Computational and Structural Biotechnology Journal 21 (2023) 3024–3031

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research article

Benefiting from the intrinsic role of epigenetics to predict patterns of CTCF binding

Camilo Villaman [a,b], Gianluca Pollastri [c], Mauricio Saez [d,e,*], Alberto J.M. Martin [b,**]

[a] Programa de Doctorado en Genómica Integrativa, Vicerrectoría de Investigación, Universidad Mayor, Santiago, Chile
[b] Laboratorio de Redes Biológicas, Centro Científico y Tecnológico de Excelencia Ciencia & Vida, Fundación Ciencia & Vida, Escuela de Ingeniería, Facultad de Ingeniería, Arquitectura y Diseño, Universidad San Sebastián, Santiago, Chile
[c] School of Computer Science, University College Dublin, Dublin, Ireland
[d] Centro de Oncología de Precisión, Facultad de Medicina y Ciencias de la Salud, Universidad Mayor, Santiago, Chile
[e] Laboratorio de Investigación en Salud de Precisión, Departamento de Procesos Diagnósticos y Evaluación, Facultad de Ciencias de la Salud, Universidad Católica de Temuco, Chile

ARTICLE INFO

Article history:
Received 19 December 2022
Received in revised form 11 May 2023
Accepted 11 May 2023
Available online 12 May 2023

Keywords:
CTCF
Binding Prediction
Histone Marks
Random Forests

ABSTRACT

Motivation: One of the most relevant mechanisms involved in the determination of chromatin structure is the formation of structural loops that are also related with the conservation of chromatin states. Many of these loops are stabilized by CCCTC-binding factor (CTCF) proteins at their base. Despite the relevance of chromatin structure and the key role of CTCF, the role of the epigenetic factors that are involved in the regulation of CTCF binding, and thus, in the formation of structural loops in the chromatin, is not thoroughly understood.

Results: Here we describe a CTCF binding predictor based on Random Forest that employs different epigenetic data and genomic features. Importantly, given the ability of Random Forests to determine the relevance of features for the prediction, our approach also shows how the different types of descriptors impact the binding of CTCF, confirming previous knowledge on the relevance of chromatin accessibility and DNA methylation, but demonstrating the effect of epigenetic modifications on the activity of CTCF. We compared our approach against other predictors and found improved performance in terms of areas under PR and ROC curves (PRAUC-ROCAUC), outperforming current state-of-the-art methods.

1. Introduction

The three dimensional structure of chromatin is one of the key elements that determines how gene expression is controlled. There are several mechanisms involved in the maintenance and determination of chromatin structure [1]. Among these, one of the most relevant is the formation of structural loops that are also related with the conservation of chromatin states [2]. Most of these loops have two copies of the CCCTC-binding factor, also known as CTCF protein, at their base [3]. Despite the relevance of chromatin structure and the key role of CTCF, very little is still known about the epigenetic factors that are involved in the regulation of CTCF binding, and thus, in the formation of structural loops in the chromatin. CTCF is a zinc finger protein expressed ubiquitously in most vertebrate tissues, initially characterized as a negative regulator of the c-myc gene [4,5]. CTCF is capable of DNA binding due to the action of its 11 zinc fingers which can bind to multiple conserved binding sites [6,7] that are located along the whole genome [8] but predominantly on intergenic regions [8]. Importantly, CTCF can also bind to less conserved DNA sequences, a fact that led to the discovery of cell-specific patterns of CTCF binding sites predominantly located inside introns [9]. CTCF is the most important insulator reported in vertebrates [10] and it has a prominent role as a chromatin architecture regulator mediating different epigenetic and molecular functions. CTCF modifies the genetic expression landscape as a repressor or as a transcriptional activator due to its architectural role, and it is capable of acting as a chromatin insulator, interfering directly between enhancers, silencers, and promoters [6,11]. CTCF is also involved in gene imprinting, chromosome X inactivation and conservation of unmethylated regions in the whole genome [12,13].

* Corresponding author at: Centro de Oncología de Precisión, Facultad de Medicina y Ciencias de la Salud, Universidad Mayor, Santiago, Chile.
** Corresponding author.
E-mail addresses: mauricio.saez@uct.cl (M. Saez), alberto.martin@uss.cl (A.J.M. Martin).

https://doi.org/10.1016/j.csbj.2023.05.012

Another functional role for CTCF is its involvement on the establishment and maintenance of topologically associated domains (TADs), genomic regions of self-interaction with increased intradomain contacts [14]. CTCF can also form chromatin loops due to the interaction of two CTCF-bound domains inside a TAD located at distant locations in the whole genome [3,15]. However, CTCF is only capable of forming loops when the bound CTCF binding sites are located in a convergent orientation, while divergent CTCF sites are unable to make loops [16]. CTCF is also capable of interacting with cohesin to allow proper loop stabilization, but both CTCF and cohesin have independent roles coordinating chromatin organization [17]. Since CTCF is associated with gene regulation, abnormal CTCF binding is related to different malignancies, such as leukemia [18], gastrointestinal cancer [19], lung cancer [20], cervical carcinoma [21], and other diseases. The role of CTCF in disease is not limited to gene regulation, as CTCF can prevent DNA methylation and the spreading of inhibitory histone marks in the promoter regions of tumor suppressor genes, and loss of CTCF binding can lead to an epigenetic silencing [22].

As changes in the CTCF landscape lead to disease, understanding the circumstances that determine CTCF-DNA binding will shed light into several disorders. There are many experimental and computational approaches that can provide new insights into how different proteins display different binding profiles on DNA, but the whole set of interrelationships between all factors involved in protein binding to specific DNA regions and promoters are not thoroughly known [23]. In this context, the determination of protein and DNA binding is an important but yet unsolved problem in computational biology and different strategies have been implemented to deal with this issue. Chromatin immunoprecipitation followed by sequencing [24] (ChIP-seq) is one of the most common experimental techniques used to obtain protein-binding profiles in the whole genome. Nonetheless, ChIP-seq experiments are expensive and experimentally complex [25] and it is impossible to test every cell type and tissue under every possible biological condition, highlighting the need for computational approaches to complement experimental results [26]. To solve this issue, machine learning algorithms have been developed and applied to the problem of in vivo prediction of transcription factor binding [27].

