Loop extrusion model predicts CTCF interaction specificity

Wang Xi[1,2], Michael A Beer[1,2]

[1]Department of Biomedical Engineering, and [2]McKusick-Nathans Department of Genetic Medicine, Johns Hopkins University School of Medicine, 733 N. Broadway, Baltimore, MD 21205

Abstract

Three-dimensional chromatin looping interactions play an important role in constraining enhancer-promoter interactions and mediating transcriptional gene regulation. CTCF is thought to play a critical role in the formation of these loops, but the specificity of which CTCF binding events form loops and which do not is difficult to predict. Loops often have convergent CTCF binding site motif orientation, but this constraint alone is only weakly predictive of genome-wide interaction data. Here we present an easily interpretable and simple mathematical model of CTCF mediated loop formation which is consistent with Cohesin extrusion and can predict ChIA-PET CTCF looping interaction measurements with high accuracy. Competition between overlapping loops is a critical determinant of loop specificity. We show that this model is consistent with observed chromatin interaction frequency changes induced by CTCF binding site deletion, inversion, and mutation, and is also consistent with observed constraints on validated enhancer-promoter interactions.

Introduction

High order chromatin structure affects various biological processes within the nucleus, ranging from gene regulation to DNA repair. The structural basis of interphase chromatin has been extensively studied by various Chromatin Conformation Capture (3C)[1–4] techniques, and has revealed functional units including chromosome compartments[1], topologically associated domains (TADs)[5] and loops[6]. Chromosomal compartments, which exhibit a checkerboard pattern on a Hi-C map, correspond to active or inactive chromatin across several megabases[1]. On the other hand, TADs and sub-TAD loops represent enriched chromatin interactions that appear at a scale of hundreds of kilobases or below[5,6]. These smaller loops shape local chromatin structure, and their disruption has been reported to lead to dramatic dysregulation of nearby gene expression[7,8]. The most prominent feature of TADs and loops is that their boundaries are usually marked by CTCF and Cohesin binding[5,6]. CTCF was initially thought to work mainly as an insulator of active chromatin marks, but since has been recognized to play a major role in chromatin organization, whereby pairs of CTCFs bind and serve as loop anchors to constrain interactions between distant regulatory elements[9,10] (Fig. 1a). It has been suggested that CTCF and Cohesin mediate TAD and loop formation through a loop extrusion mechanism, where Cohesin translocates along a nascent chromatin loop until blocked by CTCF[11,12] (Fig. 1b). Polymer simulations of a loop extrusion model successfully reconstructed TAD like structures, and predicted the impact of CTCF or Cohesin degradation on TAD strength[11]. Moreover, multiple experiments have validated *in vitro* that Cohesin is capable of moving through nucleosomal DNA[13] and generating a growing DNA loop progressively as it moves[14,15].

There are ~50,000 CTCF binding sites in normal mammalian cells, which corresponds to over 1 million possible CTCF pairs lying within 1Mb of each other. However, only about 2~5% of these are identified to be interacting by direct Hi-C or ChIA-PET measurements[6,16] (Fig. 1c). This

raises the important question about the difference between interacting and non-interacting CTCF pairs. Although it has been observed that CTCF motif orientation in loop anchors tends to be convergent[6,17], the vast majority of convergent CTCF motif pairs are not interacting with each other, therefore, a more comprehensive model of how CTCF interaction specificity is regulated remains to be elucidated. Several experiments have investigated the determinants of loop formation, such as, binding of CTCF or Cohesin[17–20], but could not explain why only a subset of available CTCF binding pairs are interacting in each cell type. While CTCF and Cohesin have been shown to play a role in determining 3D chromatin interactions overall, the process of loop extrusion has not yet been directly validated to be the molecular mechanism underlying CTCF looping interactions. Meanwhile, previous physical modeling of nuclear organization has mostly focused on the level of compartments or TADs rather than single loop[11,12]. There is one machine learning model, Lollipop, which utilized a large set of genomic and epigenomic features to predict CTCF interactions specificity with high accuracy[21]. This model provides some insight into this problem but has not fully revealed how these features play a role in the process of loop formation. Moreover, inspection of the Lollipop model shows that many of the 77 features used have substantial redundancy, making it hard to distinguish causal mechanisms, and implying that there may be simpler rules driving the specificity of CTCF interactions.

Here, we propose that CTCF interaction specificity can be predicted by a simple model based on loop extrusion. The success of this model gives indirect support for loop extrusion as an important mechanism regulating CTCF interaction specificity. We build a quantitative model to describe CTCF-mediated loop formation with only four features, CTCF binding intensity (BI), CTCF motif orientation, distance between CTCF binding events, and loop competition (LC) (Fig. 1c-1e). We show that this model can predict both ChIA-PET and Micro-C annotated CTCF loops with high accuracy. A novel component of our model, the competition between overlapping loops, is crucial to loop formation. Our model of loop competition also provides a simple mechanism by which genetic variation in CTCF binding sites directly contributes to observed differences in chromatin contact frequency. We show that our model is also predictive of cell-type specific CTCF loops. We further validate this model by CRISPRi perturbation of loop anchor binding sites, and by the predicted CTCF loops' ability to constrain enhancer promoter interactions. We expect that the insights derived from this model may also shed light onto the related important problem of enhancer-promoter interaction prediction, and the mechanisms by which the specificity of enhancer-promoter interactions are regulated.

Results

Quantitative model of loop formation by extrusion

In this loop extrusion model, the key components are CTCF, Cohesin and other loop-extruding factors[11] (Fig. 1b). The formation of a CTCF-mediated loop in mammalian cells begins when the ring-shaped Cohesin is loaded onto the DNA chromatin fiber. Through the motor activity of Cohesin or other co-factors like NIPBL, Cohesin translocates along the chromatin fiber in an ATP-dependent manner, which pushes and progressively enlarges the DNA loop. This process proceeds until Cohesin comes into contact with a DNA bound CTCF protein on each strand of the loop, which acts as a barrier that prevents further translocation. That CTCF acts as a blockade to Cohesin and acts as primary determinant of stable Cohesin locations in the genome is supported by gkm-SVM sequence analysis showing that the CTCF binding site alone is able to explain genomic binding of SMC3, a Cohesin subunit.[22] The most stable loop configuration is

