Computational Modeling of Chromatin Accessibility Identified Important Epigenomic Regulators

Yanding Zhao

Baylor College of Medicine

Yadong Dong

Baylor College of Medicine

Wei Hong

Baylor College of Medicine

Chongming Jiang

Baylor College of Medicine

Kevin Yao

Texas A&M University

Chao Cheng (✉ chao.cheng@bcm.edu)

Baylor College of Medicine

Research Article

Keywords: ENCODE, chromatin accessibility, Histone modifications, Transcription factor, Machine learning

Posted Date: September 7th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-859635/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at BMC Genomics on January 8th, 2022. See the published version at <https://doi.org/10.1186/s12864-021-08234-5>.

Abstract

Chromatin accessibility is essential for transcriptional activation of genomic regions. It is well established that transcription factors (TFs) and histone modifications (HMs) play critical roles in chromatin accessibility regulation. However, there is a lack of studies that quantify these relationships. Here we constructed a two-layer model to predict chromatin accessibility by integrating DNA sequence, TF binding, and HM signals. By applying the model to two human cell lines (GM12878 and HepG2), we found that DNA sequences had limited power for accessibility prediction, while both TF binding and HM signals predicted chromatin accessibility with high accuracy. According to the HM model, HM features determined chromatin accessibility in a cell line shared manner, with the prediction power attributing to five core HM types. Results from the TF model indicated that chromatin accessibility was determined by a subset of informative TFs including both cell line-specific and generic TFs. The combined model of both TF and HM signals did not further improve the prediction accuracy, indicating that they provide redundant information in terms of chromatin accessibility prediction. The TFs and HM models can also distinguish the chromatin accessibility of proximal versus distal transcription start sites with high accuracy.

Introduction

Chromatin accessibility is the extent to which nuclear molecules, including TFs, chromatin remodelers and histones, could physically interact with chromatinized DNA[1]. During the development, timely opening of specific genomic regions is essential for transcription activation and maintaining the regular differentiation state in cells[2, 3]. The chromatin remodeling either happens at the promoter locus to directly initiate the gene expression, or in the regulatory regions such as enhancer and silencer to impact the gene expression[4, 5]. To measure the opening genomic regions in cells, many sequencing-based techniques have been developed, with ATAC-seq now being the most widely used technique[6, 7]. Histone modifications (HMs) and transcription factors (TFs) are important determinants in chromatin accessibility[1, 8–10]. HMs regulate the chromatin accessibility by modulating the nucleosome affinity for active chromatin remodelers or changing the local chromatin structure[1, 11]. The combination of HMs, known as histone code, could dramatically change the state of the chromatin, allowing for the bindings of TFs or other active chromatin remodelers[12]. On the other hand, TFs dynamically compete with histones or interact with other chromatin binding proteins to promote the access to DNA[1]. According to the Encyclopedia of DNA elements (ENCODE), TFs play important roles in regulating chromatin accessibility as more than 90% of the accessible genome were bound by TFs[13]. In addition to TFs and HMs, other factors, such as the sequence composition or the enrichment of TF binding motifs, were also found to be associated with chromatin accessibility[14]. Though all associated with chromatin accessibility, there is a lack of quantitative, systematic studies that evaluate the different contributions of each factor during chromatin remodeling.

Many methods have been developed to predict the chromatin accessibility using genomic and epigenomic features. However, most of them only used DNA sequence information as the feature for

prediction, leaving the contribution of other genomic and epigenomic features being un-answered[15–18]. For the limited studies that integrated DNA sequence with TFs for prediction[14], the different contributions between these factors in chromatin accessibility prediction was not examined. In addition, though over 20 HMs have been identified to be associated with chromatin accessibility[11, 19], which HM functions as the most important regulator remains unknown. It's also unclear if HMs perform similarly with TFs or DNA sequence in regulating chromatin accessibility and if HMs and TFs have similar patterns in regulating chromatin accessibility among different cell lines. Therefore, a quantitative chromatin accessibility prediction model that integrates different epigenomic and genomic features is needed.

In order to build an integrative chromatin accessibility prediction model, we used the ChIP-seq data of 11 HMs and over 100 TFs from GM12878 and HepG2 cell lines. These datasets included TF binding profiles, HM patterns across the whole genome, DNA sequence information and 687 TF binding motif counts in accessible and non-accessible regions. The prediction performance of those 4 features was systematically studied and we found the HMs and TFs were the major regulators. Further applying the model revealed that different cell lines shared similar HMs that account for chromatin accessibility, yet only a few common TFs were identified between different cell lines that could predict the chromatin accessibility. In summary, we developed a computational framework for modeling the association between chromatin accessibility and different epigenomic and genomic features. This framework and the models introduced in this work could be applied to different datasets or species to study the regulation of chromatin accessibility.

Results

The integrative model for predicting chromatin accessibility

Our hypothesis was that the sequence, TF motifs, TF binding, and HMs were associated with chromatin accessibility. We validated the hypothesis by examining the signal difference between accessible and non-accessible regions. In both GM12878 and HepG2 cell lines, most of the TFs and HMs presented distinctive patterns between accessible and non-accessible regions. The binding signals for most TFs were enriched in the accessible regions while HM signals were enriched in either accessible or non-accessible regions (Figure 1A-B). In addition, the enrichment score of TFs motifs validated that TF signals were associated with accessible chromatin regions. Compared to the background, most TFs motifs had enrichment score higher than 1, indicating the association between TF motifs and chromatin accessibility (Figure 1A-B).