Machine Learning involves the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyze and draw inferences from patterns in data. Machine Learning algorithms are capable of generating mathematical models using datasets ("training sets") to make predictions or decisions without being specifically coded to implement a task. Different Machine Learning approximations have been used to predict the union of a transcription factor to DNA and there is no specific CTCF binding predictor at the date of elaboration of this manuscript, however, different CTCF loop predictors have been developed. Kai et al. [28] trained a Random Forest (RF) predictor integrating sequence and epigenetic features to predict CTCF-mediated loops from ChIP-seq data. Zhang et al. [29] used word2vec to predict if two convergent CTCF binding sites were able to form a loop using only sequence-related features. Lv H et al. [30] developed a neuronal convolutional model that integrates k-tuples of nucleotide frequencies, position conservation, position score and natural vector features to predict CTCF loops. Wang al [31] used a two-step RF model to predict CTCF loops. However, each method is dependent on earlier CTCF occupancy information, requiring ChIP-seq information or similar to generate predictions.

Here we describe a CTCF binding predictor based on RF that employs different epigenetic data and genomic features. Importantly, given the ability of the RF to determine the relevance of features for the prediction, our approach also shows how the different types of descriptors impact the binding of CTCF, confirming previous knowledge on the relevance of chromatin accessibility and DNA methylation, but demonstrating the effect of epigenetic modifications on the activity of CTCF.

2. Methods

2.1. Dataset collection and processing

The bisulfite sequencing data of DNA methylation, the ChIP-Seq data of histone modifications (H3K9ac, H3K27ac, H3K4me3, H3K4me2, H3K4me1, H3K79me2, H3K9me3, H3K27me3, H3K36me3, H4K20me1 and H2AFZ), the data from assays for transposase-accessible chromatin using sequencing (ATAC-Seq), and ChIP-Seq data of CTCF for all four cell lines (GM12878, K562, HeLa and SK-N-SH) were downloaded from the ENCODE project (http://genome.ucsc.edu/ENCODE/downloads.html). We used FIMO [32] (default values, –alpha 1.0 –max-stored-scores 100000 –motif all –motif-pseudo 0.1 –thresh 1e-4) to identify CTCF binding motifs in the whole genome using the JASPAR [33] CTCF matrix profile (MA0139.1) and the GRCh38 XY Human Reference Genome. All data types were used as available at the ENCODE web page or processed using exactly the same protocol reported by them if needed.

2.2. Division of genome regions around CTCF motif

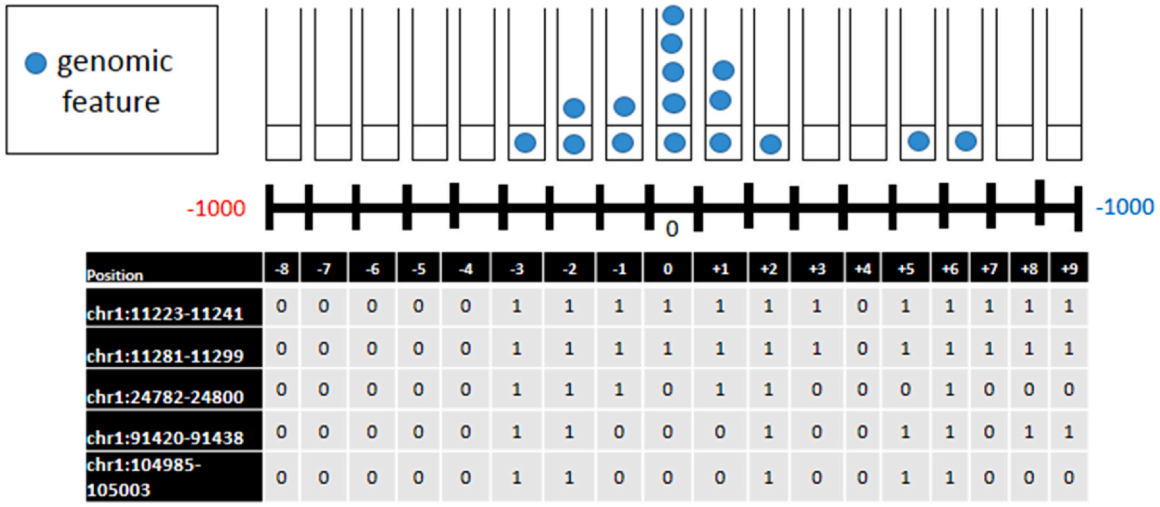
We next selected a window centered on each FIMO-predicted CTCF binding site. This window was divided into bins using different sizes. Bins were encoded as binary vectors that indicated the presence or absence of a feature overlapping with it, for example another CTCF binding motif, each of the histone modifications, DNA methylation or if the bin was annotated as being accessible (Fig. 1). Vectors for all bins for a CTCF binding site were joined into a single vector describing the site and its surroundings up and downstream. We tested different window and bin sizes (Supplementary Methods) and found that in terms of performance, a window size of 2kbp around the CTCF binding site with a bin size of 25 bp yielded better performance in terms of F-Score on the GM12878 cell line (Fig. S1).