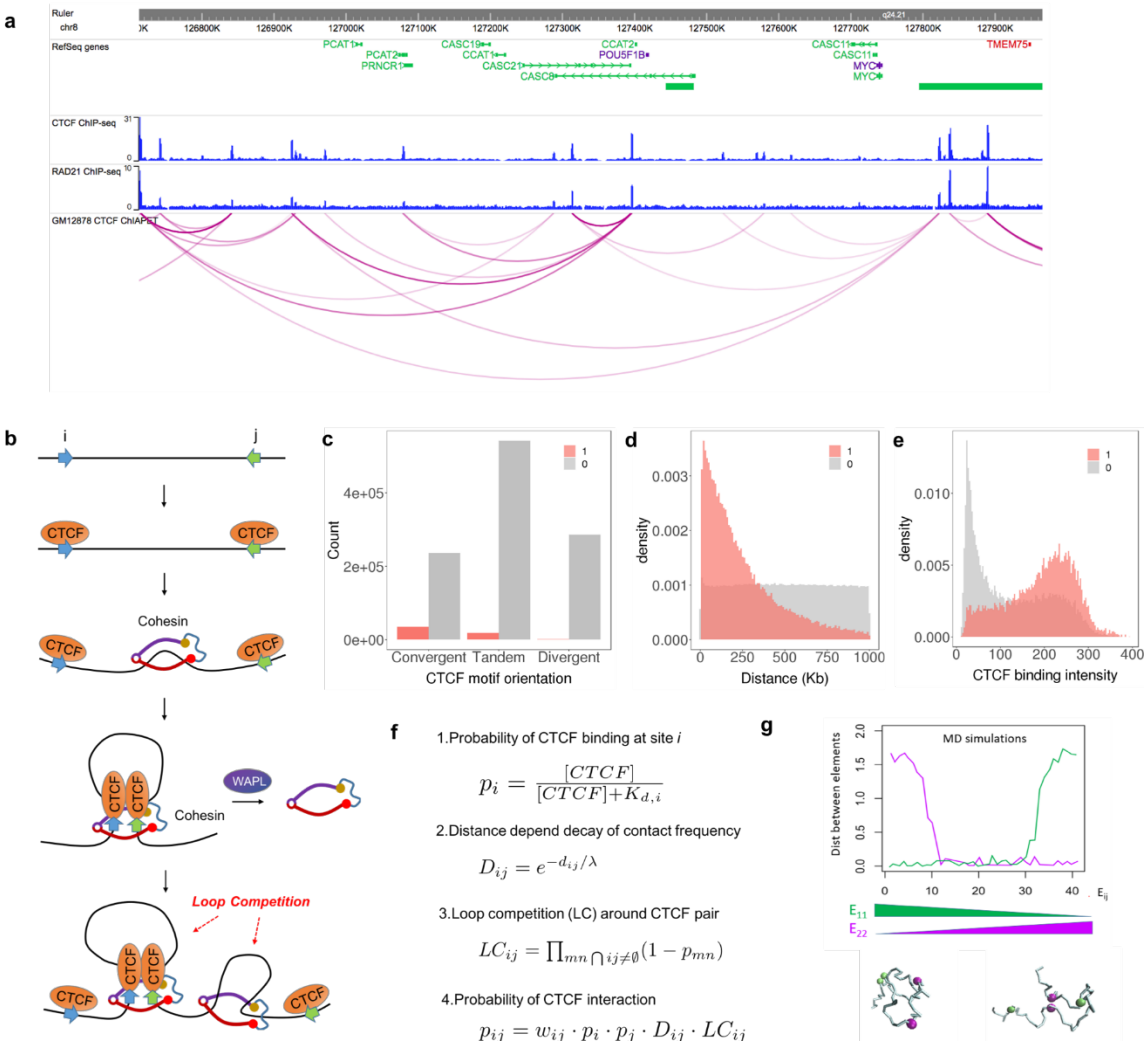

Figure.1 **Mathematical formulation of a loop extrusion model.** (a) We use CTCF ChIA-PET to train our model: the contact profile of the Myc locus in GM12878 is shown. (b) Loop extrusion model: Cohesin is loaded between (typically convergent) CTCF pairs, and the loop forms progressively as Cohesin translocates along the chromatin fiber. The extrusion process stops when Cohesin is stalled by CTCF. WAPL unloads Cohesin from chromatin. An existing loop could block movement of another Cohesin protein, leading to loop competition. (c) While measured loops prefer convergent CTCF pairs, other orientations also interact with significant frequencies and many neighboring (<1Mb) convergent CTCF motifs do not form loops: shown are interacting pair counts (1, red), and non-interacting pair counts (0, grey). (d) Distance/ distribution for interacting and non-interacting CTCF pairs. (e) CTCF binding intensity distribution for interaction and non-interaction CTCF pairs. (f) Mathematical model of loop interaction probability formed by this extrusion process. (g) Polymer simulation of competing loops. Two overlapping CTCF pairs are simulated under different interaction energies E_{11} and E_{22}.

thus a Cohesin bound DNA loop with a CTCF bound at each base of the loop, and there is a notable preference for these CTCF binding sites to be in a convergent orientation.

We built a simple model which predicts the probability of formation for all possible loops by quantitatively combining the contribution of each step in this process. (Fig. 1f). First, the probability of CTCF binding at each genomic binding site is described by the chemical equilibrium:

$$p_i = \frac{[CTCF]}{[CTCF]+K_{d,i}}$$

where [CTCF] is the concentration of CTCF to be inferred, $K_{d,i}$ is the local dissociation constant at site i.[23] We use the CTCF ChIP-seq signal to estimate this local $K_{d,i}$. We can combine the unknown $K_{d,i}$ and [CTCF] to write $p_i = \frac{1}{1+\frac{K_{d,i}}{[CTCF]}} = \frac{x}{x+a}$, and we assume the local ChIP-seq signal x is inversely proportional to $K_{d,i}$/[CTCF], with a scaling factor of a. We will learn the best value of the parameter a from the ChIA-PET data. These binding probabilities contribute independently to a loop forming between CTCF site i and CTCF site j. In addition to the binding probability at each potential loop anchor site, we account for the contribution of CTCF motif orientation on loop stability with a scalar, w_{ij}, and this term takes three different values, 1, 1/w, and 1/w^2, for convergent, tandem or divergent CTCF motifs[17,19]. The extrusion process adds an exponential term which decays with the distance, d_{ij}, between CTCF site i and site j, $D_{ij} = e^{-d_{ij}/\lambda}$, reflecting the probability of CTCF stochastically dissociating from the DNA fiber while translocating along it. This leads to decreased loop interaction frequency when the distance between two CTCF bound regions gets larger. The parameter λ can be interpreted as the processivity of Cohesin, or equivalently, the average CTCF loop length, which been estimated to be about 300kb[11]. The final notable component of our loop extrusion model is the effect of the loop competition. The mechanism of loop extrusion implies that one Cohesin bound loop could block additional Cohesin procession. This prevents all other CTCF pairs that overlap with a formed loop from interacting, since other Cohesins could not pass through, no matter where they load[11]. Therefore, the formation of one loop excludes all other overlapping loops, so the contribution of loop competition is:

$$LC_{ij} = \prod_{mn \cap ij \neq \emptyset}(1 - p_{mn}),$$

where p_{mn} is the probability of loop formation of all possible CTCF pairs overlapping site i and j. Consistent with this idea, we performed a set of polymer simulations of a DNA fragment bounded by two pairs of CTCF to show that overlapping loops are likely to be mutually exclusive, as they are in potential competition with each other (Fig. 1g). We assume that the probability of Cohesin loading is constant along genome, for the moment ignoring any non-uniformity or nuclear compartmentation. Thus in our complete model, the probability of a loop forming between CTCF binding sites i and j is given by:

(1) $$p_{ij} = w_{ij} \cdot p_i \cdot p_j \cdot D_{ij} \cdot LC_{ij}$$

Parameter determination for loop extrusion probabilistic model

We used publicly available CTCF ChIA-PET data[16] in GM12878 and HeLa cells to determine the values of the parameters [CTCF], w and λ in our model. Long read ChIA-PET data was processed with ChIA-PET2 software under standard protocols to identify significant loops[24]. The high resolution and quality of this ChIA-PET data makes it suitable for predicting CTCF-mediated loops and training our model. First, the average anchor length of ChIA-PET loop is around 1kb, which is close to the size of single CTCF binding site (Supplementary Fig. 2a). Second, comparison of CTCF ChIP-seq peaks with overlapping ChIA-PET anchors shows that they are relatively centered around each other (Supplementary Fig. 2b). We will use the CTCF ChIP-seq signal at each site as $K_{d,i}$ to infer the local CTCF binding probability. CTCF motif annotation is performed with STORM[25].

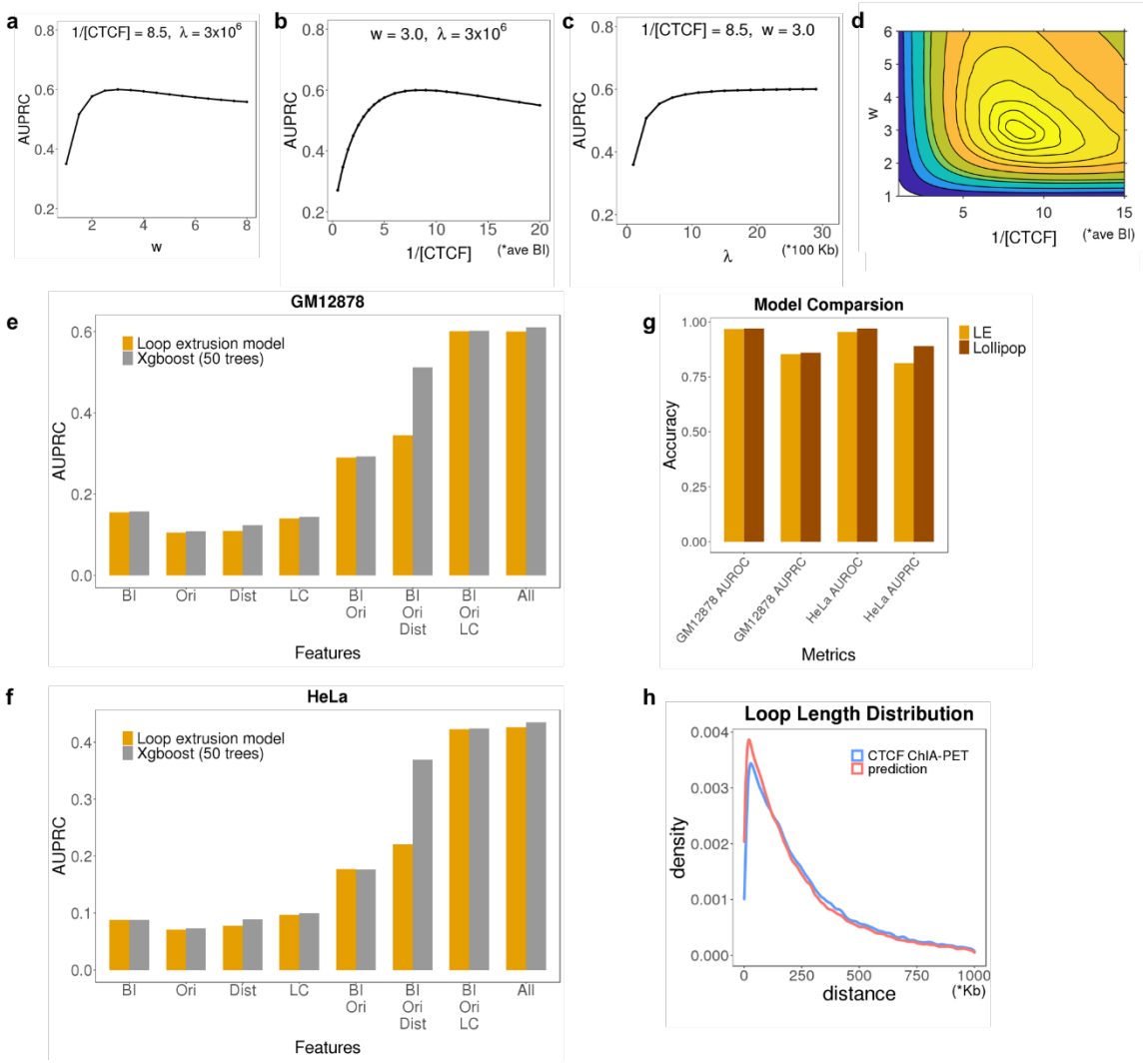

Figure 2. **Parameter optimization, performance evaluation, and feature importance.** (a)-(d) Model performance is evaluated by area under precision-recall curve (AUPRC) as parameters are varied individually (a)-(c), and by grid search (d). (e)-(f) Performance of model with different combination of features for GM12878 and HeLa. BI – CTCF binding intensity; Ori – CTCF motif orientation; Dist – distance; LC – loop competition. Performance is compared against xgboost model with 50 trees. (g) Model comparison against Lollipop under class ratio 1:5 (positive vs. negative). (h) Loop length distribution of ChIA-PET annotated loops and predicted interacting loops.