We then developed a two-layer integrative model to integrate different types of features for predicting chromatin accessibility (Figure 1C). The chromatin accessibility measured by ATAC-seq was used as the training reference, and the peak region was further binned using 100 bp as the window size so that each of the open region had the same length. In the first layer, we used the DNA sequence in the accessible and non-accessible region to train a Convolutional Neural Network (CNN) model. Taking advantage of the CNN model, we converted any given DNA region into the probability of accessibility for sequence feature

extraction. In the second layer, we integrated the derived probability of accessibility from the CNN model (first layer) with TF motif counts, HM, and TF binding signals for chromatin accessibility prediction. Specifically, TF motif counts were calculated based on the occurrences of TF motifs in the accessible and non-accessible region. HM and TF binding signals were derived from the bigwig files, which contained the signal coverage information for the whole genome. A Random Forest model was used to integrate those four categories of features. Meanwhile, we also built chromatin accessibility prediction solely based on TF binding features (TF model), HM features (HM model) or motif features (Motif Model). These models enabled us to quantify the power of each feature category in predicting chromatin accessibility.

HMs and TFs served as major determinants the chromatin accessibility

We first quantified the AUC of each category of features in predicting chromatin accessibility using Random Forest models that incorporate just one specific set of features (TF, HM, TF motif, or DNA seq) (Figure 2A-B). In GM12878 cell line, the highest AUC was obtained through the TF model (AUC=0.84, Figure 2A), followed by the AUC of the HM model (AUC=0.78, Figure 2A). The similar pattern was observed in HepG2 cell line, where TF model and HM model achieved an AUC of 0.84 and 0.79 respectively (Figure 2B). This indicated that HM and TF had comparable power in predicting chromatin accessibility. On the contrary, a much weaker prediction power was observed for models based on sequence or motif features.

Then we examined the prediction performance of using different combinations of features. Our result indicated that the sequence and motif features did not further improve the prediction accuracy of the TF model and the HM model (Figure 2C-D). In addition, when TF and HM features were combined, the resulting models achieved similar accuracy with the TF model (AUC=0.84) and the HM model (AUC=0.78), with the Seq+Motif+TF+HM model having an AUC of 0.86 in GM12878 (Figure 2C). Similar results were observed in HepG2 (Figure 2D). These results suggested that TF and HM features possessed redundant information in terms of chromatin accessibility prediction.

Core histone modifications associated with chromatin accessibility

HM functioned as a major individual determinant of the chromatin accessibility, so it is of interest to determine the contribution of each of the HM and to identify the important HMs for chromatin accessibility. Taking advantage of the Random Forest model, we derived the relative importance of each HM in the GM12878 and HepG2 cell lines respectively. Overall, the function of the HMs is very homogeneous in both cell lines. For example, H2AZ.1, a variant of histone H2A[20,21], was found to be the most important HM in GM12878 cell line and the 4th important HM in HepG2 cell line (Figure 3A-B). In contrast to H2AZ.1, the importance of H3K27me3 was consistently low in both cell lines. It ranked the second least important HM in GM12878 cell line and the least important HM in HepG2 cell line (Figure 3A-B). More importantly, 5 HMs were found to have high relative importance in both two cell lines (Figure 3A-B), suggesting that a core group of HMs was pivotal to the chromatin accessibility. To illustrate the essential role of those 5 HMs, we utilized a breakdown strategy to calculate the AUC after taking out the

most important HMs. The prediction accuracy of the model kept consistently high in both cell lines (AUC > 0.7) when at least one of the 5 core group of HMs were still included in the model (Figure 3C-D). However, after removing all H2AFZ, H3K4me2, H3K27ac, H3K9ac and H3K4me3 from the model, the prediction power of the model was heavily weakened (AUC=0.63, GM12878; AUC=0.67 HepG2, Figure 3C-D).

Different HMs could have very similar function across different cell lines[11,22]. For example, H3K4me3 is consistently associated with transcription start site (TSS) while H3K27ac is associated with enhancers[11]. We hypothesized that HMs also had conserved functions in regulating chromatin accessibility between different cell lines. To test this hypothesis, we first trained the HM model in the HepG2 cell line and achieved an AUC=0.79 by 10-fold cross validation. Then we applied this validated HepG2 model to predict the chromatin accessibility in the GM12878 cell line using the HM signals of GM12878. Shown in Figure 3E, the HepG2 model in the GM12878 cell line achieved similar prediction performance with GM12878 self-prediction performance and outperformed the random constructed model. Similar results can be shown with the same procedure of training the model in the GM12878 cell line dataset and testing on the HepG2 cell line dataset (Figure 3F).

Strong association of H2AZ.1 with chromatin accessibility

In both the HepG2 and GM12878 cell lines, we found H2AZ.1 ranked at the top of the five core HMs regulating chromatin accessibility (Figure 3A-D). H2AZ.1 is a variant of histone H2A and has found to be associated with active promoters and enhancers[20,21]. Its function in regulating the global chromatin accessibility has not been well studied. We quantified the different H2AZ.1 signals in accessible chromatin regions (ACR) verses non-accessible chromatin regions (Non-ACR, Figure 4A-B). The chromatin accessible regions had significantly higher H2AZ.1 signal compared to the inaccessible regions ($P=2.7\text{e-}159$, GM12878; $P=2.1\text{e-}99$, HepG2, Figure 4A-B). And this significant higher pattern was consistent after further separating the accessible regions into proximal and distal regions (Suppl Fig 2). Using H2AZ.1 signal as the predictor could stratify the accessible and inaccessible chromatin regions with high accuracy in both cell lines (AUC=0.74, GM12878; AUC=0.69, HepG2, Figure 4C). These results revealed that H2AZ.1 impacted the chromatin accessibility in a global manner and was not limited to specific regulatory elements region. We further examined the potential synergized HMs with H2AZ.1 by examining the colocalization of H2AZ.1 signals with other HMs. The signals of H2AZ.1 colocalized with the rest of the core HMs in both cell lines with relative high correlation (Figure 4D), indicating the synergistic effect among those 5 core HMs. In addition to those 5 core HMs, there was no strong negative association with other HMs' signals (Figure 4D).

