Published online 20 November 2015

Nucleic Acids Research, 2016, Vol. 44, No. 5 ***e49***
doi: 10.1093/nar/gkv1281

Modeling co-occupancy of transcription factors using chromatin features

Liang Liu, Weiling Zhao and Xiaobo Zhou*

Center for Bioinformatics and Systems Biology, Department of Radiology, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA

Received August 22, 2015; Revised October 15, 2015; Accepted November 4, 2015

ABSTRACT

Regulation of gene expression requires both transcription factor (TFs) and epigenetic modifications, and interplays between the two types of factors have been discovered. However study of relationships between chromatin features and TF–TF co-occupancy remains limited. Here, we revealed the relationship by first illustrating distinct profile patterns of chromatin features related to different binding events, including single TF binding and TF–TF co-occupancy of 71 TFs from five human cell lines. We further implemented statistical analyses to demonstrate the relationship by accurately predicting co-occupancy genome-widely using chromatin features including DNase I hypersensitivity, 11 histone modifications (HMs) and GC content. Remarkably, our results showed that the combination of chromatin features enables accurate predictions across the five cells. For individual chromatin features, DNase I enables high and consistent predictions. H3K27ac, H3K4me 2, H3K4me3 and H3K9ac are more reliable predictors than other HMs. Although the combination of 11 HMs achieves accurate predictions, their predictive ability varies considerably when a model obtained from one cell is applied to others, indicating relationship between HMs and TF–TF co-occupancy is cell type dependent. GC content is not a reliable predictor, but the addition of GC content to any other features enhances their predictive ability. Together, our results elucidate a strong relationship between TF–TF co-occupancy and chromatin features.

INTRODUCTION

Transcriptional regulation exists at both genetic and epigenetic levels. Binding of transcription factors (TFs) to specific DNA sequences is a pivotal step in the control of gene expression. Studies of sequence-associated TF binding preferences have led to the development of sequence-specific Position Weighted Matrix (PWM) (1) and position-specific affinity matrices (2) approaches for identification of TF binding sites (TFBSs).

Epigenetic regulation refers to the alteration of DNA accessibility to TFs coordinately with chemical modifications of chromatin (3). This process may involve in multiple factors, such as DNA shape, chromatin accessibility, histone modifications (HMs), nucleosome positions and other chromatin variants (4–9). Analyses of experimental data show that distinct HM patterns appear around TFBSs, and ChIP-Seq signals of TF bindings and HMs are highly predictive of each other (10–14). Specifically, previous studies depicted that chromatin features, such as HMs and DNA shape, are highly correlated with the quantitative changes of TF binding affinities (14,15).

TFs tend to work with others for accurately regulating expression of their target genes by binding to the same regulatory regions (16). These TFs can act either collaboratively or competitively (17–20), and are tightly associated with modeling of cell-specific *cis*-regulatory modules (21,22). Experimental studies of possible TF–TF interactions with either systematic assays or ChIP-Seq in various organisms, such as *Escherichia coli*, yeast, the *Drosophila* embryo and human cell lines (17,23–27), revealed a great number of co-localization hotspots (28), and co-localization patterns that are related to regulatory functions (21). For instance, CCCTC-binding factor (CTCF) is a TF and widely binds to thousands of loci in genome (29). CTCF performs myriad functions by controlling binding affinities with its partners (30), such as yin yang 1 (YY1). The cooperative role of CTCF and YY1 was originally seen in *trans*-activating *Tsix* ncRNA during X-chromosome inactivation (31). Genome-wide analysis depicted their global co-localizations in human cells (23), and their interactions are, at least in part, associated with the evolutionary stability of CTCF genomic occupancy (19). Even for the same TF, if two binding sites are close to each other, the binding of the TF to one site is likely to interfere its binding to another one (32).

DNA sequence and chemical modifications of chromatin can affect not only binding of an individual TF but also a cluster of TFs (23,33–35). Although a large amount of works have been done in investigating the associations of

*To whom correspondence should be addressed. Tel: +1 336 713 1789; Fax: +1 336 713 5891; Email: xizhou@wakehealth.edu

chromatin features with bindings of individual TFs (10–15), a few of studies have devoted to explore the relationships between chromatin features and TF–TF co-occupancy/interactions. This may shed light on a comprehensive understanding of the relationships between TF–TF interactions and chromatin features, as well as their regulatory mechanisms.

In this work, we firstly illustrated the distinct profiling patterns of chromatin features for two types of genome binding events, including the regions solely bound by an individual TF and others bound by this TF and its partners simultaneously. We aligned and compared the profiles of DNase I hypersensitivity (DNase I), HMs and TF binding events by taking advantage of the wealth of data from the ENCODE project (23). Statistical tests showed a strong correlation between binding events and chromatin features across five human cell lines, including A549, GM12878, H1-hESC, HepG-2 and K562. To further demonstrate the relationship between binding events and chromatin features, we then examined predictive ability of chromatin features for TF–TF co-occupancy through a computational model. Our results showed that chromatin features are able to accurately predict the TF–TF co-occupancy genome-widely. By constructing computational models with different chromatin features, we found that both DNase I and combined 11 HMs achieve similar predictive powers. In general, the predictive ability of a single HM is weak; 4 out of 11 HMs, including H3K27ac, H3K4me 2, H3K4me3 and H3K9ac, are more reliable predictors than others. Although GC content itself is not an accurate predictor, addition of GC content improves the predictive ability of DNase I or HMs. We consequently applied the models obtained from one cell line to other cells, and found that the prediction accuracy of the combined chromatin features, including DNase I, 11 HMs and GC content, is maintained consistent across cell lines. Prediction accuracies of the models with individual or the combined 11 HMs receive considerable variances across cell lines, indicating the correlation between HMs and TF–TF co-occupancy is cell type dependent. Models using DNase I on the other hand obtain more consistent predictions across all of cell lines. Taken together, our analyses depict a potential role of chromatin features as determinants in the prediction of TF–TF co-occupancy. This study will contribute to our understanding of the interplay between genetic and epigenetic regulations of gene expression.

MATERIALS AND METHODS

Datasets

All of the data used in this study were downloaded from the ENCODE project (http://genome.ucsc.edu/ENCODE/downloads.html) (1). The ENCODE project has generated TF binding data, by using ChIP-Seq technique (2), in both normal and cancer cell lines. Five human cell lines were selected in this study, including A549 (epithelial cells), GM12878 (B-lymphoblastoid cell), H1-hESC (embryonic stem cells), HepG-2 (hepatocellular carcinoma cells) and K562 (erythrocytic leukemia cells). TF binding profiles by ChIP-Seq data were obtained from the HAIB, and UW TFBS ENCODE groups.

Genome-wide profiles of HMs, including H3K9ac, H3K27ac, H3K4me3, H3K4me2, H3K4me1, H3K79me2, H3K9me3, H3K27me3, H3K36me3 and H4K20me1, and the histone variant, H2A.z, were generated using the ChIP-Seq technique (2). DNase I profiles of the five cell lines were generate with DNase-Seq technique (3).

DNA methylation levels were quantitatively profiled with the RRBS technique and Infinium HumanMethylation450 BeadChip array. The former covers >1 M CpG sites, while the latter measures the methylation levels for 485 577 CpG sites. The methylation level of each CpG is determined as the average of RRBS replicated experiments or HumanMethylation450 BeadChip data and ∼1.3 M CpGs were included.

Genomic locations of 40 193 genes with all information were extracted from the human genome version hg19, obtained from the RefSeq database (downloaded from UCSC Genome Browser at http://genome.ucsc.edu/).

We downloaded the RNA-Seq data that were profiled using Poly A+ protocol from the ENCODE project (1). The expression levels of all RefSeq genes were calculated according to the FPKM (Fragments Per Kilobase of transcript per Million mapped reads) definition. To reduce the redundancy, the expression levels from multiple replicates were merged by taking the mean expression level of each gene.

Chromatin state segmentation data was also downloaded from the USCS Genome Browser. The chromatin states were defined using the unsupervised machine learning technique ChromHMM (36), and available for the GM12878, H1-hESC, HepG-2 and K562 cell lines.

Determination of TF–TF co-occupying regions

Based on the uniform processing pipeline developed for the ENCODE Integrative Analysis effort (37), the binding sites or each TF were determined by peak calling using the SPP peak caller (38) and the consistency and reproducibility between biological replicates with the measurement of the Irreproducible Discovery Rate (IDR < 2%) (39), from the corresponding ChIP-Seq data. The ChIP-Seq peak summits were selected to represent TFBSs. There are various numbers of binding sites for each TF across different cell lines (Supplementary Table S1).

