seq data to systematically study the *in vivo* effect of DNA methylation on TF binding.

ChIP-seq data processing, peak calling, and peak signal quantification

We limited our analysis to ChIP-seq experiments performed in HepG2, K562, HEK293, GM12878 and HeLa-S3 cell lines, given the availability of WGBS and DNase-seq data for these cell lines. ChIP-seq and ChIP-exo raw reads were retrieved from four main sources: ENCODE [22, 36], Najafabadi et al. [37], Schmitges et al. [20], and Imbeault et al. [27]. ENCODE data were downloaded from ENCODE project website (https://www.encodeproject.org/experiments/), while the other data were downloaded from GEO (accession numbers GSE58341, GSE76494, and GSE78099). A total of 2677 ChIP-seq experiments were analyzed, covering 421 TFs and 5 cell lines.

Raw reads were aligned to the human reference genome (GRCh38) with *bowtie2* (version 2.3.4.1) using the *"--very-sensitive-local*" mode. Mapped reads with mapping quality score smaller than 30 were removed using *samtools* (version 1.9)[38]. ChIP-seq peaks were called using *MACS* (version 1.4) [39, 40] with a permissive p-value threshold of 0.01. We used this permissive p-value to obtain a range of TF binding signals, which our method uses to quantitatively model TF binding strength. We also included negative peaks, i.e. peaks obtained by swapping the treatment with the control experiments, to enable proper modeling of the background signal. In the end, for each ChIP-seq experiment, this process resulted in a list of peaks covering a wide range of pulldown or control (background) signal strengths, along with their associated read counts.

WGBS data processing and DNase-seq data retrieval

Raw reads from Whole-Genome Bisulfite Sequencing (WGBS) of six cell lines were retrieved from ENCODE and GEO (see **Supplementary Table 1** for accession numbers). Raw reads were trimmed based on their quality (phred33 ≥ 20) with *TrimGalore* (version 0.6.4) [41]. Paired reads were aligned to the human reference genome hg38 [42] using *bismark* (*bowtie2* mode, version 0.22.2), allowing one mismatch during alignment. Reads were deduplicated by removing those that aligned to the same genomic position (*bismark:deduplicate_bismark*). Methylation calls were then extracted, ignoring the first 2 bps from the 5' end of read 2 (*bismark:bismark_methylation_extractor*). A genome-wide coverage report with methylated and unmethylated read counts was then generated (*bismark:coverage2cytosine*). Finally, a bigwig file was generated for unmethylated and methylated counts (*bedGraphToBigWig*)[43].

For DNase-seq data, read depth-normalized bigwig files representing DNase-seq signal were retrieved from ENCODE (see **Supplementary Table 1** for accession numbers).

Formatting and preprocessing of data for JAMS

To retrieve the sequence, DNA accessibility, and DNA methylation to train our model, we focused on the positive and negative ChIP-seq peak regions that did not fall within endogenous repeat elements, since the homology of repeat elements can confound the modeling of ChIP-seq data based on sequence [37]. This was done by removing peaks that overlapped any repeat regions, as defined by RepeatMasker [42, 44].

To model the effect of sequence and epigenetic factors on TF binding using our method, it is necessary to align the peaks based on the position of the most likely TF binding site. To do so, we used the known motif of each TF, in the form of position frequency matrices (PFMs), to search for the most likely TFBS within the 100 bp range of the peak summit. PFMs were obtained from CIS-BP [45], and were augmented by *de novo* motifs identified by RCADE2 [46, 47] for the C2H2-ZF family of TFs as described in later sections. CISP-BP contains more than one PFMs per TF, as they are derived from different experimental techniques. We selected PFMs exclusively derived from *in vitro* experiments, in order to avoid the confounding effects present *in vivo*. We prioritized, in descending order, PFMs from SELEX, Selective microfluidics-based ligand enrichment followed by sequencing (SMiLE-seq), and Protein-Binding Microarrays (PBM). We used *AffiMx* [48] to identify the best motif match in each peak sequence. This process was uniformly applied to all peaks, including the negative ChIP-seq peak set.

