Ibn-Salem and Andrade-Navarro *BMC Genomics* (2019) 20:777
https://doi.org/10.1186/s12864-019-6088-0

BMC Genomics

RESEARCH ARTICLE Open Access

7C: Computational Chromosome Conformation Capture by Correlation of ChIP-seq at CTCF motifs

Jonas Ibn-Salem* and Miguel A. Andrade-Navarro*

Abstract

Background: Knowledge of the three-dimensional structure of the genome is necessary to understand how gene expression is regulated. Recent experimental techniques such as Hi-C or ChIA-PET measure long-range chromatin interactions genome-wide but are experimentally elaborate, have limited resolution and such data is only available for a limited number of cell types and tissues.

Results: While ChIP-seq was not designed to detect chromatin interactions, the formaldehyde treatment in the ChIP-seq protocol cross-links proteins with each other and with DNA. Consequently, also regions that are not directly bound by the targeted TF but interact with the binding site via chromatin looping are co-immunoprecipitated and sequenced. This produces minor ChIP-seq signals at loop anchor regions close to the directly bound site. We use the position and shape of ChIP-seq signals around CTCF motif pairs to predict whether they interact or not. We implemented this approach in a prediction method, termed Computational Chromosome Conformation Capture by Correlation of ChIP-seq at CTCF motifs (7C). We applied 7C to all CTCF motif pairs within 1 Mb in the human genome and validated predicted interactions with high-resolution Hi-C and ChIA-PET. A single ChIP-seq experiment from known architectural proteins (CTCF, Rad21, Znf143) but also from other TFs (like TRIM22 or RUNX3) predicts loops accurately. Importantly, 7C predicts loops in cell types and for TF ChIP-seq datasets not used in training.

Conclusion: 7C predicts chromatin loops which can help to associate TF binding sites to regulated genes. Furthermore, profiling of hundreds of ChIP-seq datasets results in novel candidate factors functionally involved in chromatin looping. Our method is available as an R/Bioconductor package: http://bioconductor.org/packages/sevenC.

Keywords: Chromatin interactions, Three-dimensional genome architecture, Transcription factors, ChIP-seq, 3C, 4C, 5C, Hi-C, 6C, ChIA-PET, 7C, Prediction, Chromatin loops

Background

The three-dimensional folding structure of the genome and its dynamic changes play a very important role in the regulation of gene expression [1–3]. For example, while it was well known that transcription factors (TFs) can regulate genes by binding to their adjacent promoters, many TF binding sites are in distal regulatory regions, such as enhancers, that are hundreds of kilo bases far from gene promoters [4]. These distal regulatory regions can physically interact with promoters of regulated genes by chromatin looping interactions [5–7], thus it is not trivial to associate TFs to regulated genes without information of the genome structure [8]. Such looping interactions can be measured by chromosome conformation capture (3C) experiments [9] and its variations to either study all interactions from single targeted regions (4C) [10] or multiple target regions (5C) [11], interactions between all regions genome-wide (Hi-C) [12, 13] or interactions mediated by specific proteins (6C [14] ChIA-PET [15, 16], and HiChIP [17]).

While these experimental methods have brought many exciting insights into the three-dimensional organization of genomes [1–3, 18], these methods are not only elaborate and expensive but also require large amounts of sample material or have limited resolution [19, 20]. As a consequence,

* Correspondence: j.ibn-salem@uni-mainz.de; andrade@uni-mainz.de
Faculty of Biology, Johannes Gutenberg University of Mainz, 55128 Mainz, Germany

genome-wide chromatin interaction maps are only available for a limited number of cell types and conditions.

In contrast, the binding sites of TFs can be detected genome-wide by ChIP-seq experiments, and are available for hundreds of TFs in many cell types and conditions [21–23]. Here, we propose that it is possible to use these data to detect chromatin loops.

Recent studies provide functional insights about how chromatin loops are formed and highlight the role of architectural proteins such as CTCF and cohesin [1]. CTCF recognizes a specific sequence motif, to which it binds with high affinity [24, 25]. Interestingly, CTCF motifs are present in convergent orientation at chromatin loop anchors [13, 16, 26]. Furthermore, experimental inversion of the motif results in changes of loop formation and altered gene expression [27–29]. Polymer simulations and experimental perturbations led to a model of loop extrusion, in which loop-extruding factors, such as cohesin, form progressively larger loops but stall at CTCF binding sites in convergent orientation [29–31]. According to these models, CTCF binding sites can function as anchors of chromatin loops.

Our hypothesis is, that we can use convergently aligned CTCF motifs to search for similar ChIP-seq signals at both sites of chromatin loops to predict looping interactions from the largely available ChIP-seq data in many diverse cell-types and conditions (Fig. 1a). We then developed and tested a computational method to predict chromatin looping interactions from only genomic sequence features and TF binding data from single ChIP-seq experiments. We show that our method can predict chromatin loops that were measured by Hi-C and ChIA-PET and that prediction performance depends on the ChIP-seq target, which allows screening for TFs with potential novel functions in chromatin loop formation. The predicted looping interactions might be used to (i) increase performance of other loop prediction methods to associate TF binding sites or enhancers to regulated genes for conditions where Hi-C like data is not available, and (ii) to increase the resolution of interaction maps, where low resolution Hi-C data is available. We implemented our method in the R package *sevenC*.

Results

CTCF motif pairs as candidate chromatin loop anchors

In order to predict chromatin looping interactions from ChIP-seq data, we first analyzed which features at looping anchors correlate with interacting and non-interacting anchor pairs. As a starting point for all analyses we used 38,316 CTCF motif sites in the human genome as potential chromatin loop anchors. We built a dataset of all CTCF motif pairs located within a genomic distance of 1 Mb to each other. This resulted in 717,137 potential looping interactions; we expect that only a minority of these motif pairs will be in physical contact for a given cell type and condition. To label motif pairs as true loops, we used chromatin loops from published high-resolution in-situ Hi-C data and ChIA-PET data for CTCF and Pol2 in human GM12878 cells [13, 16]. If a motif pair was measured to interact in one of the data sets, we labeled it as true interaction (Additional file 4: Figure S1). Overall 30,025 (4.19%) of CTCF motif pairs were considered as true loops using these data sets.

Similarity of ChIP-seq signals at looping CTCF motifs

The ChIP-seq protocol involves a cross-linking step, in which formaldehyde treatment results in covalent bonds between DNA and proteins [32]. This allows the pull-down and detection of sites directly bound by the targeted protein. However, cross-linking occurs also between proteins, which results in detection of sites that are indirectly bound through protein-protein interactions or chromatin looping interactions [33, 34].

We hypothesized that if any given protein binds directly to a genomic region that is in chromatin contact with other genomic regions, DNA from both loci might be pulled out in the cross-linking and DNA-purification step of ChIP-seq protocols. As a result, we expect ChIP-seq signals (e.g. mapped reads) at both genomic regions: the directly bound one and the chromatin loop interaction partner locus (Fig. 1a). Some proteins, like CTCF and potentially also RAD21, might act as homo-dimer at loop anchors. If both loop anchors are bound directly by dimerizing proteins, we expect the ChIP-seq signal at a similar distance to the CTCF motif. Thereby we assume the loop forming complex to be symmetric, that is, that the distance of the direct binding site to the CTCF motif center is the same on both anchors. Although proteins can theoretically dimerize while binding at different distance to the CTCF motif, we assume that the loop anchor complex is symmetric with the two CTCF motifs facing each other directly. For cohesin, for example, it was shown that it binds slightly upstream of CTCF motifs at both loop anchors [16].

To test our hypothesis, we used CTCF motif pairs as anchors and compared the ChIP-seq signal from one anchor to the (reversed) signal of the corresponding anchor. Using ChIP-seq data for several TFs, we found similar ChIP-seq coverage patterns around CTCF motifs more often when the two sites perform looping interactions than when they do not (Fig. 1b). To quantify the similarity of ChIP-seq coverage from any two CTCF sites, we correlated their ChIP-seq signals at ±500 bp around the CTCF motif (Fig. 1c) (see Methods for details). Measuring ChIP-seq profile similarity by correlation has the advantage that the correlation can be high even if the anchor that is not bound directly has a much lower ChIP-seq signal (which is often the case).