The BEDTools intersectBED function (40) was used to determine whether two TFs, such as CTCF and YY1, were co-localized in the same genomic regions (18–20,41,42). Here we named, for example, the genome regions co-occupied by CTCF–YY1 as CTCF–YY1 co-occupancy, if at least a 30% overlap of CTCF peak by YY1 peak, and *vice versa*, was observed in this region. In contrast, we defined genome regions bound by CTCF but not YY1 as CTCF-only events, and regions bound by YY1 not CTCF as YY1-only events. In such way, all binding sites can be classified into three binding event categories, CTCF-only, CTCF–YY1/YY1–CTCF and YY1-only (Supplementary Figure S1A). Of note, using different overlapping criteria, such as 40% overlapped ChIP-Seq peaks, or other tools, such as IntervalStats (43), only resulted in the change of numbers of TFBSs in each binding category, but not the following prediction analysis or association study (data not shown).

Sequence and chromatin features at TFBSs

We examined the sequence features among binding sites by testing the binding motifs of each TF. Taken CTCF and YY1 as an example, for the different binding events, including CTCF-only, CTCF–YY1 and YY1-only, we analyzed DNA sequences surrounding binding sites of CTCF or YY1. We used the top 1000 binding sites in each type of binding events to identify the motifs, and discovered the *de novo* motifs using MEME tool (44).

For profiles of chromatin features including DNase I and HMs, we first selected the 6k-bp genome regions centered at peak summits of ChIP-Seq data for each TF to analyze the differences related to binding events. We calculated the profiles of tag density of chromatin features at a resolution of 100 bp, and quantified tag density in RPKM. The HM patterns at TFBSs were characterized by 11 types of histone methylation and acetylation, each of which has been associated with transcriptional activation, suppression or both (7,36,45).

We also selected the 100-bp region centered at each TFBS, and calculated the normalized RPKM values of chromatin features in this small region to represent their densities. Then Student's *t*-tests were performed between the profiles of TF–TF co-occupancy and TF-only binding sites. Similar analysis was performed to sequence features, such as GC content.

For DNA methylation, we selected the methylation level of CpG site(s) mapped into the 100-bp region to compute methylation level at this TFBS. For 100-bp regions centered at TFBSs with more than one CpG site, the average of methylation levels over these mapped CpG sites was selected to represent the methylation level.

Predicting TF binding events

We have examined the HMs, DNase I and GC content in the 100-bp region centered at TFBSs (ChIP-Seq peak summits), by counting and normalizing the number of reads mapping to this region to calculate RPKM values. The TF–TF and TF-only binding sites (e.g. CTCF–YY1 versus CTCF-only) were selected as 'positive' and 'negative' datasets, respectively.

We used two non-linear classifiers and two linear methods to build the chromatin models for studying the correlations between TF–TF co-occupancy and chromatin features (Figure 1). The two non-linear classifiers were support vector machine (SVM) (46) and Random Forest (RF). The linear methods included Naïve Bayes (NB) and Linear Discriminant Analysis (LDA). For the SVM classifier, LIBSVM software (47) implemented in the R package 'e1071' and non-linear radial basis kernel were selected. R packages 'randomForest', 'e1071', and 'MASS' were used for the RF, NB and LDA, respectively.

In each cell line, we randomly selected two-third of positive and negative datasets as training and the rest as testing. The ability of chromatin model to distinguish TF–TF co-occupancy from TF-only binding events was assessed by examination of receiver operator characteristic (ROC) curves, plotting the true-positive rate versus the false-positive rate. To test the stability of these predictions, the above procedure was repeated 10 times and the means of the area under ROC curves (AUC) and prediction accuracy (ACC) values were computed to represent the prediction accuracy.

The learned models can be applied to different cell lines for the cross-cell type testing purpose. During this process, the model learned for one pair of TFs was used to other cells for the same TF–TF pair. The prediction accuracies were evaluated by the calculations of AUC and ACC values.

RESULTS

Analysis of sequence and chromatin features for individual TF binding and TF–TF co-occupying events

TFs account for ∼10% of proteins encoded by human genes (48) and their bindings are depended on both genome sequence and chemical alternatives to the sequence (10–14). Based on the assessed TFs in the ENCODE project (23), a large number of TF–TF binding partners have been identified (17,27). In this study, we used CTCF and SP1 as the key experimental TFs, and analyzed their co-occupancy with other TFs. We also selected a set of TFs, including ATF3, GABP, NRSF, POL2, USF1 and YY1, whose ChIP-Seq data were available in all of the five human cell lines, to further demonstrate the relationships between chromatin features and binding events. Of note, the analyzing approach presented in this work can be feasibly applied to other TFs.

A TF may share its binding regions with its partner, namely TF–TF co-occupancy. For instance, CTCF–YY1 co-occupancy refers to the regions co-occupied by CTCF and YY1. Consequently, we defined the genomic regions bound by CTCF but not YY1 as CTCF-only events, and the regions bound by YY1 but not CTCF as YY1-only events (Supplementary Figure S1A). For each TF–TF pair, there are various numbers of co-occupying and solely binding events (see Supplementary Materials; Supplementary Figure S1B and Supplementary Table S1). The co-occupying TFs preferably bind at specific genomic regions (Supplementary Figure S2A). Genome-wide analyses of TFBSs revealed that CTCF intends to bind at gene bodies; while its partner, YY1, as an example, prefers to regulate its target genes by binding at proximal promoters. When considering co-localized CTCF and YY1, their co-occupied regions are mainly distributed in the gene promoters (Supplementary Figure S2B). In our analysis, we considered the genome-wide TFBSs.

By aligning the profiles of DNA sequences, DNase I and HMs, we were able to examine the chromatin features for TF solely and TF–TF co-occupying binding events. *De novo* binding motifs analyses of each TF–TF pair (see 'Materials and Methods' section) illustrated a similar sequence preference for TFs involved in both co-occupying and solely binding regions (see Supplementary Materials; e.g. motifs of CTCF and YY1 shown in Supplementary Figure S3), indicating that DNA sequence may not be a determinant for the TF–TF co-occupancy.

Analyses of GC content (see 'Materials and Methods' section), which dictates nucleosome depletion at mammalian promoters with GC-richness benefiting TF binding (17), showed that CTCF–YY1 co-occupying regions are significantly associated with GC content than CTCF-only binding sites (Student's *t*-test $P < 1e\text{-}15$; see Supplementary Materials; Figure 2 and Supplementary Figure S4), suggest-

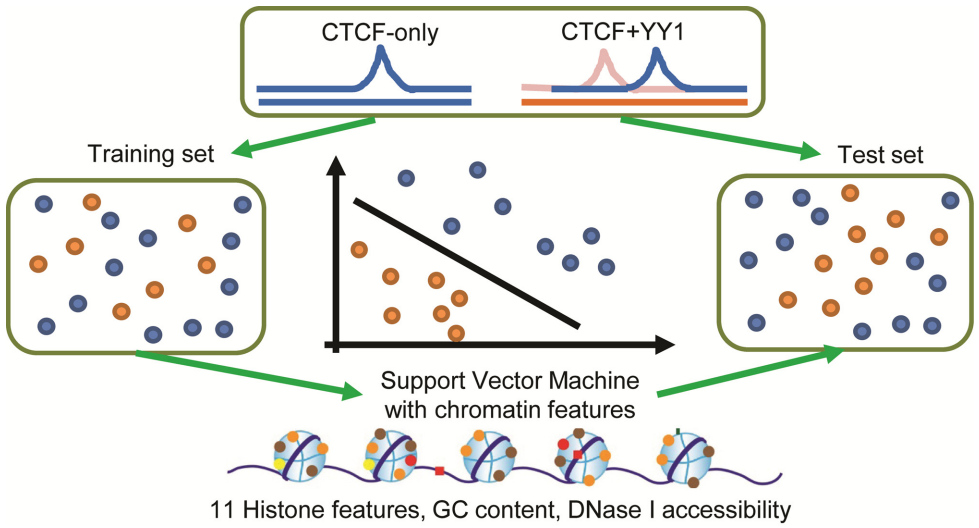

Figure 1. Schematic of the SVM approach used to predict TF–TF and TF-only, or TF–TF and TF-only binding events. YY1 was selected as a representation of CTCF binding partners for illustration. All binding events, including CTCF-TF co-occupying and CTCF-/TF-only binding regions, were separated into training and testing datasets. Then SVM classifier was trained using all or a subset of 11 HMs, DNase I and GC content. The trained model was applied to the test dataset for prediction accuracy valuation with ROC curves, AUC and ACC values. The model was also applied to the same TF–TF pair in different cell types for cross-cell type tests.