2.3. Random Forests

We used Scikit-learn [34] to split each cell line matrix into train and test sets as follows: 2/3rds of the data were used as a training set, while the remaining 1/3rd of the data was used as a test set. We tested each cell line against itself, and used the CTCF ChIP-Seq data from each cell line as truth values. We also used Scikit-learn to generate the RFs and reported Precision, Recall and F-Score as a measure of performance (Table 1, see formulas below). Afterwards, we picked 3 of the 4 cell lines as a training set and tested against the remaining cell line, while using the same performance measures (Table 2). We also calculated ROC and PR curves, and reported the AUC (Fig. 2). To properly understand which features were relevant, we obtained the relative importance of each feature and plotted it using R [35] and ggplot2 [36] (Fig. 3).

2.4. Benchmarking

To compare the performance of the RF classifier we tested it against other methods to predict protein binding [37,38] using default values for each tool, and the same accessibility and CTCF ChIP-seq results for each cell line. We tested the GM12878 cell line using chr20 as a testing set and the remaining chromosomes as a training set. We calculated ROC and PR curves, and reported the AUC of each predictor (Fig. 4).



Position	-8	-7	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6	+7	+8	+9
chr1:11223-11241	0	0	0	0	0	1	1	1	1	1	1	1	0	1	1	1	1	1
chr1:11281-11299	0	0	0	0	0	1	1	1	1	1	1	1	0	1	1	1	1	1
chr1:24782-24800	0	0	0	0	0	1	1	1	0	1	1	0	0	0	1	0	0	0
chr1:91420-91438	0	0	0	0	0	1	1	0	0	0	1	0	0	1	1	0	1	1
chr1:104985-105003	0	0	0	0	0	1	1	0	0	0	1	0	0	1	1	0	0	0

Fig. 1. Representation of the feature matrix. We predicted every CTCF binding site (CTCFBS) with FIMO, and for each predicted site we generated a window of 1000 bp upstream and downstream. After generating windows, we generated sub-bins of 25 bp inside each window. If a feature overlaps with a bin, the value of said feature in said position is 1, if not, the value for the bin for said feature is 0. We considered DNA accessibility, DMA methylation, and different histone marks as features. The generated matrixes were used as input for a random forest classifier, and the importance of each feature in each position was also reported.

Table 1.-
Precision, Recall and F-Score of 4 cell lines being tested against themselves using a random forest classifier.

Cell line	Precision	Recall	F-Score
GM12878	0.86	0.88	0.87
K562	0.87	0.85	0.86
HeLa	0.86	0.85	0.86
SK-N-SH	0.90	0.83	0.86

Table 2.-
Precision, Recall and F-Score of 4 cell lines using 3 of them as a training set and testing against the remaining one using a random forest classifier.

Predicted	Precision	Recall	F-Score
GM12878	0.86	0.89	0.87
K562	0.87	0.85	0.86
HeLa	0.88	0.84	0.86
SK-N-SH	0.90	0.81	0.85

2.5. Evaluation metrics

We employed classic classification metrics [39] commonly employed to evaluate the performance of binary classification models. The metrics employed are Precision (P), Recall (R) and F1-score (F1), calculated as follows:

$$P = TP/(TP+FP)$$

$$R = TP/(TP+FN)$$

$$F1 = 2PR/(P+R)$$

where a true positive (TP) is a CTCF binding site occupied in a CTCF ChIP-Seq experiment properly predicted; a false positive (FP) is a, unbound site predicted as bound; and a false negative (FN) is a bound site predicted as unbound.

2.6. Availability and Implementation

We used python3 with the scikit-learn library [34] to perform all experiments reported in this work. All code with example files is freely available under GNU v3 license at https://github.com/network-biolab/RF_CTCF_BP.

3. Results

Using FIMO we were able to identify 57896 possible CTCF binding motifs in the whole genome. We selected 4 cell lines and downloaded CTCF ChIP-seq experiments to assign which predicted sites were bound in each cell line. The sites with CTCF bound were considered positive examples, and the unbound sites were considered negative examples. For each cell line, we divided the dataset and used 2/3rds as a training set, while the remaining 1/3rd was used as a test set. We next used these sets to train a RF classifier to predict the labels of the test set and calculated precision, recall, and F-Score as measures of performance. In each cell line we reported a precision of over 0.86, with SH-N-SH attaining a precision of 0.9. In terms of recall, on each cell line we obtained a precision of over 0.83, with a precision of 0.88 on GM12878. In terms of F-Score, most cells had a score of 0.86, with the exception of GM12878 which had a score of 0.87 (Table 1).

To increase the amount of examples for the training set and to check if the RF predictor was able to generalize properly, we decided to use three of the four cell lines as a training set and the remaining cell line as a test set. While precision and recall improved in GM12878 and K562, in the two remaining cell lines precision and recall worsened, decreasing from 0.85 to 0.84 in HeLa, 0.83–0.81 in SK-N-SH and F-Score to 0.85 in SK-N-SH (Table 2). To compare the performance on each cell line, we obtained the class probabilities for each cell line, and plotted ROC and PR curves. GM12878 had the highest AUC in both ROC and PR curves, followed by K562, SK-N-SH and HeLa (Fig. 2).

There is evidence that different epigenetic modifications are able to determine CTCF binding, but the complete dynamics of CTCF binding are not fully understood [40]. To understand the contribution of each feature to the binding prediction, we obtained the relative importance of each feature used and plotted it for each cell line prediction. In the 4 cell lines, the most relevant feature by far is DNA accessibility, followed by histone marks and DNA methylation (Fig. 3).