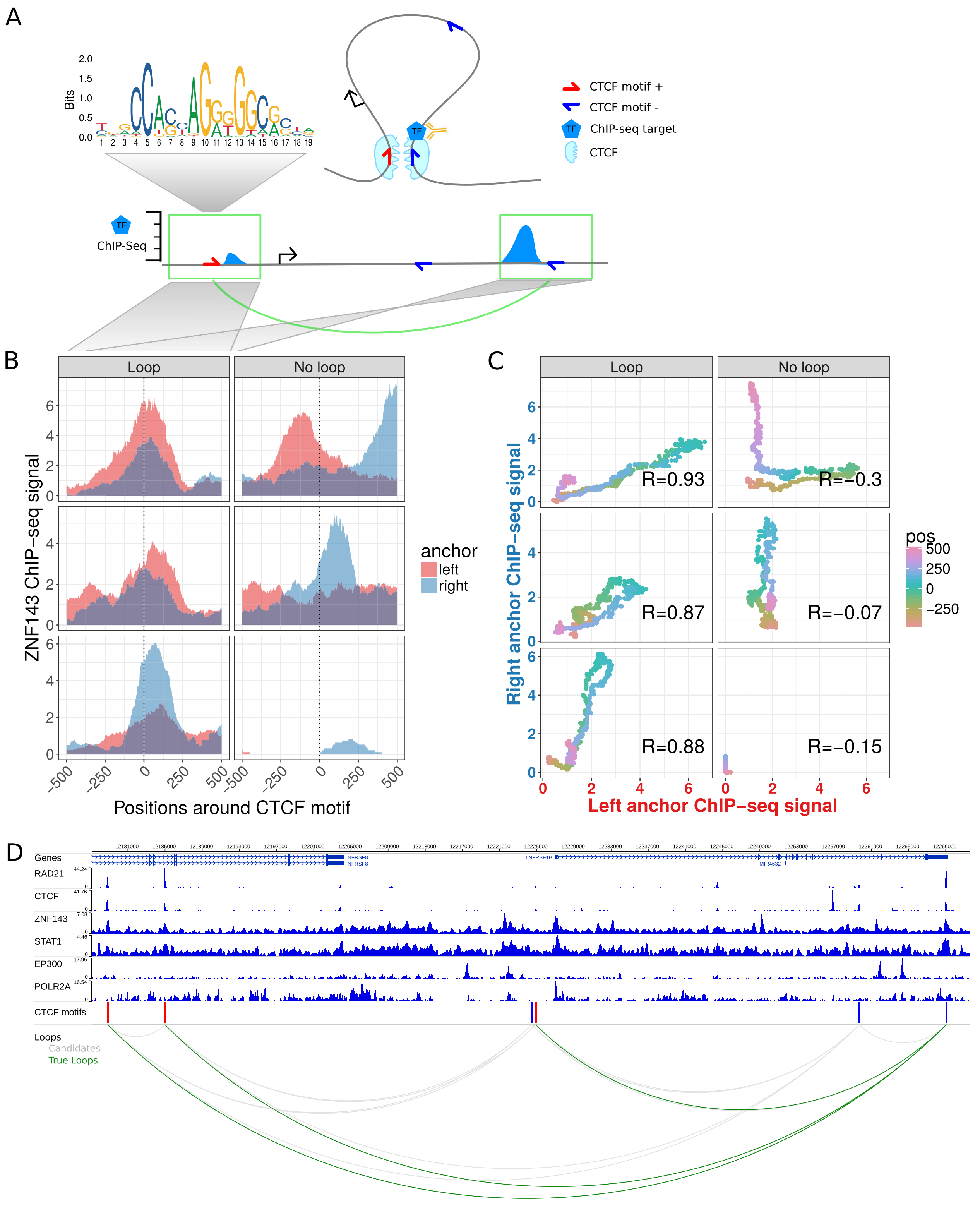

Fig. 1 (See legend on next page.)

(See figure on previous page.)
Fig. 1 Chromatin looping interactions result in ChIP-seq coverage signals at direct and indirect bound loop anchors. **a** Schematic illustration of a chromatin loop with CTCF motifs at the loop anchors (top right). A transcription factor (TF) binds directly at the right loop anchor close to the CTCF motif. This results in a ChIP-seq coverage peak at the directly bound locus (bottom right) and in a minor signal at the other loop anchor (bottom left), both at the same distance to each CTCF motif. **b** Znf143 ChIP-seq coverage at six selected example CTCF motif pairs of which the ones in the left panel interact via loops according to Hi-C and ChIA-PET data and the ones in the right panel do not interact. The ChIP-seq coverage signal for each loci pair is shown in red for the left anchor region and in blue for the right anchor region, according to the distance to the CTCF motif (x-axis). Interacting CTCF motif pairs show more similar ChIP-seq coverage signals, which are often enriched at similar distances to the CTCF motif pairs, while the profiles of non-interacting pairs are less similar. **c** The similarity of ChIP-seq profiles by correlation of the ChIP-seq coverage signals of the selected motif pairs in (**b**). For each pair, the coverage at the right anchor is plotted versus the coverage at the left anchor at the same distance (color coded) from each CTCF motif. The Pearson correlation coefficient (R) of the dots is higher for interacting loci pairs. **d** Example loci on chromosome 1 shown in the genome-browser with six ChIP-seq tracks. Red and blue bars indicate CTCF recognition motifs on the forward and reverse strand, respectively. The bottom panel shows CTCF motif pairs in gray (candidates) and actually interacting pairs in green, according to ChIA-PET and Hi-C data

Next, we compared ChIP-seq similarity at looping and non-looping CTCF motif pairs for six selected TF ChIP-seq data sets (Fig. 1d). Compared to non-interacting CTCF sites the ChIP-seq correlation is significantly higher at looping interactions (Fig. 2a). However, the overall correlation as well as the difference between looping and non-looping CTCF sites varies between TF ChIP-seq datasets (Fig. 2a). As expected, we observed a large difference for the CTCF ChIP-seq dataset but, interestingly, also for other known architectural proteins, such as Rad21 and Znf143. Moreover, other TFs, such as STAT1 have significantly higher ChIP-seq signal similarity at CTCF motifs that interact via chromatin looping. Overall, this analysis shows that ChIP-seq signals are more similar at interacting CTCF sites, indicating that this similarity can be used to predict looping interactions.

Genomic sequence features of CTCF motif pairs are associated with looping

The frequency of two genomic regions to physically interact depends on their genomic distance [12]. Consequently, we observed that CTCF motif pairs are more often in contact when they are close to each other in the genomic sequence (Fig. 2b). Recent studies on 3D chromatin structure led to an increased understanding of the molecular mechanism of chromatin loop formation and suggested a functional role of CTCF proteins, which bind specific DNA sequences [1]. The canonical CTCF

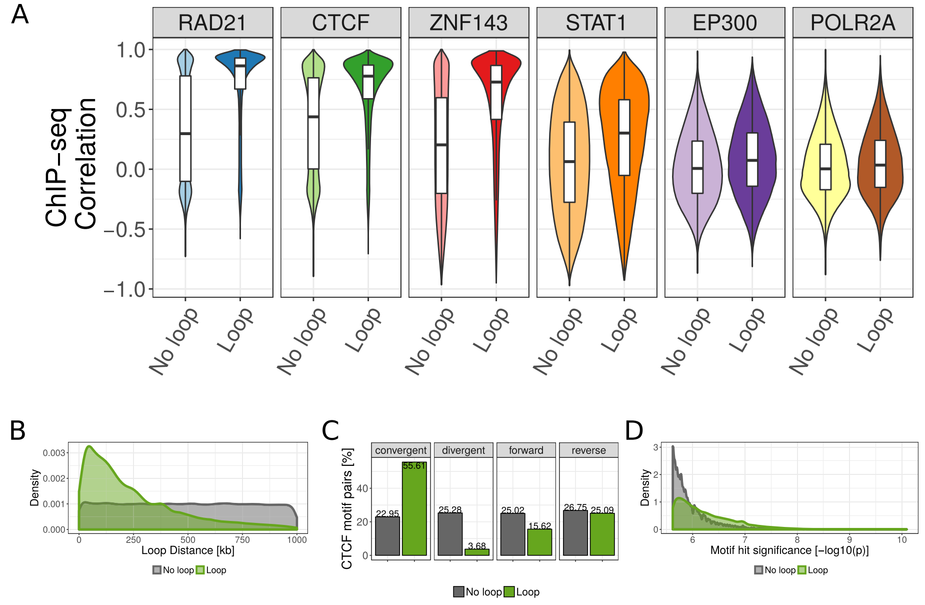

Fig. 2 ChIP-seq similarity and genomic features of looping and non-looping CTCF motif pairs. **a** Boxplot of Pearson correlation coefficients of ChIP-seq signals between CTCF motif pairs for all CTCF motif pairs within 1 Mb genome-wide. The correlation is shown separately for non-looping and looping motif pairs (according to HI-C and ChIA-PET data in GM12878 cells), and for six selected ChIP-seq data sets in GM12878 cells. **b** Distance distribution between looping (green) and non-looping CTCF motif pairs. **c** Percent of looping and non-looping CTCF motif pairs in convergent, divergent, both forward, or both reverse orientation. **d** Distribution of CTCF motif hit significance as -$\log_{10}$ transformed *p*-value for looping and non-looping CTCF motif pairs. For each motif pair only the less significant motif is considered

motif is non-palindromic and therefore occurs either in the positive or in the negative DNA strand. Importantly, it is known that CTCF motifs occur predominantly in convergent orientation to each other at chromatin loop anchors [13, 26]. Experimental inversions of CTCF motifs lead to changes of the interactions and expression of the associated genes [27, 28]. Accordingly, we observed that 55.6% of the looping CTCF pairs have convergent orientation versus only 22.9% of the non-looping pairs (Fig. 2c). We also observed that the motif match strength, as measured by the significance of a motif location to match the canonical CTCF motif [35], is higher for motifs involved in looping interactions (Fig. 2d). Together, the linear genome encodes several features, such as motif strength, orientation, and distance, that correlate with chromatin looping and can be used to predict such interactions.

Chromatin loop prediction using 7C

To make use of both the condition specific ChIP-seq signals and the genomic features of CTCF motifs to predict chromatin loops, we trained a prediction model that takes only ChIP-seq data as input. To this end, we built a logistic regression model that takes into account only four features: the correlation coefficient between the ChIP-seq signals of the paired CTCF motifs (in a window of 1000 bp around the motif), the genomic distance between motifs, the orientation, and the (minimum) motif hit significance score (see Methods for details). For each ChIP-seq data set, we trained and evaluated a separate model (Additional file 5: Figure S2A). The method is implemented as the R package ‘sevenC’, which predicts chromatin loops using as only input a bigWig file from a single ChIP-seq experiment.