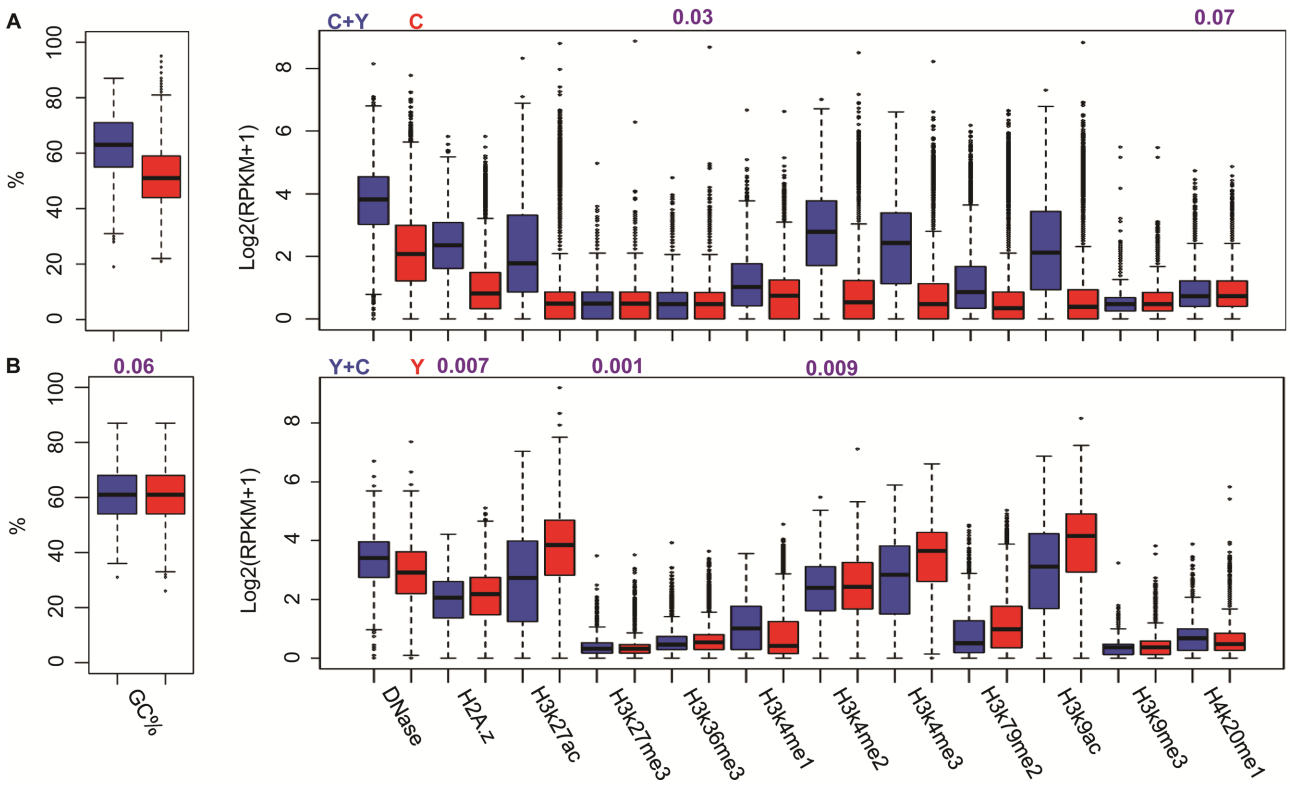

Figure 2. Comparison of chromatin feature profiles (**A**) between CTCF–YY1 co-occupying (C + Y, blue) and CTCF-only binding regions (C, red) and (**B**) between YY1-CTCF (Y + C, blue) and YY1-only (Y, red) binding regions in the K562 cell line. All tests reached *P*-values < 1.0E-5, unless values are shown in figures (numbers in purple).

ing a stronger transcriptional activities of the CTCF–YY1 co-occupied regions (19,49,50). This is consistent with previous findings that genes with CTCF–YY1 co-occupying regions are highly expressed than others solely bound by CTCF (19). It is worth to note that minor differences exist in GC-content profiles, and the chromatin feature profiles as follows, between CTCF-TF and TF-CTCF binding events, because the binding sites of co-occupied CTCF and YY1, represented by ChIP-Seq peak summits, are not located at the exactly same genome positions.

Analyses of chromatin features also revealed distinct profiling patterns for different binding events (see Supplementary Materials; Supplementary Figure S5). The HM patterns at TFBSs were characterized by 11 types of histone methylation and acetylation (see 'Materials and Methods' section), which are associated with transcriptional activation, suppression or both (7,36,45). DNase I hypersensitive sites are regions of chromatin sensitive to cleavage by DNase I. In these sites, nucleosome structure is less compacted, increasing the availability of the DNA to binding of TFs (35,51). Our results show that, HMs, except H3K27me3, are more enriched in CTCF–YY1 co-occupying regions than these in the CTCF-only sites, which is consistent with the reported association between co-localizations of CTCF and YY1 and transcription activity (19). When only considering a smaller 100-bp regions centered at TFBSs, the differences of chromatin feature enrichment are more obvious. All comparisons of individual chromatin features show significantly differences, with a few exceptions such as H3K27me3 in the HepG-2 cell line (see Supplementary Materials; Supplementary Figure S4).

We did the same analysis for other TF–TF pairs. Similar enrichment patterns were observed when comparing chromatin feature profiles between TF–TF co-occupancy and TF-only events (student's *t*-test $P < 0.05$; see Supplementary Materials; Figure 2, Supplementary Figure S4 and Supplementary Table S2). Taken together, our analyses suggest that chromatin features are strongly related to TF binding events, and encourage us to construct a computational model using chromatin features to discriminate TF–TF co-occupancy from TF-only events.

Chromatin features are predictive of TF–TF co-occupancy

We used an SVM classifier to study the direct relationship between local chromatin features and TF–TF co-occupancy, by evaluating to what extent the local chromatin features are informative of a variety of binding events. The classifier was constructed based on the normalized signals (log_2-transformed RPKMs) of chromatin features within the 100-bp window centered at TF peak summits (see 'Materials and Methods' section; Figure 1), and tested on its predictive ability by examination of ROC curves, together with the means of AUC and ACC values after 10-time repetition. The chromatin features includes DNase I, 11 HMs and GC content.

Starting with CTCF and YY1 binding events, the chromatin feature-based model enabled accurate predictions of co-occupancy, with AUC = 0.92 and 0.88 for CTCF–YY1 (distinguished from CTCF-only) and YY1–CTCF (distinguished from YY1-only) binding events, in the GM12878 cell line (Figure 3). High prediction accuracies were also achieved in the A549, H1-hESC, HepG-2 and K562 cells with AUC = 0.92 and 0.76, 0.81 and 0.85, 0.91 and 0.78, and 0.89 and 0.79, respectively (Supplementary Figure S6). The predictive ability of chromatin features were also demonstrated using ACC value estimations with an average of ACCs ∼0.80 for CTCF–YY1 and ∼0.75 for YY1–CTCF co-occupancy (Figure 3, Supplementary Figure S6 and Supplementary Table S3).

When using individual or a subset of chromatin features as predictors, the prediction accuracies obtained from DNase I and combined 11 HMs are closed to that using all features as a whole (Figure 3, Supplementary Figure S6 and Supplementary Table S3). This observation may be explained by the previously reported results indicating that both DNase I and HMs can be used to precisely estimate open chromatin (52,53) and HMs are predictive of chromatin accessibility (54).

GC content is another valuable predictor. GC content patterns are not significantly different, especially in the YY1–CTCF and YY1-only comparisons, therefore GC content itself does not enable accurate predictions. However, addition of GC content to any other features enhances their prediction capability (Figure 3, Supplementary Figure S6 and Supplementary Table S3).

We also employed the RF, NB and LDA for the prediction of CTCF–YY1 co-occupancy. High prediction accuracies were generally achieved. For instance, the RF Classifier can achieve high predictions with AUC = 0.90 and 0.76, 0.89 and 0.87, 0.79 and 0.85, 0.91 and 0.78, and 0.89 and 0.79, respectively, in the A549, GM12878, H1-hESC, HepG-2 and K562 cells, when using all chromatin features as a predictor. Similar results were observed when using different chromatin features as predictors (Supplementary Figure S7). The linear model with NB and LDA gave accurate predictions from individual chromatin features, such as DNase I, which were similar to the results obtained from non-linear models. However, the prediction accuracy from combined chromatin features, especially using 11 histone marks as a predictor, was low (Supplementary Figures S8 and S9), indicating a non-linear relationship of HMs with TF–TF co-occupancy. These results were consistent with the relationship between epigenetic modifications and individual TF binding (14). Since the SVM classifier led to better predictions, this method is selected to depict the predictive ability and consequently correlation in the following analyses.