Once the best motif hit in each peak was identified, we extracted the sequence and nucleotide-resolution methylation profile at the motif hit as well as the flanking regions (20 bp) around the motif hit. Sequences were retrieved from the reference genome hg38 using *bedtools:getfasta* [42, 49]. Methylated and unmethylated read counts at each position were retrieved from the WGBS bigwig files using *bwtool* [50].

Similarly, normalized DNA accessibility was extracted from the motif hit region and 500 bp upstream and downstream of the motif hit from the DNase-seq bigwig files. ChIP-seq read counts were extracted from the control and pull-down experiments for the +/- 400bp region surrounding the motif match using *bedtools:multicov* (MAPQ score > 30). (**Fig. 4C, bottom**) [49].

We emphasize that while a known motif of each TF was used to identify an offset for each peak and align the peak regions, this process is not expected to confound the sequence features learned by JAMS, since it is uniformly applied to all peaks regardless of the signal strength. The TF motifs themselves were also not used by

JAMS, and the sequence features that are predictive of ChIP-seq signal were learned *de novo* from the aligned peaks.

Implementation of JAMS

Our method creates a joint accessibility-methylation-sequence model (JAMS model) for each ChIP-seq experiment, in which the ChIP-seq signal of each peak is explained as a function of accessibility, methylation, and sequence at that peak. Consider the $k \times m$ matrix $\boldsymbol{X}$, which represents the value of m predictive features at k genomic positions (i.e. peaks). These m features include those related to accessibility (A), methylation (M), and sequence (S):

$$\boldsymbol{X} = [\boldsymbol{X}_A \boldsymbol{X}_M \boldsymbol{X}_S]$$

JAMS models the logarithm of TF binding strength at each of the k peaks as a linear function of the matrix $\boldsymbol{X}$:

$$\log \boldsymbol{\mu}_f = \boldsymbol{X} \times \boldsymbol{\beta}_f$$

Here, $\boldsymbol{\mu}_f$ is the vector of the binding strength for transcription factor f across k peaks, $\boldsymbol{X}$ is the $k \times m$ feature matrix described above, and $\boldsymbol{\beta}_f$ is the vector of m coefficients that describe the effect of each of the m features on the TF binding strength (matrices are denoted with bold capital letters, and vectors with bold lower-case letters).

Similarly, the background ChIP-seq signal across the peaks is also modeled as a function of $\boldsymbol{X}$:

$$\log \boldsymbol{\mu}_b = \boldsymbol{X} \times \boldsymbol{\beta}_b$$

Here, $\boldsymbol{\mu}_b$ represents the background signal strength across k peaks, and $\boldsymbol{\beta}_b$ is the vector of m coefficients that describe the effect of each of the m features on the background signal.

In a ChIP-seq experiment, the expected control (background) read counts at each peak is a function of the background signal multiplied by the library size. Therefore, the logarithm of control reads can be modeled as:

$$\log \boldsymbol{\lambda}_c = \log \boldsymbol{\mu}_b + s_c = \boldsymbol{X} \times \boldsymbol{\beta}_b + s_c$$

Here, $\boldsymbol{\lambda}_c$ is the vector of expected (average) control read counts across the k peaks, and s_c is an experiment-specific size factor that can be interpreted as the logarithm of sequencing depth for the control library.

The expected pull-down read counts in a ChIP-seq experiment, however, are a function of both the background signal and the TF binding strength, multiplied by the library size. Therefore:

$$\log \boldsymbol{\lambda}_p = \log \boldsymbol{\mu}_b + \log \boldsymbol{\mu}_f + s_p = \boldsymbol{X} \times \boldsymbol{\beta}_b + \boldsymbol{X} \times \boldsymbol{\beta}_f + s_p$$

Here, $\boldsymbol{\lambda}_p$ is the vector of expected pulldown read counts across the k peaks, and s_p can be interpreted as the logarithm of sequencing depth for the pulldown library.