Prediction performance evaluation

We used 10-fold cross-validation to assess the performance of the predictions on independent data that was not seen in the training phase. For each cutoff on the predicted interaction probability score, we computed the sensitivity, specificity and precision to plot receiver operator characteristic (ROC) and precision recall curves (PRC). Since only 4.2% of CTCF pairs are measured to interact, we mainly used the area under the PRC (auPRC) to evaluate prediction performance since, compared to ROC, the PRC gives a more accurate classification performance in imbalanced datasets in which the number of negatives outweighs the number of positives significantly [36]. Furthermore, we defined an optimal cutoff for the prediction probability p based on optimizing the f1-score. The six selected TF ChIP-seq data sets have optimal f1-scores at about $p = 0.15$ (Additional file 5: Figure S2B). For binary prediction, we provide a default prediction score threshold as the average of thresholds with optimal f1-score for the 10 best performing TF ChIP-seq datasets.

Prediction performance of sequence features and 7C with single and multiple TF ChIP-seq data sets

First, we evaluated how the sequence-encoded features can predict chromatin interactions. For this, we built logistic regression models that use only these features. Each of these features alone, CTCF motif hit significance, motif orientation or distance, were very poor predictors, and resulted in auROC between 0.67 and 0.74 (Fig. 3a) and auPRC scores between 0.08 and 0.09 (Fig. 3c). Using the three sequence features together improved prediction performance (auROC = 0.85, auPRC = 0.22).

Next, we tested the addition of ChIP-seq data as feature in the prediction model using ChIP-seq data for each of six different TFs. Three of them, CTCF, RAD21, and ZNF143, have known function in chromatin loop formation [1, 37–39], while STAT1, P300, and POL2, are to our knowledge not directly involved in chromatin loop formation. Adding any of these TF ChIP-seq datasets to the model increased prediction performance. STAT1, EP300, and POL2 only moderately increased prediction performance with auROC values between 0.86 and 0.87 (Fig. 3a) and auPRC between 0.24 and 0.26 (Fig. 3b, c). However, ChIP-seq of the known architectural proteins CTCF, RAD21, and ZNF143 resulted in markedly increased prediction performance with auPRCs of 0.31, 0.37, and 0.38 for CTCF, RAD21, and ZNF143, respectively (Fig. 3b, c). To test how much the performance depends on the actual truth set of measured loops, we trained and validated 7C on each individual Hi-C and ChIA-PET data set, as well as their intersection, and observed similar performance across data sets (Additional file 7: Figure S4A). For visual comparison, we show the actual looping interactions and 7C predictions in an example region at chromosome 11 (Fig. 3d) and an overlay of 7C predicted loops with a Hi-C interaction heatmap (Additional file 6: Figure S3).

Next, we built a full model using the sequence based features and the ChIP-seq data of all six selected TFs. This only resulted in a slight increase of prediction performance to auPRC = 0.42 (Fig. 3b, c), indicating that a single ChIP-seq experiment might be sufficient for accurate prediction of chromatin loops. We also tested if a single value of correlation of ChIP-seq signal at both loop anchors across the six different TFs is predictive. Indeed, we find high prediction performance of auPRC = 0.34 for this approach. However, this was lower than using the correlation from single TF ChIP-seq experiments for RAD21 or ZNF143 and has the disadvantage of relying on ChIP-seq data from multiple experiments.

Next, we tested whether the predictive performance of 7C comes simply from increased ChIP-seq signals at the

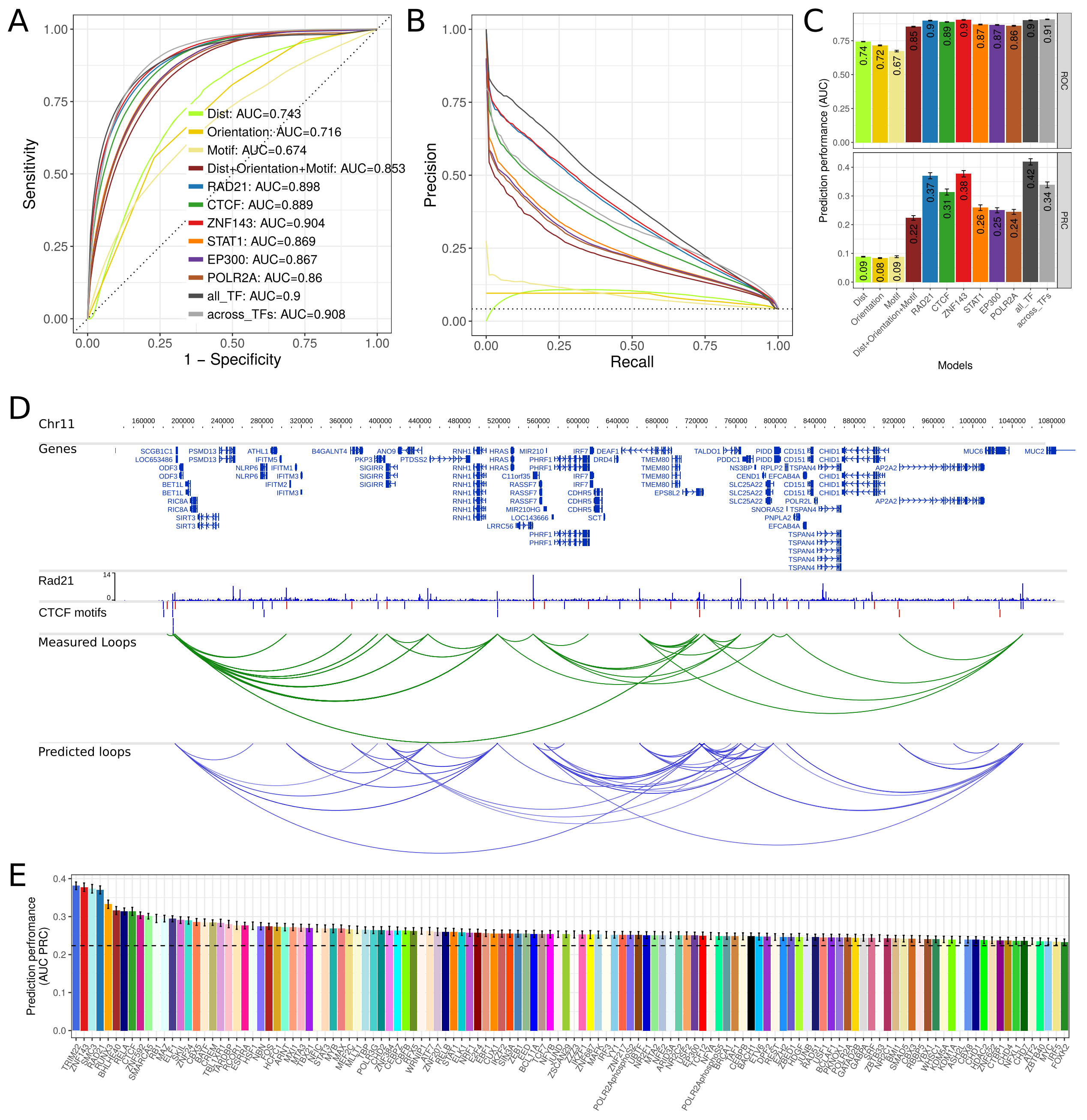

Fig. 3 Prediction performance using cross-validation. **a** ROC plot for different models to predict chromatin looping interactions. The sensitivity (y-axis) is shown against the false discovery rate (1 – specificity, x-axis) for thresholds of the prediction score. Curves show averages of 10-fold cross-validation experiments. The models “Dist”, “Orientation”, and “Motif” contain only a single feature as indicated and all three genomic features are combined in the model “Dist+Orientation+Motif”. The models “RAD21”, CTCF”, “ZNF143”, “STAT1”, “EP300”, and “POLR2A”, contain the genomic features and the ChIP-seq correlation of the indicated factor. The model “all_TF” contains the genomic features and correlation of all indicated TFs. The model “across_TFs” contains the genomic features and a single correlation feature across the six ChIP-seq datasets as described in the main text. **b** PRC plot of precision against the recall for different prediction models. Color code as in (**a**). **c** Values of the area under the ROC (top) and PRC curves (bottom) as prediction performance. Error bars indicate standard deviation in 10-fold cross-validation experiments. **d** Example region on chromosome 11 in the genome browser showing: human genes, RAD21 ChIP-seq data in GM12878, CTCF motifs, CTCF motif pairs with that interact according to Hi-C or ChIA-PET data (green arcs) and predicted chromatin loops from RAD21 ChIPseq data using 7C (blue arcs). **e** Prediction performance of 7C as auPRC values for models with 124 TF ChIP-seq data sets from ENCODE. Error bars as in (**c**). The dotted horizontal line shows prediction performance of only the combined genomic features

loop anchor region or if indeed the ChIP-seq signal similarity at both anchors is important. Therefore, we replaced the ChIP-seq correlation value as input and used instead the total signal at both anchors separately as input. Only for STAT1 we found higher prediction performance with the pure ChIP-seq signal approach and for all other selected TF we find 7C with ChIP-seq profile similarity by correlation as input more predictive than the pure signal strength at the anchors (Additional file 7: Figure S4B). This supports the hypothesis that indeed co-immunoprecipitation of DNA from both loop anchors is captured by 7C to predict interactions.

Another recently published method uses CTCF ChIP-seq peak heights together with motif orientation and distance in an iterative algorithm to predict chromatin interactions [40]. However, independent of the TF used, 7C yields higher specificity, precision and overall accuracy than this previously published method (Additional file 7: Figure S4C).

Together, these results show that sequenced based features alone have only a limited loop prediction performance, but integrating them with a single ChIP-seq experiment, 7C can predict chromatin loops with higher accuracy.

Comparison of transcription factors by prediction performance

Our results can be used to better understand the molecular mechanisms of chromatin loop formation. We hypothesize that TFs whose ChIP-seq provides high prediction performance are likely to be functionally involved in chromatin looping. These TFs would be therefore interesting targets for further investigation of their potential function in chromatin looping.