The SVM classifiers were trained and tested for other types of TFs for prediction of CTCF-TF/TF-CTCF co-occupancy. The results showed that, chromatin features are informative of binding events, with mean AUC values 0.90 and 0.72, 0.89 and 0.73, 0.83 and 0.73, 0.88 and 0.75, and 0.90 and 0.74 in the A549, H1-hESC, HepG-2 and K562 cells, for both CTCF–YY1 and TF-CTCF co-occupancy, respectively (Figure 4 and Supplementary Table S3). Consistent with the observations from CTCF–YY1 analysis, both the combined 11 HMs and DNase I enable highly accurate predictions. Predictions with individual HMs also achieve good results (Figure 3, Supplementary Figures S6 and S10, and Supplementary Table S3). In general, H3K27ac, H3K4me 2, H3K4me3 and H39Kac

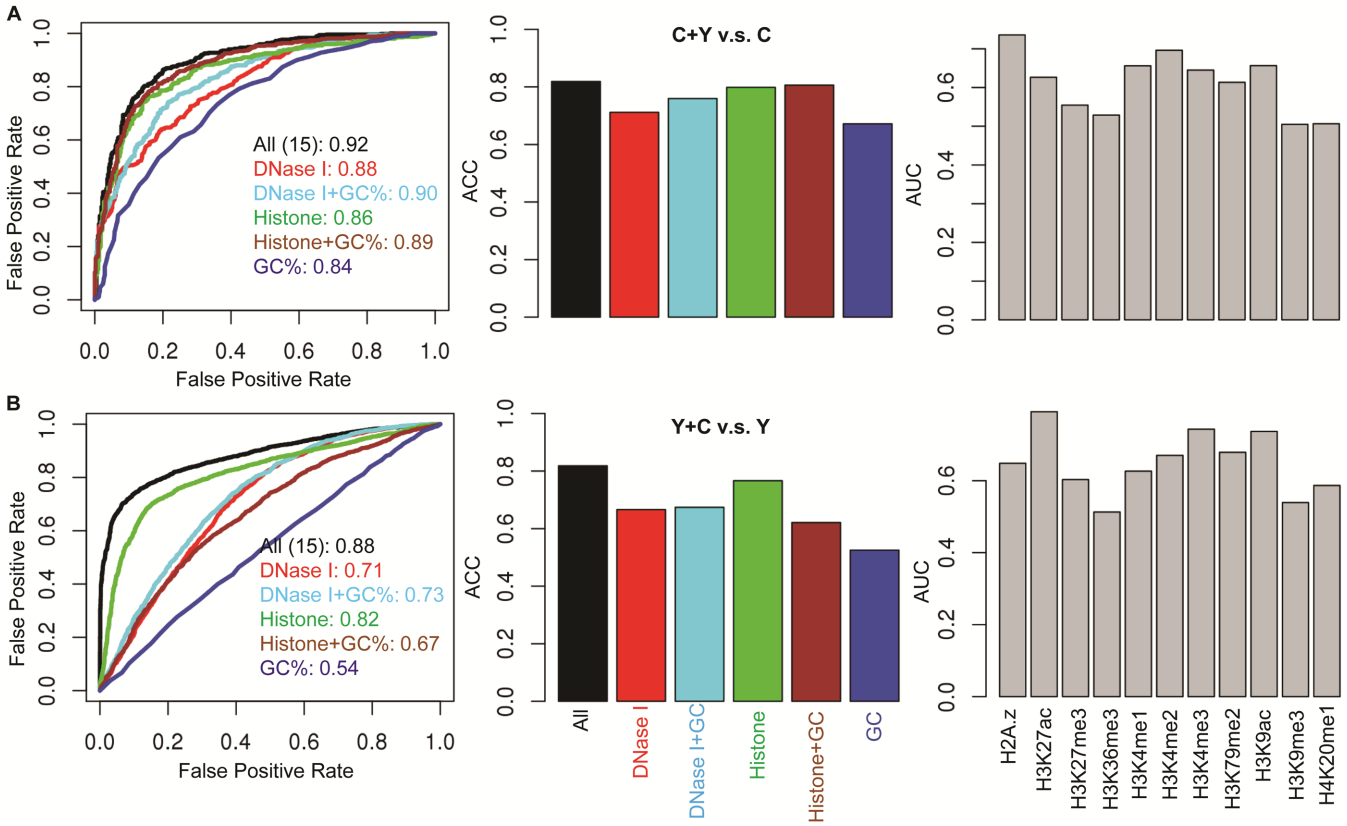

Figure 3. Chromatin features are predictive of CTCF–YY1 co-occupancy from (**A**) CTCF-only and (**B**) YY1-only events with high accuracies in the GM12878 cell line. Left: ROC curves are shown with colors representing predictions using different chromatin features and AUC values are indicated in the legend; Middle: predictions evaluated with ACCs; and Right: predictions evaluated with AUC values using individual histone features.

are more reliable predictors; however, the predictive ability of the combined or individual HMs varies across cell lines (Figures 3 and 4, and Supplementary Figure S10). This observation was further validated by cross-cell line tests in the following section. In spite that GC content alone does not achieve high prediction accuracy, addition of GC content to other features improves their prediction power.

We examined the predictive ability of chromatin features for SP1-TF co-occupancy in four cell types (ChIP-Seq data of SP1 were not available in the A549 cell), and the co-localizations of another six TFs, including ATF3, GABP, NRSF, POL2, USF1 and YY1. The latter test involved 15 binding pairs, as shown in Supplementary Table S8. Enrichment analyses of GC content, DNase I and 11 HMs showed distinct patterns between binding events, illustrated by the given examples of SP1-BCL11A and SP1-TAF1 (Supplementary Figure S11).

We consequently constructed and tested the computational models for each TF–TF pair in each cell type. High prediction accuracies were generally obtained. For instance, all chromatin features as a whole were able to accurately predict the GABP-USF1 co-occupancy with AUC values >0.86 in the GM12878 and K562 cell lines (Supplementary Figure S12A). For the prediction of SP1-TF combinational binding events, the average accuracies were ∼0.80, 0.77, 0.80 and 0.80 in the GM12878, H1-hESC, HepG-2 and K562 cells, respectively (Supplementary Figures S13A, B and Supplementary Table S5). For the 15 TF–TF co-occupancy with ATF3, GABP, NRSF, POL2, USF1 and YY1, chromatin features enable highly accurate predictions with AUC values >0.8 in all of five cell types (Supplementary Figures S14A, B and Supplementary Table S8).

The prediction abilities with a subset of or individual chromatin features for above TF–TF co-occupancy were similar to those for CTCF-TF co-occupancy. In general, DNase I and the combination of 11 HMs are able to achieve high prediction accuracy and addition of GC content enhances their performance (Supplementary Figures S13A, B and S14A, B). H3K27ac, H3K4me 2, H3K4me3 and H39Kac perform better than other HMs (Supplementary Figures S13C, D and S14C, D). This is consistent with our previous findings about the contribution of single HM to binding affinity of individual TFs (14). In summary, all of our observations demonstrate the strong relationship between chromatin features and TF–TF co-occupancy, and the former is sufficient to model the latter genome-widely.

Chromatin features enable predictions of TF–TF co-occupancy across different cell lines

Both chromatin modifications and TF binding profiles exhibit dynamic and cell-specific patterns. Given that chromatin features, together with GC content, have the ability in predicting TF–TF co-occupancy, we tested to what