We determined the optimal value of the model parameters by fitting the loop extrusion model to CTCF ChIA-PET data (Fig. 2a-2d, see Methods), by comparing measurements of actual loop formation to the probability of loop formation predicted by our model (AUPRC), using GM12878 and HeLa. The low dimensionality of our model makes overfitting highly unlikely, so we trained these three parameters on the full dataset. We did a comprehensive grid search in ($[CTCF]$, w, and λ) in GM12878, and found that the w value of best agreement with data is 3.0, which implies that a convergent CTCF pair is three times more likely to interact than a tandem CTCF pair with equivalent CTCF binding probability and distance, and nine times more likely than a divergent pair. The optimal $1/[CTCF]$ is 8.5 times the average CTCF binding intensity. The model is quite robust to parameter choices with a broad peak of high performance in the range of w (2~4) and

1/[CTCF] (5~10) (Fig 2d). Also, the optimal parameters derived from training on GM12878 and HeLa are very similar (Supplementary Fig. 3). For λ, we expected the optimal value to be around the average loop length of 300kb, as reported in previous literature[11]. However, the agreement between our model and the ChIA-PET data increases monotonically with λ, which implies that distance information is dispensable for the prediction of CTCF interactions, as larger λ reduces the variation of the exponential term with distance (Fig. 2c). Moreover, leaving the distance-associated exponential term out completely makes the agreement with data slightly better. This stands in contrast with the general view that distance regulates chromatin interaction frequency. We shall address this apparent paradox below.

Loop extrusion model accurately predicts formation of CTCF-mediated loops

We applied our quantitative model of loop extrusion (Eq. 1) to CTCF ChIA-PET data to predict CTCF interaction specificity. A total of 55,189 and 21,560 significant interactions with CTCF binding both anchors are identified for GM12878 and HeLa. All ChIA-PET detected CTCF-mediated loop interactions were labeled as positive samples, and all other (non-interacting) CTCF pairs within 1Mb were labelled as negative samples. Due to different sequencing depth and cell-type variability, the positive versus negative class ratio is roughly 1:20 for GM12878 and 1:37 for HeLa, with non-interacting CTCF pairs far outnumbering interacting pairs. A small fraction of loops had more than one CTCF binding peak at one of the anchors, when these could not be unambiguously assigned they were removed from analysis.

To assess the importance of each feature in our model, we trained on each individual feature and all combinations of features, including: CTCF binding intensity, CTCF motif orientation, distance and loop competition. An interaction probability p_{ij} was predicted for all positive and negative pairs for each model, and was then compared to the true class label. Due to the huge class imbalance of CTCF interaction datasets, we employed area under the precision-recall curve (AUPRC) to evaluate model performance. For both GM12878 and HeLa cell lines, we observed that none of the four features alone could accurately predict interaction specificity of CTCF (AUPRC 0.2~0.3) while combining them increased the performance significantly (Fig. 2e-2f, Supplementary Table. 1). The best performance is given by the model constituted of CTCF binding intensity (BI), CTCF motif orientation (Ori) and loop competition (LC), with AUPRC = 0.601. To make sure that our model indeed captures the underlying mechanism of loop formation, rather than simply leveraging information associated with interacting CTCF pairs, we also constructed a more complicated machine learning model using boosted trees, with exactly the same features to compare with our model. Surprisingly, the boosting model, with no constraints on the form of the nonlinearity among features, is only marginally better than our model (AUPRC = 0.602). This performance increases confidence in the validity of the mathematical formulation of our model and the loop extrusion hypothesis. Furthermore, adding distance (Dist) as a feature does not significantly increase performance in either our loop extrusion model or the boosting model (AUPRC = 0.611). This confirms our earlier observation (from the insensitivity of performance to λ) that distance is weakly informative and seems to be redundant for our model in this task. Notably, even without distance information, the distance distribution of interacting CTCF pairs predicted from loop extrusion model is still quite close to experimental data (Fig. 2h). Results in HeLa cell line (Fig. 2f) are qualitatively consistent with GM12878, with reduced AUPRC attributable to the larger HeLa in class ratio difference.

We then compared our model with a previously published machine learning model, Lollipop, which successfully predicted CTCF-mediated loop with 77 different sequence and epigenomic features (Fig. 2g). Under the same class ratio 1:5, we found that in both cell lines, our loop extrusion model is nearly as accurate as Lollipop in terms of both AUROC (area under the receiver operator characteristic curve) and AUPRC, which indicates that the information contained in our model is quite comprehensive, relatively more compact, and more easily interpretable.

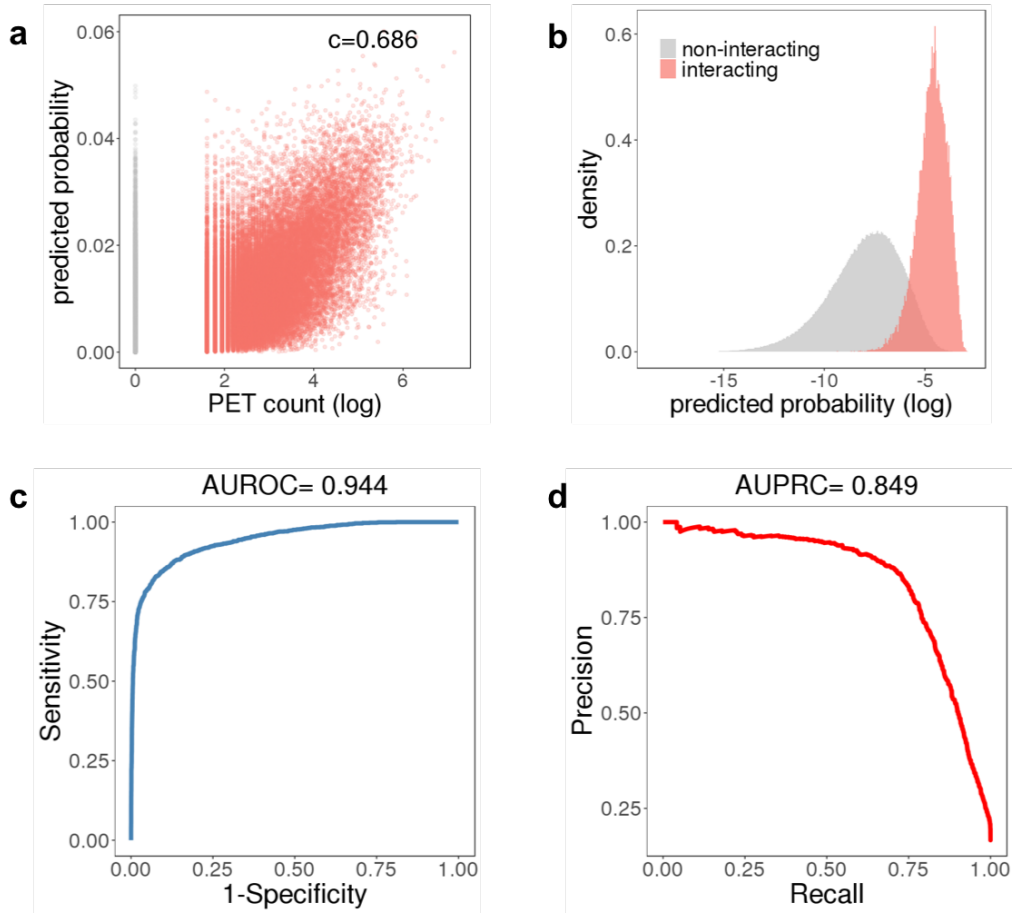

Figure 3. **Model validation with external Micro-C dataset.** (a) Distribution of PET count (log scale) against loop extrusion model predicted interaction probability. Red dots are interacting CTCF pairs while grey dots are non-interacting CTCF pairs. (b) Distribution of loop extrusion model predicted interaction probability. (c)-(d) Validation of model prediction performance on Micro-C CTCF loops with AUROC and AUPRC.

To evaluate the quantitative predictions of our model, we compared the predicted interaction probability of CTCF pairs, conditioned on their quantitative labels, to the PET counts from the ChIA-PET experiment. The model probabilities are highly correlated with PET count (C = 0.686 for GM12878 and 0.531 for HeLa) (Fig. 3a). In addition, positive and negative CTCF pairs are clearly separated by predicted interaction probability (Fig. 3b).