Cell line specific and non-specific TFs predicts chromatin accessibility

Compared with the HM model, the TF model had even better power in predicting chromatin accessibility (Figure 2C-D). Unlike HMs, the roles of TFs are largely heterogeneous. Pioneer factors could be shared among different cell types and directly bind to condensed chromatin to establish accessibility, while other TFs are recruited to the opened chromatin regions and act in a cell line specific manner. To investigate the

roles of different TFs in opening chromatin, we ranked the relative importance of each TF in the GM12878 and HepG2 models. TRIM22, a T cell activation associated transcription factor[23], had the highest relative importance in the GM12878 cell line but not in HepG2, revealing its cell line specific role in regulating the chromatin accessibility (Figure 5A). In addition to the cell line specific TF, we also identified YY1, a structure regulator of enhancer-promoter interactions[24–27], whose relative importance was high in both GM12878 and HepG2 cell lines (Figure 5A). We further performed the similar breakdown analysis and sought to find a group of core TFs. Unlike the pattern of HMs breakdown analysis, the prediction accuracy dropped very smoothly. The model collapsed and presented a notable drop after removing most of the TFs in both cell lines (Figure 5B). In addition, the TFs that caused the drop of the prediction were different in both cell lines, though some shared TFs were identified. For example, in GM12878 cell line, TRIM was removed in the very beginning in the model while it was removed at the very end in the HepG2 cell line. In contrast, YY1 was removed at the early stage in the breakdown analysis in both cell lines (Figure 5B). These results revealed that both shared TFs and cell line specific TFs participate in regulating the chromatin accessibility.

In order to identify the cell line-specific and cell line shared TFs, we plotted the importance of different TFs in the GM12878 and HepG2 cell lines on a scatterplot (Figure 5C). A weak correlation was observed between these two cell lines, supporting the hypothesis that TFs are largely heterogenous in regulating chromatin accessibility in different cell lines (Figure 5C). However, there was a group of TFs (YY1, CREM, ARID3A and BHLHE40) that had high relative importance in both cell lines and therefore were defined as the common TFs (Figure 5C). Additionally, we revealed two groups of TFs with high relative importance but limited to their own cell origins. For example, the TRIM22's high relative importance was only observed in GM12878 cell line. Driven by this, we defined 3 GM12878 specific TFs, 3 HepG2 specific TFs and 4 cell line shared TFs to train the GM12878 specific, HepG2 specific and shared TF models. The predictive power of those TFs was validated through examining if those models could be transferred across the two cell lines. For GM12878 specific TF model, the prediction accuracy dropped from 0.82 to 0.5 when applying the model to HepG2 data (Figure 5D). Similar finding could be observed when applying HepG2 specific TF model to the GM12878 data (Figure 5D). In contrast, the shared TF models could be transferred between HepG2 and GM12878 cell lines (Figure 5E). Either GM12878 or HepG2 common TF models could achieve comparable prediction accuracy in those two cell lines (Figure 5E).

HM and TF features are redundant for predicting chromatin accessibility

Epigenetic studies suggested that HMs play a central role in transcriptional regulation and revealed substantial overlaps between TFs binding and HMs[28]. In our analysis, both the histone model and the TF model achieved good performance in predicting chromatin accessibility (Figure 2-4), leaving an open question whether HMs and TFs binding provide complementary or redundant information to each other in regulating the chromatin accessibility. We first examined the level of colocalization between HMs and TF binding signals through spearman correlation. On the genome of both the GM12878 and HepG2 cell lines, the average HM signals were highly correlated with the average TF binding signals ($\text{Rho}=0.62$, GM12878; $\text{Rho}=0.71$, HepG2, Figure 6A-B). Therefore, we hypothesized that HM signals and TF binding signals

provided redundant information in predicting chromatin accessibility. HMs only, TF binding only and TF-HM combined models were built and compared. In GM12878 cell line, the HMs only model achieved an AUC=0.78 and the TFs only model achieved an AUC=0.84. The AUC didn't increase remarkably when integrating HMs with TFs (AUC=0.86, Figure 6C). The same finding was observed in HepG2 cell line (AUC=0.79, HMs; AUC=0.84, TFs; AUC=0.85, HMs+TFs, Figure 6D). In conclusion, TFs colocalized with HMs when regulating chromatin accessibility and provided redundant information to each other in chromatin accessibility prediction (Figure 6).

Accessible TSS proximal and distal regions presented distinctive epigenomic signals

We have validated the efficacy of our integrative model in predicting the chromatin accessibility. The position of open chromatin varies, which could be further separated into TSS proximal and distal regions where both TFs and HMs were expected to present distinctive signals. Principal component analysis (PCA) showed a separation of TFs and HMs signals in the TSS proximal and distal region (Figure 7A-B). This clear separation pattern introduced a chance of using our current model to predict the accessibility of a specific region. We thus applied our GM12878 TF and HM model defined in the beginning to predict the TSS proximal and distal regions' accessibility. The TF model had AUC=0.97 in predicting both the TSS proximal and distal regions' accessibility (Suppl. Figure 3A-B). Comparable performance was observed using the HM model (AUC=0.92, proximal, AUC=0.85, distal, Suppl. Figure 3A-B). Similar results were observed in the HepG2 cell line (Suppl. Figure 3C-D).