While these equations describe the expected control and pulldown read counts, the actual observed read counts are probabilistic observations that may deviate from these expected values. Here, we model the read counts as observations from negative binomial distributions [51] whose mean is given by the equations above, with a shared dispersion parameter across the peaks:

$$\boldsymbol{n}_c = NB(\boldsymbol{\lambda}_c, \varphi)$$

$$\boldsymbol{n}_p = NB\left(\boldsymbol{\lambda}_p, \varphi\right)$$

Here, $\boldsymbol{n}_c$ and $\boldsymbol{n}_p$ are the vectors of observed control and pulldown read counts across the *k* peaks, respectively, and φ is the dispersion parameter. The equations above allow us to jointly model the control and pulldown experiments as a function of $\boldsymbol{X}$. We use the glm.nb function in R for this purpose and fit a model of the form *n~XX+t+XX:t*, where *n* is an R vector that concatenates the observed control and pulldown read counts (with length 2*k*), *XX* is the result of duplicating matrix $\boldsymbol{X}$, i.e. *XX=rbind(X,X)*, and *t* is a binary vector of length 2*k* indicating whether the observed read count comes from the control experiment (0) or from the pulldown experiment (1). The coefficients returned by the glm.nb function for *XX* correspond to $\boldsymbol{\beta}_b$ in the equations above, and the coefficients for *XX:t* correspond to $\boldsymbol{\beta}_f$. The glm.nb also returns the standard error of mean and a p-value for each of these coefficients, which we use to determine the statistical significance.

Constructing the matrix $\boldsymbol{X}$: Sequence, DNA methylation and DNA accessibility are used as the predictor variables, which are included in the matrix $\boldsymbol{X}$. We used one-hot encoding for the sequence over the TFBS. Methylated and unmethylated read counts over the motif were used to calculate the methylation percentage at each position. If the average coverage of methylation and unmethylated reads over the motif is less than 10 counts, the peak is removed. Average DNA accessibility was calculated for bins of 100 bp (10 bins) plus one bin for the TFBS region itself, and then logarithm of DNA accessibility was calculated; a pseudocount equivalent of 1% of the smallest value was used to allow for log transformation of the data. Average methylation percentage and sequence composition of the flanking regions were also used as predictors.

JAMS is available at https://github.com/csglab/JAMS.

Differential binding analysis

To calculate differential TF binding between cell lines, we first identified CTCF ChIP-seq experiments from ENCODE that had at least two biological replicates per cell line (**Supplementary Table 1**), and retrieved the pull-down and control experiment data. After aligning and peak calling, we defined a unified list of peaks that were present in at least one sample. Peaks that were present in more than one sample and had summits within 100 bp of each other were merged, as they likely represent the same CTCF binding site. Then, the best motif match within 100 bp of each summit was identified [48]. We extracted ChIP-seq read counts present within a 400bp range from the motif hit in the pull-down and control experiments and created a count matrix.

We used DESeq2 [52] to compare the pulldown-to-control ratio between pairs of cell lines. The DESeqDataSetFromMatrix function from DESeq2 was used to create a DESeqDataSet object, followed by fitting a model of the form $\sim s+c{:}t$, where s is a categorical variable representing the sample/replicate (shared between pairs of control and pulldown experiments), c is a binary variable representing the two different cell lines, and t is a binary variable denoting whether the read count corresponds to the control experiment (0) or the pulldown experiment (1). After fitting the DESeq2 model, the coefficient for $c{:}t$ corresponds to the log2 fold changes. Significant differentially bound peaks ($FDR < 0.1$) were identified for every pair of cell lines, excluding cell line pairs whose ChIP-seq experiments were done in different laboratories. The pair of cell lines (GM12878 and HeLa-S3) with the highest number of significantly bound peaks were selected for further analysis.