To validate our model on an additional external dataset, we predicted CTCF loops identified from a recently published high resolution Micro-C dataset.[26] In total, 15,945 significant loops at 1kb resolution were detected in this dataset with HICCUPS.[6] For purposes of predicting CTCF-mediated loops, we sampled positive loops with CTCF binding at both ends, and generated a

five times larger negative set by sampling from non-interacting CTCF pairs. We applied our model on this dataset and achieved (AUROC = 0.944, AUPRC = 0.849) (Fig. 3c-3d), indicating that we are able to accurately predict CTCF interaction at a similar performance to those detected by ChIA-PET. Taken together, the analysis of CTCF ChIA-PET and Micro-C data shows that CTCF interaction can be successfully predicted from the loop extrusion model, and only requires information of local CTCF binding intensity, CTCF motif orientation and loop competition throughout the local neighboring region (up to 3Mb). We tested adding additional features to the boosting model, e.g. Cohesin ChIP-seq and DNase-seq signal, but found that these did not improve performance significantly (Supplementary Table. 2).

Loop competition is a more powerful predictor than distance

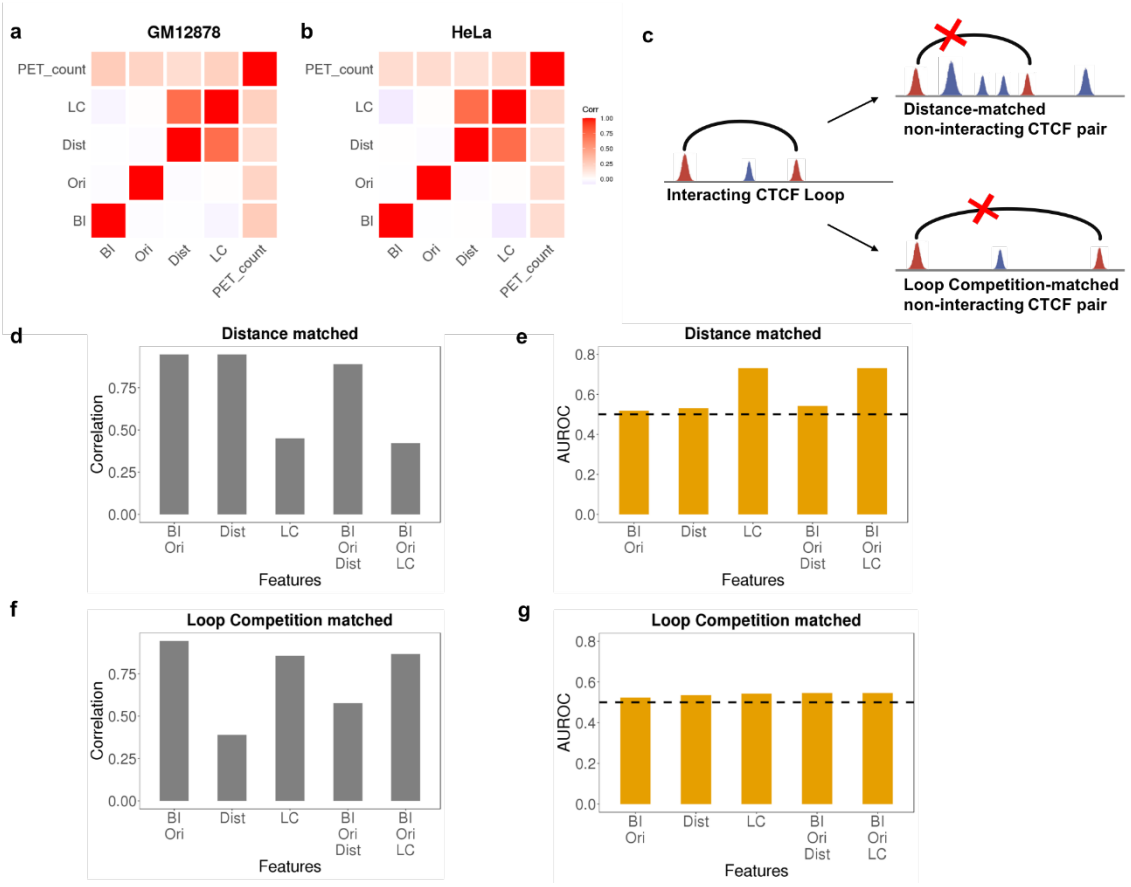

Figure 4. **Loop competition is a more crucial determinant than distance.** (a)-(b). Correlation between CTCF binding intensity, CTCF motif orientation, distance, loop competition and PET count (log scale). (c). We can evaluate the relative informative value of distance and loop competition by testing the model on distance-matched and loop competition-matched subsets of the full data. (d),(f). Correlation between positive and negative set for different combination of features in both settings. (e), (g). AUROC of loop extrusion model with different combination of features in both settings.

Because of the simple formulation of our model, we can evaluate the relative importance of each component to the loop formation process. First, we calculated the correlation between all pairs of features and PET count (Fig. 4a-4b). The only two features highly correlated with each other are distance and loop competition (Dist and LC). This correlation is to be expected, because the more distant two CTCF binding sites are, the more likely the existence of a competing loop becomes. But which of these correlated features is more predictive of CTCF

interactions by itself, distance or loop competition? Almost all studies of genome-wide chromosomal conformation capture experiments, including Hi-C, ChIA-PET and Micro-C, have reported that a longer distance between two regions is associated with reduced interaction frequency[6,16]. Intuitively, distant regions contact less frequently by diffusion in three-dimensional space, but the precise mechanism of the observed loop distance dependence has not yet been supported by much direct experimental evidence. It is possible that the distance dependence is associated with some other factor which determines loop formation.

To determine the relative importance of distance and loop competition, we generated distance-matched and loop-competition-matched test sets by sampling ChIA-PET data to distinguish their contributions (Fig. 4c). In distance-matched sampling, for each positive loop, we selected one negative loop with similar CTCF binding intensity, CTCF motif orientation, and distance (within a factor of 2 for BI and Dist). In other words, every feature except loop competition is matched between this negative set and the positive set. Compared to the full dataset, it should be harder to distinguish the positives and negatives in this set because loop competition is the only unmatched feature. By evaluating our model with on this distance matched set with different subsets of features, we find, as expected, CTCF binding intensity, CTCF motif orientation or distance are not useful for prediction on this subset (Fig. 4e). In contrast, the model including loop competition (LC) reached AUROC = 0.730, indicating that loop competition alone is predictive in this context and carries unique information about loop formation that doesn't exist in distance alone. We next performed loop-competition-matched sampling in a similar fashion, selecting positive and negative loops with similar levels of loop competition (within a factor of 2) but unmatched distance. In contrast to the distance matched subset, in the loop-competition-matched subset, distance is not predictive of CTCF loop formation, showing that distance itself cannot explain CTCF interaction specificity (Fig. 4g). The fact that loop competition is predictive in a distance matched context, while distance is not predictive in a loop-competition-matched context, indicates that loop-competition is the more informative feature. This test suggests that distance can be a predictive feature because it can serve as a proxy for loop competition when loop competition is not an explicit feature of the model. Our results show that the negative correlation between distance and contact frequency is likely to be mediated by the effect of loop competition. Consistent with this interpretation, distance has the weakest correlation with the PET count of loops among the four features (Supplementary Fig. 5). These computational experiments confer support for loop competition as an important determinant of CTCF interaction specificity.

Testing loop competition by CTCF disruption in population Hi-C data

Our model makes novel predictions about how a single CTCF binding site disruption would be expected to impact the interaction strength of multiple CTCF loops in a genomic locus. Since loop competition is a dominant feature in our model, attenuation of one loop would in turn facilitate or strengthen flanking and overlapping loops. Specifically, our model predicts that if a given CTCF binding site is disrupted by sequence variation or mutation, it will be less likely to form a loop[17], and consequently other CTCF pairs spanning the disrupted site would be more likely to interact. To test this hypothesis, we used genetic information and previously published Hi-C data in lymphoblast cells from 20 individuals[27]. In this population sample, 49 CTCF binding motifs were disrupted by genetic variation (SNPs) at key binding positions in one or more of these individuals. For each site, we separated individuals into two groups (strong or weak CTCF motif), and calculated the ratio of average contact frequency of 40kb bins in neighboring 800kb windows in the two groups (Fig. 5a). After aggregating this data for all 49 CTCF sites, we observed clear evidence that for individuals with the weak motif, all pairs of bins across the

CTCF motif exhibit a higher normalized interaction frequency than those for individuals with strong motif (100/100 bins higher for weak motif individual) (Fig. 5b). In addition, neighboring pairs on the same side of the CTCF binding site have much weaker differences, and their direction of change is much more random (52/90 bins higher for weak motif individual). The increased local interaction frequency in the weaker CTCF genotypes is consistent with reduced loop-competition from the disrupted CTCF site, so this data supports the role of loop competition in loop formation and provides an interesting mechanism of how genetic variation could affect chromatin conformation. It is also consistent with a recent report that subtle quantitative changes in CTCF loop strength could lead to phenotypic variation in gene expression[28].