We further designed a detailed model for distinguishing the accessible TSS proximal regions with distal regions using HMs and TF binding signals. Both the TF and HM models obtained decent predicting power in both the GM12878 and HepG2 cell lines (AUC=0.88, TF model in GM12878, AUC=0.89, histone model in GM12878, AUC=0.89, TF model in HepG2, AUC=0.90, histone model in HepG2, Figure 7C-D).

Discussion

In this work, we here pioneered a two-layer quantitative integrative model for the prediction of chromatin accessibility. Two recent studies either used sequence feature or integrated sequence with TF expression or only integrated HMs with TFs signals for chromatin accessibility prediction[11, 14, 29]. Compared with the previous studies, our analysis had such advantages in the following aspects. First, we provided a generalized computational framework for epigenomic data integration. Through the development of the model, we integrated sequence features, which include sequence and TF motifs, with the epigenomic features, which include ChIP-seq profiles of HMs and TFs, for chromatin accessibility prediction. Second, we systematically evaluate the contribution of different groups of features in predicting chromatin accessibility and identify HMs and TFs as the major determinants. Third, we generated feature specific prediction model to explore the mechanism that HMs and TFs played in regulating chromatin accessibility.

As early in 1960, Allfrey et al. identified the acetylation of histones, leading to the hypothesis that acetylation is associated with transcription activity[30]. In the past decades, tremendous efforts have

been made to illustrate the cellular functions of HMs[11]. TFs, generally being classified into pioneer factors or non-pioneer factors[31], have also been thought to be essential to activate and maintain the chromatin accessibility[1]. Though the association between individual HM and TF with open chromatin has been studied for decades[11], the relative importance of each HM and TF in regulating chromatin accessibility was not addressed in a global manner. In our analysis, we first identified TFs and HMs served as the major determinants of chromatin accessibility (Fig. 2). Additionally, TFs and HMs cooperated together to regulate the chromatin accessibility (Fig. 6). It was well known that TFs interacts with HMs for regulating gene expression[32, 33]. Our analysis indicated this interaction happened beyond the TSS range by observing the high correlation structure in the genome wide level (Fig. 6). And it is not surprising to observe the prediction redundancy between HMs and TFs (Fig. 6). The mechanisms behind TFs and HMs regulation are complicated and multiple models have been proposed. For example, the TFs could interact with chromatin remodeler directly to create an accessible region whereas HMs are simply the subsequent readout. However, the power of our analysis was limited by the number of ChIP seq files provided in the ENCODE dataset. With more ChIP seq files released in the future, we could further investigate the relationship between pioneer TFs, secondary TFs, chromatin remodeler and HMs in regulating chromatin accessibility.

We identified 4 important HMs and 1 histone variants which were all reported to be associated with active gene transcription or enhancers[1] (Fig. 3). Interestingly, the histone H2A variant H2AZ.1, which was found to be enriched in active promoters and enhancers[20, 21], was similar or more important in regulating chromatin accessibility compared to other well-known HMs (Fig. 4). This supports the theory that histone variants, though very similar in their major canonical histone counterparts, could have strong functional impacts[21, 34]. With more ChIP-seq files of histone variants being available, we could systematically compare the importance of histone variants with HMs in chromatin accessibility regulation. Interestingly, the importance of these 5 core HMs in regulating chromatin accessibility is shared between GM12878 and HepG2 cell lines. The prediction model that trained either in GM12878 or HepG2 cell line could be easily transferred across both cell lines, supporting the existence of a common histone code for chromatin accessibility[12] or at least those two cell lines shared the similar histone code.

Unlike the core HMs, different TFs regulate chromatin accessibility in different manners among cell lines. In our analysis, there was not a group of core TFs that were important for chromatin accessibility in both cell lines. We therefore developed a computational framework to identify pioneer TFs and cell line specific TFs. Cell line specific TFs such as IKZF1 only predict the chromatin accessibility in one cell line, though the ChIP-seq profiles of these cell line specific TFs were also measured in the other cell line. Shared TFs between different cell lines also existed as the prediction model could be transferred (Fig. 5). These shared TFs could function as potential pioneer factors due to their similar power in predicting chromatin accessibility across different cell lines (Fig. 5). However, only YY1 has been reported as the global regulator for promoter-enhancer interactions[24–27]. To our knowledge, the function of BHLHE40, ARID3A and CREM has not been reported as the global regulator yet. It is possible that those 3 TFs might

serve as the novel pioneer TFs. To further validate this, in the future, we could examine other cell lines which contained the ChIP-seq profiles of those 3 TFs for chromatin accessibility prediction.

Though ATAC-seq is the current golden standard for directly measuring chromatin accessibility, there were numerous chromatin accessibility data that were generated from other sequencing techniques[35–37]. Our model was not only able to predict chromatin accessibility measured by ATAC-seq, but also effectively predicted the accessibility that was measured by DNase- and FAIRE-seq data (Suppl. Figure 4). With more data available, for example, the MNase-seq data[38], we could further validate the efficacy of our model. Notably, our model provided a novel way of re-utilizing the traditional ChIP-seq data. In ENCODE project, over 70 cell lines have the matched ChIP-seq data while only 11 cell lines provide ATAC-seq data. With the application of our framework, the chromatin accessibility of those cell lines could be inferred instead of performing additional ATAC-seq experiments.

Conclusions

In summary, we developed a computational framework for chromatin accessibility modeling. Several important HMs and TFs were found as the major regulators of chromatin accessibility, providing novel mechanistic insights in chromatin accessibility regulation. The computational framework and the introduced integrative model in our study can be generally applied to for modeling different genomic and epigenomic data.