To investigate this for as many TFs as possible, we used all available 124 TF ChIP-seq datasets from ENCODE for the human cell line GM12878 and compared transcription factors by their prediction performance. Notably, nearly all TF ChIP-seq data sets could increase the prediction performance of sequence-based features alone (Fig. 3e). However, there was a large variance in performance between TFs and a subset of TFs with high predictive power could be identified. These include for example the known architectural proteins mentioned above, CTCF, cohesin (RAD21 and SMC3), and ZNF143, but also factors, such as TRIM22, RUNX3, BHLHE40, or RELA, which might be interesting candidate factors with functional roles in chromatin loop formation.

Prediction performance in other cell types and for different TFs

Next, we wanted to test if 7C is general enough to predict looping interactions in a different cell type than the one used to train it. To test this, we used the models presented above (trained with data from human GM12878 cells) to predict loops using as input ChIP-seq data from human HeLa cells. The prediction performance was assessed using as positives 12,480 loops (1.74% of all motif pairs) identified in HeLa cells [13, 16]. Compared to the cross validation within GM12878 cells, the prediction performance is only slightly lower when 7C was trained on GM12878 cells and applied to the data in HeLa cells as evaluated with ROC curves (auPRC up to 0.91, Fig. 4a) and PRC curves (auPRC up to 0.27, Fig. 4b,c). However, when we use the experimentally measured loops in GM12878 cells as predictor for loops in HeLa cells, we observe an even higher performance (sensitivity = 0.77, specificity = 0.97, precision = 0.32). While this indicates that 7C alone is not sufficient to call cell-type specific interactions in HeLa more accurately than just taking interactions from a different cell line, it also indicates that the chromatin architecture is remarkably similar across these cell types. Indeed, 77.6% of the 12,480 measured HeLa loops are also loops in GM12878 cells. (Additional file 7: Figure S4D).

In this analysis, we compared the prediction performance of models that were trained using a specific TF as input.. However, in a real use case, one might not be able to train the model for a specific TF of interest and the model should idealy predict loops for TFs that were not used in the training. Therefore, we built default 7C models by either averaging model parameters from all 124 TF models or by averaging across the model parameters of only the 10 best performing TFs. While all three approaches result in good prediction performance for the six selected TFs (Fig. 4c), the model averaging parameters across all TFs performs worse than the ones of only the best 10 models, which are nearly as good as the specific TF models. This is consistent with similar results from cross-validation analysis in GM12878 data (Additional file 7: Figure S4E). Furthermore, we visually inspected chromatin loop predictions from RAD21 ChIP-seq data in HeLa at an example loci on chromosome 21 (Fig. 4d). In summary, these results show that 7C can predict chromatin looping interactions in different cell types that were not used to train it. Similarly, the 7C default prediction model performs nearly as good as a TF specific model. This makes 7C applicable for ChIP-seq data from diverse TFs in many different cell types and conditions.

The high resolution of ChIP-nexus improves prediction performance

We wondered if other genomic measurements along the linear genome could provide similar signals at loop anchors potentially indicating looping interactions. Therefore, we used ChIP-seq for histone modifications and other genomic assays, such as DNase hypersensitivity (DNase-seq), ChIP-nexus and only ChIP-seq input control as input to our prediction methods (Fig. 5).

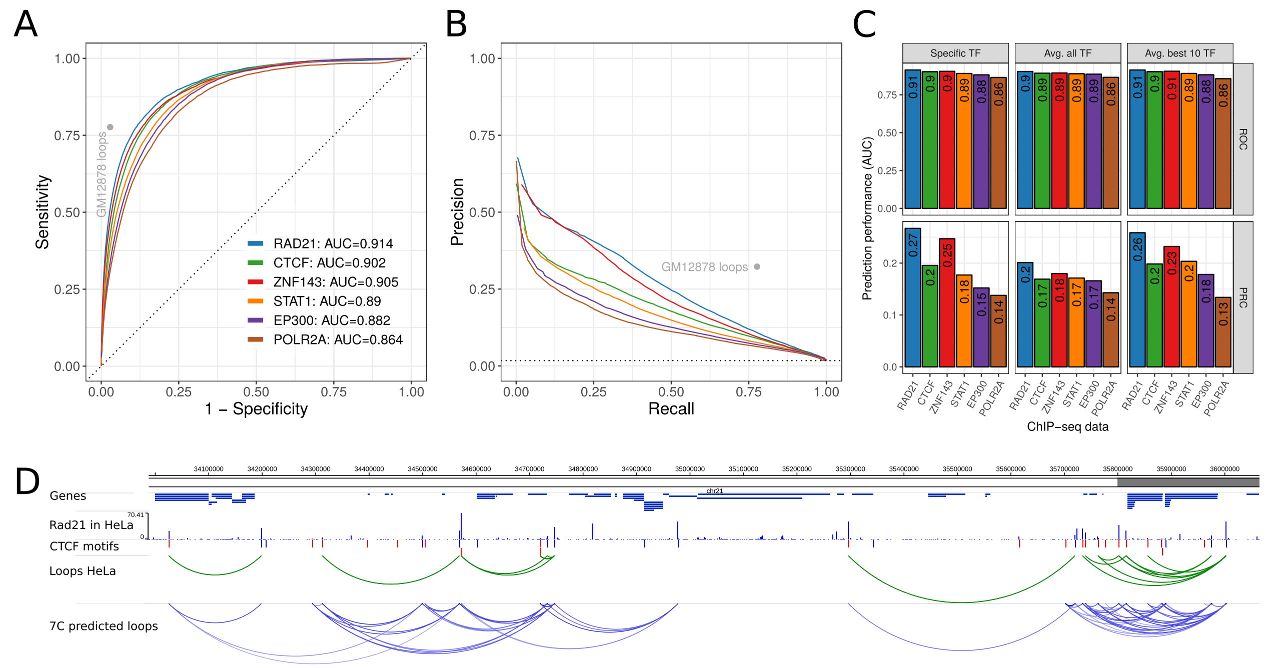

Fig. 4 Prediction performance in HeLa cells using 7C trained in GM12878 cells. **a** ROC curve of prediction performance of six selected TF ChIP-seq data sets. The 7C model was trained using ChIP-seq and true loop data in human GM1287 but loops were predicted using ChIP-seq data of the same TFs in HeLa cells and true loop data in HeLa cells. **b** Precision-Recall curves for the same analysis as in (**a**). **c** Prediction performance as auPRC (top) and auROC (bottom) in HeLa for the six TF ChIP-seq data sets (x-axis) and 7C models trained for the specific TF (left), 7C with parameters averaged across all 124 TF models (center), and 7C with parameters as average of the 10 best performing TF ChIP-seq data sets (right). **d** Example region on human chromosome 21 with genes, RAD21 ChIP-seq data in HeLa, CTCF motifs, true loops in HeLa cells according to Hi-C and ChIA-PET (green arcs) and predicted chromatin loops from RAD21 ChIP-seq data in HeLa (blue arcs)

Furthermore, we compared different signal types of ChIP-seq. During computational processing of ChIP-seq raw data, reads are shifted in 5′ direction by the estimated average fragment size [41, 42]. The coverage of these shifted reads is then compared to coverage of input control experiment (fold change over control). Furthermore, a recent study quantified read pairs (qfrags) in a specific distance to each other as estimate for the actual fragment numbers detected by ChIP-seq [42]. For most of the TFs tested here, we observed that the ChIP-seq signal types ‘shifted reads’ and ‘qfrag’ have better loop prediction performance than the ‘fold change over control’ (Fig. 5). Interestingly, even the combination of ChIP-seq control with sequence features improves the prediction performance over using sequence features alone, indicating that cross-linking efficiency and density of chromatin itself is specifically distributed at chromatin loop anchors (Fig. 5). Also, DNase-seq, which measures chromatin accessibility, predicts looping interactions with accuracy similar to ChIP-seq input control (Fig. 5). This is consistent with specific open-chromatin profiles at TF binding sites [43, 44]. We also tested if the ChIP-seq signal for the histone modifications H3K4me3, H3K4me1, H3K27me3, and H3K27ac can be used as input for 7C, but did not find a marked improvement over genomic features alone (Fig. 5). This indicates that the histone mark signal profile, which is characterized by broader enriched domains compared to sharp peaks for TF ChIP-seq, is in general not suitable for the correlation-based approach in 7C.

However, using ChIP-nexus data for RAD21 and SMC3 [16], we could markedly improve chromatin loop predictions using 7C (Fig. 5). ChIP-nexus and ChIP-exo are variations of the ChIP-seq protocol, in which additionally, an exonuclease digestion step is applied to trim the DNA from the 5′ end until the actual bound protein [45, 46]. These signals result in high-resolution binding footprints that can be used to identify different TF binding modes and cooperation with co-factors [34]. Therefore, we conclude, that the high-resolution binding profiles from ChIP-nexus allow to compute a more predictive binding signal similarity at chromatin loop anchors.

In summary, the comparison of different genomic signal types shows that (i) cross-linking effect and chromatin density at chromatin anchors are predictive signals for long-range chromatin interactions and that (ii) higher resolution TF binding assays, such as ChIP-nexus, result in improved prediction performance.