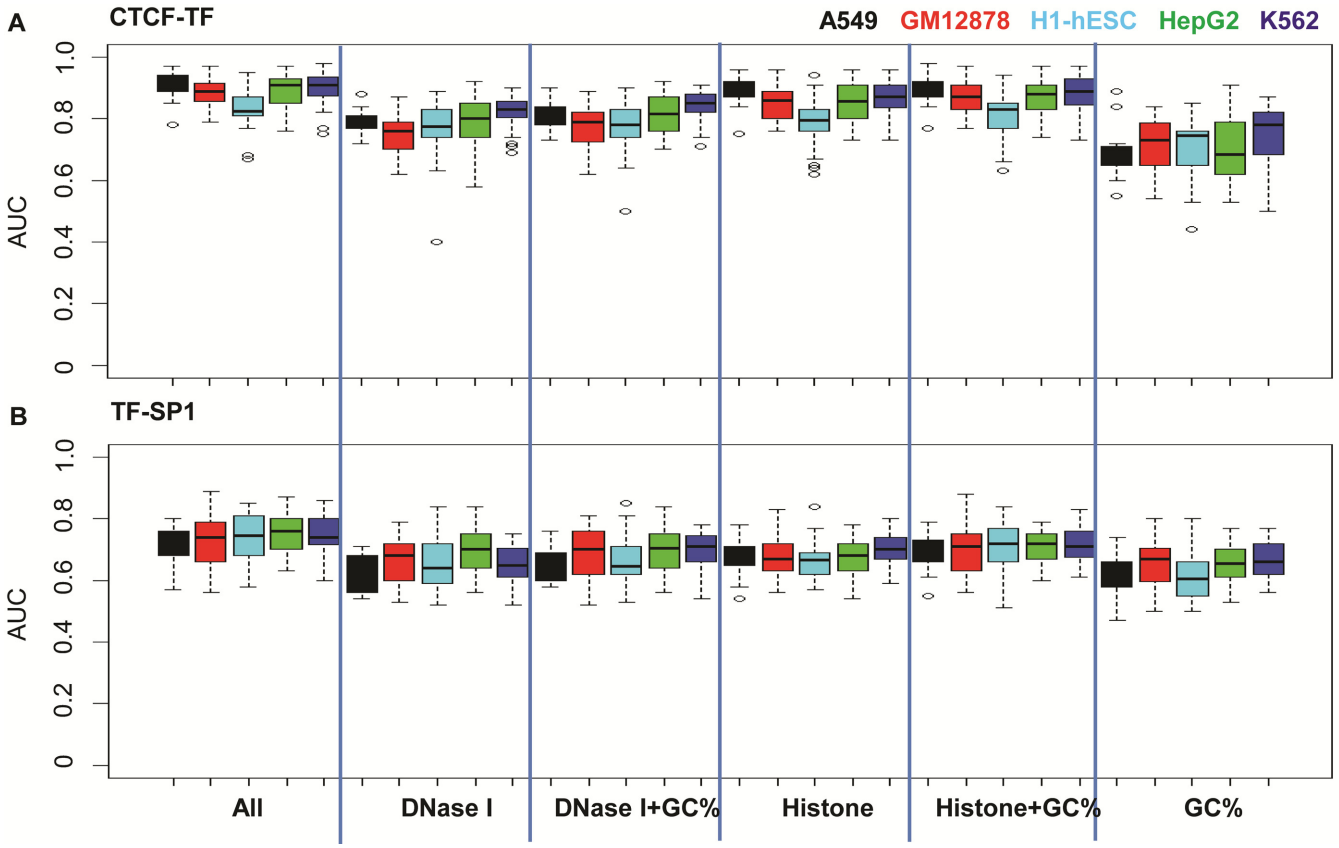

Figure 4. Chromatin features are predictive of (**A**) CTCF-TF and (**B**) TF-CTCF co-occupancy with high accuracies. Computational models were trained and applied to the same TF–TF pair in the same cell line, labeled by colors: black, A549; red, GM12878; cyan, H1-hESC; green, HepG-2; and blue, K562. Models were trained using different sets of chromatin features.

extend the chromatin-feature models could be generalized from one cell line to others.

We have constructed prediction models for all CTCF-TF pairs with 11 HMs, DNase I and GC content in the five cell lines. As shown in the Figure 3, and Supplementary Figures S6 and S10, or diagonal figures in the Figure 5 and Supplementary Figure S15, these models are able to accurately identify genome-wide binding events in the cell types they were trained on.

The models were trained on each cell line and then applied to the other four cell lines to test their prediction abilities. The results indicated that the cross-cell line applications of classifiers do not reduce their performances in predictions. For example, when we applied the models trained in the GM12878, H1-hESC, HepG-2 and K562 cell lines to the A549 cell for CTCF–YY1 predictions, the average prediction accuracies were 0.86, 0.77, 0.88 and 0.89, respectively, compared to 0.91 using the model trained by the A549 cell itself (Figure 5A and Supplementary Figure S15). The largest changes were observed with accuracies as 0.78, 0.76, 0.78 and 0.81 when models were trained within other four cells and applied to the H1-hESC cell, compared to 0.84 using model trained by the H1-hESC cell line, and *vice versa,* with accuracies as 0.77, 0.76, 0.81 and 0.79 when model was trained in the H1-hESC cell line and applied to the A549, GM12878, HepG-2 and K562 cells, compared to 0.91, 0.91, 0.91 and 0.90 when models were trained and applied to the same cell line (Supplementary Figure S15).

We did the same analyses to other CTCF-TF, SP1-TF and ATF3-/GABP-/NRSF-/USF1-TF co-occupancy. Due to the fact that not all TF binding profiles have been generated by the ENCODE project, we included various numbers of TFs in each type of cells (Supplementary Table S1). Consistent with our results from CTCF–YY1 studies, crossing-cell line applications achieved satisfactory accuracies (Supplementary Figures S12B, S16, S17 and S18). The biggest changes were seen when models from other four cell lines were applied to H1-hESC and *vice versa*. Overall, the results from our cross-cell type analyses support the generalizing associations of TF–TF co-localizations with chromatin features.

The relationships of TF–TF co-occupancy with DNase I are more conserved

We observed that the same features had different prediction powers across cell lines (e.g. Figures 3 and 4 for CTCF–YY1), especially when comparing the H1-hESC cell with others. This was further illustrated in cross-cell line predictions (e.g. Figure 5 and Supplementary Figure S16 for CTCF–YY1). Next, we examined the conservative relationships between TF–TF co-occupancy and individual chromatin features.

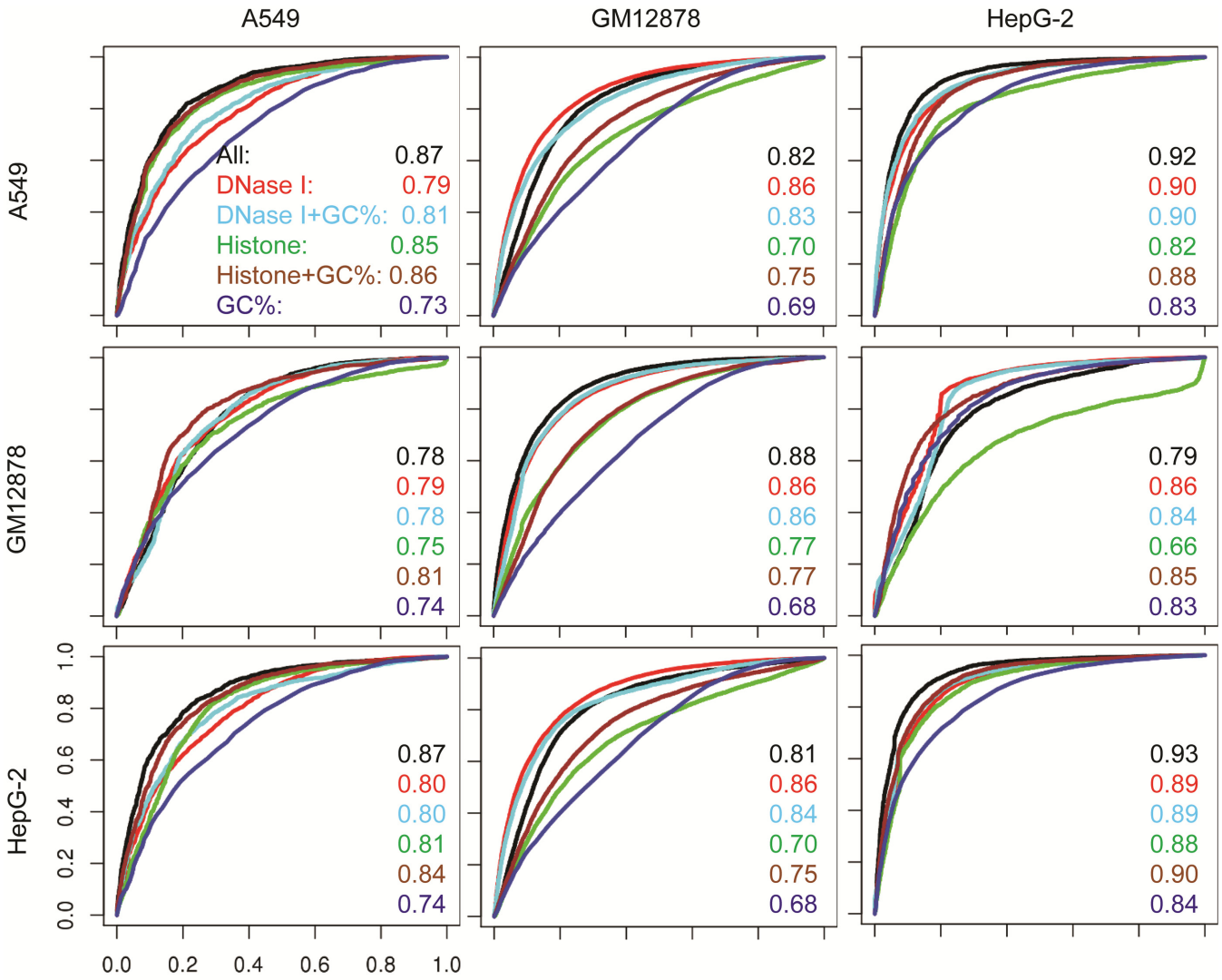

Figure 5. Chromatin features enable predictions of CTCF-TF co-occupancy across cell lines. Shown are ROC curves with colors representing the predictions using different chromatin features. Models were trained in the cell line indicated by row and tested on each of the five cell lines indicated by column. The AUC values are indicated on the plot as legend (see complete figures in Supplementary Figure S15).