Abbreviations

HM
Histone modification
TF
Transcription factor
TSS
Transcription start site
ACR
Accessible chromatin region
Non-ACR
Non-accessible chromatin region
CNN
Convolutional Neural Network
ENCODE
the Encyclopedia of DNA elements

Declarations

Ethics approval and consent to participate: Not Applicable

Consent for publication: Chao Cheng, Yadong Dong, Wei Hong, Chongming Jiang, Kevin Yao and Yanding Zhao

Availability of data and materials: Not Applicable

Competing interests: No

Funding: This study is supported by the Cancer Prevention Research Institute of Texas (CPRIT) (RR180061 to CC) and the National Cancer Institute of the National Institutes of Health (1R21CA227996 to CC). CC is a CPRIT Scholar in Cancer Research.

Authors' contributions: Conception and design: Yanding Zhao and Chao Cheng

Development of methodology: Chao Cheng and Yanding Zhao

Analysis and interpretation of data: Chao Cheng, Yadong Dong, Wei Hong and Yanding Zhao

Writing, review, and/or revision of the manuscript: Chao Cheng, Yadong Dong, Wei Hong, Chongming Jiang, Kevin Yao and Yanding Zhao

Acknowledgements: Not Applicable

Methods

Datasets Collection and processing

To construct integrative models to predict chromatin accessibility, we utilized the GM12878 and HepG2 cell lines, for which high quality chromatin accessibility, TF binding and histone modification have been generated by the Encyclopedia of DNA elements (ENCODE) project[39]. The chromatin accessibility data were obtained by using the ATAC-seq, DNase-seq and FAIRE-seq experiments and downloaded from the ENCODE database as bedgraph files. Specifically, for the GM12878 cell line, the accession IDs are ENCFF172DEA, ENCFF235KUD, and ENCFF001UYE for ATAC-seq, DNase-seq, and FAIRE-seq, respectively. For the HepG2 cell line, the accession IDs are ENCFF356TXH, ENCFF422EDI, and ENCFF001UYN for ATAC-seq, DNase-seq, and FAIRE-seq, respectively.

The ChIP-seq profiles of histone modifications and TFs were also downloaded from ENCODE as bigwig files. The GM12878 data contain 53 files with replicates for 11 different histone modifications; and 361 files with replicates for 97 different TFs. The HepG2 data contain 52 files with replicates for 11 different histone modifications; and 763 files with replicates for 221 different TFs. All files were mapped to the hg19 genome. The annotation of these ChIP-seq files is summarized in Suppl. Table 1.

Identification of chromatin accessible and non-accessible regions

The ATAC-seq bedgraph files were used to define the **chromatin accessible** (positive) and **non-accessible** (negative) regions in the genome. Peaks in the ATAC-seq represent regions in the chromatin that are highly accessible to Tn5 transposase and correlate with transcription activity. The positive region was defined as the genomic regions covered by ATAC-seq peaks listed in the bedgraph files. The negative region was defined as the neighbor region at both sides of each ATAC-seq peak with the same lengths. Namely, the positive and negative regions had exactly the same bases length with number of positive-to-negative ratio of 1:2. In addition, the negative regions shared similar sequence features such as GC content with the corresponding positive region due to genomic proximity, making them ideal negative controls to the chromatin accessible regions. The sizes of different regions vary dramatically and therefore we divided each region (both positive and negative) into small bins of 100 bp in size, which is around the half-size of DNA covered by a single nucleosome[40]. The models were constructed at the bin level.

Calculation of HM and TF binding signal profiles

For each bin, we calculated the average TF binding and HM signals based on the corresponding bigwig files, which provided the signals at nucleotide level, represented as the log ratios (the immunoprecipitation sample with respect to the control). The Deeptools software was used for calculating bin-specific HMs and TFs binding signals in both the positive and negative regions[41]. Specially, the *multiBigwigSummary* function was used with parameter “-bs 100” to calculate the average signal of each ChIP-seq file in a 100 bp window across the genome[41].

Identification of TF motifs in genomic regions

DNA sequences for positive and negative regions were retrieved from the hg19 reference genome sequence downloaded from Ensembl website[42]. The occurrences of TF binding motifs were determined by using the FIMO program from the MEME Suite[43], which identified motifs by matching with positional weighted matrices (PWMs) of TFs. A total of 687 PWMs were used, which were downloaded from JASPAR and TRANSFAC database[44,45].

A two-layer integrative model for chromatin accessibility prediction

A two-layer classification model was constructed to integrate sequence, TF motifs, binding of TF, and HMs to predict chromatin accessibility at the bin level (100-bp bins). The inputs to this model are DNA sequence of each bin, the presence of 687 TF binding motifs (1 and 0 indicate presence and absence, respectively), TF binding signals (derived from ChIP-seq data), and HM signals (derived from ChIP-seq data). The output of this model is the accessibility of bins, with 1 and 0 indicating accessible and non-accessible, respectively.

In the first layer, a convolutional neural network (CNN) model was constructed to predict the baseline accessibility at the bin level solely based on the DNA sequences (100 bp) of bins since the CNN model has been widely used for modeling sequence based features[46]. The configuration of the CNN model is

shown in Suppl. Fig 1. Each of our convolutional layer contained 320, 480, and 640 hidden neurons, and the output of each convolutional layer was activated by the ReLU function before propagating to the next maxpooling or convolutional layers. We implemented a fully connected layer with ReLU activation on top of the three convolutional layers, which was further propagated to the output sigmoid layer to compute the probability of a given input sequence having an open chromatin state. The raw DNA sequences were encoded using “One-hot” encoding scheme as the input to the model, where each base pair in the sequence is mapped to the four possible nucleotides, with the corresponding nucleotide set as 1 and the other three set as 0. The output of this model is a score ranging between [0,1], indicating the probability of the bin to be accessible at the baseline level (i.e., fully based on its 100-bp DNA sequence). In the second layer, the baseline accessibility score was combined with TF motifs, binding of TF and HMs to predict the two-class bin labels (accessible vs. non-accessible bins). These features are integrated by using the Random Forest classification model. Many TFs and HMs have multiple ChIP-seq replicates and we chose the one with the largest sequencing depths as the predicting feature so that for each TF/HM a single feature was used for easy model interpretation.