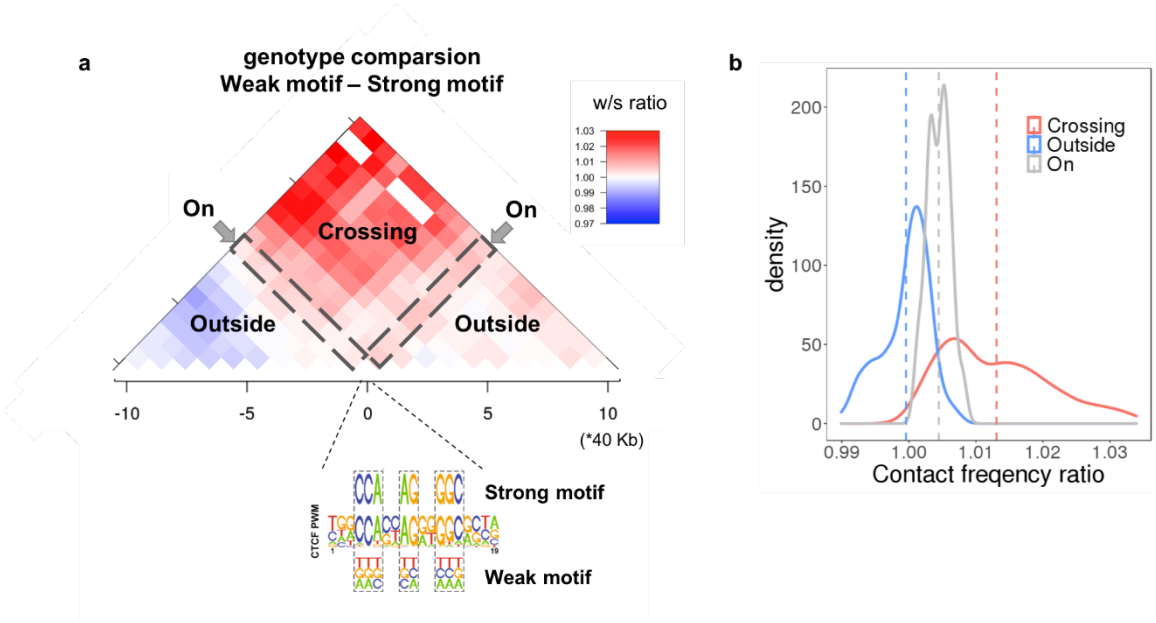

Figure 5. **Loop competition predicts the effect of population variation in CTCF binding site strength on flanking chromatin interactions.** (a) Differential Hi-C contact frequency ratio for weak vs. strong motif genotype individuals, flanking variable CTCF sites in a population of 20 individuals[27]. The heatmap is partitioned into 40kb bin pairs, and loops which are crossing, on top of, or outside the CTCF motif mutated by the SNP, the CTCF PWM shown is from[27]. (b) Contact frequency ratio distribution for the three classes of bin pairs.

Loop extrusion model predicts effect of CTCF binding perturbation and WAPL knockout

Many in vivo perturbation experiments have been carried out to study the role of CTCF in loop formation and gene regulation[29]. In addition to knocking out CTCF, many studies have deleted or inverted the CTCF binding motif, revealing a great preference of convergent CTCF motif orientation for chromatin loops[17,19,30]. These studies provide important additional contexts to test our model. In one particular study, the effect of CRISPR targeted deletion or inversion of a CTCF binding motif in mouse embryonic stem cells (mESC) was measured with 4C[17]. To make predictions in the three loci tested, we used CTCF ChIP data measured before and after the perturbation, modified w for inversions, and we calculated the corresponding loop interaction probability from our model. Before CRISPR editing, the predicted interaction probabilities matched the 4C loop measurements very well (Fig. 6a-6c, only the strongest 4C loop corresponding to the target site is shown). Moreover, after CRISPR editing, our model successfully predicts the loss of the wild-type loop induced by both deletion and inversion of CTCF binding motif for Malt1, Sox2 and Fbn2 loci (Fig. 6d-6f). Although inversion of the CTCF binding site does not change CTCF binding dramatically, inversion affects loop formation

through the parameter w, and the reduced interaction probability is consistent with the observed reduction in 4C signal.

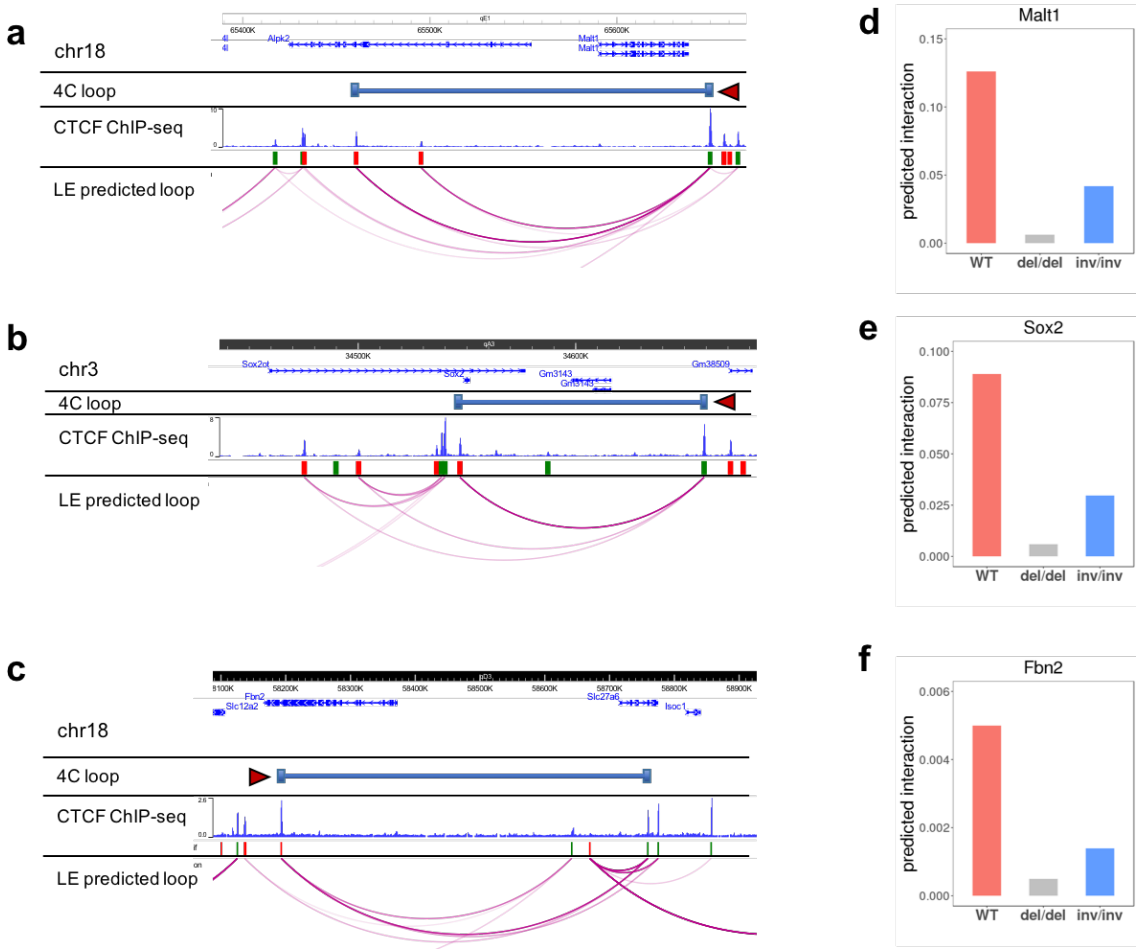

Figure 6. **Loop extrusion model predicts the effect of targeted CTCF disruption and inversion on chromatin interactions.** (a)-(c). Comparison of contact profiles of 4C-seq measurements and our loop extrusion model at the Malt1, Sox2 and Fbn2 loci. Only the strongest loop of the targeted CTCF binding site (indicated by dark red triangle) from 4C-seq is shown. The orientations of flanking CTCF motifs are indicated by red (forward) and green (reverse) bars. Our loop extrusion model predicted interacting CTCF pairs are shown, with darker color corresponding to higher interaction probability. (d)-(f). Loop extrusion model predicted probability of looping for wild-type and after CRISPR deletion or inversion of the targeted CTCF binding site.

Alternatively, the activity of Cohesin can be modulated through the Cohesin unloading factor WAPL[31,32]. It has been reported that upon WAPL knockout the overall chromatin structure transforms into a more condensed state, with an increase in loop number and size. Although it is known that WAPL knockout increases Cohesin residence time on chromatin[33], the means by which this changes loop interactions under the same set of CTCF boundary locations remains unclear. Since our original model was derived under the normal assumption of constant WAPL activity, we modified our model slightly to predict the effect of WAPL knockout on CTCF-mediated loops. In this WAPL-KO modified model (Supplementary Method), following previous work,[11] we assume that CTCF loop anchors are permeable. WAPL knockout increases the residence time of Cohesin, which consequently has a greater chance of sliding through boundary CTCFs. With enhanced permeability, the effect of loop competition is reduced

because Cohesin is moving more freely in this case. Through testing the WAPL-KO corrected model, we found permeability is positively correlated with total loop number and average loop size. At permeability around 30%, we faithfully reproduced experimental results from WAPL knockout in HAP1 and Hela cell lines[31,32] (Supplementary Fig. 6).