We constructed chromatin-feature models for CTCF-TF predictions using individual or a set of chromatin features, and then applied the models obtained from one cell type to others. We found that models with DNase I, GC content and DNase I plus GC content give more consistent predictions across all of human cells (Figure 6A and Supplementary Figure S19). The overall accuracy changes (|ΔAUC|) were ∼0.02 among five cell lines, when models were built with DNase I (Figure 6B and Supplementary Figure S20). In contrast, prediction accuracies with individual or combination of 11 HMs varied across cell lines with accuracy changes ∼0.15, especially when prediction models were exchanged between H1-hESC and one of other four cell lines (Figure 6B and Supplementary Figure S20).

The above analyses were also conducted in other two sets of TF–TF pairs, and similar trends were observed (Supplementary Figures S21 and S22). Of note, for some TF–TF pairs, such as SP1-ATF2, DNase I did not enable highly accurate predictions. In summary, all results suggest that the correlations between TF–TF occupancy and DNase I/GC content are more conserved than HMs across cell lines.

DISCUSSION

The accurate regulation of gene expression involves in a complicated interplay between TF, histone modifying enzymes and other factors. The relative importance of epigenetic modification and TF bindings in the regulation of gene expression is still under debate. Statistics analysis has revealed that these two factors regulate gene transcription in a highly coordinate manner (34), and are redundant for predicting gene expression (55). Several studies have described direct interactions between histone modifying enzymes and TFs (56). Co-occupancy of a binding site by multiple TFs plays a critical role in fine regulation of gene expression (57). Certain patterns of histone marks have been observed around co-binding sites of some TF pairs, such as FOXA1-FOXA3 (57), YY1-MYC and CTCF-NF-Y (17).

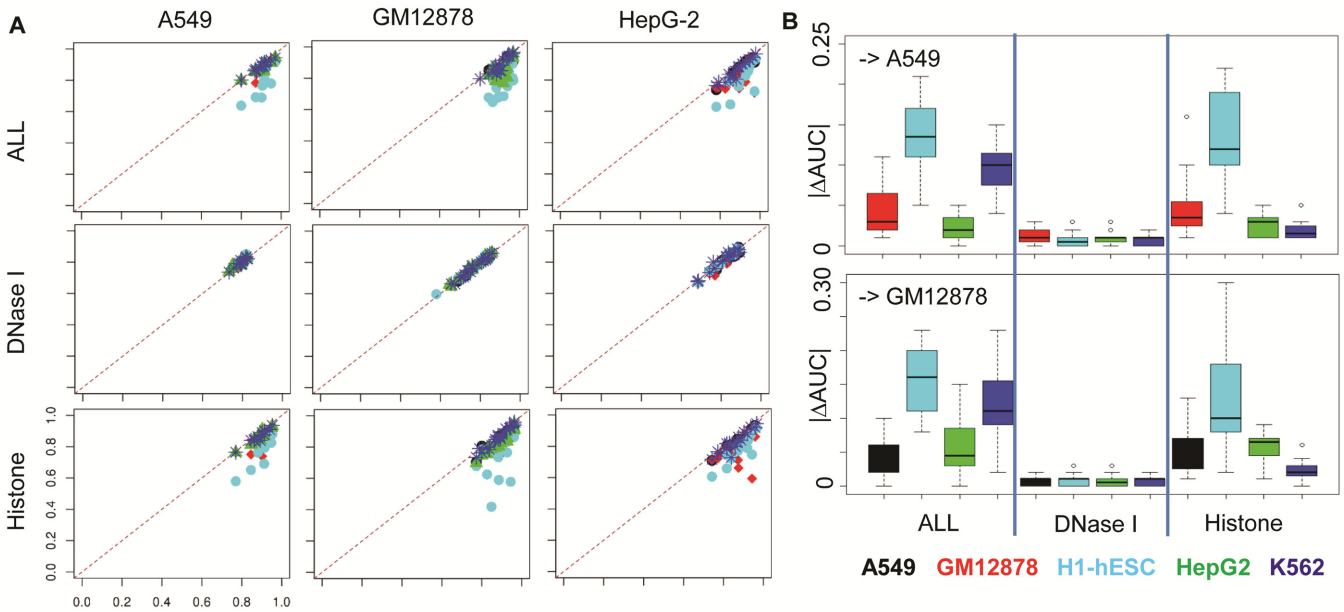

Figure 6. Cross cell predictions of CTCF-TF co-occupancy. (**A**) Comparisons of cross cell predictions (y-axis) to predictions using models obtained and trained in the same cell (x-axis). Test cell lines are shown by column and training cell line are indicated by colors: black, A549; red, GM12878; cyan, H1-hESC; green, HepG-2; and blue, K562. Different chromatin features were used indicated by rows. (see complete figures in Supplementary Figure S19). (**B**) Boxplots of prediction differences indicated by |ΔAUC|, in the A549 and GM12878 cell lines. |ΔAUC| was calculated by subtracting cross cell predictions using models from other cell lines, indicated by colors, from predictions using models trained from the same cell line. Different features were used in each test (see complete figures in Supplementary Figure S20).

However, there are no approaches available currently for computationally identifying TF–TF co-occupying sites using chromatin features and quantitatively modeling the correlation between them. We here introduced chromatin features, that are always cell-type specific, to refine the description of TF–TF co-occupancy, and observed a strong correlation of TF–TF co-occupancy with chromatin features. This relationship was further confirmed by quantitative predictions of TF–TF co-occupancy using multiple or individual chromatin features.

Experimental techniques such as ChIP-Seq have been used to identify TF–TF co-occupancy (23,32,58–60). Unfortunately, these experimental methods are always expensive and time-consuming. Meanwhile, computationally predicting models using sequence-based PWM methods (61–63) or combing ChIP-Seq data and PWMs (64) have also been developed to detect the co-occurring TFs (63) and their binding motifs (64–66). However, prediction of the putative TFBSs using the predefined PWM suffers from a high rate of false positive discovery (67). Moreover, these approaches ignore the influence from cell conditions, which are crucial for epigenetic modifications, chromatin accessibility, TF binding and consequently gene regulation (11,14,17,34,42,68). As a result, the prediction accuracy could vary greatly across cell types. For instance, the *cobindR* software (69) uses PWMs to identify the co-occurring TFs and their binding sites. This sequence-based approach can detect 6,444 CTCF–YY1 co-occupying sites (PWMs obtained from http://jaspar.genereg.net (70)), which cover 49, 20, 31, 19 and 26% sites obtained from the ChIP-Seq data (Supplementary Table S1) in the A549, GM12878, H1-hESC, HepG-2 and K562 cells, respectively. As a comparison, our method used cell type-specific chromatin features as predictors, which largely improve predictive accuracy for cell type-specific TF–TF co-occupying sites (Supplementary Table S3). Of noting, the most importance is that our study illustrated the correlation between chromatin features and TF–TF co-occupancy, which is the main aim of this study and can improve our understanding of the interactions between epigenetic and genomic regulation.

The co-occupied TFs may have different regulation functions from solely bound TF. For example, co-localizations of CTCF and YY1 can enhance transcriptional activity of genes that they are co-occupied (19). Analysis showed that the binding intensities of CTCF at regions co-occupied by YY1 are significantly greater than those bound by only CTCF (Student's *t*-test $P < 1e\text{-}17$ in the five cell lines; Supplementary Figure S23), indicating their differentially functional effects. By profiling co-occupancy of CTCF–YY1 and other TF–TF pairs with chromatin features, we demonstrated the important roles of chromatin modifications in gene regulation and the strong associations between genetic and epigenetic regulations.

Our analysis further illustrated the generalization of this correlation across cell types, which led to the possible application of a prediction model trained from one cell line using combination of chromatin features to other cells for accurate predictions of TF–TF co-occupancy. When applied individual chromatin features in our models, the associations of DNase I with TF–TF co-occupancy were very conserved, and the cross-cell type applications of models with DNase I did not result in dramatically changes of prediction accuracy. This observation suggested that, although DNase I

shows distinct profiling patterns in different cell types (68), these patterns may coordinately change with TF–TF interactions regardless of cell conditions. In contrast, the associations of individual or the combined 11 HMs are less conserved among cell lines. This may be explained by the reported cell-specific correlations between HMs and individual TF binding affinities (14).

TF–TF co-occupancy may have an effect on transcriptional output. Comparisons of the expression levels for RefSeq genes showed that the genes overlapped YY1-only binding events are significantly more highly expressed, followed by genes overlapped with CTCF–YY1 binding events, in contrast to genes overlapped CTCF-only binding events (Wilcoxon rank-sum test, *P*-values < 10e-8; Supplementary Figure S24A), consistent with the previous findings (19). This observation indicates the different functions of CTCF–YY1 co-occupying regions compared to others.

We further tested whether transcriptional output has relationship with both TF–TF co-occupancy and chromatin features, such as DNase I. Comparisons showed that, although either CTCF–YY1 co-occupancy or DNase I is associated with transcriptional activity (17,19,68), the combination of CTCF–YY1 and DNase I did not necessarily lead to higher transcriptional outputs (Supplementary Figure S24B). Indeed, even if CTCF–YY1 and DNase I occurred in the same genomic regions, such as gene promoters, the transcriptional output varied from gene to gene (Supplementay Figures S24C, D). This may be explained by the complicated correlation between DNase I and gene expression. For instance, Wang *et al*. showed that, even for the similarly expressed genes, the distribution of DNase I may differ among different chromosomes (71).