To train the model, we used the ATAC-seq data for GM12878 and HepG2 cell lines. The positive (accessible) and negative (non-accessible) bins are defined as described in section **“Identification of chromatin accessible and non-accessible regions”**. The model was evaluated by 10-folds cross-validation using AUC (Area under ROC curve) score as the metric, which summarizes the specificity and sensitivity of classification models at different thresholds. Specifically, the whole data were divided into 10 subsets of equal sizes. Each time, 9 subsets were selected and combined as the training data and the remaining subset was used as the test data. The model was trained in the training data and then applied to the test data to predict the probability of bins to be accessible. This procedure was iterated until every bin had been in a test data and predicted exactly once. The predicted values are then used to calculate the AUC to evaluate the model performance.

Model optimization and stability analysis

To optimize the second layer RF model, we applied a backward feature selection method. Initiated from the full model including all features, the feature with the least relative importance was removed iteratively. In each iteration, the performance of the model (AUC) was calculated based on cross-validation. The optimized model was determined by visualizing the “number of features” versus AUC plot. A similar procedure was used to determine the minimum number of features required to achieve fairly high prediction accuracy. Features were removed from the full model one by one, but each time the one with the largest relative importance was removed until a sudden decrease of cross-validation AUC was observed.

Model for classifying TSS-proximal versus distal accessible chromatin regions

It is known that both TF binding and HM signals have spatial patterns. Some HMs are enhancer associated while others are promoter associated. Therefore, we constructed models to classify transcription start site (TSS) proximal versus distal accessible chromatin regions. ChIPseeker package

was used to categorize the accessible regions[47] into TSS-proximal and TSS-distal regions. The former was defined as accessible chromatin regions near the +/- 3000 bp of TSS, while the latter as the regions having no overlap with TSS proximal regions or exon regions. Using the TF binding and HM signals as the predictors, a Random Forest model was constructed to classify the two categories of accessible regions.

References

- [1] Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* 2019;20:207–20. <https://doi.org/10.1038/s41576-018-0089-8>.
- [2] Tsompana M, Buck MJ. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin* 2014;7:33. <https://doi.org/10.1186/1756-8935-7-33>.
- [3] Li D, Shu X, Zhu P, Pei D. Chromatin accessibility dynamics during cell fate reprogramming. *EMBO Rep* 2021;22:e51644. <https://doi.org/10.15252/embr.202051644>.
- [4] Sundaramoorthy R, Owen-Hughes T. Chromatin remodelling comes into focus. *F1000Research* 2020;9. <https://doi.org/10.12688/f1000research.21933.1>.
- [5] Clapier CR, Cairns BR. The biology of chromatin remodeling complexes. *Annu Rev Biochem* 2009;78:273–304. <https://doi.org/10.1146/annurev.biochem.77.062706.153223>.
- [6] Sun Y, Miao N, Sun T. Detect accessible chromatin using ATAC-sequencing, from principle to applications. *Hereditas* 2019;156:29. <https://doi.org/10.1186/s41065-019-0105-9>.
- [7] Shashikant T, Etensohn CA. Genome-wide analysis of chromatin accessibility using ATAC-seq. *Methods Cell Biol* 2019;151:219–35. <https://doi.org/10.1016/bs.mcb.2018.11.002>.
- [8] Esmaeili M, Blythe SA, Tobias JW, Zhang K, Yang J, Klein PS. Chromatin accessibility and histone acetylation in the regulation of competence in early development. *Dev Biol* 2020;462:20–35. <https://doi.org/10.1016/j.ydbio.2020.02.013>.
- [9] Andersson R, Sandelin A. Determinants of enhancer and promoter activities of regulatory elements. *Nat Rev Genet* 2020;21:71–87. <https://doi.org/10.1038/s41576-019-0173-8>.
- [10] Friman ET, Deluz C, Meireles-Filho AC, Govindan S, Gardeux V, Deplancke B, et al. Dynamic regulation of chromatin accessibility by pluripotency transcription factors across the cell cycle. *ELife* 2019;8. <https://doi.org/10.7554/eLife.50087>.
- [11] Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. *Cell Res* 2011;21:381–95. <https://doi.org/10.1038/cr.2011.22>.

