PROBABILISTIC MODELING OF CHROMATIN INTERACTIONS

by
Wang Xi

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
August, 2022

Abstract

Higher-order chromatin architecture plays an important role in mammalian transcriptional regulation. However, understanding the mechanisms and impact of complex chromatin contacts remain challenging. In the past decade, breakthroughs in experimental techniques like Chromatin Conformation Capture (3C) assays enable genome-wide detection of chromatin interactions at high resolution. This provides great opportunities of computational modeling to predict functional interactions and identify target genes of disease variants.

In this thesis, I started with critical assessment of an existing method for predicting enhancer-promoter interactions. I reported severe overfitting issues of it resulting from improper machine learning experimental design. I also found the limitation of resolution in their training datasets hinder accurate assignment of single regulatory element at interaction boundaries.

In the second part, I developed a novel mathematical model to predict CTCF-mediated chromatin loops, which is the most prominent class of chromatin interactions, based on the biological hypothesis of loop extrusion. I showed that this model is capable of predicting CTCF loops measured by CTCF ChIA-PET data with high accuracy, using CTCF ChIP-seq alone as input. Furthermore, this model consistently predicts chromatin interaction frequency due to changes of CTCF binding site by genetic perturbation and looping-related protein factor degradation events.

In the last part, I applied this computational framework to a greater set of ChIA-

PET data. The analysis result inspired the development of a simple and interpretable for predicting enhancer-promoter interactions. I showed that this model outperforms existing methods on predicting CRISPRi hits that regulated gene expression.

Overall, these approaches are applicable to diverse datasets to advance our understanding of chromatin interaction mechanisms as well as their implication in gene regulation and diseases.

Thesis Readers

Dr. Michael A. Beer (Primary Advisor)
Professor
Department of Biomedical Engineering and Genetic Medicine
Johns Hopkins University

Dr. Taekjip Ha
Professor
Department of Biophysics and Biophysical Chemistry
Johns Hopkins University

Dr. Reza Kalhor
Assistant Professor
Department of Biomedical Engineering
Johns Hopkins University

Dedication

This thesis is dedicated to my parents, my wife and my dog for their eternal love, trust and support.

Acknowledgements

First of all, I would like to express much gratitude to my advisor Dr. Michael Beer for his continuous support and patience during my PhD study. He inspires me to always pursue the most important and exciting scientific problems. I would also like to acknowledge my thesis committee members, Dr. Taekjip Ha and Dr. Reza Kalhor for their comments, suggestions and feedback.

I thank all of my lab members, Dustin Shigaki, Jin-Woo Oh and Milad Razavi-Mohseni for their help and discussion. Besides that, I thank the ENCODE consortium, which is an amazing scientific community that brings teams with different expertise together to collaboratively push the boundary of human knowledge on regulatory genomics.

Finally, I want to thank my wife, Yuting, for her tremendous love and support.

Contents

List of Figures

Chapter 1

Introduction

1.1 Overview

The three-dimensional organization of chromatin profoundly affects transcriptional regulation. In the last few decades, researchers have identified huge numbers of functional DNA elements that can regulate cell-type specific gene expression. These *cis*-regulatory elements (CREs) are often located in the non-coding genomic regions, which make up of over 98% of the human genome [1]. Together with the *trans*-acting factors that bind to them, they control when and how the protein-encoding information of the genome is expressed to direct cell fate decisions during development and differentiation. Promoter, enhancers and insulators are three major classes of CREs that control context-dependent gene expression. Promoters drive gene expression adjacent to the transcription start site (TSS), and enhancers act over long distance to interact with promoters and activate transcription regardless of orientation [2]. On the other hand, CTCF proteins act as insulators to separate enhancers and promoters, and constrain their activity. Many studies have observed that enhancer-promoter interactions are established in concordance with gene expression. However, the underlying mechanism of how these CREs regulate target gene expression is largely unknown.

Technological advances in high-throughput methods for chromatin profiling provide

rich experimental data sets of genome organization and dynamics [3, 4]. These data provide unique opportunities to development new models and algorithms to address the questions above. In this thesis, I started with critical assessment of existing methods for predicting enhancer-promoter interactions (EPI) and revealed that many leading approaches often dramatically overestimated their performance [5]. In particular, a well-cited machine learning model, TargetFinder, had been reported to predict EPI with high accuracy using epigenetic features. Careful examination of their data revealed that a lot of EP pairs share information at the promoter and the intervening genomic window. When proper CV design separates training and testing set by chromosome, the accuracy dropped dramatically to only slightly better than random.

I then focused on predicting CTCF loops, which are the most prominent chromatin organization unit that can be detected by CTCF ChIA-PET at high resolution. Different from the two predominant computational approaches - machine learning and polymer simulation, I built a probabilistic model based on a biological hypothesis of loop formation, namely, loop extrusion [6]. I transformed steps of the loop extrusion process into components of my model with the language of mathematics. Various evaluations showed that this model can accurately predict CTCF loops, while remaining a lot simpler and more interpretable than previous methods.

Domains within CTCF loops are enriched for interacting EP pairs, but whether CTCF loops are causal for constraining EP interaction remains unknown. By analyzing CTCF ChIA-PET data from dozens of cell lines and applying the loop extrusion model, I obtained a high confidence set of CTCF loops. I further developed a model using enhancer activity and CTCF looping information as features to construct genome-wide maps of EPI in a given cell type. Taken together, by predicting chromatin interactions and target genes of active enhancers, my work will help to interpret the function of thousands of disease variants in the non-coding genome.

1.2 Thesis Organization

I provide background information about chromatin organization in Chapter 2. I start with general description of molecular basis of chromatin architecture, with emphasis on compartments, topologically associated domains (TADs) and loops. Then, I provide a short review of human diseases connected to chromatin organization disruption, as well as experimental techniques to map chromatin organization, including high-throughput sequencing, imaging and more. In the second part, I focus on computational approaches of chromatin organization modeling.

In Chapter 3, I look into overfitting issues of previous machine learning methods (i.e. TargetFinder) for enhancer-promoter interaction prediction. TargetFinder's high performance is only achieved with gradient boosting using a very large number of trees. As the authors fail to separate those EP pairs in cross-fold validation (CV) test sets, these shared features lead to incorrect evaluation of generalization performance. It study shows that EPI prediction cannot be predicted from local epigenetic information alone.

In Chapter 4, I introduce a new mathematical model of loop extrusion, and demonstrate how it can be used to predict CTCF looping specificity. This model reveals new insights into loop formation mechanisms, such as the importance of competition among loops. Thorough analysis shows this model also consistently predicts chromatin contact change caused by CTCF binding site perturbation in many cases.

In Chapter 5, I expand the loop extrusion mathematical model onto CTCF ChIA-PET datasets from a large number of cell lines. I also propose strategies to evaluate quality of ChIA-PET data. I describe a new approach that combines CTCF looping constraint and local chromatin activity to predict enhancer-promoter interaction. This model is shown to outperform other state-of-the-art methods on predicting EP pairs

identified by CRISPR perturbation of enhancers.

Chapter 2

Background

2.1 Chromatin Organization

The large amount of DNA that constitutes a genome is packaged into an organized way to facilitate information communication and retrieval of cellular activities. Multiple levels of DNA folding generate extensive genomic contacts that affects gene expression and other cellular functions. In this section, I first provide an overview of mammalian 3D chromatin architecture studies (2.1.1) followed by discussion of their role in development and diseases (2.1.2). I then discuss current experimental methods to map 3D chromatin architecture in detail (2.1.3).

2.1.1 Mammalian 3D Chromatin Architecture

The 2-meter length of DNA in a mammalian cell is organized into chromosomes, which are packaged and folded through various mechanisms and occupy distinct positions in the nucleus. Recent observations indicate that there could be two orthogonal mechanistic principles underlying the formation and maintenance of 3D chromatin organizations: compartmental domains and Cohesin-mediated CTCF loops.

In 2009, contact maps of Hi-C data showed a genome-wide map of mammalian chromatin interactions at 1Mb resolution for the first time [3]. This study identified a segregation of the genome into two types of compartments, named A and B, defined

by the eigenvector or first component of a principal component analysis (PCA). Sequences in the A compartment are enriched with transcribed genes and active histone modifications, while B compartment contains inactive genes and repressed histone modifications more frequently. A and B compartments preferentially interact with sequences in other A or B compartment regions. They are also found to frequently correspond to euchromatin and heterochromatin [7].

With decreased sequencing cost and advanced experimental techniques, the size of Hi-C data sets increased from 10 million reads to 200-300 million reads in 2012, allowing the partitioning of data into 40kb bins [8]. Using this smaller bin size, computational algorithm identified topologically associating domains (TADs) in the 0.2-1.0Mb scale through measuring the directionality index of interactions. Unlike compartments, TADs represent sequences that preferentially interact with themselves in a compact domain. Given the population nature of Hi-C data, TADs can be interpreted as either a physical structural unit that exists in individual nucleus or aggregated signals across a large number of cells. A prominent feature of TADs boundaries is the enrichment of CTCF binding and actively transcribed genes.

The first high-resolution Hi-C data of a mammalian genome published in 2014 contained 5 billion paired reads, making it possible to bin reads at 5-10kb resolution [9]. A set of point-to-point interactions emerges between two sequences bound by CTCF, ranging from 10 kb to a few megabases. Many of these loops can be observed as strong punctate signals at the summit of TADs. Again, population Hi-C data cannot answer if a CTCF loop occurs in all cells or only a subpopulation cells. Remarkably, 92% of CTCF loops identified by Hi-C, or 65% identified by CTCF ChIA-PET, occur between pairs of convergently orientated CTCF motifs. CTCF loops with tandem or divergent motif configuration are with comparatively weaker signal in these experiments. This asymmetric phenomenon implies that the loop formation mechanism must differ in some way from random diffusion in an unrestricted 3D space, which would not result

in a motif orientation preference. Since 2015, emerging evidence suggests that the formation of CTCF loops is likely to take place via an extrusion process mediated by Cohesin rings [10–13].

In the meantime, contacts between chromatin regions are increasingly thought to have functions in gene regulation. In particular, it has become clear that a class of genetic element named enhancers can regulate transcription by acting over long distance to interact with gene promoter [14]. CTCF also plays a role in this process by establishing architecture loops with Cohesin and promoting interactions between distally located sequence [8].

2.1.2 Chromatin Organization and Human Diseases

Deciphering the genetic and molecular basis of human traits and disease is a fundamental problem in biology and personalized medicine. Genome-wide association studies (GWAS) over the past two decades have uncovered that more than 90% of disease-associated variants lie in the non-coding DNA [1]. However, it is unclear whether and how these sequence variants affect gene expression and what their target gene may be. The general observation is that risk variants accumulate in putative enhancers and often lead to a modulated activity of the regulatory network between enhancers and their target genes. The activity of enhancers highly depends on their genomic context, as gene or enhancer relocation can both cause dramatic change in expression. More recent experiments have highlighted the complexity and unpredictability of position effects in the mammalian genome.

Structural features of the genome provide explanations for observed regulatory activity in a given genomic context. The most well-known example is the key developmental signaling gene, Shh, and a cis-acting enhancer element, ZRS, that regulates its expression during limb development [15, 16]. ZRS is located approximately 850kb away from the Shh promoter, but deletion of ZRS ablates Shh expression in the mouse

limb bud and results in severely truncated limbs. Point mutations in human ZRS can also cause limb malformation such as polydactyly [17].

In other cases, larger chromosomal rearrangements can disrupt TAD boundaries and merge the TADs with chromosomal breakpoints. As a consequence, new genes and enhancers enter each other's search space to form new regulatory circuits. For example, in many acute myeloid leukaemia (AML), inversions in chromosome s (inv(3)/t(3;3)) are associated with aberrant expression of the stem cell regulator EVI1 [18]. The inversion re-directs an enhancer from the GATA2 tumor suppressor gene to the EVI1 oncogene. Besides that, duplications of TAD boundary at the SOX9 locus causes neo-TAD formation and is associated with Cooks syndrome, short digits and nail aplasia [19]. Also, deletions, inversions and duplications at the WNT6, IHH, EPHA4 and PAX3 locus in brachydactyly, polydactyly and F-syndrome disrupt the boundaries of a TAD with limb-specific enhancers causing different congenital limb malformation depending on the gene that are being placed under the control of the limb enhancer landscape [20]. Balanced translocation at the MEF2C locus cause a regulator loss of function and are associated with anomalies of the brain and developmental delay [21]. Last but not least, the progression of cancers is accompanied by a dysregulation of transcriptional programs [22–24]. Change of gene regulation usually involves the re-organization of the human genome architecture are multiple levels. For example, perturbed CTCF function and DNA binding can cause the activation of oncogenes in cancer cells, mostly through a process of enhancer hijacking. More examples are well-documented in several review articles [25, 26].

2.1.3 Experimental Methods for 3D Chromatin Architecture Mapping

Advances in our understanding of chromosome folding depends on approaches that can map chromatin contacts genome-wide. Until recently, 3D genome studies have

focused on two main technologies: chromosome conformation capture (3C), namely Hi-C (high-throughput chromosome conformation capture); and imaging, particularly fluorescence in situ hybridization of DNA (DNA-FISH).

3C was invented as a general method to study chromosome organization in eukaryotic cells [27]. Its derived technologies have uncovered hierarchical chromatin structures, such as compartments, topologically associating domains (TADs), sub-TADs, insulated domains and chromatin loops [3, 8, 9]. It combines protein crosslinking and proximity ligation of DNA to detect long-range chromatin interactions between pairs of chromatin loci. 3C-based techniques involve the preparation of chromatin after mild formaldehyde crosslinking, followed by sonication and digestion with restriction enzymes, and ligation of digested DNA fragments attached to chromatin remnants. The ligation products are captured by a variety of approaches and amplified by PCR or sequenced using unbiased next-generation sequencing methods [28].

As the 3C methods focus on interactions between two loci ('one versus one'), circular chromosome conformation capture (4C) [29] and chromosome conformation capture carbon copy (5C) [30] were developed to map all contacts at a single locus ('one versus all') or all contacts within a large genomic region ('many versus many') with higher resolution. In 2009, high-throughput chromosome conformation capture (Hi-C) allowed genome-wide mapping of chromatin interaction for the first time ('all versus all') [3].

To explore contacts that coincide with chromatin occupancy of specific proteins, Hi-C libraries can be enriched by chromatin immunoprecipitation (ChIP) before ligation. Chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) includes sonication of the nuclei to enable efficient precipitation of chromatin [4, 31]. Other approaches such as Hi-ChIP [32] and proximity ligation-assisted chromatin immunoprecipitation sequencing (PLAC-seq) [33] perform in situ Hi-C and proximity ligation before sonication and immunoprecipitation.

On the other hand, direct visualization of nuclear structures and specific genomic sequences can be the key to understand chromatin organization. The most commonly used imaging technique for detecting chromatin contacts in fixed cells is DNA-FISH [34]. FISH uses fluorescently tagged DNA sequences (such as oligonucleotides) as probes to hybridize to complementary target regions of interest. DNA-FISH is typically used to measure the physical distance between two or a few differentially labelled genomic regions of interest. A chromatin contact is often defined by a distance threshold between 50nm to 1m [35]. DNA-FISH can also be used to visualize chromatin compaction or positioning of genomic regions with respect to nuclear structures, such as nuclear lamina [36].

More recently, improvements in imaging techniques have increased the number of loci that can be analyzed in parallel and have beed extended to live cells. Orthogonal ligation-free approaches have also emerged, namely genome architecture mapping (GAM) [37], split-pool recognition of interactions by tag extension (SPRITE) [38] and chromatin-interaction analysis via droplet-based and barcode-linked sequencing (ChIA-Drop) [39], which have started to reveal novel aspects of chromatin organizations [40].

2.2 Computational Methods of Chromatin Organization Modeling

Technological advances in next-generation sequencing and microscopy have lead to an explosion of new methods to probe chromatin organization. These data have spurred the development of new models and algorithms to interpret genome architecture and its function. Models predicting genome architecture can be useful in different ways. First, such models can test hypotheses or provide new insights into molecular mechanisms underlying 3D genome folding and chromatin interactions. Second, such models can be used to predict functional impact caused by variations or perturbations in chromatin

organization. Third, such models can make in silico prediction of 3D genome contacts in new cell types. In this section, I reviewed two distinct but interconnected approaches: biophysical modeling (2.2.1) and statistical learning that infers chromatin contact from epigenetic data (2.2.2).

2.2.1 Physical Modeling of Chromatin Organization

The physical properties of chromatin have been the subject of intense research for decades [41]. Many mechanistic models of chromatin are derived from molecular dynamics simulation. Starting with Netwon's second law, the aim is to simulate the position of all the atoms in a system over time by calculating the force on each atom. However, this process can be computationally intensive and not applicable to large systems. More recently, methods have been applied from polymer physics which represent the chromatin fiber as a connected chain of interacting units [42]. Most work of chromosomal modeling has used a much lower level of detail, whereby a chromatin fiber is represented by a connected chain of beads, with each bead representing several thousand base pairs of DNA. Such coarse-grained models allow large-scale simulations to be performed and have been informative for understanding chromatin organization.

The development and validation of physical models evolves with advances in experimental techniques measuring genome architecture. The presence of chromosome territories as well as measure of mean distance between specific loci by FISH disagree with basic swollen coil or random coil polymer properties. The fractal globule has been the most accepted attempt to improve upon these disagreements. Predictions from this model fit well with experimental scaling of Hi-C contacts with distance [43].

Another class of locus-specific model considers the formation of chromatin domains from packaging of different chromatin types. This can be imagined as a polymer composed of several distinct epigenetic states, which are typically assigned by histone modification or transcription factor binding information. The behavior of such a

copolymer could be modeled from the interaction potential between units or blocks. Once developed, these models could be used to predict chromatin architecture if epigenetic data is available. One example is known as the 'strings and binders switch' model [44]. 'Binders' are single beads representing chromatin-binding protein complexes that can form bridges between different chromatin regions. Based on this model, a polymer-based recursive statistical inference method (PRISMR) was developed that used Hi-C data obtained from wild-type cells to define the number of binder types and their affinities [45]. This approach can further model chromatin conformation change of deletion or duplication events. To further extend copolymer models, additional physical features of chromatin interactions such as loop extrusion and phase separation are also implemented in other state-of-the-art methods [46].

The examples mentioned above show that physical modeling could be a powerful tool for both validation of proposed molecular mechanism underlying chromatin architecture and predicting spatial interactions based on epigenetic data [47]. Meanwhile, several limitations of this approach should be noticed. First, physical models reply on a set of explicit rules of polymer behavior. However, our understanding of the biophysical processes involved in chromatin organization is still far from complete. Therefore, it is not surprising that none of the current models can accurately explain all the key features of genome architecture. For example, it is usually difficult to evaluate if the 'binders' inferred in the PRISMR model have the same properties of real regulatory protein such as PRC1 or Mediator. The missing correspondence might hinder quantitative prediction of chromatin contacts with this methods. Second, inferring biophysical parameters essential for modeling may be challenging. Parameters include concentration and affinity constants of binders, the position of boundaries, processivity of loop extruders and many others. This problem can sometimes be solved by fitting these models directly using Hi-C data. Third, physical modeling is usually computationally intensive.

2.2.2 Statistical Learning for Predicting Chromatin Interactions

It is known that different epigenetic marks and transcription factors correlate with various regulatory elements, chromatin states and other genomic features. For example, the histone mark H3K9me3 correlates well with constitutive heterochromatin and the B compartment, and TAD boundaries are enrich for CTCF and Cohesin protein binding [8, 9, 48]. One can develop simple statistical approach such as regression to predict some 3D genome features from epigenetic data. Oftentimes, non-linear relationship exists, for example between average chromatin contact frequency and genomic distance, which could be well described by a power law. In other cases, complex non-linearitoes among histone modification, transcription factor binding and chromatin interactions can be modeled by accounting for more epigenetic features and using more sophisticated machine learning algorithms.

Machine learning algorithms operate with a numerical representation of input information called features. Features can usually be DNA sequence, genomic distance, epigenetic marks, and experimentally determined target values such as contact frequency between loci and existence of chromatin loops. The main result of machine learning training is a function that maps input features into predicted target values. At each training step, the mapping function is applied on a training subsample to minimize the difference between predictions and experimental data defined by a loss function. This mapping function typically contains a number of adjustable parameters. Performance of the model is evaluated on held-out data. In some case, numerous parameters allow the model to fit both detail and noise in training data, making it non-generalizable over unseen samples. This problem is well known in the machine learning field as 'overfitting' [5, 49]. Therefore, it is essential that the validation data set does not contain samples presented in the training set. One should notice that different genomic objects may not be independent from a mathematical point of view,

as indirect sharing of biological information could occur between overlapping samples. An overestimation of prediction accuracy case due to improper design of training and testing set in enhancer-promoter interaction prediction is discussed in detail in Chapter 3.

The strength of machine learning-based algorithms lies in their flexibility to find complex non-linear patterns when fitting the model, enabling the prediction of structures ranging from two-point interactions to whole Hi-C maps. Several algorithms employ these methods for enhancer-promoter interaction prediction, including TargetFinder [50], JEME [51] and DeepTACT [52]. Other methods like CTCF-MP [53] or Lollipop [54] focus on predicting CTCF-mediated loops. 3D predictor [55], HiC-Reg [56], Akita [57] and DeepC [58] predict all interactions within in an 1-3Mb window to generate the whole Hi-C contact map in silico. Furthermore, these approaches have the potential to reveal biological features underlying 3D genome folding.

The development of a successful machine learning model of 3D genome architecture faces several challenges. The issue of overfitting has already been discussed above. Besides that, the definition of biological features one wants to predict is not trivial. As examples mentioned above, the definition of interacting enhancer-promoter pair is not always clear. Commonly, enhancers and promoters from genomic bins with high contact frequency observed by Hi-C are paired as interacting. The accuracy of this indirect approach depends highly on the quality and resolution of the experimental data. Accordingly, direct functional tests based on target enhancers or CRISPR-interference approaches are sometimes used as a gold standard, but usually with constrained data size. Another example is that loop or TAD prediction algorithms depend on loops or TADs called from other approaches. It is very important to consider the nature and biological properties of target features. Second, there are lots of flexibility in choosing mapping function. Differences and biases in this process could significantly affect prediction accuracy. Another caveat is that, as Hi-C contact

maps aggregates chromatin interactions across millions of cells, it does not mean that the model prediction matches a real cell even if the average matches the population Hi-C data. One solution is to treat chromatin interactions as dynamic processes and model the interaction probability across time and space.

Chapter 3

Critical Assessment of TargetFinder: A Published Enhancer-Promoter Interaction Prediction Model

We report an experimental design issue in recent machine learning formulations of the enhancer-promoter interaction problem arising from the fact that many enhancer-promoter pairs share features [50]. Cross-fold validation schemes which do not correctly separate these feature sharing enhancer-promoter pairs into one test set report high accuracy, which is actually arising from high training set accuracy and a failure to properly evaluate generalization performance. Cross-fold validation schemes which properly segregate pairs with shared features show markedly reduced ability to predict enhancer-promoter interactions from epigenomic state. Parameter scans with multiple models indicate that local epigenomic features of individual pairs of enhancers and promoters cannot distinguish those pairs that interact from those which do with high accuracy, suggesting that additional information is required to predict enhancer-promoter interactions.

3.1 Introduction

While quantitative modeling of cell-specific enhancer activity and variant impact is progressing rapidly, predicting the promoter and gene targets of enhancers remains challenging. We therefore read with great interest the paper by Whalen, et al,[50] which reported a computational model (TargetFinder) using epigenomic features that could predict enhancer-promoter (EP) interactions with high accuracy. In this brief note we report that because many EP pairs share features, the random cross-fold validation scheme used in[50] fails to properly evaluate generalization error and produces inflated test set accuracy. Proper cross-fold validation schemes predict EP interactions with much lower accuracy, due to the fact that many of the models used exhibit significantly lower performance on a reserved test set than on the training set data. If a test set is not fully reserved, these classifiers will fail to generalize to data outside the training data.