We compared the RF classifier to assess its performance against other DNA-protein binding predictors [37,38], and plotted ROC and PR curves to evaluate performance (Fig. 4). We predicted CTCF binding sites in the chromosome 20 of the GM12878 cell line, using the remaining chromosomes as a training set. The RF classifier

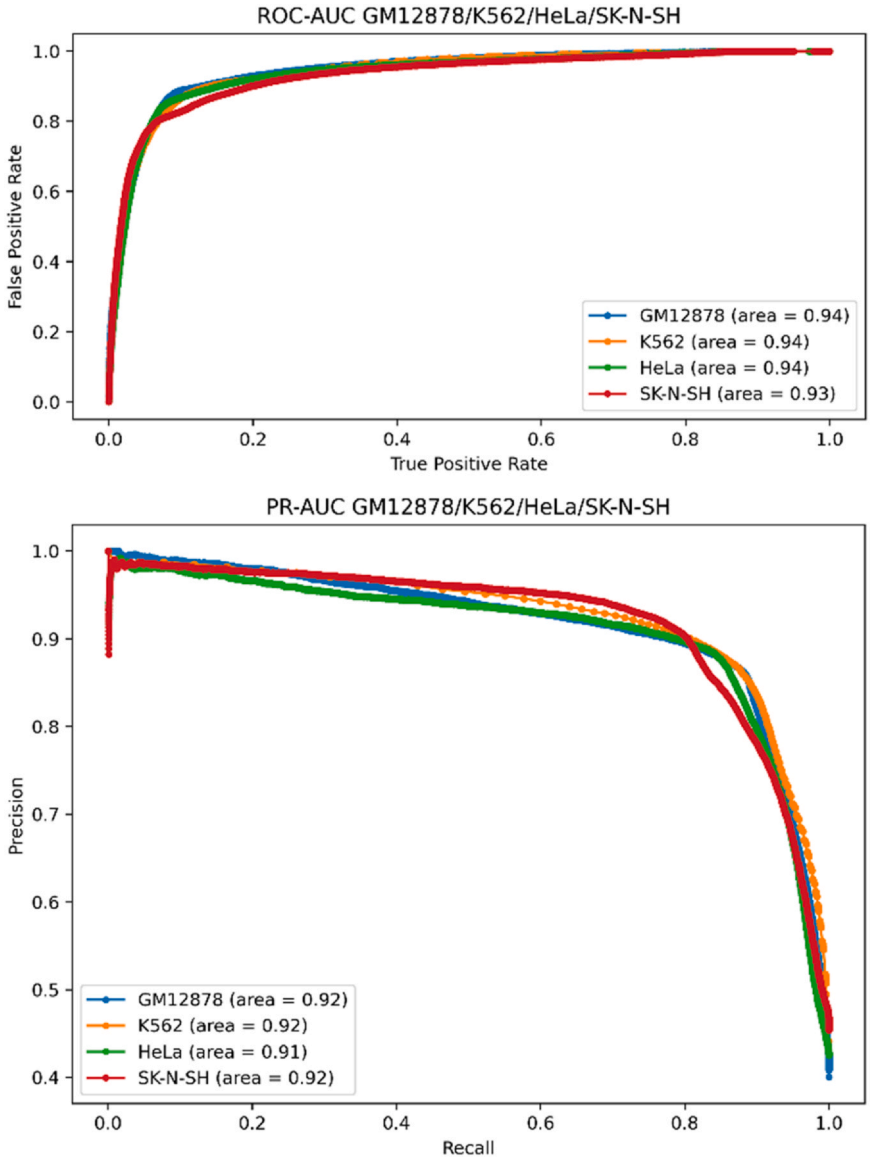


Fig. 2. ROC/PR-AUC curve for each CTCF binding site prediction in each cell line tested. From the 4 evaluated cell lines, we selected 3 as a training set and predicted the remaining cell line. The cell line name shown is the cell line predicted. We compared the results against CTCF ChIP-Seq experiments from said cell line, calculated precision, recall, and F-Score, and plotted ROC and PR curves, reporting the area under each curve.

attained a competitive performance in terms of PR-AUC and a better performance in terms of ROC-AUC than other tools.

4. Discussion

In this article we described the development of a new CTCF binding site predictor based on genomic and epigenomic features using a RF classifier algorithm. The main idea behind this predictor is its application to different research fields allowing the prediction of CTCF binding in the absence of ChIP-seq data using different related features that were not directly assessed. In this context, the fact that accessibility was an important feature to predict CTCF binding could allow us to predict binding without the need of a ChIP-seq experiment. This method could also be applied in the evaluation of changes in CTCF binding associated with disease, as methylation is an important factor that is also related with CTCF binding and the etiology of different diseases. While most CTCF binding sites are conserved [8], there is concrete evidence of different effects that may affect CTCF binding on different cell types and contexts [9]. There are also reports of aberrant CTCF binding mediated by epigenetic marks [18,41], and disease-specific CTCF binding [42]. Thus, CTCF binding identification may provide new insights about the role of CTCF and its contributions in genome regulation and disease. Importantly, the role of epigenetic marks and epigenomic information on the binding of CTCF is not yet fully understood. Moreover, even if it is possible to determine the binding sites actually occupied by CTCF experimentally, using epigenetic information to do so allows to provide distinct knowledge about the biological landscape of the cell derived from other aspects of chromatin regulation such as different activation states [43], and promoter [44] and enhancer [44] activity. The addition of epigenomic features resulted in improvements in the prediction of CTCF binding in different cell lines, allowing the RF predictor to outperform state-of-the-art binding predictors [37,38] by highlighting underlying CTCF binding patterns that could not be identified without the consideration of these features. As most CTCF sites have common binding patterns, the addition of extra predictive features becomes increasingly relevant to identify cell-specific effects or disease-specific effects, like aberrant CTCF binding derived from abnormal methylation in gastrointestinal cancers and gliomas [44,45]. The inclusion of these features allows resolution of context-