- [12] Strahl BD, Allis CD. The language of covalent histone modifications. *Nature* 2000;403:41–5. <https://doi.org/10.1038/47412>.
- [13] Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. *Nature* 2012;489:75–82. <https://doi.org/10.1038/nature11232>.
- [14] Nair S, Kim DS, Perricone J, Kundaje A. Integrating regulatory DNA sequence and gene expression to predict genome-wide chromatin accessibility across cellular contexts. *Bioinforma Oxf Engl* 2019;35:i108–16. <https://doi.org/10.1093/bioinformatics/btz352>.
- [15] Wang C, Li J. A Deep Learning Framework Identifies Pathogenic Noncoding Somatic Mutations from Personal Prostate Cancer Genomes. *Cancer Res* 2020;80:4644–54. <https://doi.org/10.1158/0008-5472.CAN-20-1791>.
- [16] Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* 2015;47:955–61. <https://doi.org/10.1038/ng.3331>.
- [17] Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 2015;12:931–4. <https://doi.org/10.1038/nmeth.3547>.
- [18] Zhou J, Park CY, Theesfeld CL, Wong AK, Yuan Y, Scheckel C, et al. Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat Genet* 2019;51:973–80. <https://doi.org/10.1038/s41588-019-0420-0>.
- [19] Zhao Y, Garcia BA. Comprehensive Catalog of Currently Documented Histone Modifications. *Cold Spring Harb Perspect Biol* 2015;7:a025064. <https://doi.org/10.1101/cshperspect.a025064>.
- [20] Giaimo BD, Ferrante F, Herchenröther A, Hake SB, Borggrefe T. The histone variant H2A.Z in gene regulation. *Epigenetics Chromatin* 2019;12:37. <https://doi.org/10.1186/s13072-019-0274-9>.
- [21] Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. *Nat Rev Genet* 2016;17:487–500. <https://doi.org/10.1038/nrg.2016.59>.
- [22] Koch CM, Andrews RM, Flicek P, Dillon SC, Karaöz U, Clelland GK, et al. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res* 2007;17:691–707. <https://doi.org/10.1101/gr.5704207>.
- [23] Obad S, Olofsson T, Mechti N, Gullberg U, Drott K. Regulation of the interferon-inducible p53 target gene TRIM22 (Staf50) in human T lymphocyte activation. *J Interferon Cytokine Res Off J Int Soc Interferon Cytokine Res* 2007;27:857–64. <https://doi.org/10.1089/jir.2006.0180>.
- [24] Weintraub AS, Li CH, Zamudio AV, Sigova AA, Hannett NM, Day DS, et al. YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* 2017;171:1573-1588.e28. <https://doi.org/10.1016/j.cell.2017.11.008>.

- [25] Han J, Meng J, Chen S, Wang X, Yin S, Zhang Q, et al. YY1 Complex Promotes Quaking Expression via Super-Enhancer Binding during EMT of Hepatocellular Carcinoma. *Cancer Res* 2019;79:1451–64. <https://doi.org/10.1158/0008-5472.CAN-18-2238>.
- [26] Wang J, Wu X, Wei C, Huang X, Ma Q, Huang X, et al. YY1 Positively Regulates Transcription by Targeting Promoters and Super-Enhancers through the BAF Complex in Embryonic Stem Cells. *Stem Cell Rep* 2018;10:1324–39. <https://doi.org/10.1016/j.stemcr.2018.02.004>.
- [27] Beagan JA, Duong MT, Titus KR, Zhou L, Cao Z, Ma J, et al. YY1 and CTCF orchestrate a 3D chromatin looping switch during early neural lineage commitment. *Genome Res* 2017;27:1139–52. <https://doi.org/10.1101/gr.215160.116>.
- [28] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;489:57–74. <https://doi.org/10.1038/nature11247>.
- [29] Cai Y, González JV, Liu Z, Huang T. Computational systems biology methods in molecular biology, chemistry biology, molecular biomedicine, and biopharmacy. *BioMed Res Int* 2014;2014:746814. <https://doi.org/10.1155/2014/746814>.
- [30] Allfrey VG, Faulkner R, Mirsky AE. ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS. *Proc Natl Acad Sci U S A* 1964;51:786–94. <https://doi.org/10.1073/pnas.51.5.786>.
- [31] Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev* 2011;25:2227–41. <https://doi.org/10.1101/gad.176826.111>.
- [32] Zhang L, Xue G, Liu J, Li Q, Wang Y. Revealing transcription factor and histone modification co-localization and dynamics across cell lines by integrating ChIP-seq and RNA-seq data. *BMC Genomics* 2018;19:914. <https://doi.org/10.1186/s12864-018-5278-5>.
- [33] Cheng C, Gerstein M. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res* 2012;40:553–68. <https://doi.org/10.1093/nar/gkr752>.
- [34] Henikoff S, Smith MM. Histone variants and epigenetics. *Cold Spring Harb Perspect Biol* 2015;7:a019364. <https://doi.org/10.1101/cshperspect.a019364>.
- [35] Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res* 2007;17:877–85. <https://doi.org/10.1101/gr.5533506>.
- [36] Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc* 2010;2010:pdb.prot5384. <https://doi.org/10.1101/pdb.prot5384>.

- [37] Minnoye L, Marinov GK, Krausgruber T, Pan L, Marand AP, Secchia S, et al. Chromatin accessibility profiling methods. *Nat Rev Methods Primer* 2021;1:10. <https://doi.org/10.1038/s43586-020-00008-9>.
- [38] Cui K, Zhao K. Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-Seq. *Methods Mol Biol Clifton NJ* 2012;833:413–9. https://doi.org/10.1007/978-1-61779-477-3_24.
- [39] Zhao Y, Schaafsma E, Cheng C. Applications of ENCODE data to Systematic Analyses via Data Integration. *Curr Opin Syst Biol* 2018;11:57–64. <https://doi.org/10.1016/j.coisb.2018.08.010>.
- [40] McGinty RK, Tan S. Nucleosome structure and function. *Chem Rev* 2015;115:2255–73. <https://doi.org/10.1021/cr500373h>.
- [41] Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 2016;44:W160-165. <https://doi.org/10.1093/nar/gkw257>.
- [42] Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. *Nucleic Acids Res* 2021;49:D884–91. <https://doi.org/10.1093/nar/gkaa942>.
- [43] Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinforma Oxf Engl* 2011;27:1017–8. <https://doi.org/10.1093/bioinformatics/btr064>.
- [44] Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* 2020;48:D87–92. <https://doi.org/10.1093/nar/gkz1001>.
- [45] Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 2003;31:374–8. <https://doi.org/10.1093/nar/gkg108>.
- [46] Liu Q, Xia F, Yin Q, Jiang R. Chromatin accessibility prediction via a hybrid deep convolutional neural network. *Bioinforma Oxf Engl* 2018;34:732–8. <https://doi.org/10.1093/bioinformatics/btx679>.
- [47] Yu G, Wang L-G, He Q-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinforma Oxf Engl* 2015;31:2382–3. <https://doi.org/10.1093/bioinformatics/btv145>.