3.2 Methods

3.2.1 Input Genomic Features

Genomic features are adapted from [50]. Briefly, functional genomics data for each cell line were downloaded from ENCODE, Roadmap Epigenomics, or GEO. TSS-containing promoter regions and strong and weak enhancer regions were identified using combined ENCODE Segway and ChromHMM annotations for K562, GM12878, HeLa-S3, and HUVEC cells and Roadmap Epigenomics ChromHMM annotations for NHEK and IMR90 cells. Enhancers closer than 10 kb to the nearest promoter were discarded to focus the model on distal interactions. Promoters were retained if actively transcribed. Peaks were intersected with promoter, enhancer, extended enhancer, and window regions. The strength of all peaks in a region or the counts of methylated bases in a region were summed and divided by the length of the region in

base pairs to generate features.

3.2.2 Chromatin Interactions

As reported in[50], interacting enhancer–promoter pairs were annotated using genome-wide Hi-C data. Non-interacting enhancer–promoter pairs were sampled from distance-matched genomic bins, using 20 negatives per positive.

3.2.3 The TargetFinder Model

TargetFinder was implemented in Python using the scikit-learn machine learning library. Specifically, they authors used LinearSVC for a linear SVM, DecisionTreeClassifier for a single decision tree, and GradientBoostingClassifier for a decision tree ensemble. The boosting classifier was fit with parameters n_estimators = 4,000, learning_rate = 0.1, max_depth = 5, and max_features = "log2". Models were fit with sample weights inversely proportional to class balance. In our case, we compared performance of gradient boosting with performance of non-linear SVM from the SVC package.

3.2.4 Evaluation of Model

In[50], models were evaluated using tenfold cross-validation (CV) where data were randomly divided into ten non-overlapping training and test sets. Performance was measured using multiple metrics, like F1-score, and the average over all test sets are reported. In our case, we compared this random CV with chromosomal segregated CV, for which training and testing sets were split based on different chromosomes the enhancer-promoter pairs came from.

3.3 Results

3.3.1 Significant Sharing of Information Occurs Between Training And Testing Datasets

Whalen, et al, [50] used an F1-score performance measure, F1 = 2PR/(P+R), where P = precision and R = recall, and reported that their average F1 across 6 cell lines was 0.83. This typically implies P,R>0.8 and AUPRC,AUROC∼0.9, and that most EP pairs are being correctly classified as interacting or not according the epigenomic state of the enhancer (E), promoter (P), and the intervening genomic interval, or window (W), between the enhancer and the promoter. Interestingly, they found that window features (W) were often selected as being most predictive in the model, and that lack of some epigenomic marks in the genomic interval were correlated with positive EP interactions, a compelling hypothesis. Curiously, this high performance was only achieved with gradient boosting using a very large number of trees, and not with a linear SVM, which achieved F1-score∼0.2. In the course of examining their training data to find direct evidence in support of this hypothesis, we noticed that partly due to the coarser resolution of the Hi-C interaction data, multiple contiguous enhancers often positively interact with the same promoter, but are all labelled independent positive EP interactions for training, even though they potentially share P and W features. If samples with identical features are not in the same cross-fold validation test set, these shared features could lead to incorrect evaluation of generalization performance. There are two distinct mechanisms by which features could be shared. The simplest is that many EP pairs share a promoter. In the K562 and GM12878 training data from Whalen, et al, [50] over 92% of EP pairs share a promoter with another EP pair and have a 2-fold or greater positive/negative class imbalance, and are thus subject to significant test set contamination through the shared P features. In Fig 3-1 we show how many EP pairs share a promoter with a given number of positive and negative enhancers (positive and negative EP pairs connected to a single

promoter). The EP pair counts in Fig 3-1 are relevant for understanding the degree of potential test set contamination. The greater the class imbalance, the more accurately this set of EP pairs can be predicted from training set promoter features alone. This is a serious problem for all pairs labelled red or blue in Fig 3-1. For example, at position (7,2) in Fig 3-1A, there are 27/(7+2) = 3 distinct promoters in the training data which each interact positively with 7 enhancers and negatively with two enhancers (two negative EP pairs) in the training data. When one of these positive pairs is in the training set, a specific but non-generalizable rule on that promoter's features would predict the other pairs with accuracy of 78% when they are in the test-set.

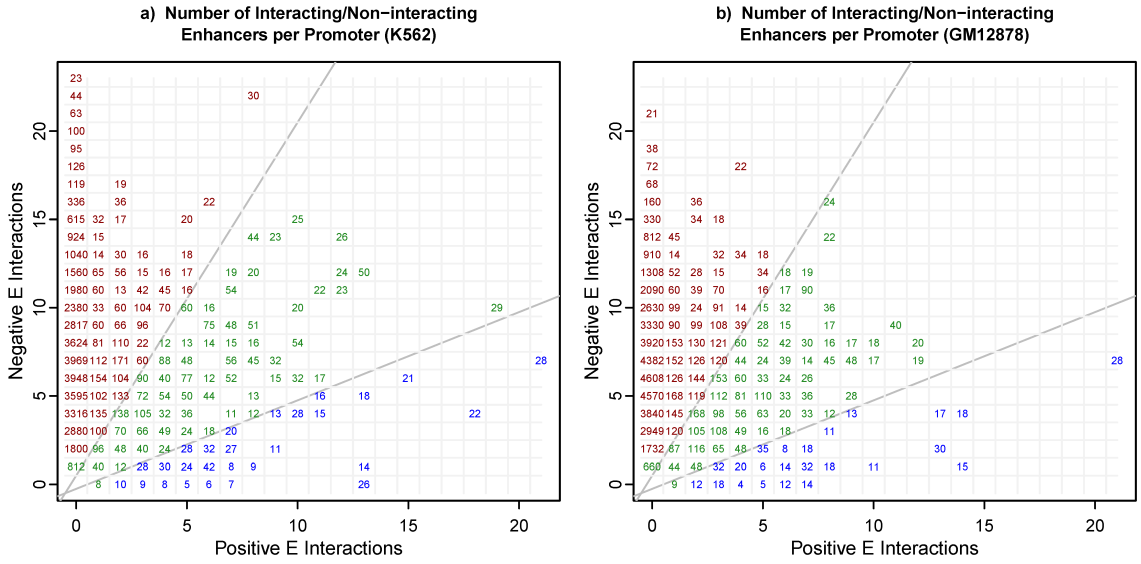

Figure 3-1. Most of the EP pairs used for training share promoters and promoter features with multiple EP pairs

a) For cell line K562, and b) for cell line GM12878, for each promoter we show the number of positive and negative EP pairs sharing a promoter. In a) for example, there are 27 interactions in position (7,2), which means three distinct promoters each interact positively with 7 enhancers and negatively with two enhancers in the training data. When the number of negative and positive EP pairs interacting with a promoter are imbalanced, they can be predicted correctly from training set promoter features. The blue interactions have a 2-fold or greater imbalance (pos>neg) and the red interactions have a 2-fold or greater imbalance (neg>pos), and both red and blue interactions (92% of the data) can lead to test set contamination by shared promoter features.

This potential problem becomes especially problematic for all EP pairs near the x

and y axes in Fig 3-1, and most of the negative samples are particularly susceptible. The second mechanism by which EP pairs could share features is that the window (W) features are defined to be the epigenomic signals in the interval between the enhancer and promoter elements, so enhancers on the same side of the promoter can share window features, or more generally, any overlapping EP pair can share some common W signal features. We used cross-fold validation (CV) test sets sorted by chromosomal position so that no EP pairs with either P or W shared features could be in the same test set, and test set performance could be correctly evaluated. We also tested a cross-fold validation scheme where all EP pairs sharing a promoter were constrained to be in the same test set (promoter segregated). In this case, promoter features cannot lead to test set contamination, but window features can when using EPW features or W features to train the model.

3.3.2 Improper Experimental Design Lead to Overestimation of Model Performance

We reproduced the published results from TargetFinder using the training data for the K562 cell line as shown in Fig 3-2A), and when we further observed that a nonlinear SVM (rbf with default C and) could not achieve the high performance of gradient boosting, and that high test-set F1 is only achieved with 4000 trees, we suspected that test set contamination was important. When we instead used cross-fold validation (CV) test sets sorted by chromosomal position so that no EP pairs with shared features could be in the same test set, we found that test-set F1 dropped dramatically, as shown in Fig 3-2B. In fact, with sorted chromosomal CV test tests the predictive performance is only slightly better than random (F1 = 1/11 for this 1:20 ratio of positive to negative interactions). When using the promoter segregated CV test set scheme, P features cannot lead to test set contamination, but window features can, and do (EPW and W), when a gradient boosting classifier is used, Fig

3-2C. A similar but slightly less dramatic reduction in test-set predictive power is seen when removing or reducing the potential for test set contamination in the GM12878 cell line (Fig 3-2D, E and F).

3.3.3 Local Epigenomic State Cannot Predict Enhancer-Promoter Interactions with High Accuracy

To examine generalization performance of these models detail, we directly compared training-set and test-set performance with multiple models, feature sets, and parameters, on both the random and chromosomally sorted CV test-set schemes. Test and training set F1 are shown in Fig 3-2A and B as the number of trees are varied for gradient boosting (used in TargetFinder) and random forest methods, using EPW features for K562. Training set F1 is much greater than test set F1 and increases with the number of trees. With random CV test sets this leads to an artificially inflated test set F1 = 0.84 for $N_{tree} = 4000$ (as reported in [50] and similar methods), while with chromosomal CV test sets the correct test set performance is F1 = 0.13. Full scans of training and test set AUROC, AUPRC, and F1 performance metrics for gradient boosting and random forests shows high training set performance with large number of trees and low test set performance for all tree models with chromosomal CV test sets.

We next varied parameters for the nonlinear RBF SVM model. In the RBF SVM the key parameters are C, which controls the weighting for misclassification error, and γ, which describes the scale of the RBF kernel function and thus controls the smoothness of the decision boundary. For large C and γ the SVM is able to fit the training data with a convoluted decision boundary. The training and test set F1 of the RBF SVM are shown in Fig 3-3C and D for C = 1 and C = 1000 as is varied, and as expected, training set F1 approaches 1 for large γ. The default sklearn package defaults for the rbf SVM are C = 1 and $\gamma = 1/n_{features}$. For EPW, EP, and W feature

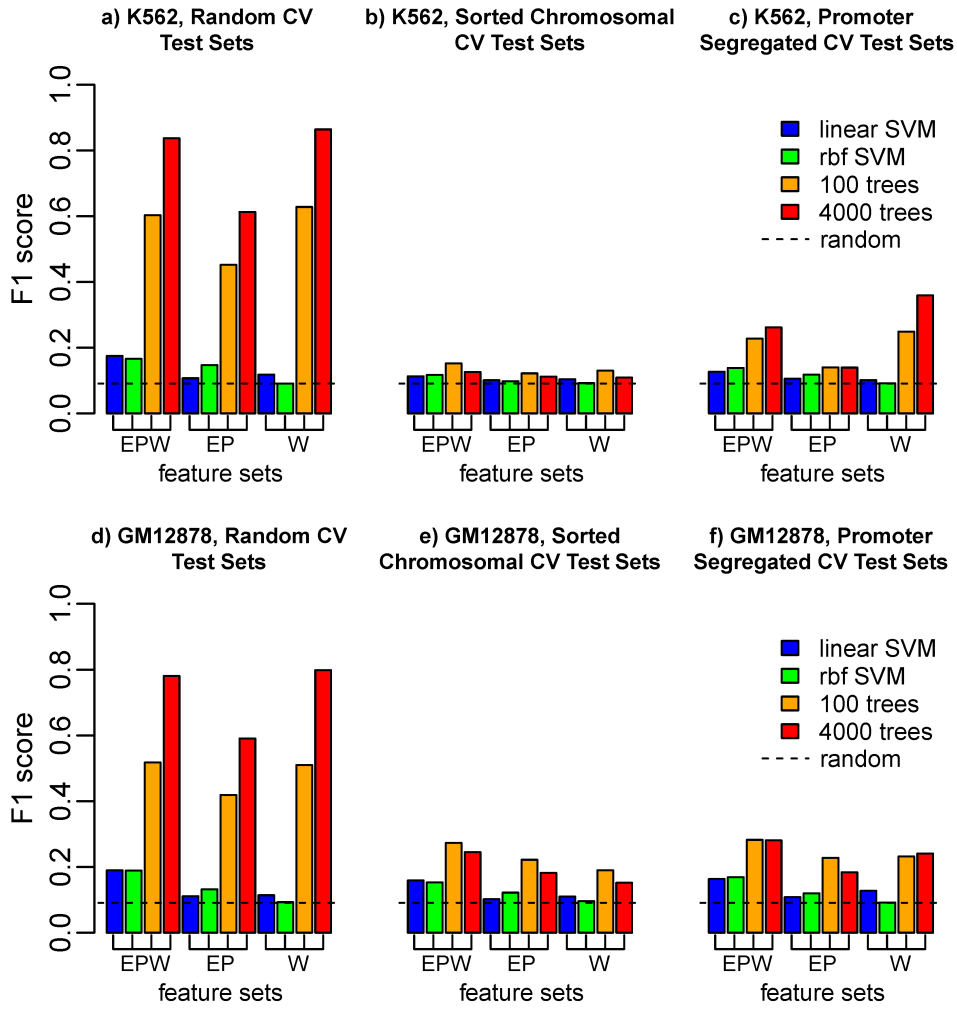

Figure 3-2. The apparently high predictive power of epigenetic state to identify enhancer-promoter (EP) interactions is largely due to incorrect evaluation of generalization performance

a) For cell line K562, using training data from Ref. 1, with random assignment of positive and negative EP pairs to cross-fold validation (CV) test sets, a gradient boosting classifier with a very large number of trees (4000) is able to achieve F1>0.8 with EPW or W features. b) If EP pairs with shared epigenomic features are properly forced to be in the same CV test set by sorting by chromosomal position, the predictive power is only slightly better than random, F1 = 1/11. c) If EP pairs sharing the same promoter are forced to be the same test sets, test set contamination through EP features is eliminated, but some can still occur through shared window features (W). d,e,f) A similar reduction in performance is observed when removing test set contamination in cell line GM12878 with sorted chromosomal CV test sets.

sets $n_{features}$ = (408, 272, and 136), so the default RBF SVM parameters are in the regime where there is little difference between training and test set accuracy (red points in Fig 3C and 3D). Full test set and training set performance for C = 1 and C = 1000, and test set F1 for a complete C and grid scan for K562 and GM12878 shows SVM performance was slightly better for GM12878. Consistent with our results from the tree models, no choice of RBF SVM parameters or features yielded performance significantly better than random with proper chromosomal CV test sets.

3.4 Discussion

We have shown that the high accuracy of TargetFinder [50] in predicting EP interactions is mostly due to improper evaluation of generalization performance. We consider it extremely likely that subsequent publications using similar or identical training data are also subject to this problem, and that reports of accurate predictive modelling of EP interactions from local epigenomic state should be re-evaluated [51, 59, 60]. On the bright side, there is now considerable room for improvement in models of EP interactions. Our results showing that no model performs much better than random guessing strongly suggests that local EP (and W) epigenomic state features alone are insufficient to distinguish interacting and non-interacting EP pairs. We suspect that EP interactions indeed may be specified by epigenomic state, but that additional features need to be considered, which are distinguishable from the feature set used in [50] by their "non-local" nature. These additional features may include relative genomic position, the presence or absence of competition or interactions between E or P elements, conformational constraints, or subtle epigenomic state differences between enhancers competing for the same promoter or multiple promoters. Finally, we note that the inability to correctly evaluate test set performance is likely to be less of a problem in approaches which learn interactions between larger scale domains in non-overlapping chromosomal bins [61].

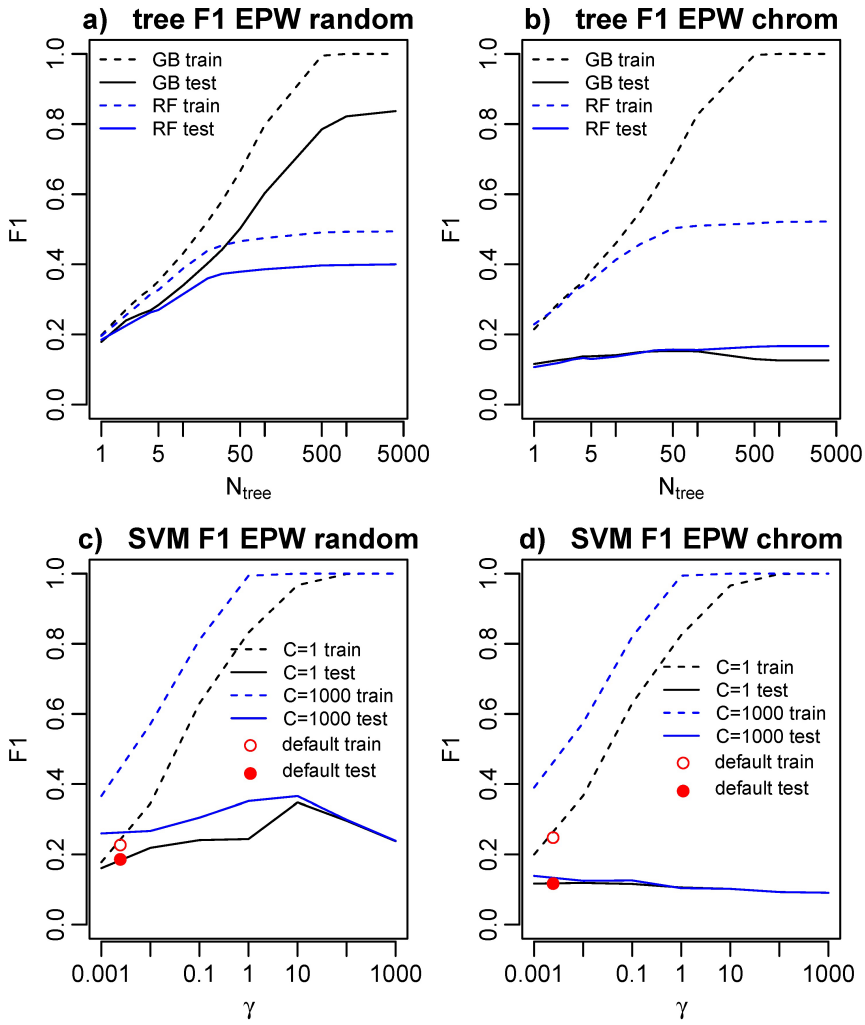

Figure 3-3. Parameter variation in gradient boosting (GB, black), random forest (RF, blue), and SVM models show significantly higher training set accuracy than test set accuracy for large number of trees or large C and γ

a) For cell line K562, with random CV test sets, test set performance appears high, but comparison with chromosomal CV test sets shows that this is due to test set contamination and failure to properly evaluate generalization performance. b) With chromosomal CV test sets, training set accuracy approaches 1 with large number of trees but now test set performance is correctly evaluated and test set F1 is very low. c) Nonlinear RBF SVM performance for incorrect random, and d) proper chromosomal CV test sets. For large , SVM training set accuracy approaches 1 and test set accuracy is low. For typical default RBF parameters C = 1 and $\gamma = 1/n_{features}$ (red) test set performance is still not significantly better than random.

Chapter 4

Loop Competiton and Extrusion Computational Model: a Probabilistic Model for CTCF Interaction Prediction

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.

4.1 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 techniques [3, 31, 62], and has revealed functional units including chromosome compartments, topologically associated domains (TADs) and loops. Chromosomal compartments, which exhibit a checkerboard pattern on a Hi-C map, correspond to active or inactive chromatin across several megabases. On the other hand, TADs and sub-TAD loops represent enriched chromatin interactions that appear at a scale of hundreds of kilobases or below. These smaller loops shape local chromatin structure, and their disruption has been reported to lead to dramatic dysregulation of nearby gene expression [36, 63]. The most prominent feature of TADs and loops is that their boundaries are usually marked by CTCF and Cohesin binding. 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 [7, 64]. It has been suggested that CTCF and Cohesin mediate TAD and loop formation through a loop extrusion mechanism, where Cohesin translocation generates a nascent chromatin loop until blocked by CTCF [10, 11] (Fig 4-1). Polymer simulations of a loop extrusion model successfully reconstructed TAD-like structures, and predicted the impact of CTCF or Cohesin degradation on TAD strength [10, 11] . Moreover, multiple experiments have validated in vitro that Cohesin is capable of moving through nucleosomal DNA [65] and generating a growing DNA loop progressively as it moves [12, 13].

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,

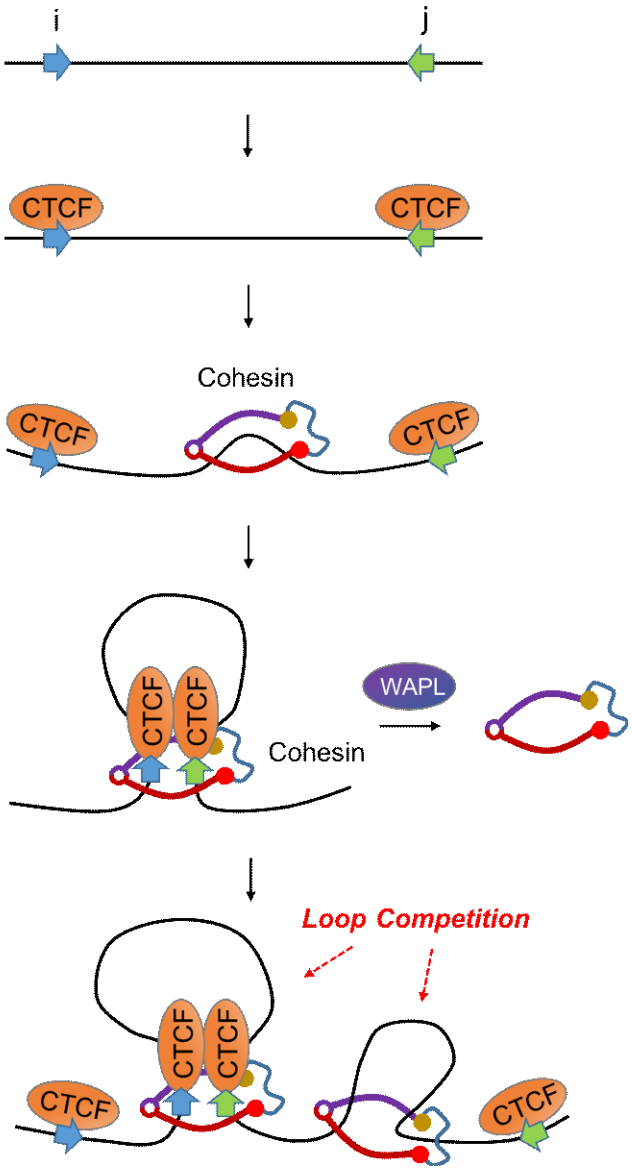

Figure 4-1. Loop extrusion model
Cohesin is loaded between (typically convergent) CTCF pairs, and the loop forms progressively as Cohesin translocates along the chromatin fiber. The extrusion process stops when Cohesin is stalled by CTCF. WAPL unloads Cohesin from chromatin. An existing loop could block movement of another Cohesin protein, leading to loop competition.

only about 2∼5% of these are identified to be interacting by direct Hi-C or ChIA-PET measurements (Fig 4-2). 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 [66], 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 [66–68], 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. Previous physical modeling of nuclear organization has focused more on general principles associated with the formation of TADs and loops but has not explored the variation in the strength of such features observed across different loci in real datasets. Additionally, polymer physics-based models treating the chromatin fiber as a connected chain of interacting units have shown encouraging global correspondence with measured contract frequencies, but their ability to predict individual CTCF loops has not been systematically evaluated [44, 45, 69, 70]. In contrast, one machine learning model, Lollipop, utilized a large set of genomic and epigenomic features to predict specific CTCF interactions with high accuracy [54]. This model motivated our approach, and provides some insight into this problem, but did not fully reveal 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

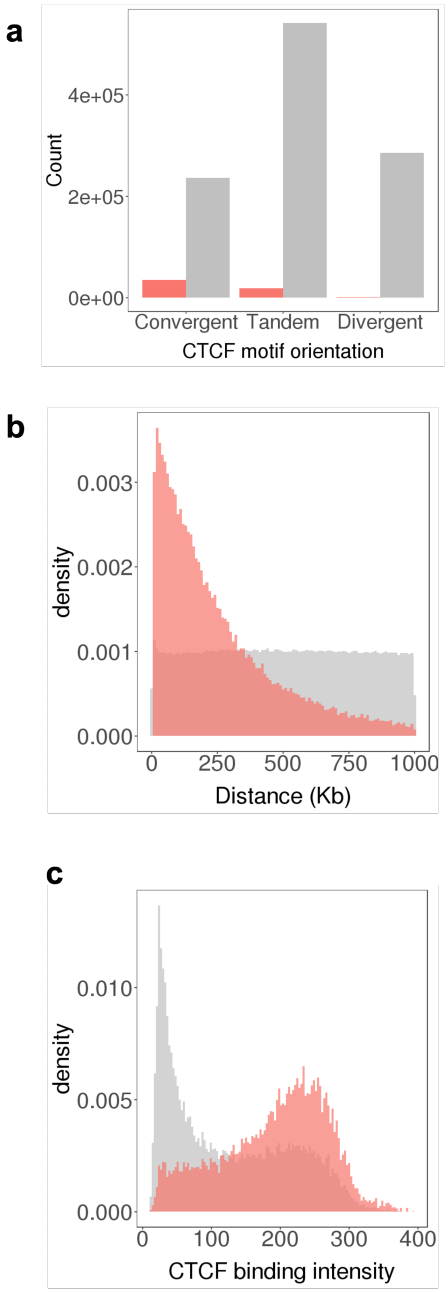

Figure 4-2. Distribution and features of CTCF loops
a) While measured loops prefer convergent CTCF pairs, other orientations also interact with significant frequencies and many neighboring (<1Mb) convergent CTCF motifs do not form loops: shown are interacting pair counts (red), and non-interacting pair counts (grey). Here interacting and non-interacting loops are defined by ChIA-PET interaction data. b) Distance distribution for interacting and non-interacting CTCF pairs. c) CTCF binding intensity distribution for interaction and non-interaction CTCF pairs.