Inference of PFMs for C2H2-ZF proteins using RCADE2

We inferred position frequency matrices (PFMs) for canonical C2H2 zinc finger proteins using RCADE2 [46, 47]. RCADE2 uses the protein sequence, the DNA sequence of the ChIP-seq peaks, and a previously computed machine learning-based recognition code to predict the DNA-binding preferences of C2H2-ZFPs. The protein sequences for these TFs were retrieved from UniProt [53]. We focused on the top 500 ChIP-seq peaks (sorted by p-value) that did not fall within endogenous repeat elements (EREs) [42, 44]. The DNA sequence of the +/- 250 region around the peak summits for the top 500 non-ERE peaks along with the protein sequence was provided as input to RCADE2, and the optimized motif was used to augment the CIS-BP motifs.

ACKNOWLEDGEMENTS

We thank Senthilkumar Kailasam for assisting with analysis of WGBS data. This work was supported by funds from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2018-05962) and resource allocations from Compute Canada to HSN. HSN holds a Canada Research Chair funded by the Canadian Institutes of Health Research.

REFERENCES

1. Watt F, Molloy PL: Cytosine methylation prevents binding to DNA of a HeLa cell transcription factor required for optimal expression of the adenovirus major late promoter. *Genes Dev* 1988, 2:1136-1143.

2. Mann IK, Chatterjee R, Zhao J, He X, Weirauch MT, Hughes TR, Vinson C: CG methylated microarrays identify a novel methylated sequence bound by the CEBPB|ATF4 heterodimer that is active in vivo. *Genome Res* 2013, 23:988-997.

3. Zhu H, Wang G, Qian J: Transcription factors as readers and effectors of DNA methylation. *Nat Rev Genet* 2016, 17:551-565.

4. Lin QXX, Rebbani K, Jha S, Benoukraf T: ZBTB33 (Kaiso) methylated binding sites are associated with primed heterochromatin. *bioRxiv* 2019:585653.

5. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, Das PK, Kivioja T, Dave K, Zhong F, et al: **Impact of cytosine methylation on DNA binding specificities of human transcription factors.** *Science* 2017, **356**.

6. Du Q, Luu PL, Stirzaker C, Clark SJ: **Methyl-CpG-binding domain proteins: readers of the epigenome.** *Epigenomics* 2015, **7:**1051-1073.

7. John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA: **Chromatin accessibility pre-determines glucocorticoid receptor binding patterns.** *Nat Genet* 2011, **43:**264-268.

8. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al: **The accessible chromatin landscape of the human genome.** *Nature* 2012, **489:**75-82.

9. Cusack M, King HW, Spingardi P, Kessler BM, Klose RJ, Kriaucionis S: Distinct contributions of DNA methylation and histone acetylation to the genomic occupancy of transcription factors. *Genome Res* 2020, 30:1393-1406.

10. Domcke S, Bardet AF, Adrian Ginno P, Hartl D, Burger L, Schubeler D: **Competition between DNA methylation and transcription factors determines binding of NRF1.** *Nature* 2015, **528:**575-579.

11. Wan J, Su Y, Song Q, Tung B, Oyinlade O, Liu S, Ying M, Ming GL, Song H, Qian J, et al: **Methylated cis-regulatory elements mediate KLF4-dependent gene transactivation and cell migration.** *Elife* 2017, **6**.

12. Grau J, Schmidt F, Schulz MH: Widespread effects of DNA methylation and intra-motif dependencies revealed by novel transcription factor binding models. *bioRxiv* 2020:2020.2010.2021.348193.

13. Xu T, Li B, Zhao M, Szulwach KE, Street RC, Lin L, Yao B, Zhang F, Jin P, Wu H, Qin ZS: **Base-resolution methylation patterns accurately predict transcription factor bindings in vivo.** *Nucleic Acids Res* 2015, **43:**2757-2766.