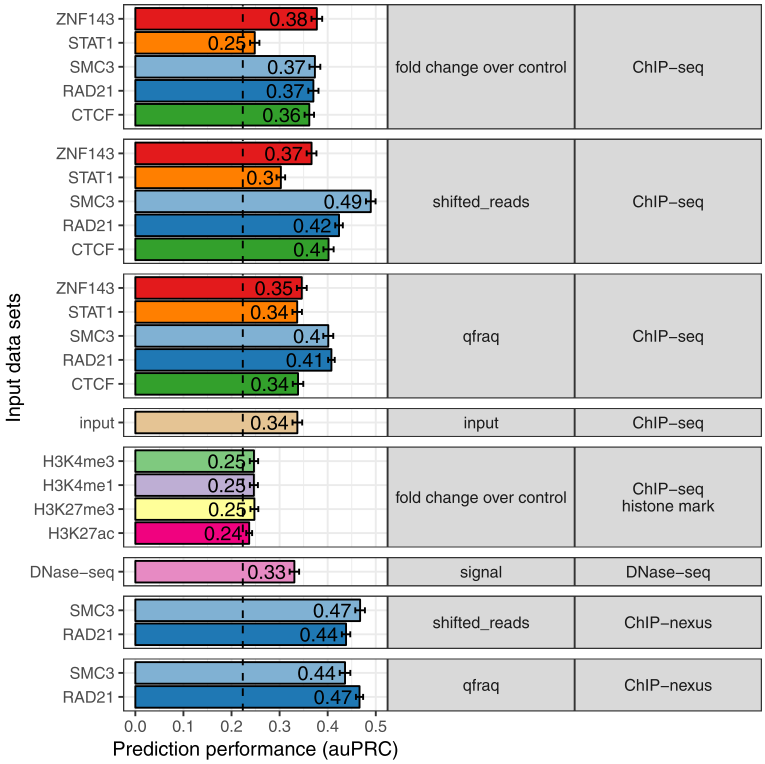

Fig. 5 Higher resolution of ChIP-nexus experiments improves prediction performance. Prediction performance as area under the precision recall curve (auPRC, x-axis) for 7C models with sequence features and different input data sets to predict chromatin looping (y-axis). Input data sets are grouped by signal-type (middle panel) and assay-type (right panel) and colored according to the TF or histone mark (if any) used as target in the experiment

Discussion

We have developed 7C to reuse ChIP-seq data, profiling the interactions of proteins with genomes, for the prediction of chromatin looping interactions between CTCF motif pairs within 1 Mb. We present this method as a complementary approach to dedicated techniques like Hi-C that directly measure genomic contacts. Since the results of ChIP-seq experiments are increasingly available for a large number of proteins, species, tissues, cell types, and conditions, our method offers an alternative when Hi-C data is not available, or cannot be produced due to cost or material limitations. Another major advantage of 7C over Hi-C is that the predictions are at a base pair resolution due to the use of paired CTCF motifs, while Hi-C only reaches resolutions of at best kilo base pairs at a high cost. Of course, the method is restricted to species where CTCF is present and forms loops (e.g. not in *Drosophila*).

7C can use ChIP-seq data from any TF. It does not make any assumptions about the mode of binding or function of the TF and the TF does not need to be involved in the loop formation. It is enough if the binding of the TF close to a CTCF-based loop results in symmetrical ChIP-seq traces. Regardless, it is possible that TFs binding very close and very often to the anchors of loops will be better predictors, and TFs with functions in loops might as well have stronger and more reproducible binding there. For example, it is interesting to note that ZNF143 has been known for a while to bind close to CTCF-cohesin in the anchors of chromatin loops (e.g. [7]), and very recently it has been reported as a regulator of loop formation [47]. Using the HIPPIE database [48], we found that while there is no experimental evidence of direct interaction between CTCF and ZNF143, there is one single protein that is reported to interact with both CTCF and ZNF143, the chromodomain helicase DNA binding protein 8 (CHD8; see [49, 50], respectively). So while no direct interaction between ZNF143 and CTCH is known, they could form a complex via CHD8. Differently, for another very good predicting factor such as TRIM22 no connection to loop architecture has been yet established; they might be none, or it could occur in an unexplored condition. TRIM22 is an antiviral protein whose expression is triggered by interferon, with cytoplasmic and nuclear localizations, and a variety of functions (see e.g. [51]); we would be careful to suggest that this protein might have yet another function controlling genomic structure.

Other computational approaches were developed to predict genomic contacts or assign regulatory regions to target genes. A commonly used approach is to compare activity signals at enhancers and promoters across many different conditions or tissues [52–56]: high correlation indicates association and potential physical interactions between enhancers and genes. However, these approaches lose the tissue specificity of the interactions. Other approaches integrate many diverse chromatin signals such as post-translational histone modifications, chromatin accessibility, or transcriptional activity [57–62], and combine them with sequence features [63], or evolutionary constrains [64]. While these methods predict enhancer-gene association with good performance, they require for each specific condition of interest a multiplicity of input datasets, which are often not available.

Further computational approaches try to directly predict chromatin interactions by using diverse sequence features [65] or multiple chromatin features such as histone modifications [66–68] or transcription [69]. One study makes use of the more recently discovered CTCF motif directionality to predict loop interactions from CTCF ChIP-seq peak locations [40], but has lower prediction performance than 7C (Additional file 7: Figure S4C). Another study combines CTCF binding locations and motif orientation with polymer modeling to predict Hi-C interaction maps [29]. However, none of these studies predicts chromatin loops from ChIP-seq signals of TFs different from CTCF by taking the CTCF motif orientation into account. Furthermore, CTCF binding sites are often only considered when the signal is strong enough for peak calling algorithms to identify binding sites. In contrast, 7C takes the distribution of ChIP-seq signals from all TFs into account without a peak-calling step. Furthermore, the other studies, except one [40], do not provide a tool for the direct prediction of pairwise interactions from single ChIP-seq experiments. Interestingly, shadow peaks in ChIP-seq data of insulator proteins in *Drosophila* were previously associated to long-range interactions [70] and used to study the contribution of sequence motifs and co-factors in loop formation [71], but not to directly predict chromatin loop interactions.

Compared to other predictive methods mentioned above, our approach has the advantage of directly predict chromatin looping interactions, and not enhancer-promoter associations, by making use of ChIP-seq signals from a single experiment with respect to CTCF motifs. However, many enhancer-promoter interactions occur in the span of interacting CTCF binding sites, which were described to form insulated neighborhoods [72, 73]. Therefore, 7C predictions can help to associate enhancers to genes when they are located between predicted loop anchors. The motifs give the prediction a base pair resolution. In fact, given several CTCF motifs within a 1 kb genomic bin, our looping prediction approach can be used to decide which of the CTCF sites is actually involved in the measured interactions and thus increase resolution even when Hi-C data is available. We showed that our approach, 7C, can work with just a single ChIP-seq experiment for many different TFs, making it usable for many diverse conditions of interest. Therefore, 7C can be used as a complement to existing enhancer-promoter association tools or can be integrated in such predictive models to improve them.

Currently, our method, by using CTCF motifs, focuses on CTCF mediated chromatin loops. It is very likely that other DNA binding proteins mediate loops: for example, recent studies suggest that other TFs are involved in enhancer promoter interactions during differentiation [74] and knockout of transcriptional repressor YY1 and other candidate factors result in loss of chromatin loops [75]. Using motifs predicted for these different transcription factors, or combinations thereof, are open avenues for the future extension of our method.

Conclusion

We demonstrated that TF binding signals of ChIP-seq experiments at CTCF motifs are predictive for chromatin looping. We provided a method, 7C, that is simple to use and integrates these signals with genomic sequence features to predict long-range chromatin contacts from single ChIP-seq experiments. 7C is freely available as R/Bioconductor package (http://bioconductor.org/packages/sevenC). The analysis of ChIP-seq experiments for 124 different TFs highlighted the role of cohesin, ZNF143 and CTCF in chromatin loop formation, but also suggested many other TFs, such as TRIM22, RUNX3, and BHLHE40, to be functionally involved in chromatin looping, likely in cooperation and protein interaction (direct or indirect) with CTCF at loop anchors.

Since our method needs only a single ChIP-seq experiment as input, it can support the analysis of chromatin interactions in diverse cell types and conditions, where Hi-C like data is not available. Therefore, 7C can be used together with other computational methods to enable condition specific associations of distal TF binding sites and enhancers to promoters of target genes. These might allow the interpretation of non-coding genetic variants by genes in physical contact with the variant loci in a specific cell type or condition of interest. Furthermore, 7C might improve the resolution of Hi-C interaction maps by facilitating base-pair specific pairing of CTCF motifs located in bins of several kb. With these applications, 7C increases the value of ChIP-seq datasets, which now can be used to improve the analysis of 3D genome folding and its dynamic changes between diverse cell types and conditions.

Methods

CTCF motifs in the human genome

The recognition motif of CTCF is well defined and available from the JASPAR database (MA0139.1) [76]. We downloaded TF binding site predictions with the CTCF motif (MA0139.1) in the human genome hg19 from the JASPAR database (http://expdata.cmmt.ubc.ca/JASPAR/downloads/UCSC_tracks/2018/hg19/tsv/MA0139.1.tsv.gz). Motif hits were filtered for p-value $\leq 2.5 \times 10^{-6}$, resulting in 38,316 highly significant CTCF motif hits genome-wide and 717,137 motif pairs within 1 Mb genomic distance that are considered as potential loop interaction anchors in this study.

Loop interaction data for training and validation

For training and validating the prediction model we used 9448 published loops derived from high-resolution in-situ Hi-C experiments [13] and 206,399 CTCF and Pol2 ChIA-PET interactions [16] in human GM12878 cells. We considered each CTCF motif pair as positive (true looping interaction) if there was at least one measured looping interaction for which each loop anchor overlapped one of the CTCF motifs. Overlaps were calculated using the R package *InteractionSet* [77]. This resulted in 30,025 (4.2%) of 717,137 candidate motif pairs that were labeled as true looping interactions in GM12878. For the prediction validation in HeLa cells we used the 3094 Hi-C loops and 402,722 ChIA-PET interactions for CTCF and Pol2 in HeLa from the same studies [13, 16] and labeled 12,480 (1.7%) of motif pairs as true loops in HeLa cells.