CTCF-binding events, and loop competition (LC) (Fig 4-2). We show that this model can predict both ChIA-PET and Micro-C annotated CTCF loops with high accuracy. Our model includes an explicit contribution from the competition between overlapping loops, which is crucial for accurate prediction of loop formation. Our model of LC 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 predicting published CRISPRi perturbations 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.

4.2 Methods

4.2.1 The Loop Extrusion Hypothesis

DNA folding is a major and well-organized process that occurs throughout cellular life cycle: during interphase, mitosis and meiosis, and in proliferating and differentiated cells. The complex process hierarchically organizes the topology of DNA at multiple levels: in most genomic regions, the 2nm long diameter DNA double helix is wrapped around histone octameters, leading to the formation of 10nm nucleosomal chromatin fibers [71]. These fibers are further folded into loops or topologically associated domains (TADs), often spanning hundreds of kilobase pairs (kb) of DNA [8, 9]. The long-range *cis*-interactions that form these structures can be detected genome-wide by techniques based on proximity-ligation and sequencing, such as Hi-C, ChIA-PET and Micro-C [3, 31, 72].

The loop extrusion hypothesis propose that long-range *cis*-interactions within a DNA molecule are generated by loop extrusion factors that bind to DNA and reel flanking regions of the same DNA molecule into a loop. According to this idea, chromatin loops would initially be small but would increase in size over time as the loop extrusion factor processively moves the chromatin into the loop. Activities of extrusion can be halted when loop extrusion factor encounters boundary factors, which also stabilizes the loop. The idea that eukaryotic DNA could be folded through loop extrusion was independently discussed multiple times, first to explain how different segments in immunoglobulin genes could be recombined [73], how enhancer sequences could find and activate gene promoters [74] and later as a possible mechanism through which Condesin complexes might fold DNA into mitotic chromosomes [75].

Cohesin was initially discovered for its ability to physically connect replicated DNA molecules [76]. It has been suggested that Cohesin might be able to connect DNA sequence not only in *trans* to mediate Cohesion but, also in *cis* to form chromatin loops [77]. This speculation was based on the observation that Choesin accumulate at genomic sites occupied by CTCF [78], which is a sequence-specific DNA binding protein that was found to mediate the expression of imprinted genes at the *H19/IGF2* locus by forming allele-specific chromatin loops [79].

Recently, the loop extrusion model has been put forward to explain chromosomal domain and loop formation based on in silico simulations [10, 11]. Specifically, it could explain how Cohesin forms chromatin loops in interphase that are anchored at CTCF sites: Cohesin would initially form a small loop that would grow until the extruding Cohesin complex encounters CTCF molecules bound to their cognate recognition sites on either side of Cohesin. At these sites, CTCF would stop the extrusion sites [80]. This simple concept could explain why and how Cohesin accumulates at CTCF sites and why both proteins are found at loop anchors and TAD boundaries. Importantly, this hypothesis also made testable predictions, which have largely been experimentally

confirmed: to be able to extrude loops, Cohesin should be mobile on DNA, as ChIP-seq experiments and time-dependent, gradual DNA compaction in yeast and mammalian cells have suggested [13, 81]; and the accumulation of Cohesin at CTCF binding sites should depend on CTCF binding, as was observed in CTCF depletion experiments [82]. More details about loop extrusion model is reviewed in [83].

4.2.2 Quantitative Model of Loop Formation by Extrusion

In this loop extrusion model, the key components are CTCF, Cohesin, and other loop-extruding factors (Fig 4-1). Our model relies on the assumption that looping interactions found in the CTCF ChIA-PET experiments are due to the blocking and localization of Cohesin at CTCF-binding sites. 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 and 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 dissociates from DNA or 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 genomic locations enriched in Cohesin 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 [84]. The most stable loop configuration is 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. First, the probability of CTCF binding at each genomic binding site is described by the chemical equilibrium:

$$p_i = \frac{[CTCF]}{[CTCF] + K_{d,i}} \qquad (4.1)$$

where [CTCF] is the concentration of CTCF to be inferred, $K_{d,i}$ is the local dissociation constant at site i [85]. We will use the local ChIP-seq signal, x, to determine $K_{d,i}$. To normalize, we let x be the local CTCF-binding intensity signal divided by its genome average. Since $K_{d,i}$ is the dissociation constant, x is inversely proportional to $K_{d,i}$, but with some unknown scaling factor. Since [CTCF] is also a constant, we can combine [CTCF] and the ChIP-seq signal scaling factor to write $x = \frac{a \cdot [CTCF]}{K_{d,i}}$ or $\frac{K_{d,i}}{[CTCF]} = \frac{a}{x}$, so the probability of binding, Eq. (4.1), can be simply written as: $p_i = \frac{1}{1+\frac{K_{d,i}}{[CTCF]}} = \frac{x}{x+a}$. The dimensionless parameter a can be thought of as an estimate of the average $\frac{K_{d,i}}{[CTCF]}$ over all the CTCF-binding sites, and turns the local ChIP-seq signal intensity into a probability of occupancy. 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. This simple one parameter orientation effect model is consistent with a more general treatment. The extrusion process adds an additional term which reflects the probability that Cohesin does not stochastically dissociate from the DNA fiber while translocating along it. A constant dissociation rate leads to an exponential decay term of the form:

$$D_{ij} = e^{-\frac{d_{ij}}{\lambda}} \qquad (4.2)$$

where d_{ij} is the distance between CTCF sites i and j. For example, if the probability of not falling off while translocating 1bp is , the probability of not falling off after

translocating n bp is α^n, and in terms of distance $e^{-\frac{1}{\lambda}} = \alpha$. This term leads to decreased loop interaction frequency when the distance between two CTCF-bound regions gets larger. The parameter can also be interpreted as the processivity of Cohesin, or equivalently, the average CTCF loop length, which has been estimated to be about 300kb [10].

The final notable component of our LC and extrusion model is the effect of LC. The mechanism of loop extrusion implies that one Cohesin bound loop could block additional Cohesin procession. This blocking prevents all CTCF pairs that overlap with a formed loop from interacting, since other Cohesins would have difficulty passing through, no matter where they load [10, 11]. We will consider both a complete blocking model, where the presence of one loop excludes the formation of all overlapping loops, and an incomplete blocking model, where Cohesin can process through existing bound Cohesins with some probability. Allowing some pass-through is motivated by emerging evidence from a structurally similar loop extrusion factor, Condensin [86], which we will also discuss in the context of a WAPL knockout. In the complete blocking model, the formation of one loop excludes all other overlapping loops, so the contribution of LC is

$$LC_{ij} = \prod_{mn \cap ij \neq \emptyset} (1 - p_{mn}) \qquad (4.3)$$

where p_{mn} is the probability of loop formation between CTCF sites m and n, as defined in Eq. (4.5) below. Specifically, LC_{ij} is an additional contribution to p_{ij} that reflects the constraint that an overlapping loop between two CTCF sites m and n is not formed. In this sense, the complete model with LC (Eq. (4.5), below) should be solved iteratively. But in the "Methods" section, we show that the full iterative solution of Eq. (4.3) is consistent with a simpler model, which just requires that all CTCF sites internal to the loop ij are unoccupied, using p_m from Eq. (4.1) for the probability of occupancy of site m. This approximate LC model can thus be written:

$$LC_{ij} = \prod_{i<m<j} (1 - p_m) \qquad (4.4)$$

In practice this approximate LC term reflects the fact that strong sites inside a loop can contribute to internal loop formation and outcompete the formation of the loop *ij*. We assume that the probability of Cohesin loading is constant along the 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

$$p_{ij} = w_{ij} \cdot p_i \cdot p_j \cdot D_{ij} \cdot LC_{ij} \qquad (4.5)$$

4.2.3 Parameter Determination of The Model

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 that 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 $\frac{<K_{d,i}>}{[CTCF]}$ 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 $\frac{<K_{d,i}>}{[CTCF]}$ is 8.5. Optimal w and $\frac{<K_{d,i}>}{[CTCF]}$ 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.

We used publicly available CTCF ChIA-PET data [87] in GM12878 and HeLa cells to determine the values of the parameters $\frac{<K_{d,i}>}{[CTCF]}$, w and λ in our model. Long read

ChIA-PET data was processed with ChIA-PET2 software under standard protocols to identify significant loops [88]. 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 open chromatin region around single CTCF-binding site. Second, comparison of CTCF ChIP-seq peaks with overlapping ChIA-PET anchors shows that they are relatively centered around each other. 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 [89].

We determined the optimal value of the model parameters by fitting the loop extrusion model to CTCF ChIA-PET data (Fig 4-3), 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, and training these three parameters on the full dataset or 5-fold cross validation both yield the same optimal values. We did a comprehensive grid search in ($\frac{<K_{d,i}>}{[CTCF]}$, w and λ) in GM12878 (Fig 4-4), 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 value of $\frac{<K_{d,i}>}{[CTCF]}$ is 8.5. K_d in vitro for CTCF binding to the H19/Igf2 CTCF-binding site has been measured to be 370nM [90] and nuclear [CTCF] is around 144nM [91–93]. This leads to $\frac{K_d}{[CTCF]}$=2.6. While our estimate of this parameter $a = \frac{<K_{d,i}>}{[CTCF]}$=8.5 is near this value, it is not unreasonable to expect that the global average of K_d at binding sites on chromatin in vivo will be somewhat higher than that measured on naked DNA in vitro at the H19/Igf2 site. The model is quite robust to parameter choices with a broad peak of high performance in the range of w (2–4) and $\frac{K_{d,i}}{[CTCF]}$ (5–10) (Fig 4-3). Also, the optimal parameters derived from training on GM12878

and HeLa are very similar.

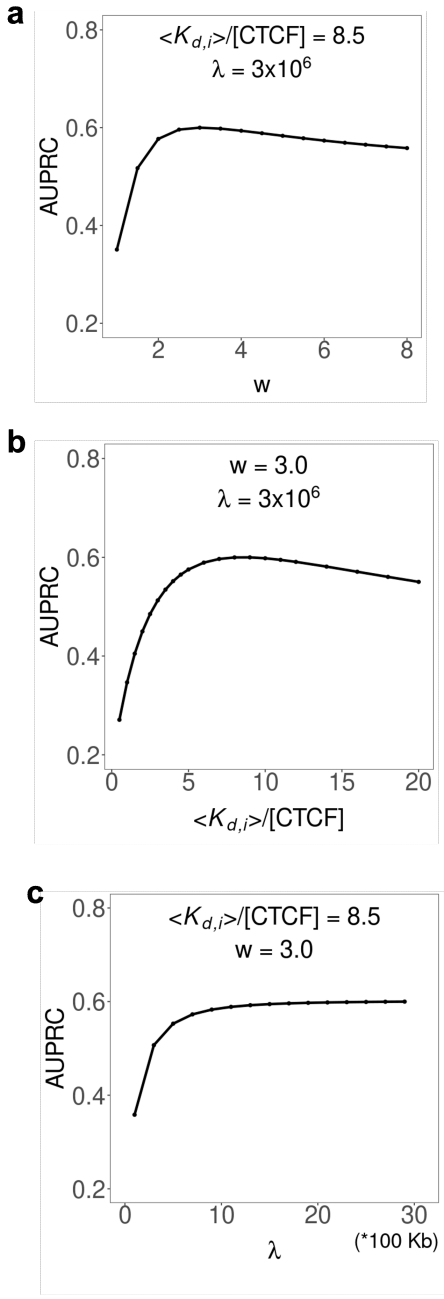

Figure 4-3. Parameter determination of loop extrusion mathematical model
a–c) Model performance is evaluated by area under precision-recall curve (AUPRC) as parameters are varied individually.

For λ, we expected the optimal value to be around the average loop length of 300kb. However, the agreement between our model and the ChIA-PET data increases

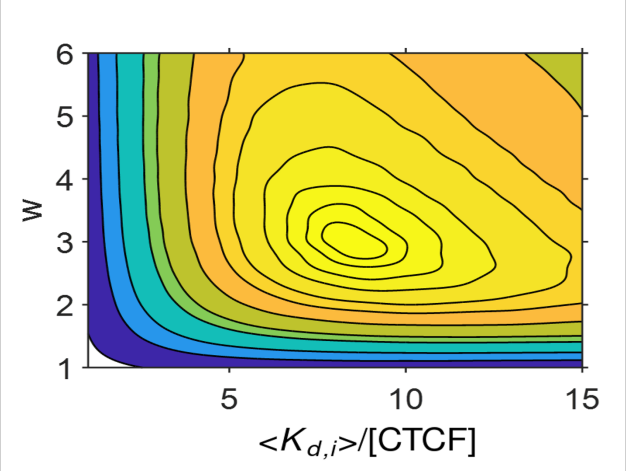

Figure 4-4. Grid search of optimal model parameter
Grid search of $\frac{<K_{d,i}>}{[CTCF]}$ and w to identify joint optimal value of parameter.

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 4-3C). 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. Previous Hi-C studies have reported a power-law decay relationship between chromatin contact frequency and genomic distance [11]. For completeness, we also compared a power-law decay with exponential decay, in the absence of LC, replacing D_{ij} with $D_{ij} = {d_{ij}}^{-k}$, but performance was slightly degraded for all choices of k relative to the exponential distance function. The distribution of CTCF loop lengths is constrained by the genomic position of CTCF-binding sites, and unlike Hi–C interactions does not follow a power-law distribution, but the loop length distribution of our predictions is in close agreement with the measured length distribution. We shall directly address the apparent paradox that predictive performance is independent of loop length in detail below.

4.2.4 CTCF ChIA-PET Data

GM12878 and HeLa CTCF ChIA-PET data were taken from a published dataset [87]. ChIA-PET2 pipeline with long read mode was used to process data and identify loops [88, 94]. 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 four PETs for GM12878 and three PETs for HeLa. We further constrained CTCF interactions to be within 1 million bp (Mb), as over 96% of loops fell into this range.

4.2.5 CTCF ChIP-seq Data

CTCF ChIP-seq of GM12878, HeLa and K562 was obtained from the ENCODE portal. Reads were aligned with BWA to the hg38 reference genome [95]. Peaks were called by MACS2 with default parameters [96].

4.2.6 CTCF Motif Analysis of ChIP-seq Data

The position weight matrix of human CTCF was download from JASPAR [97]. 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.

4.2.7 Boosting model

An ensemble-learning-based boosting model was constructed with the python Xgboost package. The model consisted of 50 trees, each with maximum depth of five 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. Xgboost is able to perform better than the loop extrusion model on the limited subset of features

(BI, Ori, Dist). We believe this is because when retrained on this subset, Xgboost is learning an appropriate distance weighting in the absence of LC, while for the loop extrusion model we used the optimal λ determined using all features. LC and distance are correlated features, and Xgboost can learn some of the effects of LC by regressing on distance.

4.2.8 Lollipop model

Lollipop is a previously published random forest model which can accurately predict CTCF interaction specificity using 77 features [54]. 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. Although the original setup of Lollipop training used random test sets, the performance was similar when we reran with chromosomal test sets, with AUPRC=0.88 (random) and 0.86 (chromosomal) for GM12878, and AUPRC=0.90 (random) and 0.89 (chromosomal), so the overfitting due to shared features is minimal for this training data set.

4.2.9 Polymer simulation of loop competition

A chain of 50 monomers were simulated under Brownian-like conditions using Langevin dynamics by LAMMPS [98]. 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 setting and parameters are the same as described in [11].

4.2.10 Micro-C Data

A total of 15,945 loops were called from 2.6B reads of mESC Micro-C dataset [72]. Chromatin loops were identified by using HiCCUPS6. Loops were called at 1kb resolutions at peak size=4kb, window size=10kb, distance to merge=2.5kb and

FDR<0.1.

4.2.11 Distance and LC-Matched Sampling of ChIA-PET Dataset

The effects of CTCF-binding intensity, orientation, distance, and LC on CTCF loop formation are quantified separately by four terms $p_i \cdot p_j$, w_{ij}, D_{ij}, and LC_{ij}. For each positive interaction loop, we define a distance matched non-interacting CTCF pair to be one with the same CTCF motif orientation, with the difference of $p_i \cdot p_j$ and D_{ij} between the two loop pairs within a factor of two. Therefore, the difference of distance between them is controlled, while the magnitude of the LC term LC_{ij} is not. Similarly, a LC-matched non-interacting CTCF pair is one with the same CTCF motif orientation, with the difference of $p_i \cdot p_j$ and LC_{ij} between them within a factor of two. These selection procedures generate two positive and negative CTCF pair sets with either matched distance or matched LC. We then evaluate our model's ability to accurately distinguish the positive and negative pairs in both sets, when including either LC or distance terms in our model.

4.2.12 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 $\frac{<K_{d,i}>}{[CTCF]} = 8.5$, $w = 3$, $\lambda = 3,000,000$. The effect of CRISPR deletion and inversion of CTCF motif on CTCF-binding intensity are taken from [66]. For 4C signal, we calculated the ratio of read counts per kilobase between 20kb bins centered around the perturbed CTCF-binding site and 200kb random genomic regions. The change of binding intensity and orientation are then integrated into model to determine the resulting interaction probability.

4.2.13 Population Hi-C Data

Normalized Hi-C contact matrices of lymphoblastoid cell lines (LCLs) were taken from [99]. 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 [95]. Raw counts of contact matrices were normalized to correct for known biases with HiCNorm [100].

4.2.14 WAPL Knockout Model

WAPL is known as a Cohesin unloading factor, as it removes Cohesin from binding with chromatin fiber. It has been reported that WAPL knockout increases cross-TAD chromatin interaction frequency and extends the size of chromatin loops. We hypothesize that this effect is due to the longer residence time of Cohesin on the chromatin fiber in the context of a WAPL knockout, which allows Cohesin to pass through existing loop boundaries (e.g. CTCF or other Cohesin) with some small probability, s. This pass-through probability attenuates the influence of LC and facilitates longer loop formation. The pass-through probability, s, thus reduces the LC effect of each overlapping CTCF loop by a factor of $1s$, and results in a larger loop interaction probability p_{ij}, which explains the increased loop numbers under WAPL knockout. The effect of Cohesin passing through is especially strong for distant CTCF pairs, as their interactions are likely to be affected by more competing loops than nearby CTCF pairs. Therefore, it also explains the experimentally observed formation of higher order loop interactions, and is consistent with the shift of CTCF loop length distribution to the higher end under large s.

4.2.15 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.

4.3 Results

4.3.1 Loop Competition and Extrusion Model Accurately Predicts Formation of CTCF-Mediated Loops

We applied our quantitative model of LC and extrusion (Eq. (4.5)) 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 the analysis.

In addition to the systematic performance evaluation by AUPRC described below, one specific example comparing our model predictions with Hi-C and ChIA-PET data is shown in the TRIM5/6 locus in Fig 4-5. In this locus our model predicts a complex pattern of CTCF interactions that closely matches the ChIA-PET interaction counts. Hi-C picks up additional interactions within each CTCF loop which are not directly due to CTCF-interactions.

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 LC. An interaction probability p_{ij} was predicted for all positive and negative pairs for each model, and was then compared to the true class

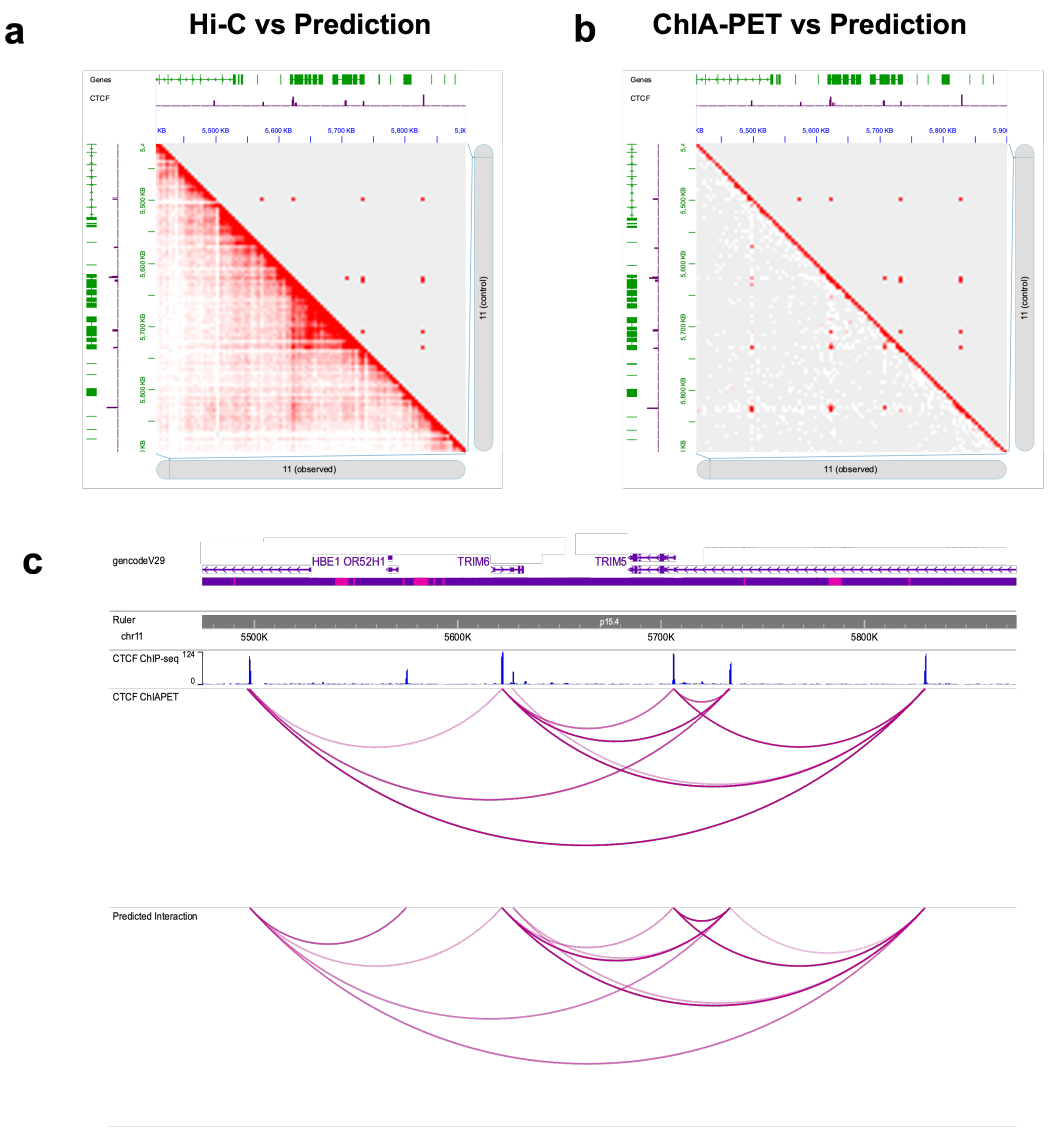

Figure 4-5. Model predictions compare favorably with Hi-C and ChIA-PET data in the TRIM5/6 locus

a) Hi-C data (bottom left) and f ChIA-PET data (bottom left). CTCF ChIP-seq signal is also shown as purple tracks in (a, b). The model predicts most of the direct CTCF ChIA-PET loop interactions, and Hi-C picks up additional contacts within the loops (or TADs) that are not the result of direct CTCF–CTCF interactions. c) In the same locus, loops called by our model and ChIA-PET data are quite similar, as visualized using the WashU epigenome browser.

label. Due to the huge class imbalance of CTCF interaction datasets, we employed area under the precision-recall curve (AUPRC) to evaluate model performance (see the "Methods" section). 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 4-6). The best performance is given by the complete model, combining CTCF-binding intensity (BI), CTCF motif orientation (Ori) and LC, with AUPRC=0.601. Performance on cross-fold validation test sets was: AUPRC=0.6005, std=0.003.