DNA methylation is another type of epigenetic modification involved in the regulation of gene expression, cell growth and disease development (6,72). Early studies reported that DNA methylation is related with TF binding (73), but it alone is not sufficient to prevent protein binding (74–76), or had a weak correlation with individual TF binding affinity (14). We examined the relationship between DNA methylation (see 'Materials and Methods' section) and TF–TF co-occupancy. We selected the methylation level of CpG site(s) mapped into the 100-bp bin centered at each TFBS to compute methylation level at that binding site. Most of TFBSs do not have methylated CpG site(s). In the GM12878 cell lines, 5583 out of 40 247 CTCF binding sites have ≥1 CpGs, including 2,415 CTCF–YY1 and 3148 CTCF-only sites. We constructed SVM classifier with DNA methylation and/or other chromatin features. The results showed that DNA methylation has very fair predication ability, with accuracies ∼0.57 in the five human cell lines. Moreover, the combinations of DNA methylation with any other chromatin features led to nearly same predictive performances (Supplementary Figure S25).

TFs prefer working together to regulate gene expression by targeting the same genomic regions, namely TF hotspots (21,22,28,77). These regions are usually cell type specific, represented by active histone marks and reflect certain chromatin states (77,78). We therefore examined the correlation between chromatin states and TF–TF co-occupancy. Since the CTCF ChIP-Seq has been used by the ChromHMM for the determination of chromatin state segmentation, we selected the SP1-TF pairs as the testing examples. TFBSs were mapped into the 15 chromatin states with different distributions and the association of SP1-TF co-occupancy with all 15 chromatin states were assessed (Supplementary Figure S26). In general, the models using the chromatin state segmentation gave predictive outcomes with accuracies lower than the ones using the combination of all chromatin features or the 11 HMs (Supplementary Figures S27A, B). This suggested that TF–TF co-occupancy is not only reflected by chromatin states or those HMs used for chromatin state determination, but other chromatin features, such as DNase I and GC components. As well, the adding of chromatin state segmentation to any other chromatin features did not significantly change the predictive outcomes (Supplementary Figure S27A, C). This may be because that, during the prediction of chromatin state segmentation with the ChromHMM model, 8 of the 11 HMs from our study have been used as the inputs, and therefore the information from the chromatin state segmentation is partially overlapped with the one from the 11 HMs.

In summary, we have presented a statistical and computational approach to investigate the complicated interplays between genetic and epigenetic regulations of gene expression. Although our analysis cannot demonstrate the relative importance or causative role of TFs or chromatin features in transcriptional regulation, we have elucidated a strong relationship between TF–TF co-occupancy and various chromatin features through a large-scale statistical and computational experiments, which will help us in understanding the mechanisms of combinational regulatory landscape.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to acknowledge the members of Center for Bioinformatics and Systems Biology at Wake Forest School of Medicine and Dr. Di Huang (Computational Biology Branch, National Institutes of Health) for valuable discussions and advices. The authors acknowledge the Texas Advanced Computing Center (TACC) at the University of Texas at Austin (http://www.tacc.utexas.edu) and the DEMON high performance computing (HPC) cluster at Wake Forest University School of Medicine for providing HPC resources.

FUNDING

National Institutes of Health [1U01CA166886 to X. Z.]; NSFC [61373105, in part]. Funding for open access charge: National Institutes of Health [1U01CA166886].
Conflict of interest statement. None declared.

REFERENCES

1. Stormo,G.D. and Zhao,Y. (2010) Determining the specificity of protein-DNA interactions. *Nat. Rev. Genet.*, **11**, 751–760.
2. Bussemaker,H.J., Foat,B.C. and Ward,L.D. (2007) Predictive modeling of genome-wide mRNA expression: from modules to molecules. *Annu. Rev. Biophys. Biomol. Struct.*, **36**, 329–347.

3. Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.
4. Berger,S.L. (2002) Histone modifications in transcriptional regulation. *Curr. Opin. Genet. Dev.*, **12**, 142–148.
5. Slattery,M., Zhou,T., Yang,L., Dantas Machado,A.C., Gordan,R. and Rohs,R. (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.*, **39**, 381–399.
6. Jaenisch,R. and Bird,A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.*, **33**(Suppl), 245–254.
7. Berger,S.L. (2007) The complex language of chromatin regulation during transcription. *Nature*, **447**, 407–412.
8. Schuettengruber,B., Chourrout,D., Vervoort,M., Leblanc,B. and Cavalli,G. (2007) Genome regulation by polycomb and trithorax proteins. *Cell*, **128**, 735–745.
9. Ho,L. and Crabtree,G.R. (2010) Chromatin remodelling during development. *Nature*, **463**, 474–484.
10. Benveniste,D., Sonntag,H.J., Sanguinetti,G. and Sproul,D. (2014) Transcription factor binding predicts histone modifications in human cell lines. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 13367–13372.
11. Arvey,A., Agius,P., Noble,W.S. and Leslie,C. (2012) Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res.*, **22**, 1723–1734.
12. Pique-Regi,R., Degner,J.F., Pai,A.A., Gaffney,D.J., Gilad,Y. and Pritchard,J.K. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res.*, **21**, 447–455.
13. Chen,C.C., Xiao,S., Xie,D., Cao,X., Song,C.X., Wang,T., He,C. and Zhong,S. (2013) Understanding variation in transcription factor binding by modeling transcription factor genome-epigenome interactions. *PLoS Comput. Biol.*, **9**, e1003367.
14. Liu,L., Jin,G. and Zhou,X. (2015) Modeling the relationship of epigenetic modifications to transcription factor binding. *Nucleic Acids Res.*, **43**, 3873–3885.
15. Zhou,T., Shen,N., Yang,L., Abe,N., Horton,J., Mann,R.S., Bussemaker,H.J., Gordan,R. and Rohs,R. (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 4654–4659.
16. Maston,G.A., Evans,S.K. and Green,M.R. (2006) Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.*, **7**, 29–59.
17. Wang,J., Zhuang,J., Iyer,S., Lin,X., Whitfield,T.W., Greven,M.C., Pierce,B.G., Dong,X., Kundaje,A., Cheng,Y. *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.
18. Sanders,D.A., Ross-Innes,C.S., Beraldi,D., Carroll,J.S. and Balasubramanian,S. (2013) Genome-wide mapping of FOXM1 binding reveals co-binding with estrogen receptor alpha in breast cancer cells. *Genome Biol.*, **14**, R6.
19. Schwalie,P.C., Ward,M.C., Cain,C.E., Faure,A.J., Gilad,Y., Odom,D.T. and Flicek,P. (2013) Co-binding by YY1 identifies the transcriptionally active, highly conserved set of CTCF-bound regions in primate genomes. *Genome Biol.*, **14**, R148.
20. Stefflova,K., Thybert,D., Wilson,M.D., Streeter,I., Aleksic,J., Karagianni,P., Brazma,A., Adams,D.J., Talianidis,I., Marioni,J.C. *et al.* (2013) Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell*, **154**, 530–540.
21. Gerstein,M.B., Kundaje,A., Hariharan,M., Landt,S.G., Yan,K.K., Cheng,C., Mu,X.J., Khurana,E., Rozowsky,J., Alexander,R. *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91–100.
22. He,A., Kong,S.W., Ma,Q. and Pu,W.T. (2011) Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 5632–5637.
23. Encode Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
24. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al.* (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.
25. Lifanov,A.P., Makeev,V.J., Nazina,A.G. and Papatsenko,D.A. (2003) Homotypic regulatory clusters in Drosophila. *Genome Res.*, **13**, 579–588.
26. Balazsi,G., Barabasi,A.L. and Oltvai,Z.N. (2005) Topological units of environmental signal processing in the transcriptional regulatory network of Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 7841–7846.
27. Xie,D., Boyle,A.P., Wu,L., Zhai,J., Kawli,T. and Snyder,M. (2013) Dynamic trans-acting factor colocalization in human cells. *Cell*, **155**, 713–724.
28. Moorman,C., Sun,L.V., Wang,J., de Wit,E., Talhout,W., Ward,L.D., Greil,F., Lu,X.J., White,K.P., Bussemaker,H.J. *et al.* (2006) Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 12027–12032.
29. Zlatanova,J. and Caiafa,P. (2009) CTCF and its protein partners: divide and rule? *J. Cell Sci.*, **122**, 1275–1284.
30. Ong,C.T. and Corces,V.G. (2014) CTCF: an architectural protein bridging genome topology and function. *Nat. Rev. Genet.*, **15**, 234–246.
31. Donohoe,M.E., Zhang,L.F., Xu,N., Shi,Y. and Lee,J.T. (2007) Identification of a Ctcf cofactor, Yy1, for the X chromosome binary switch. *Mol. Cell*, **25**, 43–56.
32. Levo,M., Zalckvar,E., Sharon,E., Dantas Machado,A.C., Kalma,Y., Lotam-Pompan,M., Weinberger,A., Yakhini,Z., Rohs,R. and Segal,E. (2015) Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res.*, **25**, 1018–1029.
33. Yip,K.Y., Cheng,C., Bhardwaj,N., Brown,J.B., Leng,J., Kundaje,A., Rozowsky,J., Birney,E., Bickel,P., Snyder,M. *et al.* (2012) Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.*, **13**, R48.
34. Cheng,C., Alexander,R., Min,R., Leng,J., Yip,K.Y., Rozowsky,J., Yan,K.K., Dong,X., Djebali,S., Ruan,Y. *et al.* (2012) Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res.*, **22**, 1658–1667.
35. Chen,H., Li,H., Liu,F., Zheng,X., Wang,S., Bo,X. and Shu,W. (2015) An integrative analysis of TFBS-clustered regions reveals new transcriptional regulation models on the accessible chromatin landscape. *Sci. Rep.*, **5**, 8465.
36. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
37. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
38. Kharchenko,P.V., Tolstorukov,M.Y. and Park,P.J. (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.*, **26**, 1351–1359.
39. Li,Q.H., Brown,J.B., Huang,H.Y. and Bickel,P.J. (2011) Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.*, **5**, 1752–1779.
40. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
41. Mullen,A.C., Orlando,D.A., Newman,J.J., Loven,J., Kumar,R.M., Bilodeau,S., Reddy,J., Guenther,M.G., DeKoter,R.P. and Young,R.A. (2011) Master transcription factors determine cell-type-specific responses to TGF-beta signaling. *Cell*, **147**, 565–576.
42. Griffon,A., Barbier,Q., Dalino,J., van Helden,J., Spicuglia,S. and Ballester,B. (2015) Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Res.*, **43**, e27.
43. Chikina,M.D. and Troyanskaya,O.G. (2012) An effective statistical evaluation of ChIPseq dataset similarity. *Bioinformatics*, **28**, 607–613.
44. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
45. Bernstein,B.E., Meissner,A. and Lander,E.S. (2007) The mammalian epigenome. *Cell*, **128**, 669–681.
46. Cortes,C. and Vapnik,V. (1995) Support-Vector Networks. *Mach. Learn.*, **20**, 273–297.
47. Chang,C.C. and Lin,C.J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 27.
48. Babu,M.M., Luscombe,N.M., Aravind,L., Gerstein,M. and Teichmann,S.A. (2004) Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.*, **14**, 283–291.