14. Ngo V, Wang M, Wang W: **Finding de novo methylated DNA motifs.** *Bioinformatics* 2019, **35:**3287-3293.

15. Viner C, Johnson J, Walker N, Shi H, Sjöberg M, Adams DJ, Ferguson-Smith AC, Bailey TL, Hoffman MM: **Modeling methyl-sensitive transcription factor motifs with an expanded epigenetic alphabet.** *bioRxiv* 2016**:**043794.

16. Prokhortchouk A, Hendrich B, Jorgensen H, Ruzov A, Wilm M, Georgiev G, Bird A, Prokhortchouk E: **The p120 catenin partner Kaiso is a DNA methylation-dependent transcriptional repressor.** *Genes Dev* 2001, **15:**1613-1618.

17. Filippova GN, Fagerlie S, Klenova EM, Myers C, Dehner Y, Goodwin G, Neiman PE, Collins SJ, Lobanenkov VV: An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol Cell Biol* 1996, 16:2802-2813.

18. Holwerda SJ, de Laat W: CTCF: the protein, the binding partners, the binding sites and their chromatin loops. *Philos Trans R Soc Lond B Biol Sci* 2013, 368:20120369.

19. Hansen AS, Pustova I, Cattoglio C, Tjian R, Darzacq X: CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *Elife* 2017, 6.

20. Schmitges FW, Radovani E, Najafabadi HS, Barazandeh M, Campitelli LF, Yin Y, Jolma A, Zhong G, Guo H, Kanagalingam T, et al: **Multiparameter functional diversity of human C2H2 zinc finger proteins.** *Genome Res* 2016, **26:**1742-1752.

21. Lakisic G, Lebreton A, Pourpre R, Wendling O, Libertini E, Radford EJ, Le Guillou M, Champy MF, Wattenhofer-Donze M, Soubigou G, et al: **Role of the BAHD1 Chromatin-Repressive Complex in Placental Development and Regulation of Steroid Metabolism.** *PLoS Genet* 2016, **12:**e1005898.

22. Consortium EP: An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012, 489:57-74.

23. Auerbach RK, Euskirchen G, Rozowsky J, Lamarre-Vincent N, Moqtaderi Z, Lefrancois P, Struhl K, Gerstein M, Snyder M: **Mapping accessible chromatin regions using Sono-Seq.** *Proc Natl Acad Sci U S A* 2009, **106:**14926-14931.

24. Maurano MT, Wang H, John S, Shafer A, Canfield T, Lee K, Stamatoyannopoulos JA: **Role of DNA Methylation in Modulating Transcription Factor Occupancy.** *Cell Rep* 2015, **12:**1184-1195.

25. Zuo Z, Roy B, Chang YK, Granas D, Stormo GD: Measuring quantitative effects of methylation on transcription factor-DNA binding affinity. *Sci Adv* 2017, 3:eaao1799.

26. Keilwagen J, Posch S, Grau J: Accurate prediction of cell type-specific transcription factor binding. *Genome Biol* 2019, 20:9.

27. Imbeault M, Helleboid PY, Trono D: KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* 2017, 543:550-554.

28. Thomas JH, Schneider S: Coevolution of retroelements and tandem zinc finger genes. *Genome Res* 2011, 21:1800-1812.

29. Ambrosi C, Manzo M, Baubec T: **Dynamics and Context-Dependent Roles of DNA Methylation.** *J Mol Biol* 2017, **429:**1459-1475.

30. de la Rica L, Rodriguez-Ubreva J, Garcia M, Islam AB, Urquiza JM, Hernando H, Christensen J, Helin K, Gomez-Vaquero C, Ballestar E: **PU.1 target genes undergo Tet2-coupled demethylation and DNMT3b-mediated methylation in monocyte-to-osteoclast differentiation.** *Genome Biol* 2013, **14:**R99.

31. Guilhamon P, Eskandarpour M, Halai D, Wilson GA, Feber A, Teschendorff AE, Gomez V, Hergovich A, Tirabosco R, Fernanda Amary M, et al: **Meta-analysis of IDH-mutant cancers identifies EBF1 as an interaction partner for TET2.** *Nat Commun* 2013, **4:**2166.