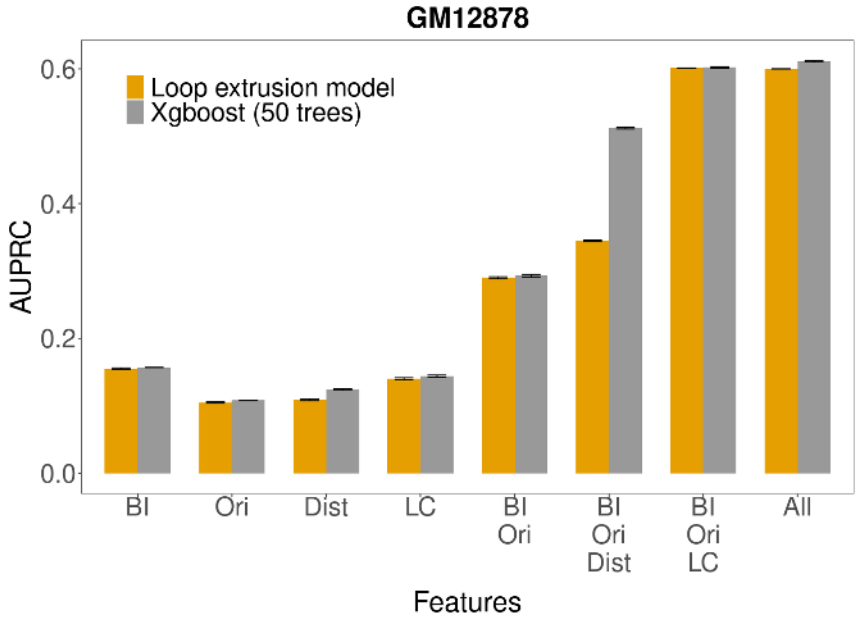

Figure 4-6. Model performance evaluation and feature importance in GM12878 Performance of our model with different combination of features for GM12878. BI—CTCF-binding intensity; Ori—CTCF motif orientation; Dist—distance; LC—loop competition. Performance is also compared against xgboost model with 50 trees. Down sampling of 10% of the data was repeated 10 times and 95% confidence intervals are shown.

This model combines these features in a functional form specific to an underlying mechanism of loop formation. To test this mechanistic assumption, we also constructed a more general 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 comparable 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 extremely close to experimental data (Fig 4-7) and matches Fig S2g of [87].

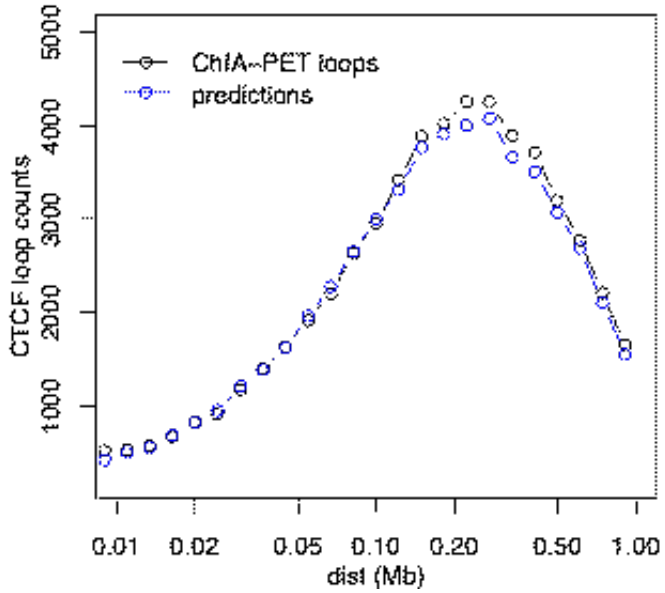

Figure 4-7. Distribution of predicted loop distance]
Loop length distribution for measured ChIA-PET loops and predicted interacting loops are quite similar.

Results in the HeLa cell line are qualitatively consistent with GM12878, with reduced AUPRC attributable to the larger HeLa 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. 4-8). 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 area under the receiver operator characteristic curve (AUROC) and AUPRC, which indicates that the information contained in our model is quite comprehensive, relatively more compact, and more easily interpretable.

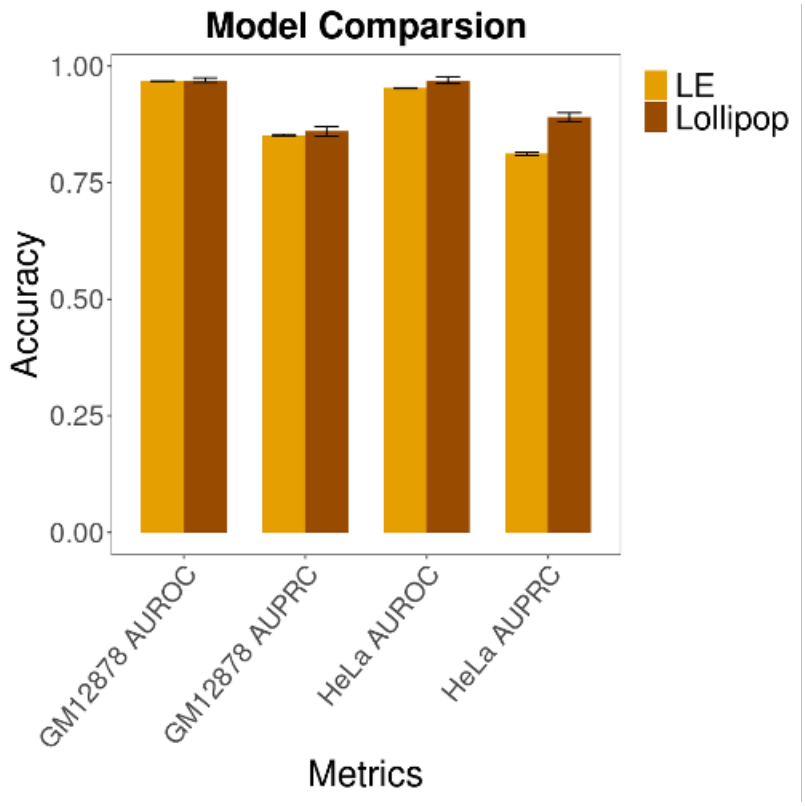

Figure 4-8. Comparing loop extrusion mathematical model vs Lollipop] Model comparison against Lollipop under class ratio 1:5 (positive vs. negative).

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 4-9A). In addition, positive and negative CTCF pairs are clearly separated by predicted interaction probability (Fig 4-9B).

To validate our model on an additional external dataset, we predicted CTCF loops identified from a recently published high-resolution Micro-C dataset [72]. In total, 15,945 significant loops at 1kb resolution were detected in this dataset with HICCUPS [9]. 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

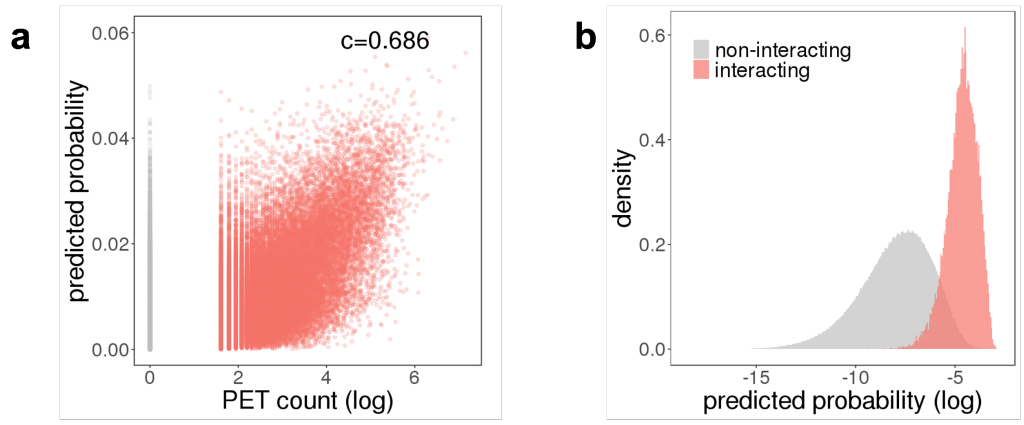

Figure 4-9. Model validation by quantitively assessing CTCF ChIA-PET dataset] a) Distribution of PET count (log scale) against loop extrusion model predicted interaction probability. Red dots are interacting CTCF pairs while grey dots are non-interacting CTCF pairs. b) Distribution of loop extrusion model predicted interaction probability.

(AUROC=0.944, AUPRC=0.849) (Fig 4-10), 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 LC 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.

4.3.2 Loop Competition is a More Powerful Predictor Than Distance

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 4-11). The only two features highly correlated with each other are distance and LC (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

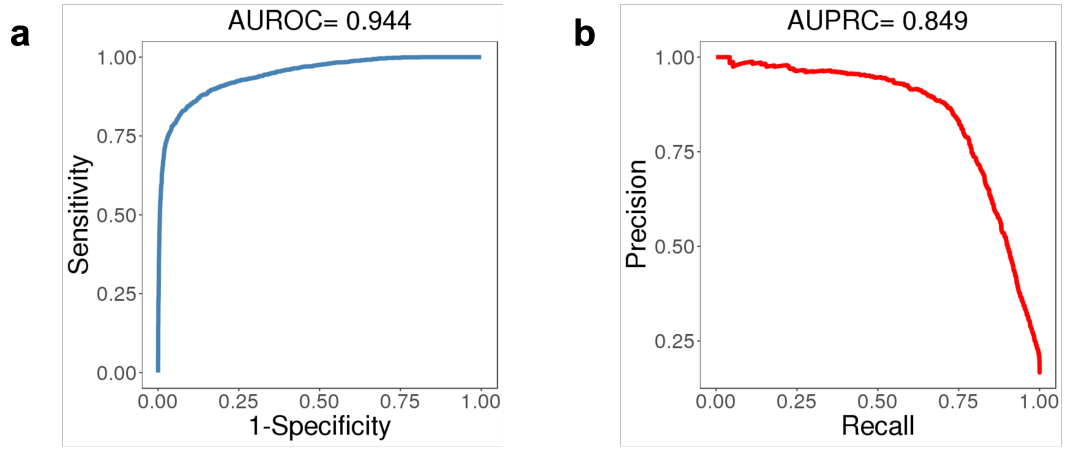

Figure 4-10. Model validation by quantitively assessing Micro-C dataset] a, b) Validation of model prediction performance on Micro-C CTCF loops with AUROC and AUPRC.

correlated features is more predictive of CTCF interactions by itself, distance or LC? 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 [3, 87]. 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 LC, we generated distance-matched and loop-competition-matched test sets by sampling the ChIA-PET data to isolate the contributions of each feature (Fig 4-12).

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 two for BI and Dist) (see the "Methods" section). In other words, every feature except LC 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 LC is the only unmatched feature. By evaluating our model on this

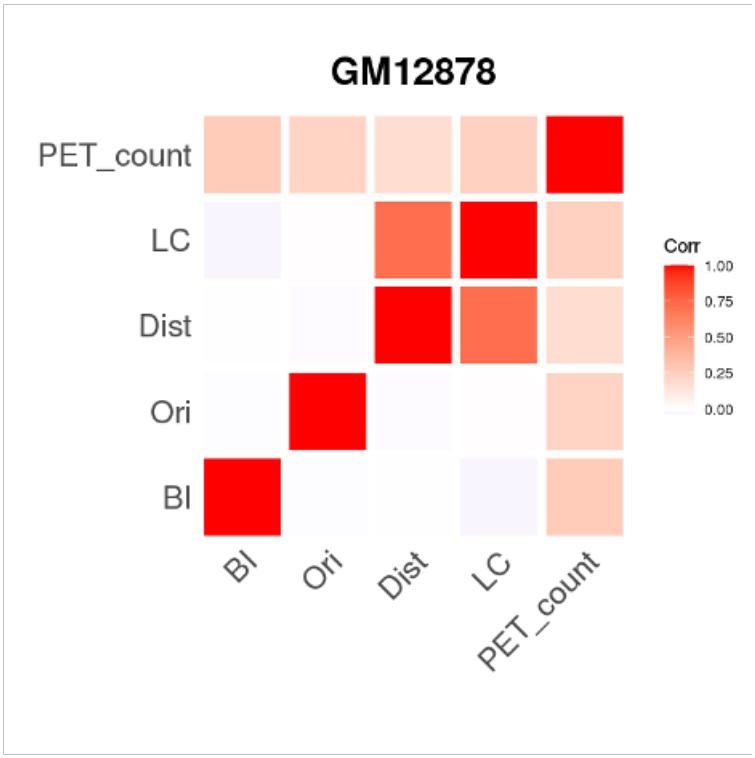

Figure 4-11. Feature correlation of loop extrusion mathematical model] Correlation of features (CTCF-binding intensity, CTCF motif orientation, distance, loop competition (LC) and PET count (log scale)) across all positive and negative pairs. Since loop competition and distance are correlated, we designed an additional experiment to isolate their relative informative value.

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 4-13B). In contrast, the model including LC reached AUROC=0.730, indicating that LC alone is predictive in this context and carries unique information about loop formation that does not exist in distance alone. We next generated a loop-competition-matched sample in a similar fashion, selecting positive and negative loops with similar levels of LC (within a factor of two) but unmatched distance, in which LC for each CTCF pair is determined by Eq. (4.4). 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 4-13D). The fact that LC is predictive in a distance matched context,

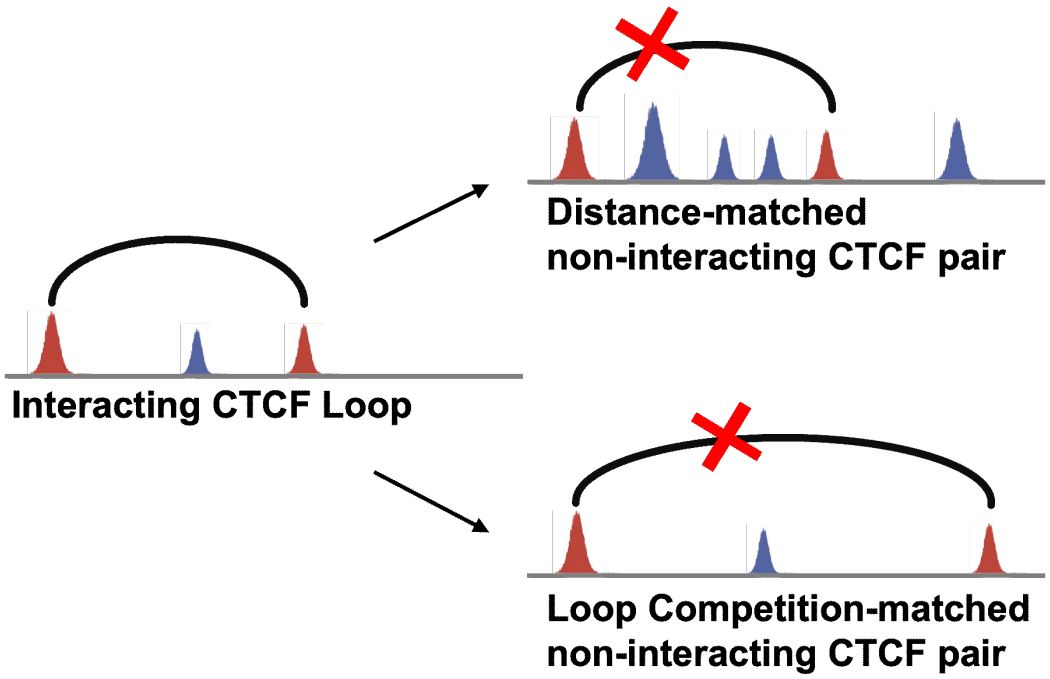

Figure 4-12. Assessing importance of loop competition and distance]
We generated distance-matched and loop competition-matched subsets of the full data by choosing a negative pair (marked with X) for each positive pair with either LC or distance matched within a factor of two.

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 LC when LC 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 LC. Consistent with this interpretation, distance has the weakest correlation with the PET count of loops among the four features. These computational experiments confer support for LC as an important determinant of CTCF interaction specificity.

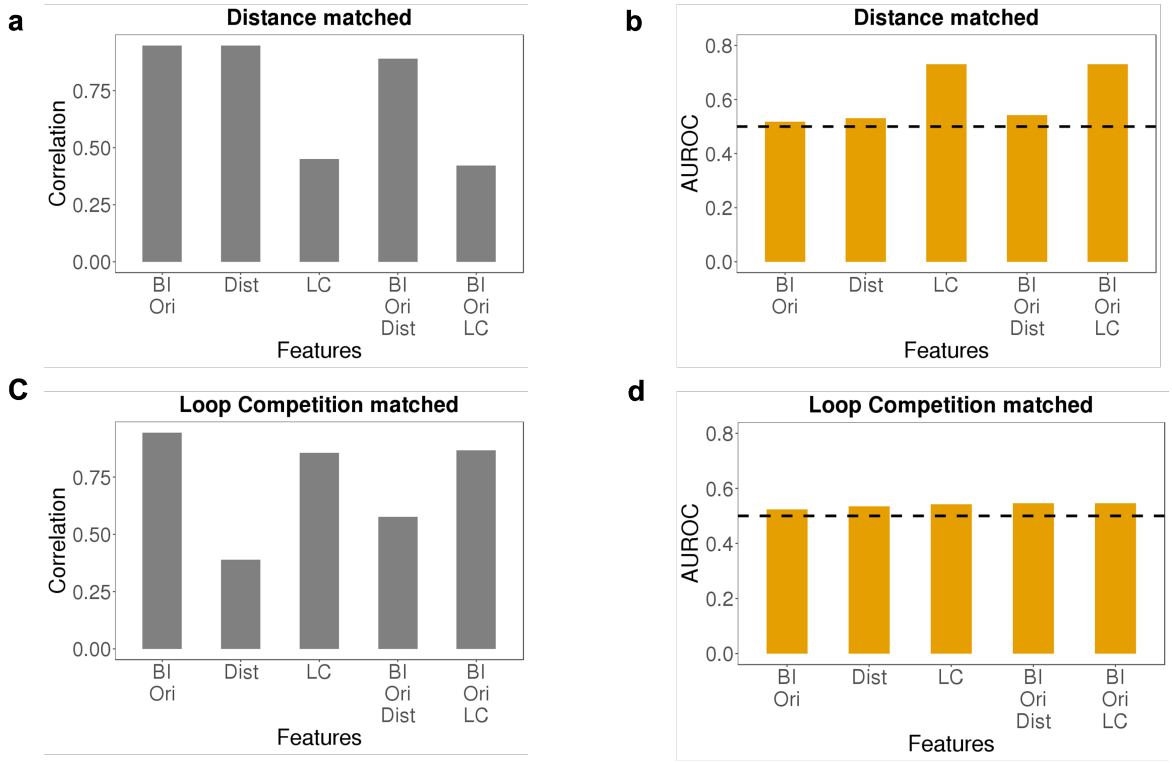

Figure 4-13. Loop competition is a more crucial determinant than distance] a, c) Correlation between positive and negative sets for different combinations of features in both matched settings. b, d) AUROC of loop extrusion model with different combinations of features in both settings. Down sampling of 10% of the data was repeated 10 times and 95% confidence intervals are shown. Since LC adds informative value in a distance matched evaluation set but the converse is not true, loop competition is the more predictive feature.

4.3.3 Loop Competition Is Validated by CTCF Disruption in Population Hi-C Data

Our model makes quantitative 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 LC 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 [66], and consequently other CTCF pairs spanning the disrupted site would be more likely to interact, as a result of reduced LC. A previously published dataset which measured Hi-C loop

interaction frequencies in lymphoblast cells derived from 20 individuals provides a direct means to test our model predictions of how CTCF disruption affects loop strength [99]. Natural genetic variation in this sample disrupted 49 CTCF-binding sites by SNPs. For each CTCF-binding site disruption, we separated individuals into two groups (strong or weak CTCF motif, as strong motif defined as those consistent with CTCF PWM at key positions in dashed boxes in Fig 4-14A, and calculated the ratio of average contact frequency in 40kb bins in neighboring 800kb windows in the two groups (Fig 4-14). After aggregating this data for all 49 CTCF sites, we observed that on average, bins that represent interactions between pairs of loci that span the CTCF motif (labeled as 'Cross') exhibit a higher normalized interaction frequency in weak vs. strong motif individuals (100/100 bins higher for weak motif individual), consistent with reduced LC in our model (model predictions shown in Fig 4-14B). In addition, interactions that do not span the CTCF-binding site (labeled as 'Outside') have much weaker differences, and their direction of change is much more random (52/90 bins higher for weak motif individual). This data supports the role of LC 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 [101].

4.3.4 The 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 [102]. 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 [9, 66, 103]. These studies provide important additional contexts to test our model. In one

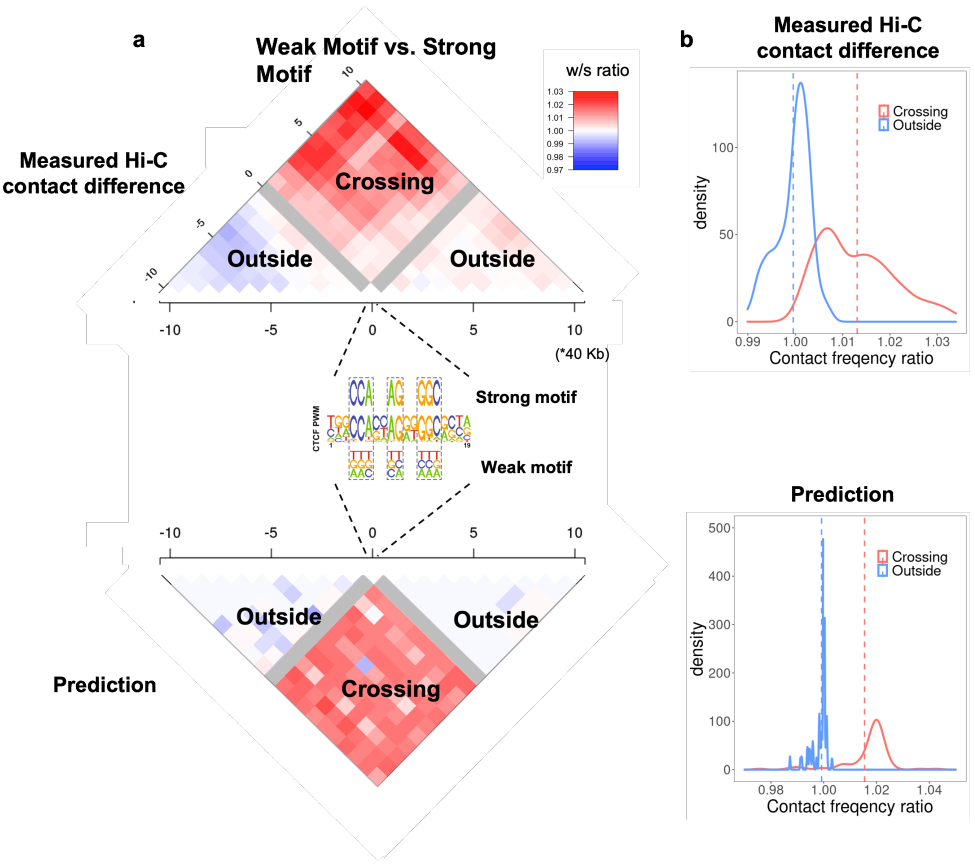

Figure 4-14. Loop competition predictions are consistent with changes in chromatin interaction frequency induced by naturally occurring CTCF-binding site disruption

a) Measured differential Hi-C contact frequency flanking SNP disrupted CTCF sites. The contact ratio for weak vs. strong CTCF motif genotype in a population of 20 individuals37 is shown. The heatmap is partitioned into 40kb bin pairs. Gray bins directly overlap the disrupted CTCF-binding site on one end, and loops which span the CTCF motif (Crossing) or do not span the CTCF motif (Outside) are indicated. Only loops which span the disrupted CTCF motif have increased contact frequency (top, red), consistent with reduced loop competition from our model predictions (bottom). Only SNPs which disrupt the indicated informative positions in the CTCF motif are used, and the strong or weak versions are labeled on the PWM from Jaspar MA0139.1. b) The same data is used to generate the contact frequency ratio distribution for the two classes (Crossing and Outside) of bin pairs for measurements (top) and our model predictions (bottom).