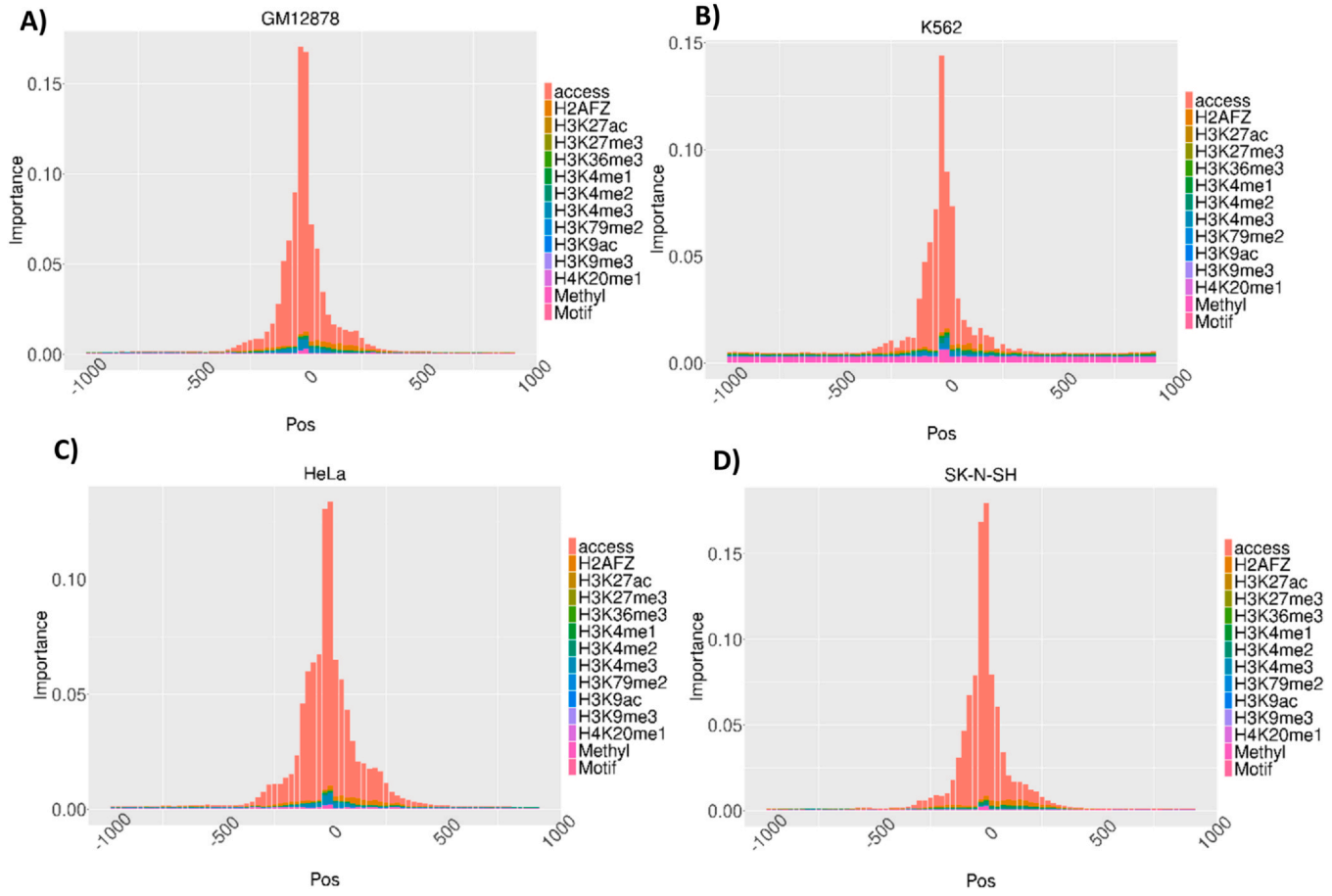


Fig. 3. Relative importance of each feature in each position around the CTCF binding motif in 4 cell lines, using 3 as a training set and predicting the remaining one. The relative importance is the mean decrease in impurity (or gini importance) attributed to each variable when used as a splitting variable accumulated over all the generated trees in the random forest. Higher importance means the feature is more relevant in classification. A) GM12878 as a test set. B) K562 as a test set. C) HeLa as a test set. D) SK-N-SH as a test set.

relevant CTCF binding sites, highlighting the relevance of this predictor in comparison to non-specific predictors. While our method is binary, other methods also consider the amplitude and the extension of each feature [46]. Following this non binary line of work, an interesting approach could be to consider the discrimination of different thresholds for CTCF binding based on epigenomic features as an alternative to our binary approach.

Since there are many approaches for protein binding prediction that can use only the presence of the binding motif and DNA accessibility for a successful binding prediction [37,38,47], we tried using only these two features to predict CTCF binding (Supplementary Methods, Fig S2). While the addition of features improved performance, DNA accessibility around the binding site remained the most important feature to assess if a site was bound in every analyzed cell line. There are certain features that could improve prediction and might deserve consideration, for instance the presence of other proteins that are known to interact with CTCF, such as YB-1 [48] or Cohesin [49], or RNA expression [50]. CTCF occupancy has been considered mostly invariable with many cell groups sharing the same occupied CTCF binding sites, nevertheless, there is evidence of different cell-specific CTCF binding patterns [9] and different epigenetic factors may be contributing to CTCF binding. In this context, there is evidence that DNA Methylation directly interferes with CTCF binding [51], and the relative importance of the features used to generate the RF predictor reflected this association. In CD4 + T-cells, approximately 26000 CTCF binding sites were classified according to their similarity with the canonical motif and CTCF occupancy. Low-occupancy sites were found to be cell specific and related with active histone marks and higher gene expression [52], high-occupancy sites were associated with repressive histone marks and greater gene co-expression inside CTCF-flanked genomic blocks [52], and while we were able to confirm some of those associations in terms of relative importance, they don't contribute to classification of bound CTCF sites as much as DNA accessibility, followed by DNA Methylation. There is evidence that methylation is related with CTCF binding and may directly interfere with its union to the DNA [53]. It is also known that CTCF is related with DNA expression and active enhancers in the same way that hydroxymethylation is associated with active enhancers [54]. In addition, CTCF also has been related to maintaining the boundaries of TADs and the loss of CTCF binding allows the spreading of methylation silencing genes, leading to the loss of the cell transcriptional landscape. There is supplementary evidence that the DNA sequence of 1 kb surrounding a CTCF binding site contains information that improves the prediction of CTCF binding [55], thus, it contains information related to CTCF activity. However, we decided to include and test longer windows surrounding the CTCF binding site. These tests were based on the fact that the RF reports which features are important in relation to their position on the sequence as determined by the feature importance. In this way, features such as DNA methylation can be important beyond the 1 kb range. Since CTCF binding is related with methylation, and so is nucleosome repositioning, it might be interesting to evaluate if nucleosome repositioning is a relevant feature while building RFs. However, we were not able to test this feature with the available ENCODE datasets. While we also believe that gene expression could be an important predictor of CTCF binding, gene expression could also be evaluated using chromatin accessibility as a proxy. This assumes that