CTCF loops constrain enhancer-promoter interactions

a

		Enhancer-Promoter link			
		FALSE	TRUE	T/F fraction	enrichment
ChIA-PET	cross	1941	43	0.0221	
	not-cross	2073	133	0.0641	2.9
LE model	cross	2966	54	0.0182	
	not-cross	1047	125	0.1193	6.6
ChIA-PET	contain	1180	70	0.0593	
	not-contain	2834	106	0.0374	1.6
LE model	contain	410	84	0.2048	
	not-contain	3603	95	0.0263	7.8

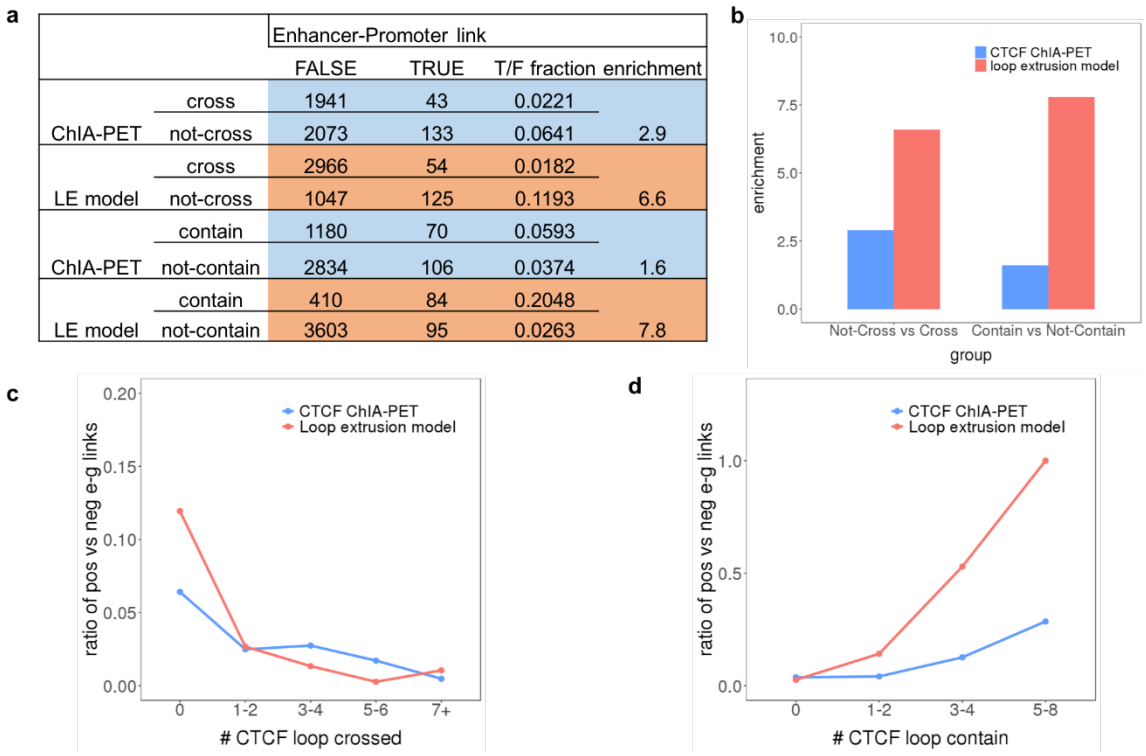

Figure 7. **CTCF loops are predicted to constrain enhancer-promoter interactions, but loop extrusion model predicted loops do so more accurately**. (a) Counts of true (interacting) and false (non-interacting) enhancer-promoter (E-P) pairs according to whether they cross, or are contained within CTCF loops. Ratios of true and false E-P links are also shown (T/F). (b) Enrichment of T/F ratio between each group are calculated and compared between CTCF ChIA-PET annotated loops and loops predicted by our loop extrusion model. Strikingly, the predicted CTCF loops are much more enriched for loops which contain (and do not cross) E-P interacting pairs. (c) T/F ratio against the number of CTCF loops each E-P link crosses is plotted. (d) T/F ratio against the number of CTCF loops containing each E-P link is plotted.

An important proposed function of CTCF loops is to shape local chromatin architecture to constrain interactions between other types of regulatory elements, especially enhancers and promoters[34]. According to this idea, enhancer-promoter interactions should preferentially occur within CTCF loops, and not to cross CTCF loops. To assess this hypothesis with our model, we took an integrated enhancer perturbation dataset consisting of 4194 enhancers and 65 gene promoters in the K562 cell line from 11 studies[35–45]. We counted the number of CTCF loops crossed by each enhancer-promoter (E-P) link and the number of CTCF loops which contain each E-P link. We then compared the fraction of interacting vs. non-interacting E-P pairs in loop-crossing and loop-containing events. Consistent with our hypothesis, based on K562 CTCF ChIA-PET measured loops, we observed a 2.9 fold enrichment of true E-P links in the group that does not cross any CTCF loop, compared to the group crosses one or more CTCF loop. Similarly, there is a 1.6 fold enrichment of true E-P links in the group that is contained by one or

more CTCF loop, compared to the group that is not contained within any CTCF loop. Strikingly, the level of enrichment of 'not cross' and 'contain' groups increased dramatically to 6.6 and 7.8, using our loop extrusion model CTCF loops instead of ChIA-PET annotated loops. Although this clearly lends support to our model, it may seem perplexing that a model trained on ChIA-PET data seems to be more consistent with expectations of E-P loop crossing than the ChIA-PET data itself. One possible explanation is that our model prediction is largely coming from CTCF ChIP-seq intensity, orientation, and loop competition, all single-point measurements, while ChIA-PET interactions are pairwise and require much more sequencing depth to achieve comparable signal-to-noise ratios. Technical considerations may contribute to false positive or negative loop interactions in the ChIA-PET data which do not constrain E-P interactions as effectively as those predicted by our model. While genomic ChIA-PET data with thousands of loops can reliably determine the parameters in our model, the model may actually be more accurate at predicting functional CTCF loops in a given locus.

CTCF binding intensity is predictive of cell-type specific loops

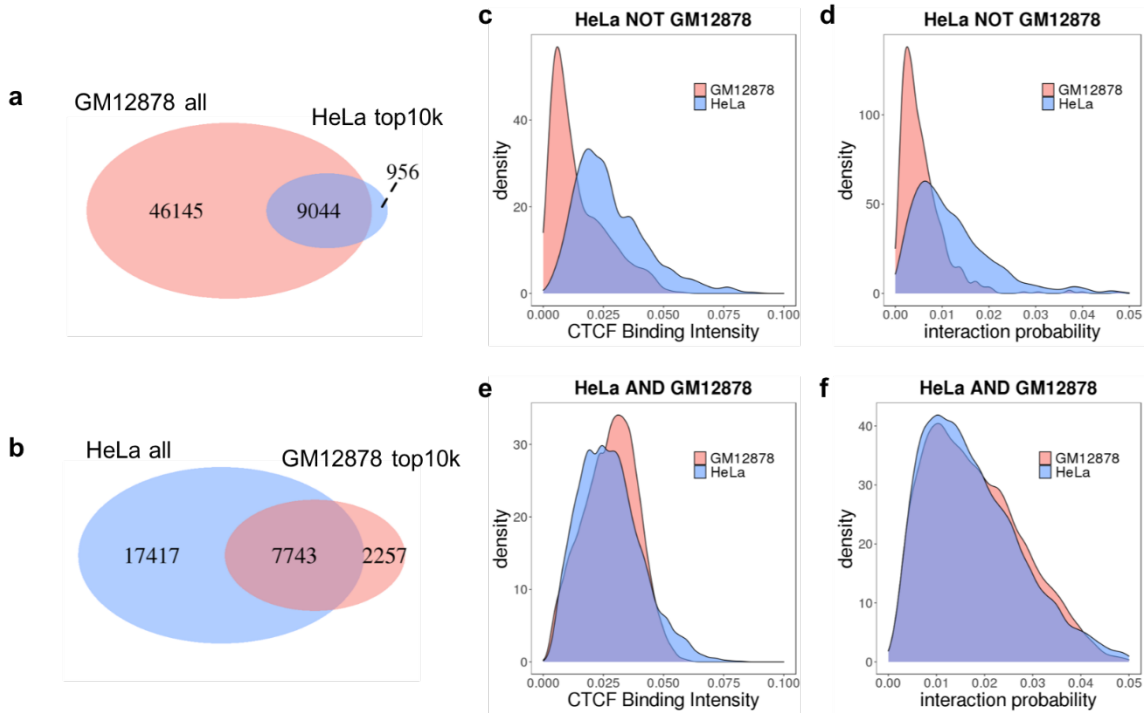

Figure 8. **CTCF binding intensity is predictive of cell type specific loops.** (a)-(b). Venn diagram of CTCF-mediated loops identified from GM12878 and HeLa ChIA-PET. Only the strongest 10,000 loops are compared against each other due to different sequencing depth. (c)-(f) CTCF binding intensity distribution and predicted interaction probability distribution for HeLa specific CTCF loops and shared loops.

Next, we investigated the cell-type specificity of CTCF loops and whether cell-type dependent CTCF loops could be predicted by the loop extrusion model. Cell-type specific chromatin interactions are of great interest because they have been demonstrated to be an important mechanism for gene regulation in lineage differentiation[34,46]. We noticed that GM12878 and HeLa ChIA-PET experiments have very different numbers of detected loops, but this is mostly due to differences in sequencing depth. To eliminate this bias, we constrained our analysis to the strongest 10,000 CTCF loops in each cell line. We find that these top loops are quite

conserved. Over 75% of them are shared between the two cell lines (Fig. 8a-8b). These cell-type specific CTCF loops can also be predicted with our loop extrusion model, because the difference in their activity is strongly associated with CTCF binding intensity in GM12878 vs. HeLa (Fig. 8c-8d).