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 [66]. 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 4-15A–C, 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 4-15D–F). 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.

Alternatively, the activity of Cohesin can be modulated through the Cohesin unloading factor WAPL [104, 105]. 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 [92], 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 (Fig 4-16 and the "Methods" section), following previous work [86], we assume that Cohesin is not completely blocked at CTCF loop anchors, but can pass through with some small probability, s. WAPL knockout increases the residence time of Cohesin, which consequently has a greater chance of passing through boundary CTCFs. With enhanced pass-through probability, the effect of LC is reduced because Cohesin is moving more freely in this case. Through testing the WAPL-KO corrected model, we found pass-through probability is positively correlated with total loop number and average loop size. At pass-through probability around 0.4, we faithfully reproduced experimental results from WAPL knockout in HAP1 and Hela cell lines41,42 (Fig 4-16B,C). We also compare Hi-C data to model

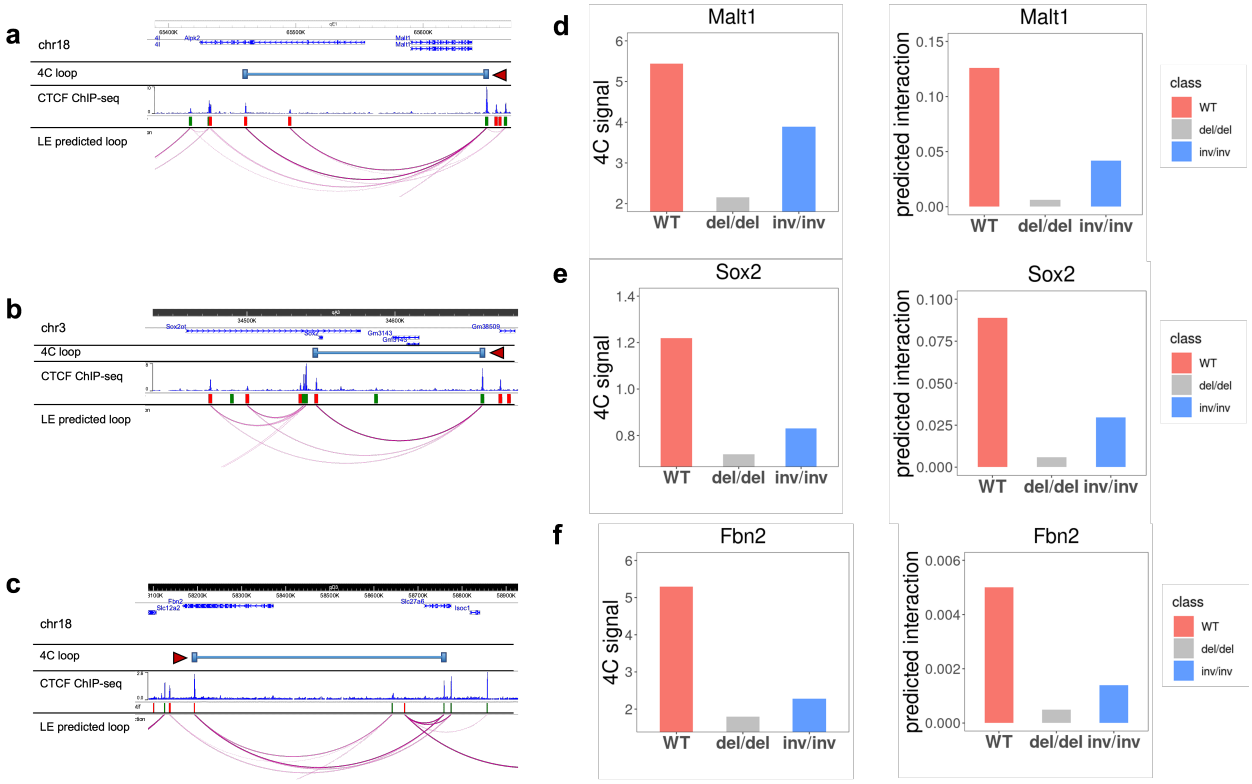

Figure 4-15. Loop extrusion model predicts the effect of targeted CTCF disruption and inversion on chromatin interactions

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) 4C-measured interaction frequency and loop extrusion model predicted probability of looping for wild-type and after CRISPR deletion or inversion of the targeted CTCF-binding site.

predictions in the context of the WAPL knockout in Hela in Fig 4-16D.

4.3.5 CTCF Loops Constrain Enhancer–Promoter Interactions

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. According to this idea, enhancer–promoter interactions should preferentially occur within CTCF loops, and not to cross CTCF loops. To

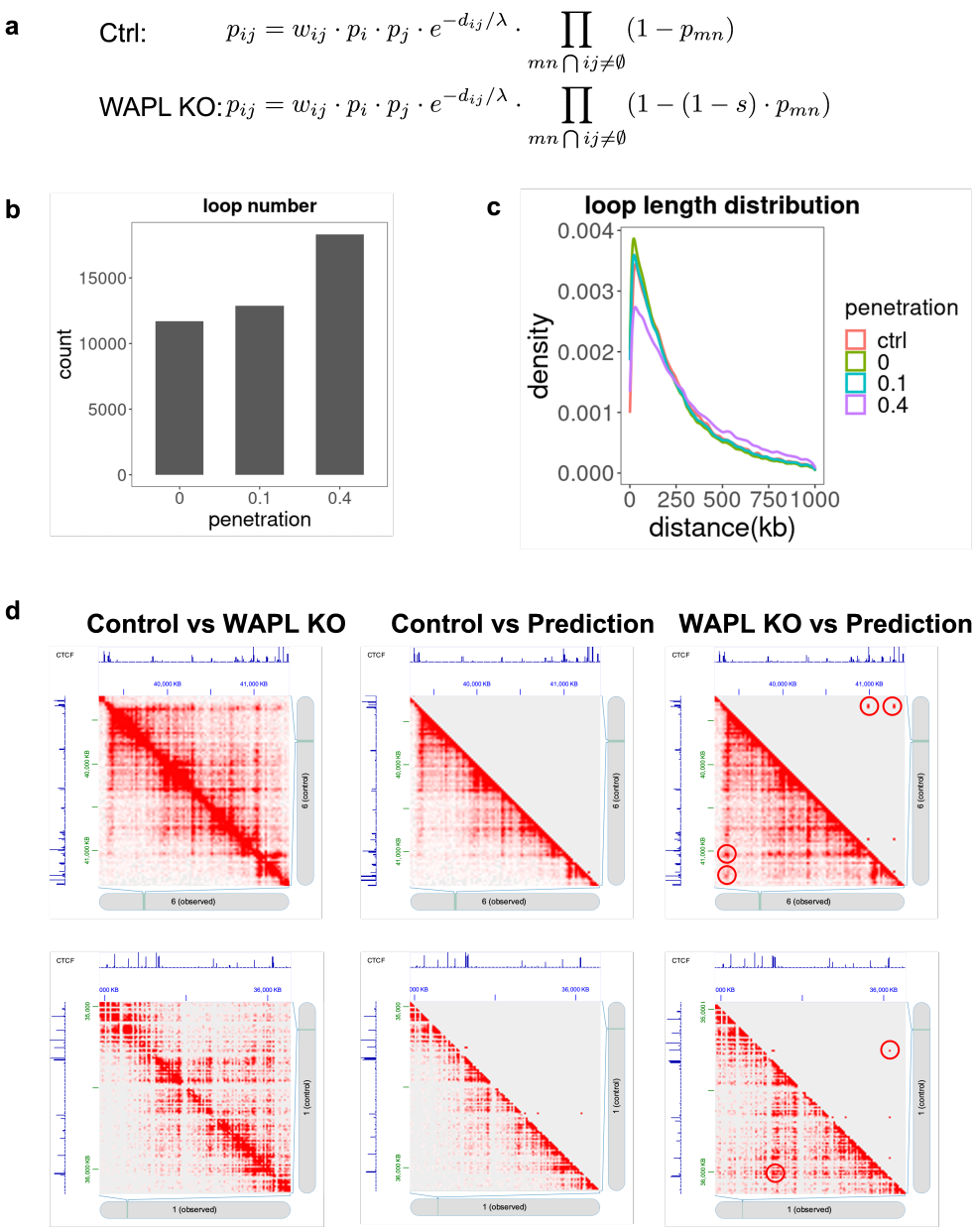

Figure 4-16. WAPL knockout increases overall CTCF loop length and number
a) Loop extrusion mathematical model accounting for WAPL knock out conditions. The parameter s indicates the probability of Cohesin passing through another bound Cohesin. b) Loop count increases with higher probability of Cohesin pass-though. c) Longer CTCF loops tend to be formed at higher probability of Cohesin pass-through. (d) Comparison between Hi-C data and model prediction before and after WAPL knockout. Our model successfully predicts formation of long distance CTCF loops.

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 [106–116]. 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 that crosses one or more CTCF loop (Fig 4-17). 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 LC, 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.

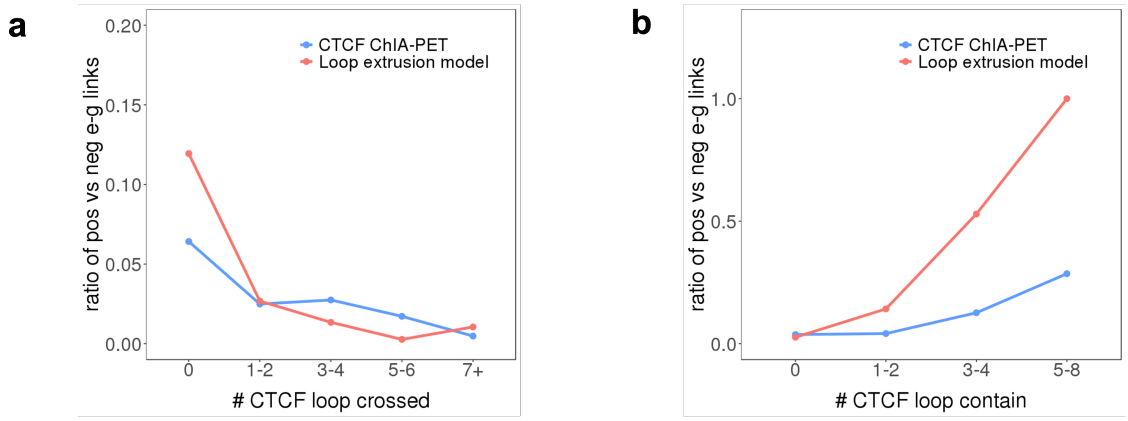

Figure 4-17. CTCF loops are predicted to constrain enhancer–promoter interactions, but loop extrusion model predicted loops do so more accurately
a) True/False ratio against the number of CTCF loops each E–P link crosses is plotted.
b) T/F ratio against the number of CTCF loops containing each E–P link is plotted.

4.3.6 CTCF-Binding Intensity Is Predictive of Cell-Type-Specific 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 [117]. 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 4-18A,B). 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 (AUPRC 0.955 for GM12878-specific loops, 0.739 for HeLa-specific loops) (Fig 4-18C,D).

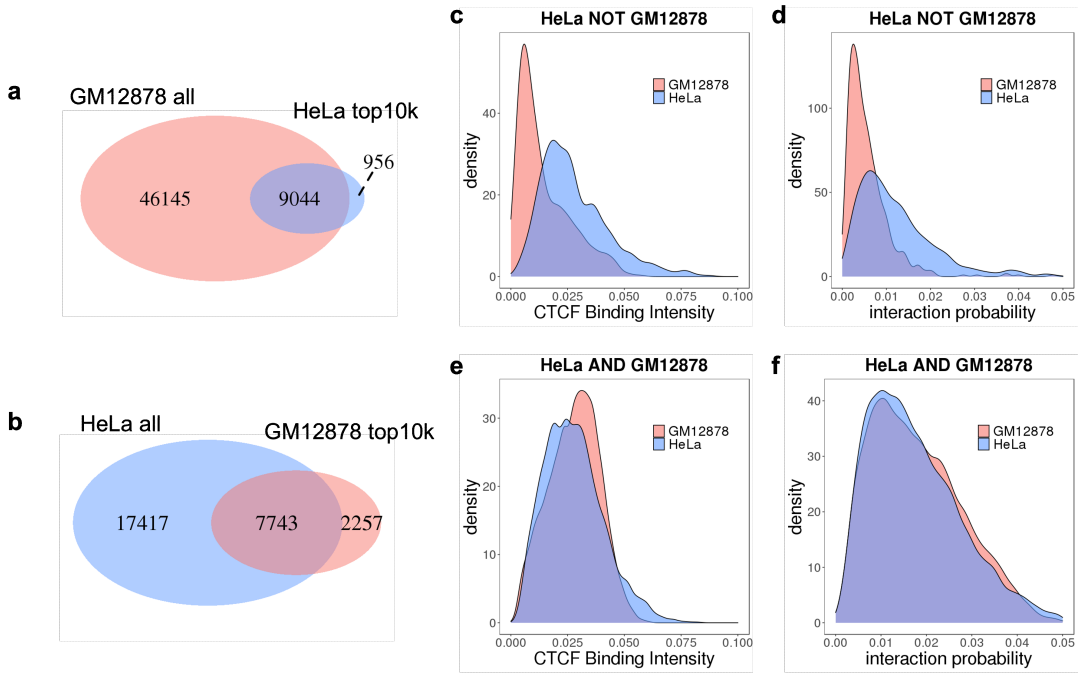

Figure 4-18. CTCF-binding intensity is predictive of cell type-specific loops
a, b) Venn diagram of CTCF-mediated loops identified from GM12878 and HeLa ChIA-PET. Only the strongest 10,000 loops are compared against each other due to different sequencing depth. c–f) CTCF-binding intensity distribution and predicted interaction probability distribution for HeLa-specific CTCF loops and shared loops.

4.4 Discussion

Recent progress in 3C techniques has enabled comprehensive annotation of higher order chromatin architecture, including CTCF-mediated loops. Predicting CTCF-mediated loops is a crucial first step toward understanding the mechanisms controlling regulatory element interactions and transcriptional regulation. While dramatic progress has been made mapping regulatory element activity in a large collection of cells and tissues [118] and detecting active TF-binding sites in these elements with machine learning [119], connecting regulatory element activity to dynamical models of gene networks and cell state transitions is in its infancy [119, 120], in large part due to our limited understanding of what controls enhancer–promoter interactions and how competitive or cooperative interactions between multiple enhancers are integrated at a target promoter. It has been shown that the interaction between enhancers

and their target gene promoters cannot be predicted solely from local epigenetic signals [5, 49]. 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 genes both inside and outside of the loop. Moreover, recent sequence-based modeling of enhancer–promoter interactions has also identified CTCF binding as the most important player [121]. We were motivated to develop a simpler model of CTCF interactions 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 [54].

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 LC. 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 [122]. 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 LC arises naturally from the loop extrusion process (Fig 4-1). The LC 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. 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 [36]. We used our quantitative predictions of LC 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. 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. Our modeling shows that these cell-specific CTCF loops are mediated by variable cell-specific activity of CTCF-binding sites.

Although our model is trained on ChIA-PET data collected from a population of cells, the probabilistic formulation of our model is consistent with quantitative measurements of the number of CTCF molecules in a single cell [91, 93], which suggest that not all CTCF-binding sites detected by ChIP-seq are consistently occupied, but that CTCF and Cohesin are popping on and off the genome as extrusion occurs. This is supported by the probabilistic form of our LC term. While in any given cell, an extrusion through a given CTCF-binding site may or may not be blocked, the time or population average of the probability of loop formation is what correlates with ChIA-PET contact frequency.

It is worth noting that the model presented here does not rule out other chromatin organization or loop formation mechanisms. For example, emerging experimental and computational evidence has suggested that phase separation could be the underlying mechanism for the larger scale A/B compartmentalization observed in Hi-C contact maps [46, 69, 123], and we envision the Cohesin/CTCF loop formation described here as operating on a shorter length scale within compartments. In addition, although Cohesin degron experiments provide compelling evidence for a model where Cohesin extrusion and CTCF blocking is the primary determinant of CTCF-mediated loop formation, this is not the only possibility. Loops could also be formed by interactions between other protein bound complexes (e.g. enhancers and promoters) modelled by "Strings-and-Binders" or SBS polymer models [44] that may dominate within CTCF loops. These polymer physics-based models have been used to predict the impact of structural variants on 3D structure [124], and to predict contact frequency variability between individual cells [125]. Our model can make a subset of these predictions, but our current formulation focuses only on CTCF-mediated loop interactions.

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 [126] would be an interesting next step, as these efforts are beginning to focus more on regulatory processes [127–129]. 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.

Chapter 5

Enhancer-promoter interaction prediction from CTCF looping constraints

Recent technological developments in chromatin interaction profiling have revealed CTCF and Cohesin as key factors for mammalian nuclear architecture. The idea that CTCF loops constrain enhancer-promoter(EP) interactions has existed for a long time, but its significance has not been systematically evaluated across the whole genome. Here, we analyze high resolution CTCF ChIA-PET data from 20 cell lines and experiments to investigate the dynamics of CTCF looping across cell types. We further show that the specificity of these CTCF interactions can largely be interpreted by the loop extrusion model presented in the previous chapter. We find strong evidence that long range EP interactions as detected by Pol ll ChIA-PET are frequently contained by CTCF contact domains (CCD), and crossed few or no CTCF looping boundaries. To systematically evaluate this observation, we built a mathematical model based on CTCF looping constraints and local chromatin accessibility that predicts EP interactions with high accuracy. The performance of our model is consistent with and sometimes superior to other state-of-the-art methods, such as the Activity-by-Contact (ABC) model in predicting distal enhancer hits identified by CRISPRi. Taken together, our analysis supports the hypothesis that CTCF interactions serve as scaffolds for

cell-type specific EP interaction formation and transcription regulation.

5.1 Introduction

Cell-type specific gene transcription in mammalian cells is regulated by enhancers. Enhancers can be located up to megabases away from their target genes, and interact with gene promoters in 3D space to activate their expression [14, 130]. The interaction between enhancers and promoters are modulated by the 3D organization of chromatin in the nucleus, which folds into loops and topologically associating domains(TADs) through a process mediated by CTCF and Cohesin proteins [8, 9]. Mechanistically, nested loops are formed through DNA extrusion activity of Cohesin, with CTCF acting as barrier and TAD boundaries [10–13]. The function of CTCF and Cohesin in establishing mammalian chromatin organization has been elucidated in several studies [8, 9, 67, 102, 105]. CTCF proteins are pervasively located at TAD boundaries and loop anchors[8, 9], and either perturbation of CTCF binding sites [66, 68] or degradation of CTCF and Cohesin [67, 102, 105] proteins can change chromatin organization dramatically[105]. Meanwhile, multiple computational approaches aiming at predicting chromatin interactions have also identified CTCF binding as the key determinant of 3D genome structure[55–58].

TADs and CTCF loops are thought to regulate gene expression by facilitating enhancer-promoter (EP) interactions contained within their boundaries and prohibiting EP interactions cross their boundaries [131, 132]. However, although Chromatin conformation capture (3C) assays have revealed that EP interactions are significantly enriched inside TADs, the functional implications of TADs and loops remain debatable. For example, while CTCF or Cohesin degradation were able to remove TAD patterns, they did not cause global change of gene expression (over a short time-scale of 6hrs) [67, 105]. Perturbation of a single CTCF binding site or TAD boundary does not always lead to dramatic expression change of nearby genes either [66]. Moreover, EP interactions

along with other classes of non-CTCF-mediated interactions also contribute to TAD formation, making it challenging to uncover a causal relationship by analysis at the level TAD. To address the role of chromatin organization, especially CTCF loops in modulating EP interactions, several studies have dissected clusters of CTCF binding sites with genetic perturbations to study their transcriptional and phenotypic outcomes [116, 133, 134]. These studies suggest that the number and orientation of CTCF binding events positioned nearby or in between enhancers and promoters are intimately associated with their contact frequencies, quantitatively aligning with gene expression changes, and predictive of developmental phenotypes. These results motivate us to systematically analyze the rules of CTCF looping constraints on enhancer-promoter interactions genome-wide by computational modeling.

5.2 Methods

5.2.1 CTCF ChIP-seq data

CTCF ChIP-seq data of 302 experiments and 34 additional ChIP-seq data for SMC3 and RAD21 were downloaded from the ENCODE portal. Data was processed using the ENCODE Transcription factor ChIP-seq processing pipeline. Reads were aligned with BWA to the hg38 reference genome [95]. Peaks were called by MACS2 with default parameters.

5.2.2 Consensus CTCF binding sites

For 291 CTCF and 34 Cohesin datasets with n>15000 IDR thresholded peaks, all peaks were trimmed to 300bp and then the top 20000 peaks from each dataset were merged, yielding 86,805 total peaks of variable length. Because most peaks overlapped very cleanly, final mean length of merged peaks was still 399.7 bp.

5.2.3 CTCF motif orientation

The position weight matrix of human CTCF was downloaded from JASPAR [97]. STORM with default parameters was used to identify the strongest CTCF motif and the corresponding strand for each CTCF-binding site [89].

5.2.4 CTCF ChIA-PET data

Long-read CTCF ChIA-PET data of 35 experiments were downloaded from the ENCODE portal. Among them, 17 are immune-related primary cells. The rest are common cell lines, including K562, GM12878, MCF7, HepG2, etc. Most experiments have two replicates, with some having one or three replicates. Additionally, two CTCF ChIA-PET datasets, named Tang_GM12878 and Tang_HeLa, were taken from a published work [87]. Pair-end read (PET) sequences were scanned for the bridge linker (sequence: 5'-CGCGATATCTTATCTGACT-3') and only PETs with the bridge linker were used for downstream processing. After trimming the linkers, the sequences flanking the linker were aligned with bwm-mem to the hg38 reference genome [95]. PCR duplicates were removed using custom code. Inter-chromosomal contacts and self ligation PETs with genomic span less than 3 kb were filtered.The remaining unique PETs were intersected with pairs of CTCF binding sites less than 1Mb to determine their contact frequency. Replicates were merged for all analysis except for correlation analysis. Pairs of CTCF binding sites supported by at least 3 PETs were called as loops.

5.2.5 Pol ll ChIA-PET data

Loops identified from Pol ll ChIA-PET data of 8 cell lines were downloaded from the ENCODE portal. Briefly, linkers were trimmed and PETs were mapped to the hg38 reference genome as described above. Then, inter-ligation PETs were extended by 500bp and clustered at their anchors into loops. Loops with PET count great than

2 were retained for analysis below.

5.2.6 The loop extrusion mathematical model

We previously developed a mathematical model of loop extrusion to predict CTCF interaction specificity [6]. It takes CTCF ChIP-seq data as input, and generates features of CTCF binding intensity, genomic distance, CTCF motif orientation and loop competition. This model depends on two parameters, dissociation constant $\frac{K_{d,i}}{[CTCF]}$ and contribution of motif orientation w. As the two parameters are likely independent of each other, their optimal values were determined separately by fitting the loop extrusion model to CTCF ChIA-PET data for different cell lines. Overall, the optimal $\frac{K_{d,i}}{[CTCF]}$ and w ranged from 4∼12 and 2∼6, which were consistent with previous study. Performances of the model were also robust to these parameter choices. The models were then applied to predict interaction probability for all pairs of CTCF binding sites across cell lines. The accuracy of predictions were evaluated by classifying the top 20k strongest CTCF loops defined by PET count from 20 fold non-interacting CTCF pairs (PET=0) with the predicted interaction probability.