ChIP-seq datasets in GM12878 cells

We downloaded publicly available ChIP-seq data from the ENCODE data portal [21, 22] by requiring the assay to be ChIP-seq, the target to be a transcription factor, the biosample term name to be GM12878, the genome assembly to be hg19, and the file-type to be bigWig. Furthermore, we filtered the data to have output type ‘fold change over control’ or ‘signal’ and to be built from two replicates. Then we selected for each TF only one unique experiment as bigWig file with either output type ‘fold change over control’ or, if unavailable, output type ‘signal’. This resulted in 124 ChIP-seq experiments for different TFs (Additional file 1: Table S1). ChIP-seq data for HeLa were retrieved analogously and filtered for the selected targets: RAD21, CTCF, ZNF143, STAT1, EP300, and ZNF143 (Additional file 2: Table S2).

ChIP-seq data types

To analyze the effect of different ChIP-seq signal types and other genomic assays on loop prediction performance, we selected five TFs (ZNF143, STAT1, SMC3, RAD21, and CTCF) and downloaded the mapped reads of ChIP-seq experiments as BAM files from the ENCODE data portal [22] and from the UCSC Genome Browser [78]. Furthermore, we downloaded signal tracks as bigWig files for ChIP-seq input control experiment, DNase-seq experiments, and ChIP-seq for the histone marks H3K4me3, H3K4me1, H3K27me3, and H3K27ac in GM12878 cells. File accession identifiers and download links are provided in Additional file 3: Table S3. We used the ChIP-seq peak caller *Q* [42] with option ‘-w’ for each human chromosome to generate signal tracks in BED format of shifted reads and qfrags. ‘Shifted reads’ are counts of mapped reads that are shifted in 5′ direction by half of the estimated fragment size. ‘qfrags’ are pairs of forward and reverse mapped reads within a given distance [42] and are shown to improve signal to noise ratio in ChIP-seq peak calling [42]. We then combined resulting BED files from all chromosomes and converted them to the bedGraph and bigWig formats using the *bedtools* [79] and *bedGrpahtoBigWig* tools from the UCSC Genome Browser [80].

ChIP-nexus data processing for RAD21 and SCM3

ChIP-nexus data for RAD21 and SMC3 in GM12878 cells were published recently [16]. We downloaded the corresponding raw reads from the Sequence Read Archive (SRA) (Run IDs SRR2312570 and SRR2312571). Reads were processed using *felxcut* for barcode removal and adapter trimming as recommended in the user guide of the *Q-nexus* tool [81]. Reads were then mapped to human genome hg19 using *Bowtie* version 2.3.2 with default settings. Duplicate reads were removed using *nexcat* [81]. Finally, we created shifted-reads and qfraq profiles using *Q-nexus* [81] with options ‘--nexus-mode’ and ‘-w’ for each chromosome and combined them to bigWig files as described above.

Similarity of ChIP-seq profiles as correlation of coverage around motifs

For each CTCF sequence motif in the human genome, we quantified the number of reads overlapping each base within +/− 500 bp around the motif center. This results in a vector $x_i = (x_{i,\ 1}, x_{i,\ 2}, \ldots, x_{i,\ n})$ where $x_{i,\ k}$ is the ChIP-seq signal at position k around CTCF motif i. ChIP-seq signal vectors for motif hits reported on the minus strand were reversed because CTCF motif sites are assumed to be symmetrically aligned to each other when cooperating at loop anchors (Fig. 1a) [13, 16, 27–29]. For all considered pairs of CTCF motifs i and j, we calculated the ChIP-seq profile similarity as Pearson correlation coefficient $r_{i,\ j}$ of the corresponding coverage vectors x_i and x_j.

Genomic sequence features of chromatin loops

Besides the correlation of ChIP-seq profiles, we used genomic features of motif pairs as features to predict interactions. The distance d is the number of bp between the two motif centers. The categorical variable orientation o is either, *convergent*, *forward*, *reverse*, or *divergent*, depending on the orientation of CTCF motifs in the pair (+−, ++, −−, and − +, respectively). The motif hit similarity s is the minimum of the two motif hit scores in each pair; we derived these motif scores from the JASPAR motif hit tracks as -$\log_{10}$ transformed p-values [35].

7C prediction model

We used a logistic regression model to predict the log-likelihood probability of CTCF motif pairs to perform chromatin looping interactions. The probability p that two sites interact is modeled as:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_{1,}\ldots\beta_k x_k$$

where β are the unknown model parameters and $x_{1,}\ldots x_k$ the features.

More specifically, for the 7C model with a single ChIP-seq experiment as input, the logistic regression model for the interaction probability p is:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 d + \beta_2 o + \beta_3 s + \beta_4 r$$

Parameters were estimated using the function 'glm()' with option 'family = binomial()' in R during model training as described below.

Training and validation of the prediction model

We used the R package *rsample* for 10-fold cross-validation. Thereby, we randomly split the dataset of CTCF motif pairs into ten equal sized subsets. For each round of cross-validation one subset is held out (test dataset) and the model parameters are trained on the remaining 90% of the samples (training dataset). The model parameters are shown for six selected TFs and combined models in Additional file 5: Figure S2A. For each split, the performance of the model is then evaluated on the test dataset. For prediction performance in HeLa cells, we trained on all motif pairs using ChIP-seq and true loops from GM12878 cells and evaluated performance on all motif pairs using the true loop data in HeLa.

Analysis of prediction performance

We quantified prediction performance using the receiver operating characteristic (ROC) and precision recall curves (PRC) as implemented in the R package *precrec* [82].

Given the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), the sensitivity is defined as TP/(TP + FN), specificity as TN/(TN + FP), precision as TP/(TP + FP), and recall as TP/(TP + FN). For each cross-validation split, the area under the curve is computed separately, and the mean across splits together with the standard deviation reported. To get binary prediction outputs, we computed the f1-score as harmonic mean of precision and recall for all prediction scores on all cross-validation folds using the R package *ROCR* [83]. Then we computed the prediction score that maximizes the f1-score as default cutoff for binary prediction output (Additional file 5: Figure S2B).

Comparison to a previously published tool

We downloaded the script provided by Oti et al. 2016 [40] from https://doi.org/10.5281/zenodo.29423. We downloaded peaks from the same CTCF ChIP-seq experiment that was analyzed with 7C from ENCODE (https://www.encodeproject.org/files/ENCFF710VEH/@@download/ENCFF710VEH.bed.gz). We filtered the CTCF motifs described above to overlap a peak region and assigned to each motif a score between 0 and 1000 according to the overlapping peak height. This data was then used as input to the script "ctcf_peaks2loops.py" to predict loops. We computed sensitivity, specificity, precision, and accuracy by overlap with the true loops described above and compared the performance to loop predictions of 7C.

Implementation of 7C and compatibility to other tools

We implemented 7C as an R package, called *sevenC*, by using existing infrastructure for chromatin interaction data from the *interactionSet* package [77] and functionality for reading bigWig files from the *rtracklayer* package [84] from the Bioconductor project [85]. Predicted loops can be written as interaction tracks for visualization in the WashU Epigenome Browser [86] or as BEDPE format using the *GenomicInteractions* package [87] for visualization in the Juicebox tool [88]. The package is freely available and easy to install from Bioconductor https://bioconductor.org/packages/sevenC. All analyses presented in this work were implemented in R and all scripts used have been made available in a separate GitHub repository: https://github.com/ibn-salem/sevenC_analysis.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-019-6088-0.

Additional file 1: Table S1. Metadata of ChIP-seq experiments from ENCODE in human GM12878 cells with accession ID and download link.

Additional file 2: Table S2. Metadata of ChIP-seq experiments from ENCODE human HeLa cells with accession ID and download link.

Additional file 3: Table S3. Accession numbers and download URLs for data sets used in data type comparisons.

Additional file 4: Figure S1. Hi-C and ChIA-PET interactions and their overlap with CTCF motif pairs. **(A)** Number of genome-wide CTCF motifs by motif hit significance cutoff. **(B)** Number of CTCF motif pairs within 1

Mb distance by motif hit significance. **(C)** Percent of CTCF motif pairs that overlap with experimentally measured Hi-C and ChIA-PET loops by the motif hit significance. **(D)** Upset plot of true loop data sets (rows) and their size (horizontal bars) with their intersections (columns, and vertical bars) based on the number of overlapping CTCF motif pairs. **(E)** Distribution of interaction span (distance between anchors) of Hi-C loops and ChIA-PET loops in GM12878 that are used as gold standard. The dotted red line indicates the distance cutoff (1 Mb) used in this study. **(F)** Number and percent of Hi-C and ChIA-PET loops that overlap with CTCF motif pairs within a distance of 1 Mb. **(G)** Number and percent of Hi-C and ChIA-PET loops that overlap with 1, 2, 3, 4, 5 or more than 5 CTCF motif pairs. The percent values are relative to all loops that overlap at least one CTCF motif pair.

Additional file 5: Figure S2. 7C model parameters and optimal cut-offs for binary prediction. **(A)** Parameter values of the logistic regression model in 7C for different features (columns), separated for different models (rows). Average of model parameters of model training in 10-fold cross-validation is shown with error bars indicating the standard deviations. While the first six rows represent the models with the indicated TF ChIP-seq data and the genomic features, “Avg. all TF” is the average across all 124 TFs analyzed and “Avg. best 10 TF” is the average across the best ten performing TF models. **(B)** Prediction performance as f1 score (y-axis) for different cutoffs on the prediction probability *p* for the six selected models.