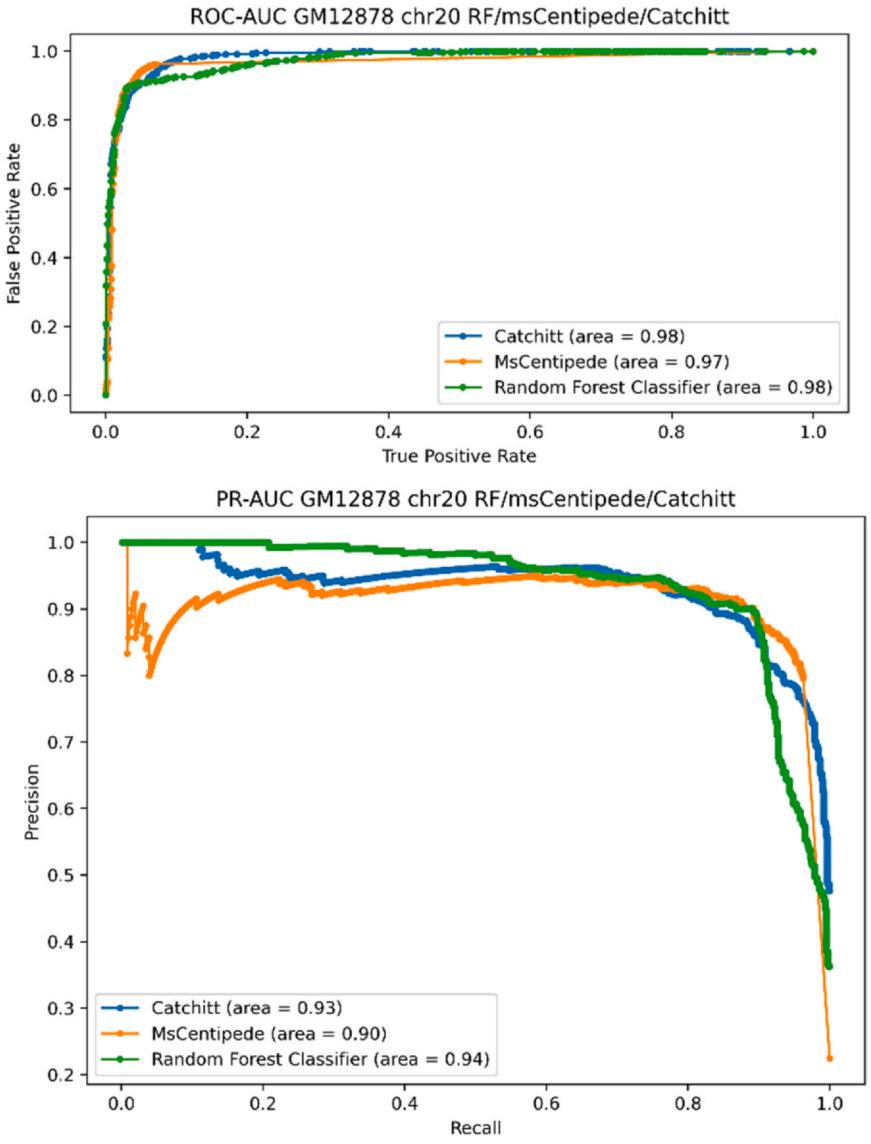


Fig. 4. ROC/PR-AUC curve for 3 different tools predicting chr20 of the GM12878 cell line and using the remaining chromosomes as a training set.

open chromatin is related with active gene expression, and there is evidence that CTCF is capable of keeping the DNA accessible by maintaining the border of TADs and acting as a barrier for methylation [53].

CTCF binding sites are depleted for H3K27me3 and enriched with the histone variant H3.3, as CTCF can open chromatin, incorporating H3.3 and removing the H3K27me3 mark [56]. CTCF is also related with the H2A.Z histone variant, as removal of this histone can enhance CTCF binding [57]. Information about histone variants was not available for the evaluated cell lines, however they could be an asset to improve CTCF binding predictions. There is evidence that histone acetylation, DNA methylation, and gene expression are closely related to TF binding and facilitate accurate prediction of TF-binding events [58]. From the features we evaluated, we could confirm a link between DNA methylation and CTCF binding. Further, there is also a connection between histone acetylation and CTCF binding, even though this is a much weaker relationship when compared with other epigenetic signatures. In our tests, accessibility was the most relevant feature, and assuming a positive relationship between active gene expression and accessibility, it is possible to employ gene expression as a proxy for DNA accessibility. By doing so, we would allow to integrate disease-related aberrant gene expression datasets and CTCF binding. RF models trained to predict the activity of transcription factors have been tested in the HeLa cell line and have been applied to cervical cancer [58]. This study found that TFs were bound to promoters of genes associated with biological processes such as cell proliferation and DNA repair [58]. Nonetheless, we only analyzed cell lines instead of complex samples. However, K562, HeLa-S3 and SK-N-SH are associated with certain specific types of cancer and it would be interesting to expand the functionality of this predictor to check if we are able to identify distinct CTCF binding patterns in different diseases. There is also evidence that H3K9ac and H3K27ac are important features in TF-binding prediction models. We were able to confirm these findings, however our most relevant signal was by far DNA accessibility, followed by H2A.Z, DNA methylation, H3K4me3, and H3K4me2. H3K9ac, H3K27ac, H3K4me3, H3K9ac and H3K27ac are related with activation of transcription, and H3K4me2 is also directly related with both transcriptionally active genes and genes primed for future expression [61]. Moreover, our results also support the idea that repressive marks, such as DNA methylation, are also related with CTCF binding, highlighting the role of CTCF in the overall control of gene expression.