Discussion

Recent progress in 3C techniques has provided comprehensive annotation of higher order chromatin architecture, including CTCF-mediated loops. Predicting CTCF-mediated loops is crucial to understand the mechanisms controlling regulatory element interactions and transcriptional regulation. It has been shown that the interaction between enhancers and their target gene promoters cannot be predicted solely from local epigenetic signals[47,48]. The missing element is very likely to be the spatial organization of chromatin, as disruption of CTCF-mediated loops have been confirmed to be able to change the expression of gene both inside and outside of the loop. Moreover, recent sequence-based modeling of enhancer-promoter interactions have also identified CTCF binding as the most important player[49]. We were motivated to develop a model of CTCF interactions based on a simple process after a machine learning approach showed that CTCF interactions in ChIA-PET data could be predicted with high accuracy using a large set of epigenomic features[21].

Our model correctly distinguishes interacting CTCF pairs from a vast number of non-interacting CTCF pairs. This could not be achieved using only convergent CTCF motif orientation as a feature, as many convergent CTCF motifs do not interact, and some true interactions are tandem. Our model is easily interpretable, as the contribution of each component is independently modelled by its corresponding probability. We validate our model on a wide range of complementary datasets: ChIA-PET, Micro-C, Hi-C, genetic variation in CTCF binding sites, CRISPRi perturbation of loop anchor binding sites, and by the predicted CTCF loops' ability to constrain enhancer promoter interactions.

Our analysis reveals that the distance between two CTCF pairs, previously thought to be important for constraining chromatin interactions, actually becomes unimportant when we explicitly calculate the contribution from loop competition. This raises the question of whether this is specific to CTCF-mediated loops or a broader class of 3D chromatin interactions. A recent study from E.coli proposed an interesting 'small world' hypothesis that because the bacteria genome is so small and compact, different parts of the genome, regardless of their linear position, are all equally likely to randomly collide with each other[50]. This is unlikely for the human genome given its huge size and partitioning into chromosomes, but may be true within single TADs.

The concept of loop competition arises naturally from the loop extrusion process (Fig. 1b). The loop competition hypothesis is that CTCF pairs across an existing loop are less likely to be formed, while those within or outside it are unaffected. This idea is supported by observations that strong CTCF corner peaks prohibit cross TAD interactions[5]. Disruption of CTCF binding sites and rearrangement of corresponding CTCF loops facilitates ectopic interactions between enhancers and gene promoters over long distances and could potentially give rise to severe pathogenic phenotypes like polydactyly[8]. We used our quantitative predictions of loop competition to predict the consequences of CTCF motif sequence variation on neighboring chromatin interactions, and showed that the impact is significant, consistent with our modeling, and detectable over several hundred kilobases. Importantly, this result shows that chromatin

architecture should not be viewed simply as a combination of independent structural units, since there can be extensive interplay between adjacent elements.

We found that CTCF-mediated loops are rather stable across cell lines, consistent with previous studies[46]. However, although less common, when cell-specific CTCF loops do occur, they can be consequential, as cell-type specific loops are often accompanied by gene activation or repression[46]. Our modeling shows that these cell-specific CTCF loops are mediated by variable cell-specific activity of CTCF binding sites.

In summary, we constructed a mathematical framework to predict single loop level chromatin architecture based on a loop extrusion model. We validated our model by showing that the model predictions are in agreement with four diverse experimental datasets, which in turn provides substantial support for the loop extrusion hypothesis. Although we have extensively tested our model on existing data, prediction of CTCF looping interactions in blind computational assessment challenges such as CAGI[51] would be an interesting next step, as these efforts are beginning to focus more on regulatory processes[52]. We expect our loop extrusion model to be useful for further exploration of both the features and mechanisms of chromatin packaging and its impact on gene regulation, and as a component of more comprehensive models of enhancer-promoter interactions.

Code and Data Availability

Source code and training data are available for download from https://github.com/wangxi001/Loop-Extrusion-Model.

Acknowlegements

We thank the following members of the Beer Lab for discussion and useful comments on the manuscript: Dustin Shigaki, Jin-Woo Oh, and Milad Razavi Mohseni. We thank J. Nasser and J. Engreitz for kindly providing a summary table of the enhancer-promoter interactions from references[35–45]. This work was supported by NIH grants HG009380 and HG007348 to MB.

Author Information

Johns Hopkins University School of Medicine,
733 N. Broadway, Baltimore, MD 21205
Wang Xi & Michael A Beer
Correspondence to: mbeer@jhu.edu.

Author Contributions

XW and MB designed the study, XW analyzed the data, and XW and MB wrote the paper.

Methods

Loop extrusion model

The loop extrusion model is a hypothesis that describes the formation of CTCF-mediated loops via Cohesin movement. The probability of CTCF loop formation is determined by four components.

1) **CTCF binding intensity**. The occupancy of CTCF is characterized by the standard calculation of chemical equilibrium[23].

$$p_i = \frac{[CTCF]}{[CTCF]+K_{d,i}}$$

[CTCF] corresponds to the concentration of CTCF, and is represented by the normalized read count in the window of the binding site. $K_{d,i}$ is the equilibrium dissociation constant for each CTCF binding site. This dissociation is not necessarily simply due to the strength of the CTCF binding motif, as local chromatin context and interactions with flanking factors may contribute to CTCF binding. Therefore we will estimate this local $K_{d,i}$ from the CTCF ChIP-seq signal. We can combine the unknown $K_{d,i}$ and [CTCF] to write $p_i = \frac{1}{1+\frac{K_{d,i}}{[CTCF]}} = \frac{x}{x+a}$, and we will further assume that the local ChIP-seq signal x is inversely proportional to $K_{d,i}$/[CTCF], with a scaling factor of a. We will learn the best value of the parameter a from the ChIA-PET data. The precise form of the ChIP-seq signal scaling with $1/K_{d,i}$ is not critical, as we have also tried a different parameterization of the binding probability using $p_i = \tanh(ax)$, which yields almost equivalent performance. With the assumption that CTCF binding at each site are independent, joint probability of CTCF binding at two sites at the base of a loop is given by their product $p_i \cdot p_j$.

2) **CTCF motif orientation**. CTCF-mediated loops have strong motif orientation preference, with convergent motifs being the most favored configuration and divergent motifs being the least favored. To model this difference, we modeled the relative stability of convergent, tandem, and divergent loops as 1, $1/w$, and $1/w^2$, where w is a scalar, $w > 1$. This can be interpreted as an orientation dependent stability of the CTCF-Cohesin complex at the base of a loop, where each "non-inward" CTCF motif decreases the stability the complex by a factor of w.

3) **Distance.** A strong anti-correlation has been found between chromatin contact frequency and the distance between the interacting regions in genome-wide 3C experiments. Various probabilistic distributions have been used to fit this relationship, and we selected an exponential distribution due to its conciseness and power to approximate the contact frequency distribution.

$$D_{ij} = e^{-d_{ij}/\lambda}$$

The parameter λ in this distribution is the average CTCF loop length. This exponential distribution is consistent with a constant dissociation probability of Cohesin as it translocates down the chromatin fiber.

4) **Loop competition**. The process of loop extrusion implies a competition between two Cohesins translocating along the same linear chromatin segment. Since the final state of the extrusion is Cohesin contacting a CTCF barrier pair, this further implies a competition between CTCF pairs which overlap each other. 'Overlapping' here is defined with regard to the window between CTCF binding sites. As Cohesin cannot move across another Cohesin on a pre-formed loop, a prerequisite of loop formation would be that no overlapping loops exist, therefore

$$LC_{ij} = \prod_{mn \bigcap ij \neq \emptyset}(1 - p_{mn})$$

describes this probability. When we compute this term, we actually use the requirement that there is no CTCF binding event in the current window, as this is a sufficient condition that guarantees no overlapping loop exists.

The final probability of loop formation is the joint probability or product of these five terms (two CTCF binding probabilities, one from each of the two sites), initially assuming they are independent.

Parameter determination

To find optimal parameter values, we fit the loop extrusion model to CTCF ChIA-PET data by fixing two of the three parameters and varying the remaining one. The best-fitting parameter is defined to be the one reaches maximum AUPRC. This method is effective since the nonlinearly in this model makes it hard to perform a maximum likelihood estimation by canonical methods like logistic regression. Taking GM12878 as an example, by fixing dissociation constant $1/[\mathrm{CTCF}]$ (a) and Cohesin processivity λ, we found w value of the best agreement with data is 3. By fixing w and λ, we found the optimal $1/[\mathrm{CTCF}]$ is 8.5. Optimal w and $1/[\mathrm{CTCF}]$ for HeLa is quite similar, 2.8 and 8. For λ, the performance of our model monotonically increases when λ is larger, and asymptotically approaches to the performance of model without this distance-associated exponent term (D_{ij}=1). We also performed a grid search over these three parameters and found high performance in a broad range around this single optimal set of values.

Polymer simulation of loop competition

A chain of 50 monomers were simulated under Brownian-like conditions using Langevin dynamics by LAMMPS[53]. Two different pairs of monomers have stronger binding energy with each other, ranging from 1 to 40, while all other monomers are identical with binding energy 0.1. All other settings and parameter are the same as described in[12].

CTCF ChIA-PET data processing

GM12878 and HeLa CTCF ChIA-PET data were taken from a published dataset[16]. ChIA-PET2 pipeline with long read mode was used to process data and identify loops[24,54]. One mismatch was allowed in identifying reads with linkers in linker filtering step. Default parameters were used for other steps. Loops are required to be supported by at least 4 PETs for GM12878 and 3 PETs for HeLa. We further constrained CTCF interactions to be within 1 million bp (Mb), as over 96% of loops fell into this range.