As the distribution of motif orientation depends on distance, we thought to improve the loop extrusion model by introducing additional free parameters to capture this dependency. We proposed that the contribution of motif orientation to loop stability increases with log(distance), which can be characterized by a sigmoid function. Instead of a single scalar w, the motif orientation term was determined by transition distance, slope and w max. We performed grid search to find their joint optimal value. Results showed that the model with w scaled by distance consistently outperforms the model with fixed w in all cell lines.

5.2.7 Defining CTCF contact domains (CCDs)

The idea of CTCF contact domains (CCD) was first introduced in [87]. Here we proposed a different strategy to identify CCD and interpret clusters of CTCF loops. As CTCF binding sites in the genome are highly connected, we computed the CTCF loop coverage by aggregating CTCF loop PET count across the whole genome. We then applied the Hidden Markov Model (HMM) with n states (n=2, 3, 4) from python sklearn.hmm package on this 1D coverage track to define CCD with different levels of CTCF contact frequency. We compared the CCDs to TADs called from Hi-C data in GM12878 by computing the distance between their boundaries. For each promoter in gencode V24, we found its relative position to the nearest CCD boundary and aggregated them. We also sampled random pairs of genomic regions inside or outside strongest CCDs and compared their Hi-C contact frequency.

5.2.8 CTCF-loop Constrained Inter-Action (CIA) model

We proposed the regulation of enhancers on target genes can be inferred by combining chromatin activity of enhancers and the 3D contact between enhancers and promoters.

$$CIA_{E,P} = A_E \cdot \frac{CC_{E,P}}{CC_P + 1} \tag{5.1}$$

Under this model, the activity of enhancers (A_E) were computed from the geometric mean of normalized DNase-seq and H3K27ac ChIP-seq signal, as defined in the ABC model [Fulco 2019]. As spatial interactions between enhancers and promoters are largely constrained by CTCF loops outside, the 3D contact were characterized by CTCF Constraint ($CC_{E,P}$), i.e. the difference between total PET count of CTCF loops that contain the enhancer-promoter pair ($\sum_{CTCF\ loop\ i\ contains\ E-P} PET_i$) and the total PET count of CTCF loops that cross the enhancer-promoter pair ($\sum_{CTCF\ loop\ j\ crosses\ E-P} PET_j$).

$$CC_{E,P} = \sum_{CTCF\ loop\ i\ contains\ E-P} PET_i - \sum_{CTCF\ loop\ j\ crosses\ E-P} PET_j \tag{5.2}$$

This term was normalized by the sum of PET count of CTCF loops that span the promoter pair ($\sum_{CTCF\ loop\ i\ contains\ P} PET_i$) to allow cross-gene comparison. We filtered out CTCF loops greater than 500kb as most of them were weak and can be noisy.

$$CC_P = \sum_{CTCF\ loop\ i\ contains\ P} PET_i \tag{5.3}$$

5.2.9 Activity-by-Contact (ABC) model

The ABC model developed by Fulco et al. [107] predicts the contribution of an element on a gene's expression based on its enhancer activity and contact frequency with the gene promoter. The ABC score, Activity and Hi-C Contact are all derived from Nasser and Engreitz, personal communication, 2020.

5.2.10 CRISPRi-FlowFISH data

CRISPRi-FlowFISH is an assay that quantifies the effect of regulatory elements on gene expression [107]. Guide RNAs targeting candidate promoters or enhancers direct KRAB-dCAS9 to repress them in a population of cells. Then, RNAs of a gene of interest are fluorescently labeled with FISH. Cells are sorted by fluorescence label with FACS into bins and gRNA abundance are sequenced to infer the effect of perturbation on expression. We evaluated the predictions of the CIA model, ABC model and a simple activity-over-distance baseline model on CRISPRi-FlowFISH data at CCDC26, FTL, FUT1, GATA1, HNRNPA1, JUNB, KLF1, NFE2, PLP2, PQBP1, PRDX2, RAD23A. Data is taken from [107]. We filtered out genes with less than 2 distal hits to ensure that the evaluation (AUPRC) will not be biased by genes with few enhancers.

5.2.11 HCR-FlowFISH data

HCR-FlowFISH, developed by Reilly et al., also employs CRISPRi to perturb thousands of genomic loci [135]. It leverages hybridization chain reaction (HCR) to amplify transcripts for accurate detection. We evaluate the CIA model and other methods on HCR-FlowFISH at FADS1, FADS2, FADS3, MYC, PVT1, MEF2C, NMU, CD164, LMO2, HBG1, HBG2, HBE1. Data is taken from the ENCODE portal. Genes with more than 2 distal hits are retained in the analysis.

5.3 Results

5.3.1 CTCF loop annotation and integration

Here we present a framework that includes high-resolution CTCF loop detection, integration and interpretation, with a focus on their role of constraining enhancer-promoter interactions. In this study, we obtained CTCF ChIA-PET data of 35 cell lines generated during ENCODE phase 4 (18 non-immune progenitor, B or T cell lines), which is the most deeply-sequenced CTCF interaction dataset to our knowledge. Compare with other 3D genome mapping methods like Hi-C [3] or Micro-C[72], chromatin interactions associated with specific protein factors are enriched by ChIA-PET through immunoprecipitation [31]. Two additional experiments on GM12878 and HeLa are taken from an earlier work [87]. To analyze this data, we first generated a list of consensus CTCF binding sites of 86,805 as pre-defined loop anchors, through merging 291 CTCF ChIP-seq and 34 Cohesin experiments available on ENCODE (Methods 5.2.1). CTCF peak intensity from different cell types and experiments are in general highly consistent (median correlation = 0.82, Fig 5-1).

For CTCF ChIA-PET data, after mapping to the reference genome and removing duplicates, all remaining pair-end reads (PETs) are uniformly mapped onto the pairs of CTCF binding sites within 1Mb from the consensus CTCF binding list, as the

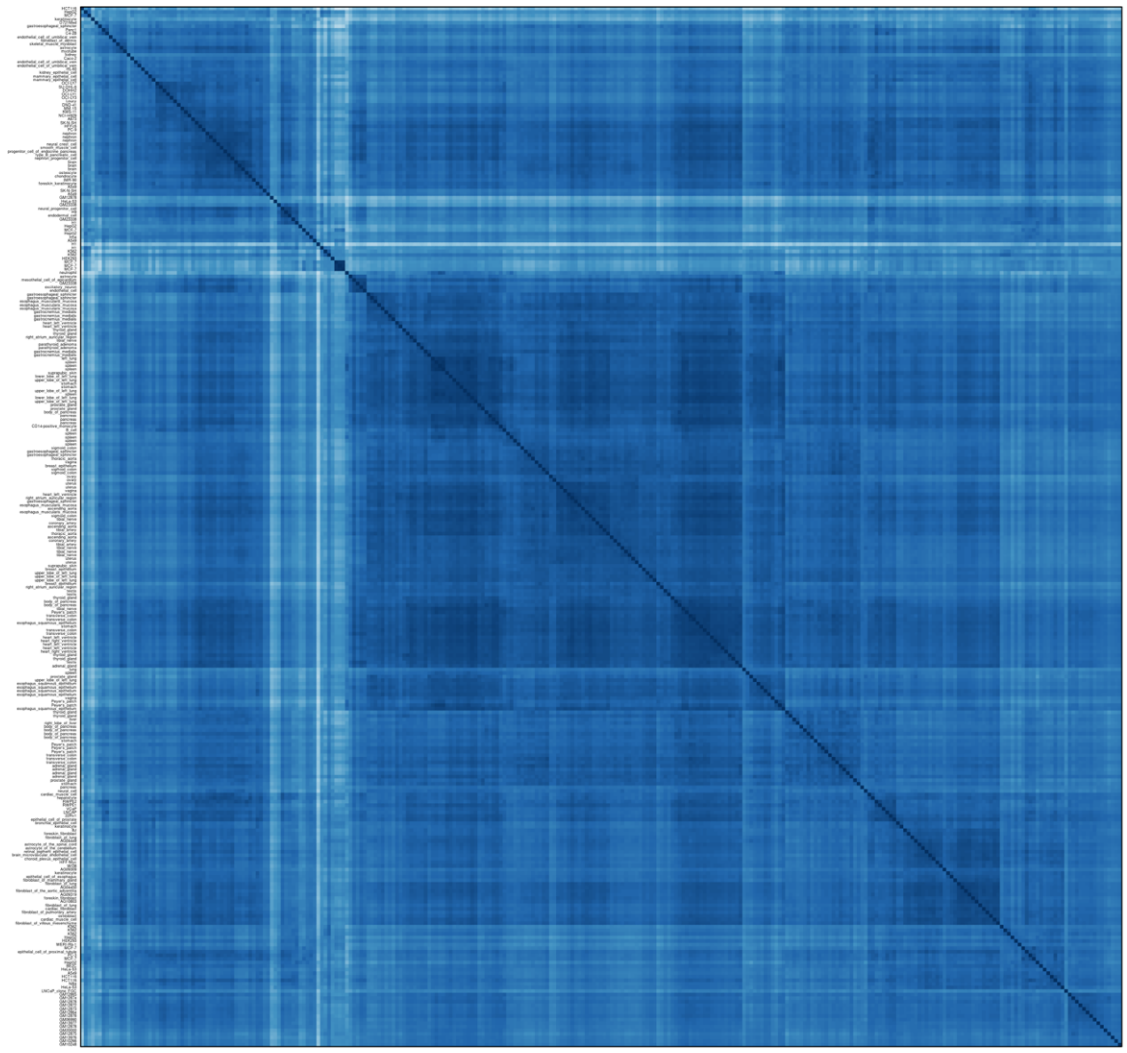

Figure 5-1. Clustering of CTCF ChIP-seq data
Hierarchical clustering of 291 ChIP-seq experiments. CTCF ChIP-seq read count for each peak is used as input.

majority of CTCF loops fall within this range [6]. This allows convenient cross cell-type comparison of loops as loop anchors are all shared, and CTCF binding events outside our consensus list are most likely false positives. CTCF loops between replicates are highly correlated. Clustering of CTCF loops across all experiments reveals distinct patterns, although the overall correlation is high (median correlation = 0.52, Fig 5-2).

We further constructed a classification task, by using PET count from one cell type to predict the strongest 20k CTCF loops in another cell type (Fig 5-3A), with

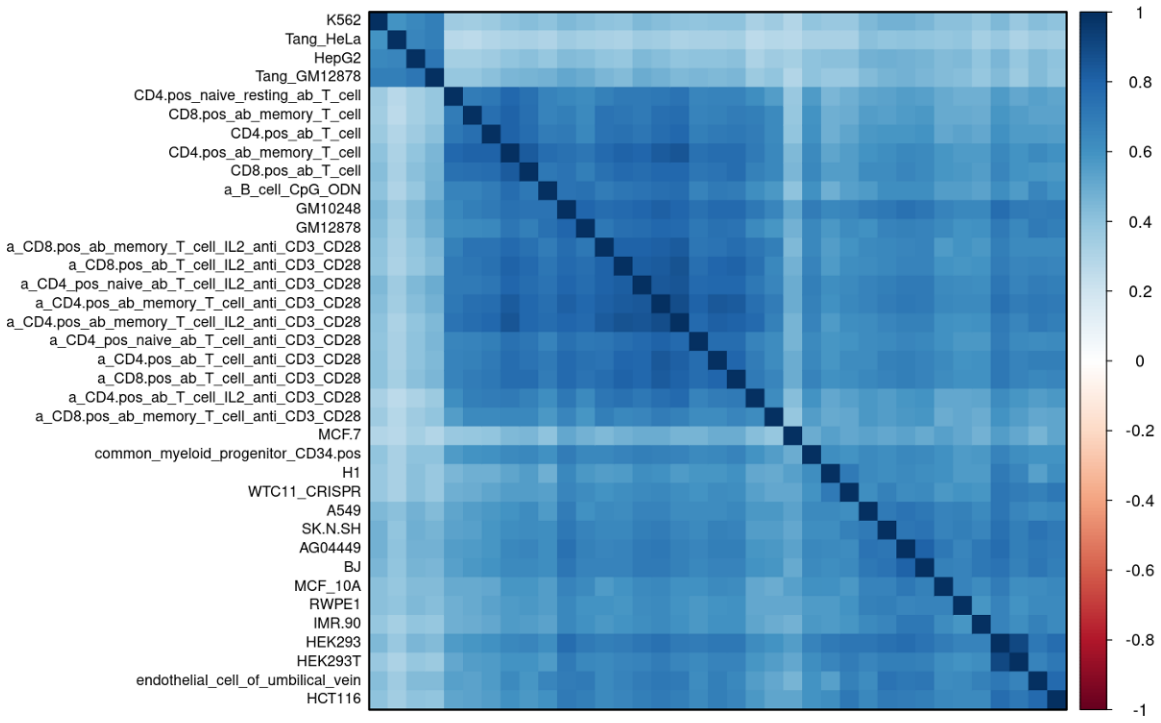

Figure 5-2. Clustering of CTCF ChIA-PET data
Hierarchical clustering of 37 ChIA-PET experiments at CTCF loop level. PET counts of CTCF binding site pairs are taken as input. All loops with distance less than 3kb are excluded. Replicates are merged before clustering.

immune cell lines that are over-represented excluded. Specifically, the results show that HepG2, K562 and the two old datasets GM12878 and HeLa from [87] are more predictive of each other than other experiments (Fig 5-4).

To understand the technical and biological reasons underlying this variation, we first asked about the ratio of PETs mapped to CTCF binding sites at both sides in each experiment. As CTCF ChIA-PET is designed to target at chromatin interactions involving CTCF binding, a higher CTCF PET/unique PET ratio would likely indicate better fragment extraction specificity. Interestingly, we found that this ratio is significantly higher in HepG2, K562, Tang_GM12878 and Tang_HeLa than other experiments. We then looked at the distribution of motif orientation for CTCF loops. Consistent with previous studies, convergent is the predominant orientation.

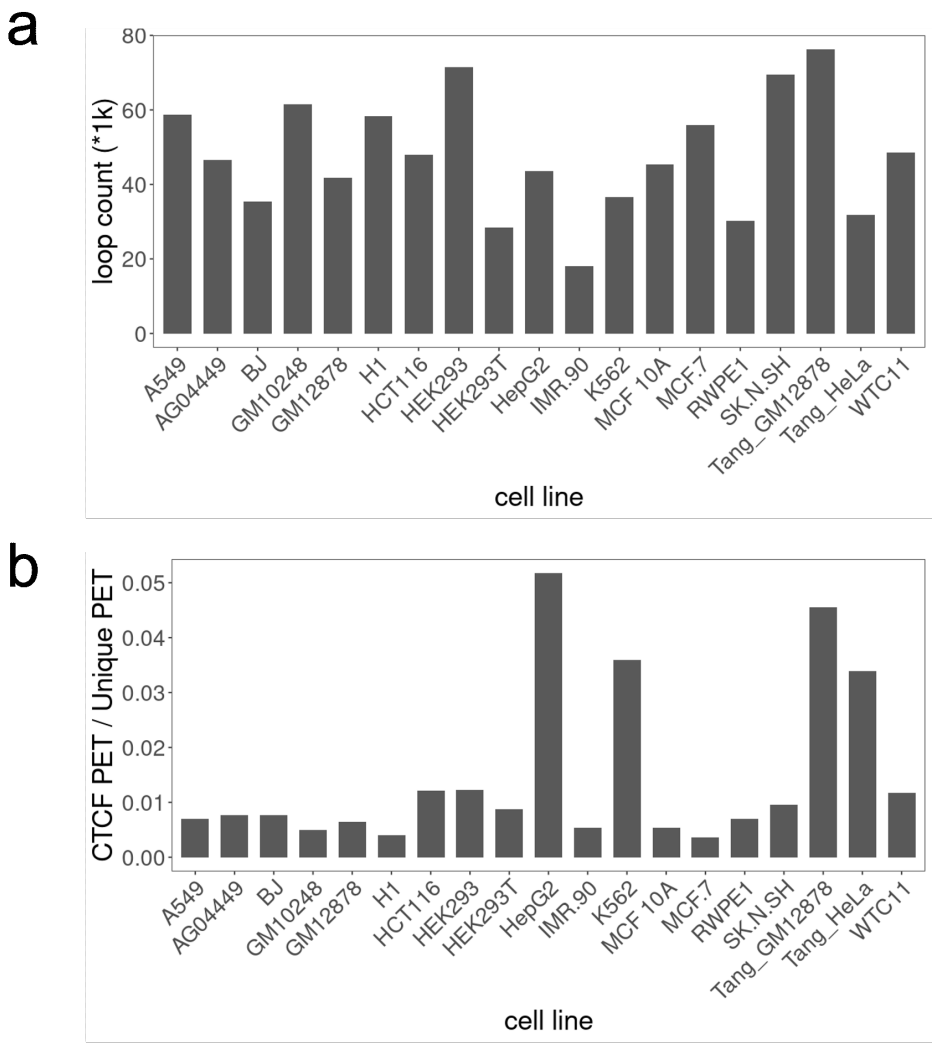

Figure 5-3. CTCF loop count and CTCF PET ratio across ChIA-PET experiments a) Number of CTCF loops detected in each ChIA-PET experiment. b) Ratio of CTCF PET and unique PET in each ChIA-PET experiment.

Convergent loops are usually supported by more PETs and of higher confidence. We found that the ratio between convergent and tandem loop was significantly higher in Tang_GM12878 and HepG2 at all distance intervals (the third is K562, Fig 5-5). This evidences suggest that HepG2, K562, Tang_GM12878 and Tang_HeLa data could be of higher quality.

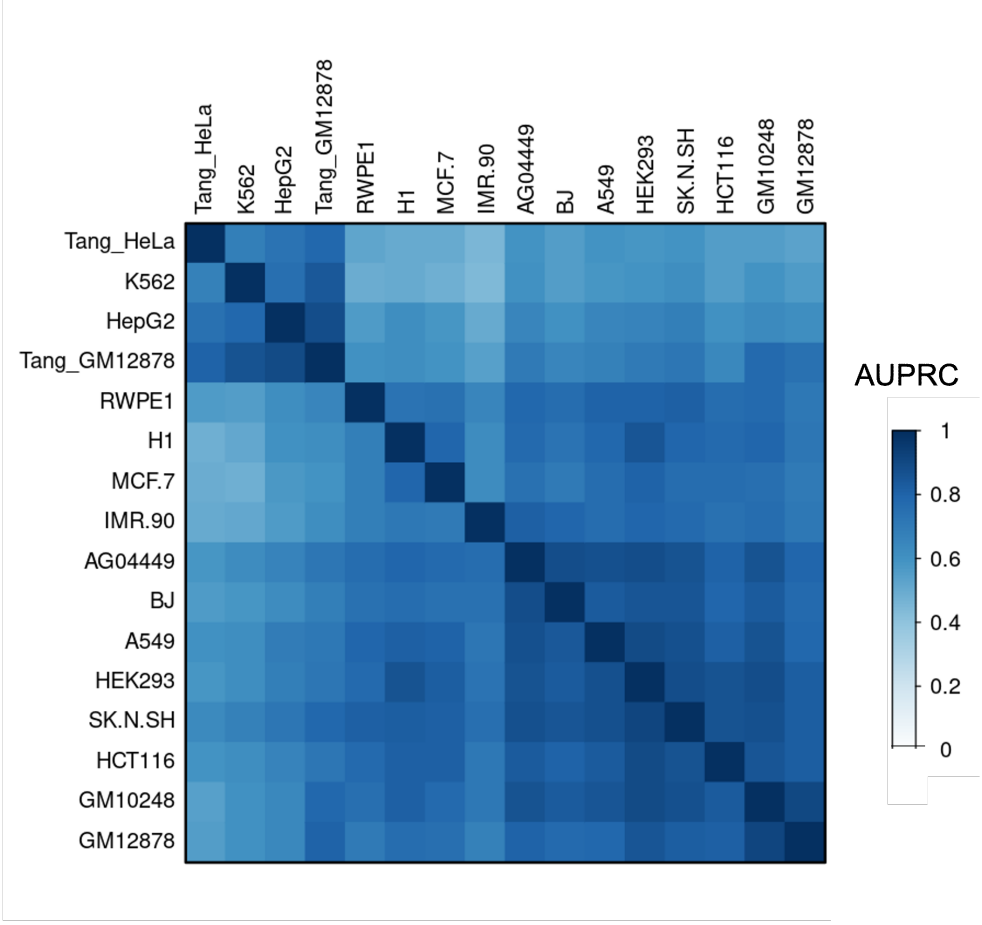

Figure 5-4. Cross cell-type prediction of CTCF loops
Using PET count of CTCF pairs in one cell types to predict CTCF loops in another cell type. Positive set is composed of top 20k CTCF loops with highest PET count. Negative set of 20 fold is sampled from CTCF pairs with PET count equal to 0.

5.3.2 CTCF loops can be accurately predicted from loop extrusion computational model across cell types

Next, we examined the mechanisms to predict this large collection of CTCF loops. Loop extrusion is the predominant hypothesis for CTCF loop formation, which states that Cohesin translocates in between a pair of convergently oriented CTCF motifs to reel flanking DNA and form a loop [10, 11]. Previously, we have developed a mathematical model of loop extrusion to understand CTCF interaction [6]. This model takes CTCF ChIP-seq data as input and computes CTCF binding intensity, CTCF motif orientation, genomic distance and loop competition. The output of

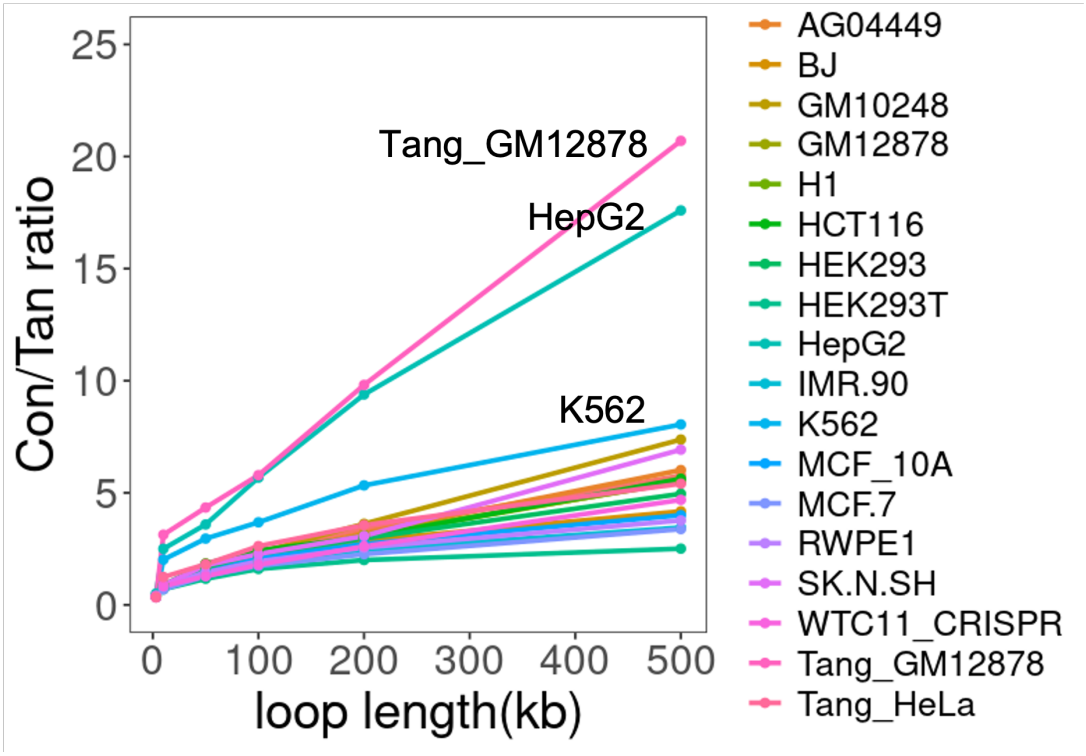

Figure 5-5. Distribution of loop orientation by distance
Ratio of convergent CTCF loop to tandem CTCF loop numbers across experiments and distance intervals.

this model is the interaction probability of all pairs of CTCF binding sites within 1Mb. Testing this model on ENCODE 4 CTCF ChIA-PET datasets will allows us to evaluate how consistent the CTCF loops experimentally detected are with the hypothesis of loop extrusion. Using parameters optimized from our previous work, we found this mathematical model is capable of predicting CTCF loops from all cell types and experiments at high accuracy (Fig 5-6). Again, Tang_GM12878 and HepG2 are the two most predictable experiments by this model.