While our approach was successful, there are many aspects that could be improved. One of the greatest limitations is that we are constrained by the FIMO predictions, and FIMO may not be able to predict non-canonical CTCF binding sites using only the canonical motif reported in JASPAR as we did. There are other approaches that could be used to predict binding sites of CTCF missed by FIMO [59]. We did not include sequence-based information besides the presence of another CTCF binding site nearby, however there is a correlation between CTCF site conservation and CTCF binding [60]. Information about several RNAs bound to DNA could also be an important feature to improve our approach considering that CTCF is known to interact with these RNAs, in spite of requiring evaluation as a predictive feature.

Our approach also blurs sequence information: we consider if a feature overlaps with the 25 bp bin, but we are unable to pinpoint exactly which base overlaps with said feature, and certain features acquire biological meaning depending on where they are located; as an example, methylation over the CTCF binding site is more relevant than methylation away from the binding site, and methylation over specific bases leads to an inability to bind CTCF. Reducing DNA Methylation bins from 25 bp to a base-resolution could improve binding prediction and it will be considered in future revisions of this method.

While our intention here was to build a CTCF-specific binding predictor, many of the ideas mentioned could be applied to other DNA-binding proteins. However, the assessment of this approach on other proteins lies beyond the scope of this project. We are planning to expand on the functionality of this method in the future by developing a CTCF loop predictor that would allow us to determine the tridimensional organization of chromatin loops.

In conclusion, we built a CTCF binding predictor to evaluate the binding state of CTCF binding sites using genomic and epigenomic features. This method is capable of outperforming other state-of-the-art non-specific binding predictors, using an ensemble of features not considered by other approaches, highlighting the importance of epigenomic features on the CTCF binding patterns on the whole human genome.

Funding

This work was funded by: FONDECYT Regular Project [1181089] and Centro Ciencia & Vida, FB210008, Financiamiento Basal para Centros Científicos y Tecnológicos de Excelencia de ANID. Powered@NLHPC: this research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02).

Availability

All the code employed in this work is available athttps://github.com/networkbiolab/RF_CTCF_BP under GNU v3 license together withexample files. Contact: alberto.martin@uss.cl mauricio.saez@uct.cl

Declaration of Competing Interest

None declared.

Acknowledgements

We thank all members in our research groups for all useful discussions, feedback and funny jokes that helped during the time spent working on this article.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2023.05.012.

References

[1] Rowley MJ, Corces VG. Organizational principles of 3D genome architecture. Nat Rev Genet 2018;19:789–800.
[2] Kadauke S, Blobel GA. Chromatin loops in gene regulation. Biochim Biophys Acta - Gene Regul Mech 1789;17–25:2009.
[3] Rao SSP, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 2014;159:1665–80.
[4] Lobanenkov VV, et al. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5′-flanking sequence of the chicken c-myc gene. Oncogene 1990;5:1743–53.
[5] Bell AC, West AG, Felsenfeld G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell 1999;98:387–96.
[6] Ohlsson R, Renkawitz R, Lobanenkov V. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. Trends Genet 2001;17:520–7.
[7] Filippova GN, et al. An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. Mol Cell Biol 1996;16:2802–13.
[8] Kim TH, et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 2007;128:1231–45.
[9] Chen H, Tian Y, Shu W, Bo X, Wang S. Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome e41374–e41374 PLoS One 2012;7. e41374–e41374.
[10] Franco MM, Prickett AR, Oakey RJ. The Role of CCCTC-Binding Factor (CTCF) in Genomic Imprinting, Development, and Reproduction. Biol Reprod 2014;91(1–9):125.
[11] Phillips JE, Corces VG. CTCF: master weaver of the genome. Cell 2009;137:1194–211.
[12] Filippova GN. Genetics and Epigenetics of the Multifunctional Protein CTCF. in. Current Topics in Developmental Biology 80. Academic Press,; 2007. p. 337–60.
[13] Ohlsson R, Lobanenkov V, Klenova E. Does CTCF mediate between nuclear organization and gene expression? Bioessays 2010;32:37–50.
[14] Razin SV, Gavrilov AA. Structural–functional domains of the eukaryotic. Genome *Biochem* 2018;83:302–12.
[15] Vietri Rudan M, et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep 2015;10:1297–309.
[16] Xi W, Beer MA. Loop competition and extrusion model predicts CTCF interaction specificity. Nat Commun 2021;12:1046.
[17] Phillips-Cremins JE, et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell 2013;153:1281–95.
[18] Mujahed H, et al. AML displays increased CTCF occupancy associated with aberrant gene expression and transcription factor binding. Blood 2020;136:339–52.
[19] Guo YA, et al. Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. Nat Commun 2018;9:1520.
[20] Eldholm V, Haugen A, Zienolddiny S. CTCF mediates the TERT enhancer-promoter interactions in lung cancer cells: identification of a novel enhancer region involved in the regulation of TERT gene. Int J Cancer 2014;134:2305–13.
[21] Velázquez-Hernández N, et al. BORIS and CTCF are overexpressed in squamous intraepithelial lesions and cervical cancer. Genet Mol Res 2015;14:6094–100.
[22] Recillas-Targa F, de la Rosa-Velázquez IA, Soto-Reyes E. Insulation of tumor suppressor genes by the nuclear factor CTCF. Biochem Cell Biol 2011;89:479–88.
[23] Lähdesmäki H, Rust AG, Shmulevich I. Probabilistic inference of transcription factor binding from multiple data sources. PLoS One 2008;3:e1820.
[24] Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science 2007;316:1497–502.
[25] Behjati Ardakani F, Schmidt F, Schulz MH. Predicting transcription factor binding using ensemble random forest models [version 2; peer review: 2 approved]. F1000Research 2019;7.
[26] Kundaje A., Boley N., Kuffner R., Heiser L., Costello J., Stolovitzky G., Norman T., Hoff B, F.S. ENCODE-DREAM in vivo Transcription Factor Binding Site Prediction Challenge. Synapse. doi:10.7303/syn6131484.
[27] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet 2015;16:321–32.
[28] Kai Y, et al. Predicting CTCF-mediated chromatin interactions by integrating genomic and epigenomic features. Nat Commun 2018;9:4221.
[29] Zhang R, Wang Y, Yang Y, Zhang Y, Ma J. Predicting CTCF-mediated chromatin loops using CTCF-MP i133–i141 Bioinformatics 2018;34. i133–i141.
[30] Lv H, et al. A sequence-based deep learning approach to predict CTCF-mediated chromatin loop. *Brief Bioinform* 22, bbab031 2021.
[31] Wang W, Gao L, Ye Y, Gao Y. CCIP: predicting CTCF-mediated chromatin loops with transitivity. Bioinformatics 2021;37:4635–42.
[32] Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics 2011;27:1017–8.
[33] Fornes O, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res 2020;48:D87–92.
[34] Pedregosa F, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011;12:2825–30.