CTCF ChIP-seq data processing

CTCF ChIP-seq of GM12878, HeLa and K562 was obtained from the ENCODE portal. Reads were aligned with BWA to the hg38 reference genome[55]. Peaks were called by MACS2 with default parameters[56].

CTCF motif analysis of ChIP-seq data

The position weight matrix of human CTCF was download from JASPER[57]. STORM with default parameters was used to identify the strongest CTCF motif and the corresponding strand for each CTCF binding site, to select the value of the orientation parameter w.

Boosting model

An ensemble-learning-based boosting model was construct with the python Xgboost package. The model consisted of 50 trees, each with maximum depth of 5 layers. The components of the loop extrusion model are used as input features independently. We performed 10-fold cross validation on segregated chromosomes, and averaged performance to account for randomness between chromosomes.

Lollipop model

Lollipop is a previously published random forest model which can accurately predict CTCF interaction specificity using 77 features[21]. It has been evaluated on the same CTCF ChIA-PET dataset processed in a very similar method. Therefore, we directly compare the AUROC and AUPRC with Lollipop.

Micro-C data processing

A total of 15,945 loops were called from 2.6B reads of mESC Micro dataset[26]. Chromatin loops were identified by using HiCCUPS[6]. Loops were called at 1Kb resolutions at peak size = 4Kb, window size = 10Kb, distance to merge = 2.5Kb and FDR<0.1.

Predicting CRISPR perturbation effect

mESC CTCF ChIP-seq data were taken from GSE72720. The loop extrusion model was built and interacting CTCF pairs are predicted quantitatively, with K_d = 8.5, w = 3, λ = 3,000,000. The effect of CRISPR deletion and inversion of CTCF motif on CTCF binding intensity are taken from[9]. The change of binding intensity and orientation are then integrated into model to determine the resulting interaction probability.

Population Hi-C data processing

Normalized Hi-C contact matrices of lymphoblastoid cell lines (LCLs) were taken from[27]. Briefly, Hi-C was performed on LCLs of 20 individuals with previously cataloged genetic variation. Reads were aligned to hg19 reference genome with BWA-MEM as described in[46,55]. Raw counts of contact matrices were normalized to correct for known biases following[27].

Cell-type specific CTCF loop identification

Loops from two cell lines are defined to be common if both anchors overlap, if not, we classify them as cell-type specific. We compared the top 10,000 loops in HeLa with all loops in GM12878, and found 956 HeLa-specific loops. Similarly, we compared the top 10,000 loops in GM12878 with all loops in HeLa, and found 2,257 GM12878-specific loops.

1. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289–293 (2009).
2. Fullwood, M. J. *et al.* An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
3. Hsieh, T.-H. S. *et al.* Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**, 108–119 (2015).
4. Fullwood, M. J., Wei, C.-L., Liu, E. T. & Ruan, Y. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res.* **19**, 521–532 (2009).
5. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
6. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).
7. Narendra, V. *et al.* CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* **347**, 1017–1021 (2015).
8. Lupiáñez, D. G. *et al.* Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* **161**, 1012–1025 (2015).
9. Phillips, J. E. & Corces, V. G. CTCF: Master Weaver of the Genome. *Cell* **137**, 1194–1211 (2009).
10. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. *Nature Reviews Genetics* **19**, 789–800 (2018).
11. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports* **15**, 2038–2049 (2016).
12. Sanborn, A. *et al.* Chromatin Extrusion Explains Key Features of Loop and Domain Formation in Wild-type and Engineered Genomes. *The FASEB Journal* **30**, 588.1-588.1 (2016).
13. Stigler, J., Çamdere, G. Ö., Koshland, D. E. & Greene, E. C. Single-Molecule Imaging Reveals a Collapsed Conformational State for DNA-Bound Cohesin. *Cell Reports* **15**, 988–998 (2016).
14. Davidson, I. F. *et al.* DNA loop extrusion by human cohesin. *Science* **366**, 1338–1345 (2019).
15. Kim, Y., Shi, Z., Zhang, H., Finkelstein, I. J. & Yu, H. Human cohesin compacts DNA by loop extrusion. *Science* **366**, 1345–1349 (2019).
16. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163**, 1611–1627 (2015).
17. de Wit, E. *et al.* CTCF Binding Polarity Determines Chromatin Looping. *Molecular Cell* **60**, 676–684 (2015).
18. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320.e24 (2017).
19. Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**, 900–910 (2015).
20. Pugacheva, E. M. *et al.* CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. *PNAS* **117**, 2020–2031 (2020).
21. Kai, Y. *et al.* Predicting CTCF-mediated chromatin interactions by integrating genomic and epigenomic features. *Nature Communications* **9**, 4221 (2018).
22. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. *PLOS Computational Biology* **10**, e1003711 (2014).
23. Granek, J. A. & Clarke, N. D. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biology* **6**, R87 (2005).
24. Li, G., Chen, Y., Snyder, M. P. & Zhang, M. Q. ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. *Nucleic Acids Res* **45**, e4–e4 (2017).
25. Schones, D. E., Smith, A. D. & Zhang, M. Q. Statistical significance of cis-regulatory modules. *BMC Bioinformatics* **8**, 19 (2007).
26. Hsieh, T.-H. S. *et al.* Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. *Molecular Cell* **78**, 539-553.e8 (2020).
27. Gorkin, D. U. *et al.* Common DNA sequence variation influences 3-dimensional conformation of the human genome. *Genome Biology* **20**, 255 (2019).
28. Greenwald, W. W. *et al.* Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nature Communications* **10**, 1054 (2019).
29. Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944.e22 (2017).
30. Guo, Y. *et al.* CRISPR-mediated deletion of prostate cancer risk-associated CTCF loop anchors identifies repressive chromatin loops. *Genome Biol* **19**, 160 (2018).

31. Haarhuis, J. H. I. *et al.* The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell* **169**, 693-707.e14 (2017).
32. Wutz, G. *et al.* Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *The EMBO Journal* **36**, 3573–3599 (2017).
33. Hansen, A. S., Pustova, I., Cattoglio, C., Tjian, R. & Darzacq, X. CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *eLife* **6**, e25776 (2017).
34. Hnisz, D., Day, D. S. & Young, R. A. Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell* **167**, 1188–1200 (2016).
35. Fulco, C. P. *et al.* Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016).
36. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics* **51**, 1664–1669 (2019).
37. Thakore, P. I. *et al.* Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. *Nature Methods* **12**, 1143–1149 (2015).
38. Klann, T. S. *et al.* CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. *Nature Biotechnology* **35**, 561–568 (2017).
39. Liu, S. J. *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science* **355**, (2017).
40. Qi, Z. *et al.* Tissue-specific Gene Expression Prediction Associates Vitiligo with SUOX through an Active Enhancer. *bioRxiv* 337196 (2018) doi:10.1101/337196.
41. Ulirsch, J. C. *et al.* Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. *Cell* **165**, 1530–1545 (2016).
42. Wakabayashi, A. *et al.* Insight into GATA1 transcriptional activity through interrogation of cis elements disrupted in human erythroid disorders. *PNAS* **113**, 4434–4439 (2016).
43. Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Molecular Cell* **66**, 285-299.e5 (2017).
44. Xu, B. *et al.* Selective inhibition of EZH2 and EZH1 enzymatic activity by a small molecule suppresses MLL-rearranged leukemia. *Blood* **125**, 346–357 (2015).
45. Huang, J. *et al.* Dissecting super-enhancer hierarchy based on chromatin interactions. *Nature Communications* **9**, 943 (2018).
46. CTCF Promotes Long-range Enhancer-promoter Interactions and Lineage-specific Gene Expression in Mammalian Cells | bioRxiv. https://www.biorxiv.org/content/10.1101/2020.03.21.001693v1.abstract.
47. Xi, W. & Beer, M. A. Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy. *PLoS Comput Biol* **14**, (2018).
48. Cao, F. & Fullwood, M. J. Inflated performance measures in enhancer–promoter interaction-prediction methods. *Nature Genetics* **51**, 1196–1198 (2019).
49. Cao, F., Zhang, Y., Loh, Y. P., Cai, Y. & Fullwood, M. J. Predicting chromatin interactions between open chromatin regions from DNA sequences. *bioRxiv* 720748 (2019) doi:10.1101/720748.
50. Walker, D. M., Freddolino, P. L. & Harshey, R. M. A Well-Mixed E. coli Genome: Widespread Contacts Revealed by Tracking Mu Transposition. *Cell* **180**, 703-716.e18 (2020).
51. Andreoletti, G., Pal, L. R., Moult, J. & Brenner, S. E. Reports from the fifth edition of CAGI: The Critical Assessment of Genome Interpretation. *Human Mutation* **40**, 1197–1201 (2019).
52. Shigaki, D. *et al.* Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. *Human Mutation* **40**, 1280–1291 (2019).
53. Plimpton, S. *Fast parallel algorithms for short-range molecular dynamics*. https://www.osti.gov/biblio/10176421 (1993) doi:10.2172/10176421.
54. Li, G. *et al.* ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol* **11**, R22 (2010).
55. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio]* (2013).
56. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).
57. Mathelier, A. *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **44**, D110–D115 (2016).