Although the performance of the original loop extrusion mathematical model is comparable to other much more complicated machine learning methods, it has not fully explain specificity of CTCF interactions yet. One of the challenge is to interpret the mechanism underlying loops with non-convergent CTCF motif orientations. Specifically, parameter w in our model controls the contribution of motif orientation

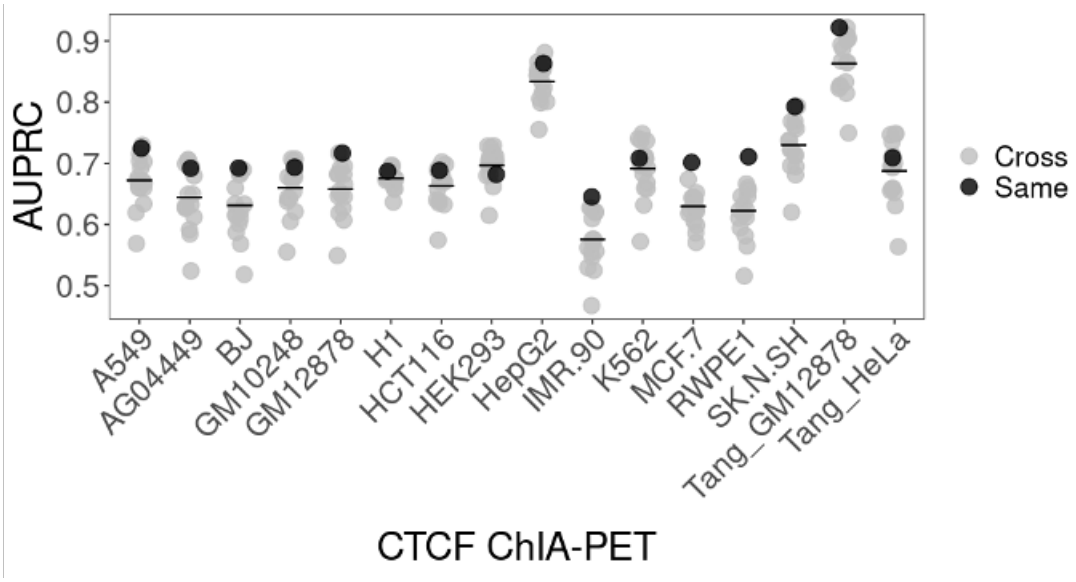

Figure 5-6. Loop extursion mathematical model predict CTCF loops identified by ChIA-PET

Applying the loop extrusion mathematical to predict CTCF loops by CTCF ChIP-seq. Cross: ChIP-seq and ChIA-PET data from different cell lines; Same: ChIP-seq and ChIA-PET data from the same cell line.

to CTCF looping probability. It represents the extent that the convergent CTCF motif orientation is preferred over tandem and divergent orientations. In the original model, w was a fixed scalar ranging from 2 to 4. However, we found that the portion of convergent loops increases dramatically with distance (Fig 5-7).

This phenomenon suggests that CTCF loops could emerge stochastically due to linear proximity at short distances (<10kb), in which different CTCF motif orientation does not contribute to stability of those loops. At longer distances, as the chance of random collision decreases, Cohesin-mediated loop extrusion becomes the dominant mechanism for loop formation. Therefore, to characterize the two orthogonal mechanisms of CTCF interaction, We introduce additional free parameters to scale w with distance (Methods, Fig 5-8). Our model indeed yields more accurate predictions of CTCF loops after this adjustment (Fig 5-9).

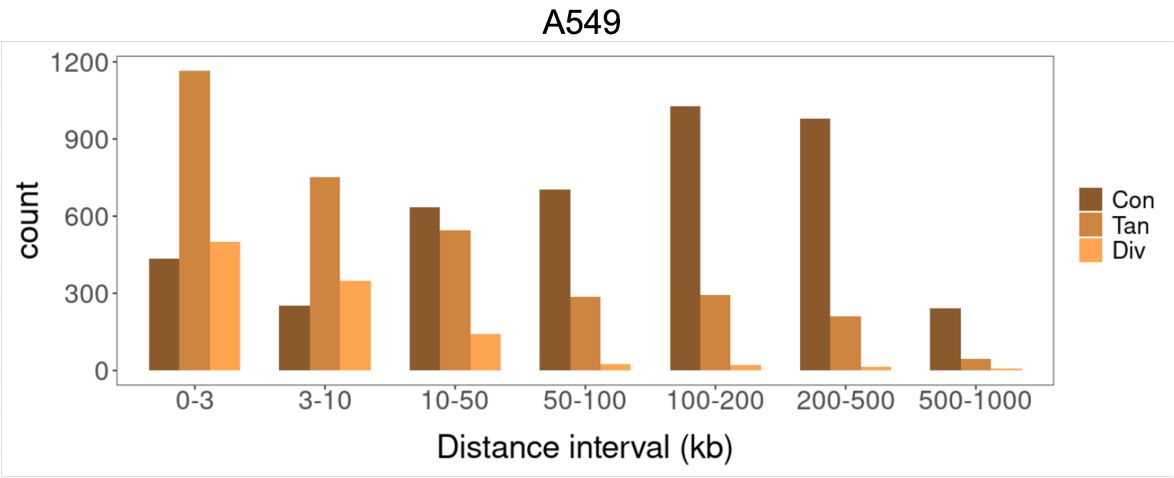

Figure 5-7. Distribution of loop orientation by distance in A549
Number of CTCF loops in convergent, tandem and divergent orientation at different distance intervals in A549.

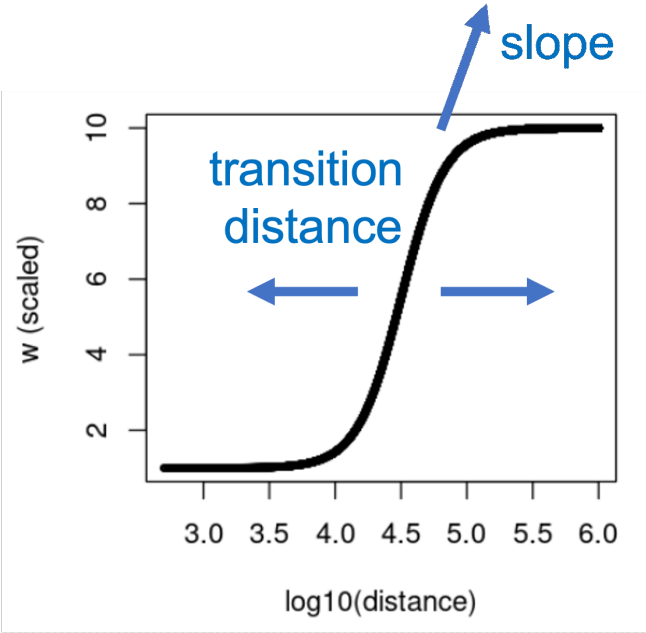

Figure 5-8. Scaling *w* by distance
w increases with distance, following a sigmoid function. This scaled *w* is determinted by parameters slope, transition distance and w_max.

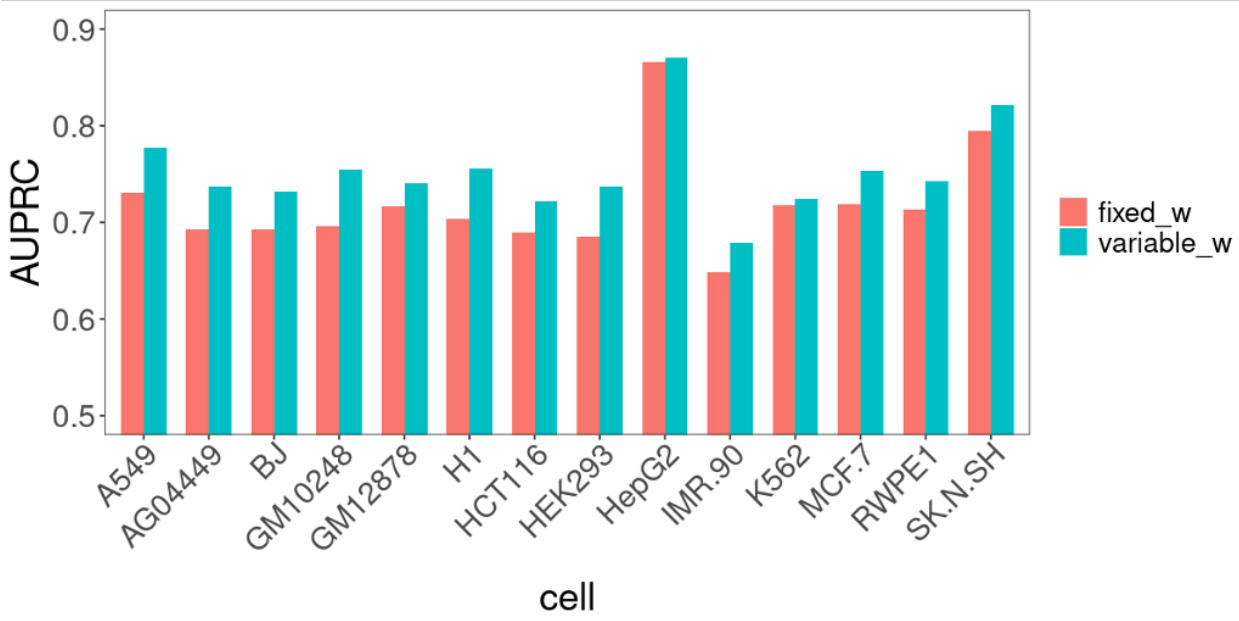

Figure 5-9. Scaled *w* improves prediction accuracy of the loop extrusion mathematical model
After using grid search to identify optimal value for parameters slope, transition distance and w_max, the model with scale *w* improved over the original model for all cell lines.

5.3.3 CTCF contact domain facilitate interpretation of CTCF loops across cell types

Overlapping and nested loops are prevalent in the genome. On average, each base pair in the human genome is covered by 5 CTCF loops(estimated by a total of 50k loops, with average length 300kb, Fig 5-10). Interpreting the pattern and functional impact of these loops remains challenging. To overcome that, we adopted the idea of topologically associated domain (TAD), as well as previous analysis of ChIA-PET data, to build a computational model to define CTCF contact domains across the genome (Methods). We first computed CTCF loop coverage for each base by summing up the PET count of all CTCF loops that cross it, to generate a 1D loop coverage track. We then applied a hidden markov model with four hidden states on this track to segregate the genome into domains with high or low CTCF interaction frequency. We call the domains with highest CTCF interaction frequency as CTCF contact domains(CCD).

In GM12878, there are a total of 2,764 CTCF contact domains in the genome, with medium size 319kb.

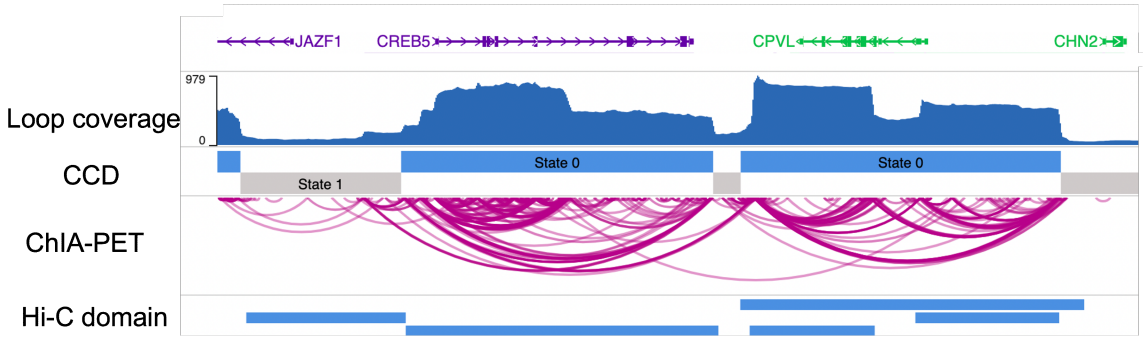

Figure 5-10. CTCF loops and CTCF contact domains in CREB5 locus
Densely connected CTCF binding sites and nested CTCF loop motivate the development of CTCF contact domains (CCD). CTCF loop coverage, CCD, CTCF loops from ChIA-PET and Hi-C domain are show for CREB5 locus in GM12878.

Most CTCF contact domains are stable across cell types (for example, IMR90 shares over 77% CCD with GM12878). CCD boundaries are also frequently coinciding with TAD boundaries identified from [9] (Fig 5-11). One thing to notice is that CCD may not be a physical structural unit that is present in cells; it is more likely to represent a genomic region enriched with higher CTCF contact frequency. Strikingly, transcription start sites(TSS) are strongly enriched at or inside CCD boundaries, suggesting their role in regulation gene expression (Fig 5-12). Enhancers are also preferentially located nearby CCD boundaries. Consistent with that, we randomly sampled pairs of genomic regions inside or outside CCDs, and found those inside CCDs have significantly higher contact frequency as measured by Hi-C (Fig 5-13).

5.3.4 Evaluating CTCF looping constraint on Pol ll ChIA-PET data

It has been discussed for a long time that one of the major roles of CTCF loops is to organize the chromatin architecture into smaller domains and regulate enhancer-promoter interactions. However, it has been difficult to test this hypothesis systemically,

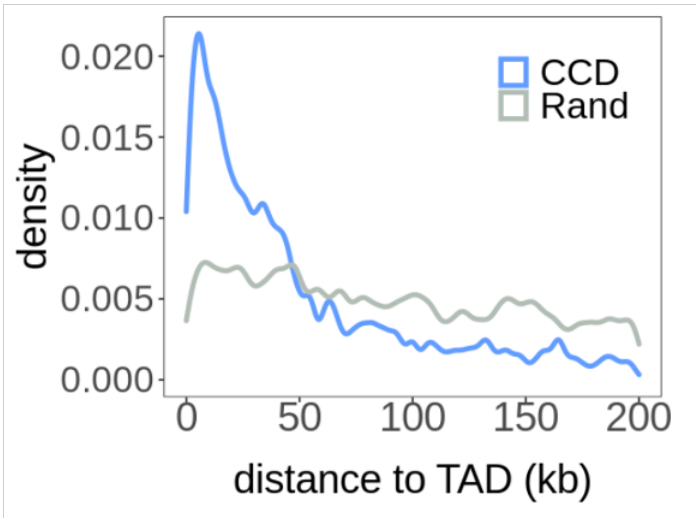

Figure 5-11. Analysis of CCD and TAD boundaries
Boundaries of CCD are enriched nearby TAD boundaries. The majority of CCD boundaries falls within 50kb of closet TAD boundaries.

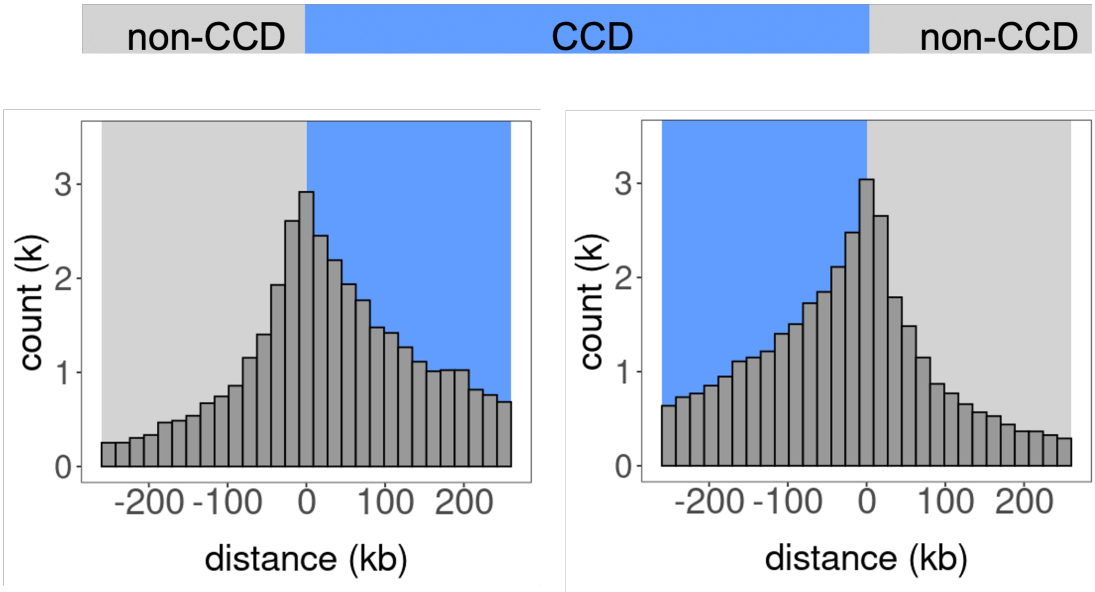

Figure 5-12. Transcription start sites are enriched inside CCD or nearby CCD boundaries
Histogram of distances of transcription start sites (TSS) to nearest CCD boundaries. This plot is aggregated for all gene TSS from gencode V24 and their nearest CCDs. Genes of distance within +/- 250kb to left or right CCD boundaries is shown. Blue: CTCF contact domains; Grey: Non-CTCF contact domains.

mostly due to limited capability in identifying either CTCF loops or EP interactions. Recent development in chromatin conformation capture (3C) assays like Hi-C, Micro-C

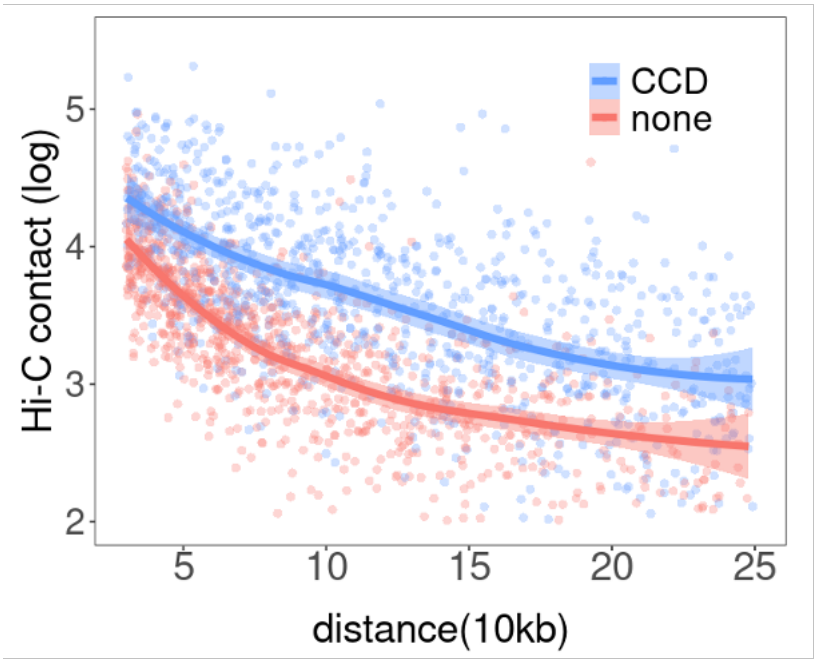

Figure 5-13. Hi-C contact frequencies inside or outside CCD
Scatter plot of Hi-C contact frequency for pairs of genomic regions inside(blue) or outside(red) CCDs.

and ChIA-PET [3, 31, 72], as well as functional characterization assays like CRISPRi provides the opportunity to overcome this challenge [107, 135]. In general, 3C assays allows genome wide detection of chromatin interactions that are either generic or associated with specific protein factors, while CRISPR perturbation is able to annotate distal enhancers for a limited number of genes at higher confidence level. So we sought to include both data types in our analysis to validate the hypothesis of CTCF constraint of EP interaction.

In principle, chromatin interactions can be stochastic and dynamic in single cells. As it is difficult to trace those contacts at high-throughput by current techniques, we focused on the statistically averaged behavior of interactions across time and cells, which can be measured by ChIA-PET. We reasoned that 3D chromatin conformation established by CTCF loops can impact contact frequency of neighboring genomic regions. Specifically, a CTCF loop spanning an EP pair is going to bring them into proximity and facilitate EP interaction. A CTCF loop inside or across an EP pair is

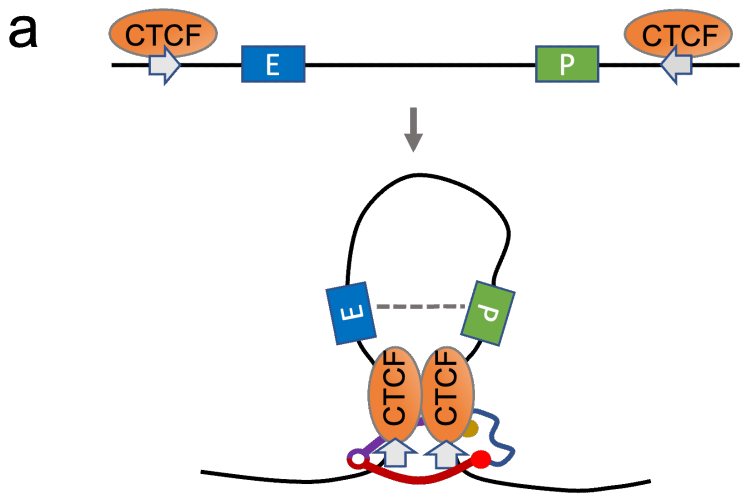

Figure 5-14. Two scenarios of CTCF loops constraining enhancer-promoter interactions

Conceptual frameworks of how CTCF loops constrain interaction between enhancer and promoters. a) For EP pairs inside CTCF loops, they are more likely to interact because CTCF loops bring the enhancers and promoters into proximity. b) For EP pairs crossed by CTCF loops, they are less likely to interact due to spatial hindrance of CTCF and Cohesin protein.

going to create spatial hurdles and prohibit EP interaction (Fig 5-14). The interacting EP pairs are derived from Pol ll ChIA-PET data also generated during ENCODE4. Loops are detected through standard ChIA-PET data processing (Methods 5.2.4). A loop is preserved as an interacting EP pair in downstream analysis if at least one of its anchors intersect with a gene TSS. The number of EP pairs is highly variable across cell types, ranging from 1,500 to 20,000, partially due to inconsistent data quality. However, this still represents one of the most comprehensive lists of EP interactions across the genome. We found the number of interactions associated with promoters is correlated with gene expression at 0.40, which is reasonable, considering promoter activity itself correlates with gene expression at around 0.6 0.65. The number of PET counts is a bit more predictive of gene expression (correlation = 0.42).

To examine the relationship between CTCF loop and EP interactions, we first showed that EP interactions are enriched in strongest CCDs by 2 fold, and depleted in weakest CCD (or non-CCD) by 3 fold (Fig 5-15A). Similarly, interacting EP pairs tend to cross significantly fewer CCD boundaries than non-interacting CCD pairs (Fig 5-15B). These results are largely consistent with existing knowledge. To further investigate the contribution of CTCF loops to gene regulation in a quantitative manner, we sought to use CTCF information to classify interacting versus non-interaction EP pairs. For each EP pair, the number of CTCF loops contain, inside and cross them is counted. As chromatin contact frequency is known to sharply decay with genomic distance [43], EP pairs with variable distance are not comparable to each other, so we split them into 5 distance intervals to mitigate this effect (Fig 5-16). At short distance, none of the three features is predictive, as short range interactions are not heavily affected by chromatin organization. At longer distances, as expected, being contained by CTCF loops increased the chance of EP interaction significantly, while having CTCF loops inside or across decreased this chance accordingly. Overall, the longer the distance is, the stronger the impact of CTCF looping constraints are. The number

of CTCF loops are correlated with distance, but distance is not predictive within each interval, indicating CTCF loops are regulating EP interactions independent of genomic distance. Furthermore, integrating CTCF looping information with cell-type specific chromatin accessibility at enhancer and promoter (Method 5.2.8) also improves prediction accuracy at large distances by over two-fold, suggesting a crucial role of 3D chromatin organization in regulating gene expression (Fig 5-16B).