Figures

Figure 1

Integrative model configuration. (A) and (B). Heatmap showing the signal difference of TFs and HMs in accessible and non-accessible regions and the enrichment of TF motifs in accessible regions in GM12878 (A) and HepG2 (B). (C). The chromatin accessibility of GM12878 and HepG2 data were acquired from ENCODE database. The DNA-sequence in the accessible region was extracted and embedded into the CNN model to convert the DNA sequence information to the probability of accessibility (Layer 1 model). The resulting probability was next integrated with ChIP-seq signal of HMs and TFs and the TFs motif occurrence in the accessible regions to train a Random Forest model.

Figure 2

Histone and TF determines the chromosome accessibility. (A) and (B). ROC curves of using DNA sequence, TF motif occurrence, HMs signal and TFs binding signal as features for chromatin accessibility prediction in GM12878 (A) and HepG2 (B) cell lines. (C) and (D). Barplots showing the AUC score of using combined DNA sequence + Motif, DNA sequence + HM, DNA sequence + TF and DNA

sequence + Motif + HM + TF as features for chromatin accessibility prediction in GM12878 (C) and HepG2 (D) cell lines.

Figure 3

Histone models identify five core histones with strong effects on chromosome accessibility. (A) and (B). Boxplots indicating the relative importance of each histone modifications in chromatin accessibility prediction in GM12878 (A) and HepG2 (B) cell lines. (C) and (D). Barplots showing the AUC score change

of taking out each HM for chromatin accessibility prediction in GM12878 (C) and HepG2 (D) cell lines. (E). Barplots showing the AUC score of HepG2 model in GM12878 HM signals and GM12878 model in GM12878 HM signals for GM12878 chromatin accessibility prediction. (F). Barplots showing the AUC score of GM12878 model in HepG2 HM signals and HepG2 model in HepG2 HM signals for HepG2 chromatin accessibility prediction.

Figure 4

H2AZ.1 serves as major histone modifications in chromatin accessibility regulation. (A) and (B). Boxplots indicating the H2AZ.1 signal difference in ACRs and non-ACRs. P value was calculated by Wilcoxon rank sum test. (C). ROC curves of using H2AZ.1 as feature for chromatin accessibility prediction in GM12878 and HepG2 cell lines. (D). Heatmap showing the correlation between H2AZ.1 signals and other HM signals. The correlation coefficient was calculated by Spearman Correlation.

Figure 5

TF models identify cell line specific and shared specific TFs with strong effects on chromosome accessibility. (A). Boxplots indicating the relative importance of each TFs binding signals in chromatin accessibility prediction in GM12878 (up panel) and HepG2 (bottom) cell lines. (B). Barplots showing the AUC score change of taking out each TF binding signal for chromatin accessibility prediction in GM12878 (up panel) and HepG2 (bottom) cell lines. (C). Scatterplot showing the consistency of TFs

relative importance between GM12878 and HepG2 cell lines. The circles indicated GM12878 specific TFs, HepG2 specific TFs and common TFs. (D). Barplots showing the AUC score of GM12878 specific TF model with GM12878 TFs binding signals for GM12878 chromatin accessibility prediction and GM12878 specific TF model with HepG2 TFs binding signals for HepG2 chromatin accessibility prediction (left panel). Barplots showing the AUC score of HepG2 specific TF model with HepG2 TFs binding signals for HepG2 chromatin accessibility prediction and HepG2 specific TF model with GM12878 TFs binding signals for GM12878 chromatin accessibility prediction (right panel). (E). Barplots showing the AUC score of GM12878 common TF model with HepG2 TFs binding signals for HepG2 chromatin accessibility prediction and GM12878 specific TF model with HepG2 TFs binding signals for HepG2 chromatin accessibility prediction (left panel). Barplots showing the AUC score of HepG2 common TF model with GM21878 TFs binding signals for GM12878 chromatin accessibility prediction and GM12878 common TF model with GM12878 TFs binding signals for GM12878 chromatin accessibility prediction (right panel).

Figure 6

Redundancy between HM and TF features in determining chromosome accessibility. (A) and (B). Scatterplot showing the correlation between histone modifications signals and TFs binding signals in GM12878 (A) and HepG2 (B) cell lines. The correlation coefficient was calculated by Spearman correlation. (C) and (D). Barplots showing the AUC score of HM model, TF model and the combined HM with TF model in GM12878 (C) and HepG2 (D) cell lines.

Figure 7

Accessible of TSS proximal and distal regions present distinctive epigenomic signals. (A) and (B). PCA plot showing the separation of accessible TSS proximal region versus accessible distal region using TFs binding signals (A) and HMs (B). (C) and (D). Barplots showing the AUC scores of using TFs binding signals and histone modifications signals for accessible proximal region versus distal region prediction in GM12878 (C) and HepG2 (D).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplFigure1.pdf](#)
- [SupplFigure2.pdf](#)
- [SupplFigure3.pdf](#)
- [SupplFigure4.pdf](#)
- [SupplTable1.xlsx](#)