32. Suzuki T, Shimizu Y, Furuhata E, Maeda S, Kishima M, Nishimura H, Enomoto S, Hayashizaki Y, Suzuki H: **RUNX1 regulates site specificity of DNA demethylation by recruitment of DNA demethylation machineries in hematopoietic cells.** *Blood Adv* 2017, **1:**1699-1711.

33. Kribelbauer JF, Laptenko O, Chen S, Martini GD, Freed-Pastor WA, Prives C, Mann RS, Bussemaker HJ: **Quantitative Analysis of the DNA Methylation Sensitivity of Transcription Factor Complexes.** *Cell Rep* 2017, **19:**2383-2395.

34. Garton M, Najafabadi HS, Schmitges FW, Radovani E, Hughes TR, Kim PM: **A structural approach reveals how neighbouring C2H2 zinc fingers influence DNA binding specificity.** *Nucleic Acids Res* 2015, **43:**9147-9157.

35. Pavletich NP, Pabo CO: Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. *Science* 1991, 252:809-817.

36. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, et al: **The Encyclopedia of DNA elements (ENCODE): data portal update.** *Nucleic Acids Res* 2018, **46:**D794-D801.

37. Najafabadi HS, Albu M, Hughes TR: Identification of C2H2-ZF binding preferences from ChIP-seq data using RCADE. *Bioinformatics* 2015, 31:2879-2881.

38. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H: **Twelve years of SAMtools and BCFtools.** *Gigascience* 2021, **10**.

39. Feng J, Liu T, Qin B, Zhang Y, Liu XS: **Identifying ChIP-seq enrichment using MACS.** *Nat Protoc* 2012, **7:**1728-1740.

40. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS: **Model-based analysis of ChIP-Seq (MACS).** *Genome Biol* 2008, **9:**R137.

41. Martin M: Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal* 2011, 17:10-12.

42. Navarro Gonzalez J, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, Powell CC, Nassar LR, Maulding ND, Lee CM, et al: **The UCSC Genome Browser database: 2021 update.** *Nucleic Acids Res* 2021, **49:**D1046-D1057.

43. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D: **BigWig and BigBed: enabling browsing of large distributed datasets.** *Bioinformatics* 2010, **26:**2204-2207.

44. **RepeatModeler** [http://www.repeatmasker.org/]

45. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al: **Determination and inference of eukaryotic transcription factor sequence specificity.** *Cell* 2014, **158:**1431-1443.

46. Dogan B, Kailasam S, Corchado AH, Nikpoor N, Najafabadi HS: A domain-resolution map of <em>in vivo</em> DNA binding reveals the regulatory consequences of somatic mutations in zinc finger transcription factors. *bioRxiv* 2020:630756.

47. Dogan B, Najafabadi HS: Computational Methods for Analysis of the DNA-Binding Preferences of Cys2His2 Zinc-Finger Proteins. *Methods Mol Biol* 2018, 1867:15-28.

48. Lambert SA, Albu M, Hughes TR, Najafabadi HS: **Motif comparison based on similarity of binding affinity profiles.** *Bioinformatics* 2016, **32:**3504-3506.

49. Quinlan AR: **BEDTools: The Swiss-Army Tool for Genome Feature Analysis.** *Curr Protoc Bioinformatics* 2014, **47:**11 12 11-34.

50. Pohl A, Beato M: **bwtool: a tool for bigWig files.** *Bioinformatics* 2014, **30:**1618-1619.

51. Venables WN, Ripley BD: *Modern applied statistics with S-PLUS.* Springer Science & Business Media; 2013.

52. Love MI, Huber W, Anders S: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014, 15:550.

53. UniProt C: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res* 2019, **47:**D506-D515.

54. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, et al: **DNA-binding specificities of human transcription factors.** *Cell* 2013, **152:**327-339.

55. Wagih O: ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* 2017, 33:3645-3647.