Additional file 6: Figure S3. High resolution Hi-C map with 7C loop predictions. The red color intensity shows Hi-C interaction frequencies at an example locus of chromosome 1. The blue squares indicate 7C loop predictions using a Rad21 ChIP-seq experiment. The figure was created using the Juicebox tool by loading the combined Hi-C data set in GM12878 from [13] with mapping quality MAPQ ≥30 at a resolution of 5 kb.

Additional file 7: Figure S4. (A) Prediction performance (auPRC) of 7C when trained and evaluated on different datasets of experimentally measured loops as gold standard. Rao_GM12878 refers to Hi-C loops from [13], Tang2015_GM12878_CTCF, and Tang2015_GM12878_RNAPII to ChIA-PET loops using CTCF or Polymerase II as the target [16]. In Union, all datasets were taken together, and in Intersection, only those CTCF motif pairs that were measured in all datasets were considered positive. **(B)** Prediction performance (auPRC) of 7C compared to a logistic regression model that uses only the the total coverage signal within +/− 500 bp around the motif center at both loop anchor sites separately. In both models the genomic signal are also included as input features. The performance is shown for six selected TFs in cross-validation on the GM12878 data set. **(C)** Prediction performance of 7C with six different TFs compared to the method by Oti et al. [40]. The figure shows from top to bottom the accuracy, precision, sensitivity, and specificity of the predictions. **(D)** Venn-Diagram of CTCF motif pairs overlapping experimentally determined chromatin interactions in GM12878 cells and HeLa cells by Hi-C and ChIA-PET. **(E)** Prediction performance as auPRC (top) and auROC (bottom) of four different models (colors) on ChIP-seq data for six selected TFs (x-axis). ‘Specific TF’ is the model fitted using the ChIP-seq data indicated on the x-axis, ‘RAD21’ is the model trained on RAD21 ChIP-seq data, ‘Avg. all TF’ is a model averaged across all 124 models of analyzed TFs, and ‘Avg. best 10 TF’ is the averaged model across the 10 best performing models.

Abbreviations

qfrags: read pair combinations; 3C: chromosome conformation capture; 7C: Computational Chromosome Conformation Capture by Correlation of ChIP-seq at CTCF motifs; auPRC: Area under the PRC; CHD8: Chromodomain helicase DNA binding protein 8; DNase-seq: DNase hypersensitivity; FN: False negative; FP: False positive; PRC: Precision recall curve; ROC: Receiver operator characteristic; SRA: Sequence Read Archive; TF: Transcription factor; TN: True negative; TP: True positive

Acknowledgments

The authors thank all members of the CBDM group for fruitful discussions, especially Katerina Taškova for ideas on modeling and cross-validation. We further thank Idan Gabdank (Stanford University) for support in understanding ENCODE data set and updating it, Takaya Saito (University of Bergen) for updating and improving the *precrec* R package, Peter Hansen (*Charité, Berlin*) for support in ChIP-seq and ChIP-nexus analysis using *Q*, Morgane Thomas-Chollier (*Ecole normale supérieure*) for discussions on genome-wide motif analysis, and Sebastiaan Meijsing (Max Planck Institute for Molecular Genetics) for initial discussions on ChIP-exo profiles at chromatin loop anchors.

Authors‘contributions

JI developed and implemented the method, conceived the study and performed all analyses. MA supervised the study. JI and MA wrote the manuscript. Both authors have read and approved the manuscript.

Funding

Not applicable.

Availability of data and materials

The method 7C is implemented as R packages sevenC and can be downloaded with documentations from Bioconductor: http://bioconductor.org/packages/sevenC (DOI: https://doi.org/10.18129/B9.bioc.sevenC). The source code is also available on GitHub: https://github.com/ibn-salem/sevenC. The source code for all analyses in this manuscript is available in a separate GitHub repository: https://github.com/ibn-salem/sevenC_analysis. All the genomic data analyzed here are freely available to be downloaded from the GEO repository or ENCODE as described in the methods section.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Received: 11 April 2019 Accepted: 9 September 2019
Published online: 25 October 2019

References

1. Merkenschlager M, Nora EP. CTCF and cohesin in genome folding and transcriptional gene regulation. Annu Rev Genomics Hum Genet. 2016;17:17–43 Available from: http://www.annualreviews.org/doi/10.1146/annurev-genom-083115-022339.
2. Krijger PHL, de Laat W. Regulation of disease-associated gene expression in the 3D genome. Nature Reviews Molecular Cell Biology. 2016;17:771–82.
3. Andrey G, Mundlos S. The three-dimensional genome: regulating gene expression during pluripotency and development; 2017. p. 3646–58.
4. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13:613–26.
5. Tolhuis B, Palstra RJ, Splinter E, Grosveld F, de Laat W. Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell. 2002;10:1453–65.
6. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489:109.
7. Heidari N, Phanstiel DHD, He C, Grubert F, Jahanbani F, Kasowski M, et al. Genome-wide map of regulatory interactions in the human genome. Genome Res. 2014;24:1905–17.
8. Mora A, Sandve GK, Gabrielsen OS, Eskeland R. In the loop: promoter–enhancer interactions and bioinformatics. Brief Bioinform. 2016;17:980–95.
9. Dekker J, Rippe K, Kleckner M. Capturing chromosome conformation. Science. 2002;295:1306–11.
10. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nat Genet. 2006;38:1348.
11. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, et al. Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. 2006;16:1299–309.
12. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93. https://doi.org/10.

1126/science.1181369 Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2858594&tool=pmcentrez&rendertype=abstract.
13. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159:1665–80. https://doi.org/10.1016/j.cell.2014.11.021 Available from: http://linkinghub.elsevier.com/retrieve/pii/S0092867414014974 http://www.ncbi.nlm.nih.gov/pubmed/25497547.
14. Tiwari VK, Cope L, McGarvey KM, Ohm JE, Baylin SB. A novel 6C assay uncovers Polycomb-mediated higher order chromatin conformations. Genome Res. 2008;18:1171–9.
15. Fullwood MJ, Liu MH, Pan YF, Xu J, Mohamed H, Orlov YB, et al. An oestrogen-receptor-α-bound human chromatin interactome. Nature. 2009; 462:58–64.
16. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, et al. CTCF-Mediated Human 3D genome architecture reveals chromatin topology for transcription. Cell. 2015;1–17. Available from: doi: https://doi.org/10.1016/j.cell.2015.11.024
17. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods. 2016;13:919–22.
18. Bonev B, Cavalli G. Organization and function of the 3D genome. Nat Rev Genet. 2016;17:661–78.
19. Sati S, Cavalli G. Chromosome conformation capture technologies and their impact in understanding genome function. Chromosoma. 2017;126:33–44.
20. Schmitt AD, Hu M, Ren B. Genome-wide mapping and analysis of chromosome architecture. Nature Reviews Molecular Cell Biology. 2016;17:743–55.
21. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74 Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3439153&tool=pmcentrez&rendertype=abstract.
22. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018;46:D794–801.
23. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30 Available from: http://www.nature.com/doifinder/10.1038/nature14248 http://www.ncbi.nlm.nih.gov/pubmed/25693563 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4530010.
24. Kim TH, Abdullaev ZK, Smith AD, Ching KA, Loukinov DI, Green RD, et al. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell. 2007;128:1231–45.
25. Nagy G, Czipa E, Steiner L, Pongor T, Barta S, Park E, et al. Motif oriented high-resolution analysis of ChIP-seq data reveals the topological order of CTCF and cohesin proteins on DNA. BMC Genomics. 2016;17:637.
26. Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, et al. Comparative hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Rep. 2015;10:1297–309 Available from: http://linkinghub.elsevier.com/retrieve/pii/S2211124715001126 http://www.ncbi.nlm.nih.gov/pubmed/25732821 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4542312.
27. Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, et al. CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. Cell. 2015;162:900–10 Available from: http://linkinghub.elsevier.com/retrieve/pii/S0092867415009150.
28. Wit ED, Vos ESM, Holwerda SJB, Valdes-quezada C, Verstegen MJAM, Teunissen H, et al. CTCF Binding Polarity Determines Chromatin Looping. Mol Cell. 2015:1–9. Available from. https://doi.org/10.1016/j.molcel.2015.09.023.
29. Sanborn AL, Rao SSP, Huang S-C, Durand NC, Huntley MH, Jewett AI, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci. 2015;112:201518552.
30. Fudenberg G, Imakaev M, Lu C, Goloborodko A, Abdennur N, Mirny LA. Formation of chromosomal domains by loop extrusion. Cell Rep. 2016;15:2038–49.
31. Schwarzer W, Abdennur N, Goloborodko A, Pekowska A, Fudenberg G, Loe-Mie Y, et al. Two independent modes of chromatin organization revealed by cohesin removal. Nature. 2017;551:51–6.
32. Orlando V, Strutt H, Paro R. Analysis of chromatin structure byin VivoFormaldehyde cross-linking. Methods. 1997;11:205–14.
33. Hoffman EA, Frey BL, Smith LM, Auble DT. Formaldehyde crosslinking: a tool for the study of chromatin complexes. J Biol Chem. 2015;290:26404–11.
34. Starick SR, Ibn-Salem J, Jurk M, Hernandez C, Love MI, Chung H-R, et al. ChIP-exo signal associated with DNA-binding motifs provide insights into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. Genome Res. 2015;25:825–35 Available from: http://www.ncbi.nlm.nih.gov/pubmed/25720775 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4448679&tool=pmcentrez&rendertype=abstract.
35. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 2018;46:D260–6.
36. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10:1–21.
37. Ye B-Y, Shen W-L, Wang D, Li P, Zhang Z, Shi M-L, et al. ZNF143 is involved in CTCF-mediated chromatin interactions by cooperation with cohesin and other partners. Mol Biol. 2016;50:431–7 Available from: http://link.springer.com/10.1134/S0026893316030031.
38. Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, et al. ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. Nat Commun. 2015;2:6186.
39. Rao SSP, Huang S-C, Hilaire BGS, Engreitz JM, Perez EM, Kieffer-Kwon K-R, et al. Cohesin loss eliminates all loop domains. Cell. 2017;171:305–320.e24.
40. Oti M, Falck J, Huynen MA, Zhou H. CTCF-mediated chromatin loops enclose inducible gene regulatory domains. BMC Genomics. 2016;17:252 Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-2516-6.
41. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137 Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2592715&tool=pmcentrez&rendertype=abstract.
42. Hansen P, Hecht J, Ibrahim DM, Krannich A, Truss M, Robinson PN. Saturation analysis of ChIP-seq data for reproducible identification of binding peaks. Genome Res. 2015;25:1391–400 Available from: http://www.ncbi.nlm.nih.gov/pubmed/26163319 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4561497.
43. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, Pritchard JK. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. 2010. https://doi.org/10.1101/gr.112623.110 Available from: http://genome.cshlp.org/content/early/2010/11/24/gr.112623.110. [cited 2017 Dec 19].
44. Yardımcı GG, Frank CL, Crawford GE, Ohler U. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. Nucleic Acids Res. 2014;42:11865–78.
45. Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell. 2011;147:1408–19 Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3243364&tool=pmcentrez&rendertype=abstract.
46. He Q, Johnston J, Zeitlinger J. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nat Biotechnol. 2015;33:395–401. Available from. https://doi.org/10.1038/nbt.3121.
47. Wen Z, Huang Z-T, Zhang R, Peng C. ZNF143 is a regulator of chromatin loop. Cell Biol Toxicol. 2018;34:471–8.
48. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. Nucleic Acids Res. 2017;45:D408–14.
49. Ishihara K, Oshimura M, Nakao M. CTCF-dependent chromatin insulator is linked to epigenetic remodeling. Mol Cell. 2006;23:733–42.
50. Yuan C-C, Zhao X, Florens L, Swanson SK, Washburn MP, Hernandez N. CHD8 associates with human Staf and contributes to efficient U6 RNA polymerase III transcription. Mol Cell Biol. 2007;27:8729–38.
51. Vicenzi E, Poli G. The interferon-stimulated gene TRIM22: a double-edged sword in HIV-1 infection. Cytokine Growth Factor Rev. 2018;40:40–7.
52. Sheffield NC, Thurman RE, Song L, Safi A, Stamatoyannopoulos JA, Lenhard B, et al. Patterns of regulatory activity across diverse human cell types predict tissue identity , transcription factor binding , and long-range interactions predict tissue identity , transcription factor binding , and long-range interactions. Genome Res. 2013;23:777–88.
53. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database. 2017;2017:1665–80 Available from: https://academic.oup.com/database/article-lookup/doi/10.1093/database/bax028.

54. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507:455–61 Available from: http://www.ncbi.nlm.nih.gov/pubmed/24670763.
55. O'Connor TR, Bailey TL. Creating and validating cis-regulatory maps of tissue-specific gene expression regulation. Nucleic Acids Res. 2014;42:11000–10.
56. O'Connor T, Bodén M, Bailey TL. CisMapper: predicting regulatory interactions from transcription factor ChIP-seq data. Nucleic Acids Res. 2017;45:e19–e19.
57. Roy S, Siahpirani AF, Chasman D, Knaack S, Ay F, Stewart R, et al. A predictive modeling approach for cell line-specific long-range regulatory interactions. Nucleic Acids Res. 2015;43:8694–712.
58. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat Genet. 2015;48:1–10.
59. Zhu Y, Chen Z, Zhang K, Wang M, Medovoy D, Whitaker JW, et al. Constructing 3D interaction maps from 1D epigenomes. Nat Commun. 2016;7:10812.
60. Schreiber J, Libbrecht M, Bilmes J, Noble WS. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. bioRxiv. 2018;103614.
61. Dzida T, Iqbal M, Charapitsa I, Reid G, Stunnenberg H, Matarese F, et al. Predicting stimulation-dependent enhancer-promoter interactions from ChIP-Seq time course data. PeerJ. 2017;5:e3742.
62. Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, et al. Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. Nat Genet. 2017;49:1428–36.
63. Zhao C, Li X, Hu H. PETModule: a motif module based approach for enhancer target gene prediction. Sci Rep. 2016;6:30043.
64. Naville M, Ishibashi M, Ferg M, Bengani H, Rinkwitz S, Krecsmarik M, et al. Long-range evolutionary constraints reveal cis-regulatory interactions on the human X chromosome. Nat Commun. 2015;6:6904.
65. Nikumbh S, Pfeifer N. Genetic sequence-based prediction of long-range chromatin interactions suggests a potential role of short tandem repeat sequences in genome organization. BMC Bioinformatics. 2017;18:218.
66. Brackley CA, Brown JM, Waithe D, Babbs C, Davies J, Hughes JR, et al. Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. Genome Biology. 2016;17:59.
67. Chen Y, Wang Y, Xuan Z, Zhang M. De novo deciphering three-dimensional chromatin interaction and topological domains by wavelet transformation of epigenetic profiles. Nucleic Acids Res. 2016;44:e106.
68. Pierro MD, Cheng RR, Aiden EL, Wolynes PG, Onuchic JN. De novo prediction of human chromosome structures: epigenetic marking patterns encode genome architecture. Proc Natl Acad Sci. 2017;114:12126–31.
69. Rowley MJ, Nichols MH, Lyu X, Ando-Kuri M, Rivera ISM, Hermetz K, et al. Evolutionarily Conserved Principles Predict 3D Chromatin Organization. Molecular Cell. 2017;67:837-852.e7.
70. Liang J, Lacroix L, Gamot A, Cuddapah S, Queille S, Lhoumaud P, et al. Chromatin immunoprecipitation indirect peaks highlight long-range interactions of insulator proteins and pol II pausing. Mol Cell. 2014;53:672–81.
71. Mourad R, Li L, Cuvier O. Uncovering direct and indirect molecular determinants of chromatin loops using a computational integrative approach. PLoS Comput Biol. 2017;13:e1005538.
72. Hnisz D, Weintraub AS, Day DS, Valton A, Bak RO, Li CH, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science. 2016;351:1454–8.
73. Sun F, Chronis C, Kronenberg M, Chen X-F, Su T, Lay FD, et al. Promoter-Enhancer Communication Occurs Primarily within Insulated Neighborhoods. Mol Cell. 2018; Available from: http://www.sciencedirect.com/science/article/pii/S109727651830933X. [cited 2018 Dec 13].
74. Bonev B, Cohen NM, Szabo Q, Hugnot J, Tanay A, Cavalli G, et al. Multiscale 3D genome rewiring during mouse article multiscale 3D genome rewiring during mouse neural development. Cell. 2017;171:557.e1–557.e24. Available from. https://doi.org/10.1016/j.cell.2017.09.043.
75. 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.
76. Mathelier A, Fornes O, Arenillas DJ, Chen C, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2015;44:gkv1176 Available from: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv1176.
77. Lun ATL, Perry M, Ing-Simmons E. Infrastructure for genomic interactions : Bioconductor classes for Hi-C , ChIA-PET and related experiments [version 1; referees: 2 approved]. F1000Research. 2016;5:1–6 Available from: http://f1000research.com/articles/5-950/v1.
78. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 2006;34:D590–8. Available from. https://doi.org/10.1093/nar/gkj144.
79. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2 Available from: https://academic.oup.com/bioinformatics/article/26/6/841/244688. [cited 2017 Dec 21].
80. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D. BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics. 2010;26:2204–7 Available from: https://academic.oup.com/bioinformatics/article/26/17/2204/199001. [cited 2017 Dec 21].
81. Hansen P, Hecht J, Ibn-Salem J, Menkuec BS, Roskosch S, Truss M, et al. Q-nexus: a comprehensive and efficient analysis pipeline designed for ChIP-nexus. BMC Genomics. 2016;17:873 Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3164-6.
82. Saito T, Rehmsmeier M. Precrec: Fast and accurate precision-recall and ROC curve calculations in R. Bioinformatics. 2016;33:btw570 Available from: http://bioinformatics.oxfordjournals.org/lookup/doi/10.1093/bioinformatics/btw570.
83. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21:3940–1 Available from: https://academic.oup.com/bioinformatics/article/21/20/3940/202693. [cited 2017 Dec 21].
84. Lawrence M, Gentleman R, Carey V. rtracklayer: an R package for interfacing with genome browsers. Bioinformatics. 2009;25:1841–2 Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2705236/. [cited 2017 Dec 7].
85. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with Bioconductor. Nat Methods. 2015;12:115–21 Available from: http://www.nature.com/doifinder/10.1038/nmeth.3252.
86. Zhou X, Lowdon RF, Li D, Lawson HA, Madden P a F, Costello JF, et al. Exploring long-range genome interactions using the WashU Epigenome Browser. Nat Methods. 2013;10:375–6 Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3820286&tool=pmcentrez&rendertype=abstract.
87. Harmston N, Ing-Simmons E, Perry M, Barešić A, Lenhard B. GenomicInteractions: An R/Bioconductor package for manipulating and investigating chromatin interaction data. BMC Genomics. 2015;16:963 Available from: http://www.biomedcentral.com/1471-2164/16/963.
88. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, et al. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 2016;3:99–101.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.