[35]. R Core Team. R: A Language and Environment for Statistical Computing. (2017).
[36] Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag; 2016.
[37] Raj A, Shim H, Gilad Y, Pritchard JK, Stephens M. msCentipede: Modeling Heterogeneity across Genomic Sites and Replicates Improves Accuracy in the Inference of Transcription Factor Binding. PLoS One 2015;10:e0138030.
[38] Keilwagen J, Posch S, Grau J. Accurate prediction of cell type-specific transcription factor binding. Genome Biol 2019;20:9.
[39] Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiv* abs/2010 2020;1.
[40] Alharbi AB, Schmitz U, Bailey CG, Rasko JEJ. CTCF as a regulator of alternative splicing: new tricks for an old player. Nucleic Acids Res 2021;49:7825–38.
[41] De Biase I, Chutake YK, Rindler PM, Bidichandani SI. Epigenetic silencing in friedreich ataxia is associated with depletion of CTCF (CCCTC-Binding Factor) and antisense transcription. PLoS One 2009;4:e7914.
[42] Fang C, et al. Cancer-specific CTCF binding facilitates oncogenic transcriptional dysregulation. 01.17.910687 bioRxiv 2020;2020. https://doi.org/10.1101/2020.01.17.910687. 01.17.910687.
[43] Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods 2012;9:215–6.
[44] Flavahan WA, et al. Altered chromosomal topology drives oncogenic programs in SDH-deficient GISTs. Nature 2019;575:229–33.
[45] Flavahan WA, et al. Insulator dysfunction and oncogene activation in IDH mutant gliomas. Nature 2016;529:110–4.
[46] Zhang Q, et al. Computational prediction and characterization of cell-type-specific and shared binding sites. Bioinformatics 2023;39.
[47] Li H, Quang D, Guan Y. Anchor: trans-cell type prediction of transcription factor binding sites. Genome Res 2019;29:281–92.
[48] Chernukhin IV, et al. Physical and Functional Interaction between Two Pluripotent Proteins, the Y-box DNA/RNA-binding Factor, YB-1, and the Multivalent Zinc Finger Factor, CTCF *. J Biol Chem 2000;275:29915–21.
[49] Kagey MH, et al. Mediator and cohesin connect gene expression and chromatin architecture. Nature 2010;467:430–5.
[50] Saldaña-Meyer R, et al. RNA interactions are essential for CTCF-mediated genome organization. Mol Cell 2019;76:412–22.
[51] Wang H, et al. Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome Res 2012;22:1680–8.
[52] Essien K, et al. CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. Genome Biol 2009;10:R131.
[53] Damaschke NA, et al. CTCF loss mediates unique DNA hypermethylation landscapes in human cancers. Clin Epigenetics 2020;12:80.
[54] Ehrlich M, Ehrlich KC. DNA cytosine methylation and hydroxymethylation at the borders. Epigenomics 2014;6:563–6.
[55] Wiehle L, et al. DNA (de)methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries. Genome Res 2019;29:750–61.
[56] Weth O, et al. CTCF induces histone variant incorporation, erases the H3K27me3 histone mark and opens chromatin. Nucleic Acids Res 2014;42:11941–51.
[57] Wen Z, Zhang L, Ruan H, Li G. Histone variant H2A.Z regulates nucleosome unwrapping and CTCF binding in mouse ES cells. Nucleic Acids Res 2020;48:5939–52.
[58] Huang Y, et al. Prediction of transcription factors binding events based on epigenetic modifications in different human cells. Epigenomics 2020;12:1443–56.
[59] Kaplow IM, Banerjee A, Foo CS. Neural network modeling of differential binding between wild-type and mutant CTCF reveals putative binding preferences for zinc fingers 1–2. BMC Genom 2022;23:295.
[60] Khoury A, et al. Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. Nat Commun 2020;11:54.
[61] Barth TK, Imhof A. Fast signals and slow marks: the dynamics of histone modifications. Trends Biochem Sci 2010;35(11):618–26. https://doi.org/10.1016/j.tibs.2010.05.006