49. Fenouil,R., Cauchy,P., Koch,F., Descostes,N., Cabeza,J.Z., Innocenti,C., Ferrier,P., Spicuglia,S., Gut,M., Gut,I. *et al.* (2012) CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters. *Genome Res.*, **22**, 2399–2408.
50. Kudla,G., Lipinski,L., Caffin,F., Helwak,A. and Zylicz,M. (2006) High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol.*, **4**, e180.
51. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
52. Bannister,A.J. and Kouzarides,T. (2011) Regulation of chromatin by histone modifications. *Cell Res.*, **21**, 381–395.
53. Tropberger,P. and Schneider,R. (2013) Scratching the (lateral) surface of chromatin regulation by histone modifications. *Nat. Struct. Mol. Biol.*, **20**, 657–661.
54. Cui,P., Li,J., Sun,B., Zhang,M., Lian,B., Li,Y. and Xie,L. (2013) A quantitative analysis of the impact on chromatin accessibility by histone modifications and binding of transcription factors in DNase I hypersensitive sites. *Biomed. Res. Int.*, **2013**, 914971.
55. Cheng,C. and Gerstein,M. (2012) Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res.*, **40**, 553–568.
56. Schuettengruber,B., Martinez,A.M., Iovino,N. and Cavalli,G. (2011) Trithorax group proteins: switching genes on and keeping them active. *Nat. Rev. Mol. Cell Biol.*, **12**, 799–814.
57. Motallebipour,M., Ameur,A., Reddy Bysani,M.S., Patra,K., Wallerman,O., Mangion,J., Barker,M.A., McKernan,K.J., Komorowski,J. and Wadelius,C. (2009) Differential binding and co-binding pattern of FOXA1 and FOXA3 and their relation to H3K4me3 in HepG2 cells revealed by ChIP-seq. *Genome Biol.*, **10**, R129.
58. Whitington,T., Frith,M.C., Johnson,J. and Bailey,T.L. (2011) Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res.*, **39**, e98.
59. Giannopoulou,E.G. and Elemento,O. (2013) Inferring chromatin-bound protein complexes from genome-wide binding assays. *Genome Res.*, **23**, 1295–1306.
60. Tsankov,A.M., Gu,H., Akopian,V., Ziller,M.J., Donaghey,J., Amit,I., Gnirke,A. and Meissner,A. (2015) Transcription factor binding dynamics during human ES cell differentiation. *Nature*, **518**, 344–349.
61. Yu,X., Lin,J., Masuda,T., Esumi,N., Zack,D.J. and Qian,J. (2006) Genome-wide prediction and characterization of interactions between transcription factors in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **34**, 917–927.
62. Hannenhalli,S. and Levy,S. (2002) Predicting transcription factor synergism. *Nucleic Acids Res.*, **30**, 4278–4284.
63. Rohr,C.O., Parra,R.G., Yankilevich,P. and Perez-Castro,C. (2013) INSECT: in-silico SEarch for co-occurring transcription factors. *Bioinformatics*, **29**, 2852–2858.
64. Oh,Y.M., Kim,J.K., Choi,S. and Yoo,J.Y. (2012) Identification of co-occurring transcription factor binding sites from DNA sequence using clustered position weight matrices. *Nucleic Acids Res.*, **40**, e38.
65. Kato,M., Hata,N., Banerjee,N., Futcher,B. and Zhang,M.Q. (2004) Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol.*, **5**, R56.
66. GuhaThakurta,D. and Stormo,G.D. (2001) Identifying target sites for cooperatively binding factors. *Bioinformatics*, **17**, 608–621.
67. Cartharius,K., Frech,K., Grote,K., Klocke,B., Haltmeier,M., Klingenhoff,A., Frisch,M., Bayerlein,M. and Werner,T. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, **21**, 2933–2942.
68. Natarajan,A., Yardimci,G.G., Sheffield,N.C., Crawford,G.E. and Ohler,U. (2012) Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.*, **22**, 1711–1722.
69. Benary,M., Kroeger,S., Lee,Y. and Lehmann,R. (2013) cobindR: finding co-occuring motifs of transcription factor binding sites. R package version 1.6.0.
70. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C.Y., Chou,A., Ienasescu,H. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.
71. Wang,Y.-M., Zhou,P., Wang,L.-Y., Li,Z.-H., Zhang,Y.-N. and Zhang,Y.-X. (2012) Correlation between DNase I hypersensitive site distribution and gene expression in HeLa S3 cells. *PLoS One*, **7**, e42414.
72. Chen,X., Liu,L., Mims,J., Punska,E.C., Williams,K.E., Zhao,W., Arcaro,K.F., Tsang,A.W., Zhou,X. and Furdui,C.M. (2015) Analysis of DNA methylation and gene expression in radiation-resistant head and neck tumors. *Epigenetics*, **10**, 545–561.
73. Banovich,N.E., Lan,X., McVicker,G., van de Geijn,B., Degner,J.F., Blischak,J.D., Roux,J., Pritchard,J.K. and Gilad,Y. (2014) Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet.*, **10**, e1004663.
74. Weih,F., Nitsch,D., Reik,A., Schutz,G. and Becker,P.B. (1991) Analysis of Cpg methylation and genomic footprinting at the tyrosine aminotransferase gene—dna methylation alone is not sufficient to prevent protein-binding in vivo. *EMBO J.*, **10**, 2559–2567.
75. Maurano,M.T., Wang,H., Shafer,A., John,S. and Stamatoyannopoulos,J.A. (2013) DNA methylation alone does not cause most cell-type selective transcription factor binding. *Epigenet. Chromatin*, **6**, P103.
76. Maurano,M.T., Wang,H., John,S., Shafer,A., Canfield,T., Lee,K. and Stamatoyannopoulos,J.A. (2015) Role of DNA methylation in modulating transcription factor occupancy. *Cell Rep.*, **12**, 1184–1195.
77. Siersbaek,R., Nielsen,R., John,S., Sung,M.H., Baek,S., Loft,A., Hager,G.L. and Mandrup,S. (2011) Extensive chromatin remodelling and establishment of transcription factor 'hotspots' during early adipogenesis. *EMBO J.*, **30**, 1459–1472.
78. Ernst,J. and Kellis,M. (2013) Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome Res.*, **23**, 1142–1154.