5.3.5 Evaluating CTCF looping constraint on CRISPRi data

We next moved to functional characterization datasets to evaluate this hypothesis. Specifically, CRISPR perturbation has been widely-applied for studying distal regulatory mechanisms of gene expression. We selected two state-of-the-art datasets known as CRISPRi-FlowFISH [107] and HCR-FlowFISH [135], which tested inactivation of hundreds distal regulatory elements on dozens of genes in K562 cell lines. The authors of CRISPRi-FlowFISH paper also proposed a simple and powerful activity by contact(ABC) model [107], using local chromatin activity measured by DNase-seq and H3K27ac ChIP-seq, as well as 3D contact frequency between enhancers and promoters measured by Hi-C, to explain and predict enhancer-gene links identified from their experiment. Given the success of their model, we sought to understand the driving mechanism of EP contact. As CTCF looping has been shown to be important from our analysis above, we developed a model of enhancer activity and CTCF looping constraint, which we called the CTCF-loop Constrained Inter-Action (CIA) model. This model predicts an interaction score for a given EP pair based on the product of enhancer activity (A) and CTCF constraint (CC) in the corresponding cell type, normalized by promoter CTCF constraint (Fig 5-17).

The CTCF constraint is the difference between PET count of CTCF loops containing the EP pair and PET count of CTCF loops crossing or inside the EP pair. The underlying hypothesis is that differences in EP contact frequency is due to chro-

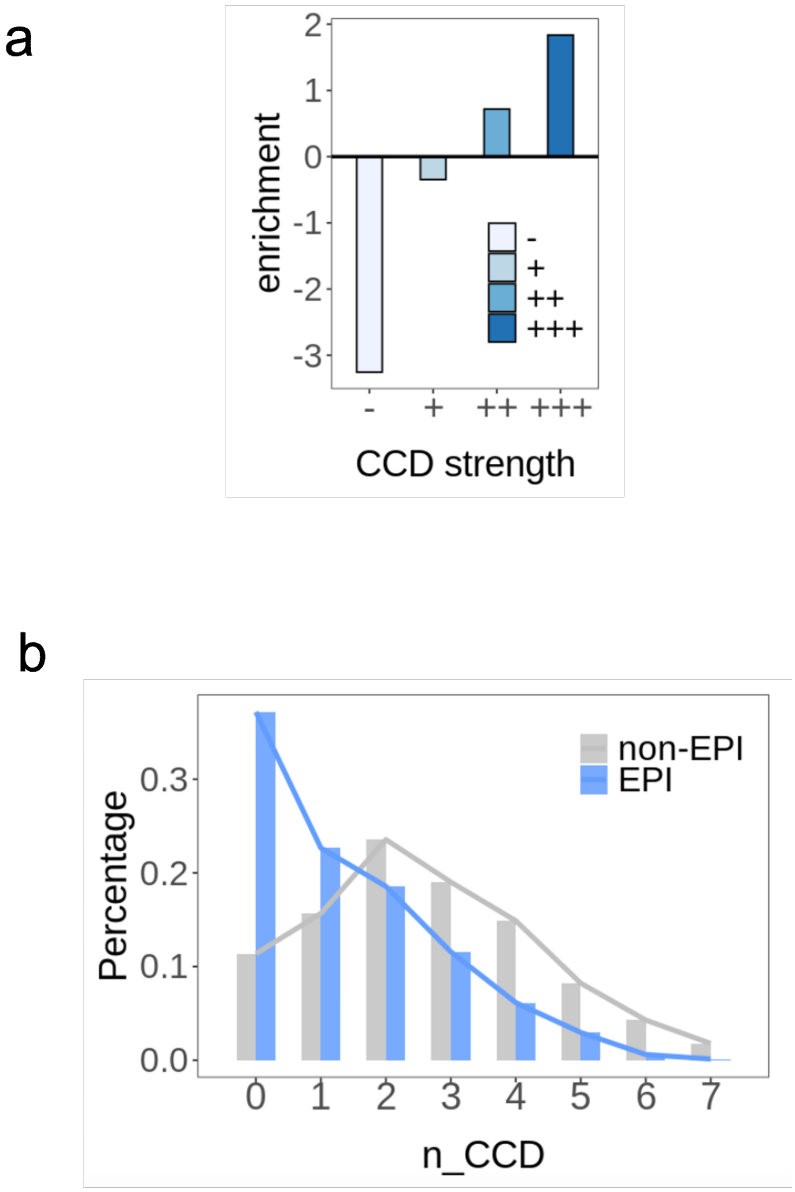

Figure 5-15. Enhancer-promoter interaction detected by Pol II ChIA-PET are constrained by CCD

a) Histogram of EP interaction enrichment in CTCF contact domains at different strength. CCDs are generated by running HMM model with 4 hidden state on CTCF ChIA-PET data. They represent genomic regions with CTCF interaction frequency from low to high level. EP pairs are identified from Pol II ChIA-PET of the same cell line. b) Histogram of number of CCD boundaries crossed by EP interaction and non-interacting EP pairs.

matin conformation established by CTCF loops, and can be quantitatively inferred by combining PET counts of CTCF ChIA-PET measurements.

We then applied the CIA model, the ABC model, as well as activity combined with

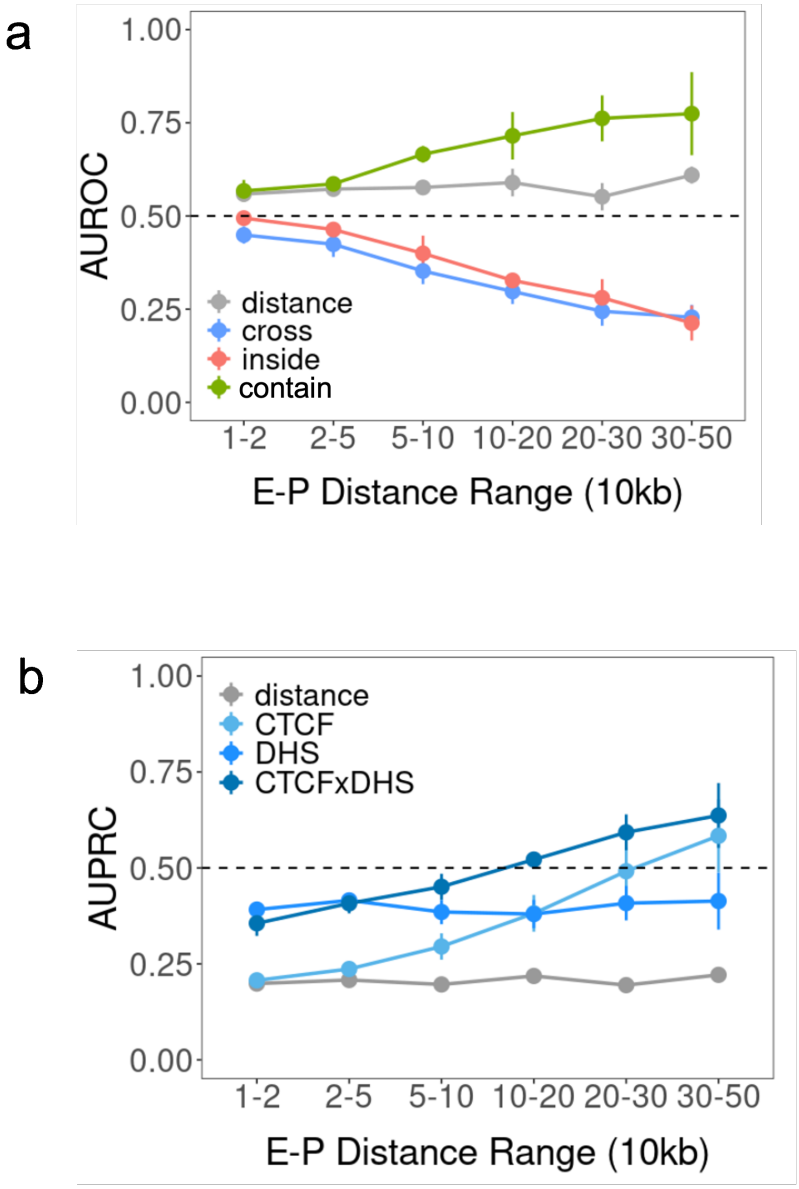

Figure 5-16. CTCF looping information are predictive of enhancer-promoter interaction detected by Pol II ChIA-PET

a) Classifying interacting and non-interacting EP pairs with CTCF looping constraint. Features includes probability of CTCF loops form outside, inside and cross EP pairs, as well as distance between EP pairs. Prediction accuracy are shown as AUROC by distance intervals. b) Integrating local activity with CTCF constraint improves prediction of EP interaction. DHS: the product of DNase-seq signal of enhancer and promoter. CTCF: see definition of CTCF Constraint (CC) below.

linear distance, to predict EP interaction hits from CRISPRi-FlowFISH and HCR-FlowFISH. In general, the prediction results from CIA and ABC are well-correlated,

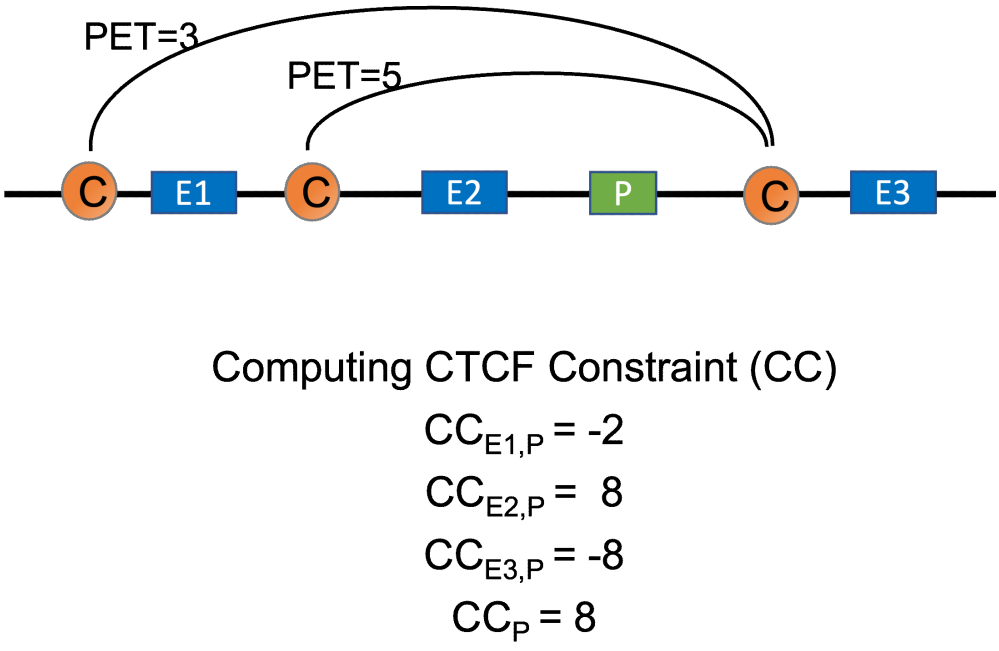

Figure 5-17. CTCF-loop Constrained Inter-Action (CIA) model for predicting enhancer-promoter interactions
An interaction score of an enhancer-promoter pair is calculated by the product of enhancer activity and difference of CTCF loops span this pair and CTCF loops cross this pair. For example, there are two CTCF loops span enhancer 2 and the promoter, and no CTCF loops cross them therefore total CTCF loop PET count for P–E2 ($CC_{E2,P}$) is 8. $CC_{E2,P}$ is then normalized by total CTCF loop PET count at the promoter(CC_P+1). C: CTCF binding sites; E1-3: enhancers; P: gene promoter. Arcs: CTCF loops with PET count indicated on top.

as the activity feature used by them is shared (Fig 5-18, Fig 5-19). On the HCR-FlowFISH dataset, the K562 CIA model outperforms ABC model remarkably in the vast majority of genes, and they both outperformed activity-over-distance model (Fig 5-18). Interestingly, the 3D contact features alone - CTCF loop PET count, Hi-C contact and distance, have similar accuracy, which means that CCA model gained more accuracy by including enhancer activity. Replacing the K562 CTCF ChIA-PET by other cell types results in slight degeneracy of CIA model performance.

The FADS gene loci is shown as an example in Fig 5-20. In this region, HCR-FlowFISH clearly detects promoter and distal hits for all three genes, FADS1, FADS2 and FADS3, which are all contained inside strong CTCF loops. However, the ABC

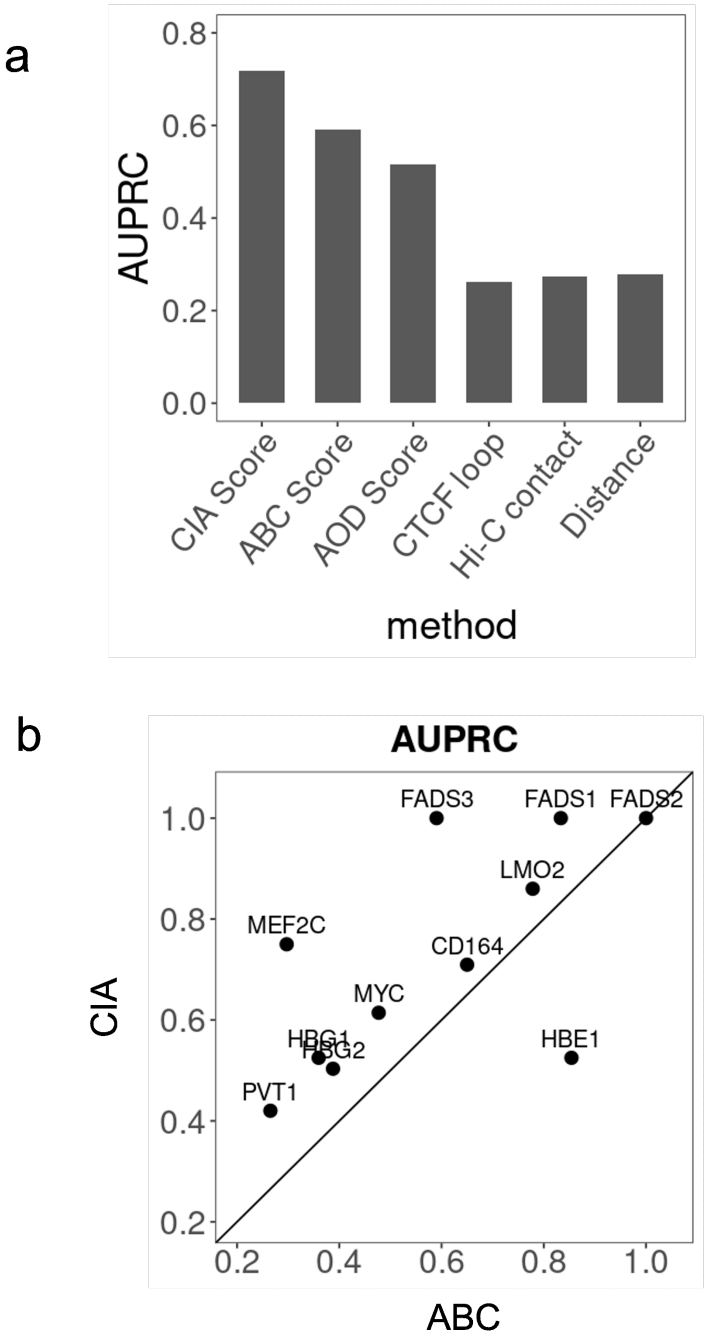

Figure 5-18. CIA model evaluation on HCR-FlowFISH dataset
Comparison of classifiers on enhancer-promoter pairs identified by HCR-FlowFISH. Positive interactions are those for which perturbation of distal element significantly changes expression of the gene. Y axis is AUPRC of predictions averaged across all genes. ABC: acitivity-by-contact model. AOD: activity-over-distance model.

score of DHS1 is higher than the ABC score of DHS5 for FADS1 and FADS2, thus making a false positive prediction for these target genes. This is due to DHS1 having

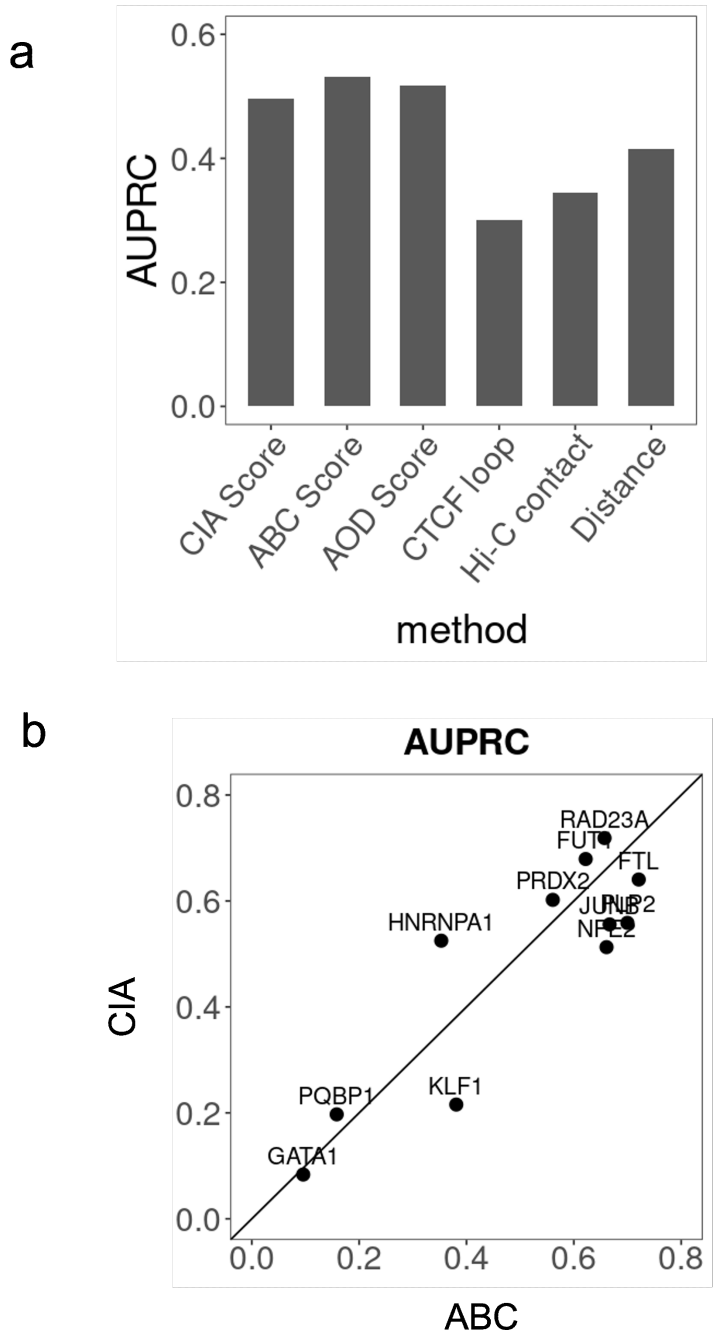

Figure 5-19. CIA model evaluation on HCR-FlowFISH dataset
Comparison of classifiers on enhancer-promoter pairs identified by CRISPRi-FlowFISH. Positive interactions are those for which perturbation of distal element significantly changes expression of the gene. Y axis is AUPRC of predictions averaged across all genes. ABC: acitivity-by-contact model. AOD: activity-over-distance model.

strong enhancer activity and medium Hi-C contact frequency with their gene promoters.

When applying the CIA model on this region, all true positive hits and their target

genes are contained by the CTCF loops, with peak 1 excluded. The CTCF Constraint score for DHS1-FADS1 ($CC_{E1,P}$) is therefore much smaller than other DHS1. On CRISPRi-FlowFISH data, the performance of the three methods are close to each other (Fig 5-19), indicating heterogeneity across CRISPRi datasets.

FADS1	1	3	4	5
Activity	24	19	15	10
Hi-C	11	32	28	24
ABC	2.79	6.18	4.19	2.48
$CC_{E,P}$	-18	18	18	14

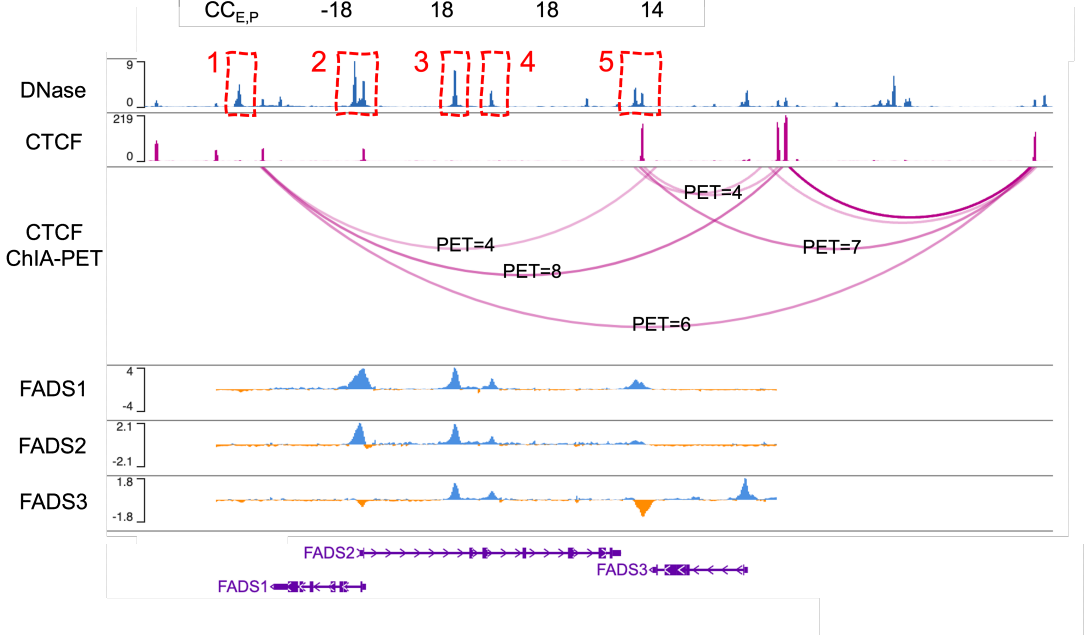

Figure 5-20. Comparison of CIA and ABC model on FADS loci
Comparison of CIA and ABC model in FADS loci. Prediction score for both models, as well as features including activity, normalized Hi-C contact and normalized PET count are shown for candadate DNase-seq hypersentivity sites (DHS) in K562. In HCR-FlowFISH, DHS1 is not a distal hit for FADS1 gene but is a false positive prediction made by the ABC model. CIA model successfully distinguished DHS1 from the 3 true positive hits (DHS2-4).

5.4 Discussion

Our work reveals inconsistency of CTCF ChIA-PET data quality across cell-types, and raised several quality metrics including ratio of CTCF PET to unique PET, proportion of convergent loops and predictability by the loop extrusion mathematical model. As CTCF looping constraints have been shown to be critical in transcriptional

regulation, this work also motivates developing more accurate assays for both indiscriminately probing all chromatin interactions, such as Hi-C, Micro-C, and specifically mapping CTCF interactions, such as ChIA-PET and Hi-ChIP [32]. We updated the useful concept of the CTCF contact domain (CCD) to understand CTCF contact strength across the genome. The definition of CCD makes them more homogenous in terms of composition than TADs. On the other hand, recent work using single-molecule imaging has demonstrated that many chromatin interactions, including CTCF loops, are highly transient. The dynamic nature of these loops has not been fully incorporated into the current CIA model yet. As some strong CTCF loops were estimated to exist in only around 5% of time [136], the exact mechanisms of how they regulate gene expression remain to be investigated.

The success of our CIA model argues that the impact of 3D contact on gene expression can largely be predicted from CTCF interactions. It is consistent with the role of CTCF as the most prominent insulator in mammalian genomes. In many cases, TAD boundaries and loop anchors are formed by clusters of CTCF binding sites. How insulation level is affected by multiple CTCF loops together is an interesting open question. In the CIA model, we assume that CTCF binding sites and loops nearby are independent of each other, as the contributions of different loops to CIA score are additive. In theory, the effect could also be synergic (super-additive) or complementary (sub-additive), which are both supported by recent works studying specific loci [116, 133]. We reason that to obtain a more general rule of CTCF insulation effect on gene expression, it is necessary to systematically place CTCF binding motifs of various numbers, strengths, orientations at different positions around enhancers and genes, and measure the transcriptional outcomes [137]. This strategy can further inform quantitative prediction of gene expression dysregulation in diseases from chromatin conformation disruption.

Conclusions and general discussion

Rich genomic and epigenomic datasets accumulated over the past 20 years have allowed the identification of regulatory elements across cell types and tissues expansively. However, our ability to predict phenotypes and disease from genetic information remains limited. To elucidate the mechanisms of transcriptional changes in cell differentiation and disease progression, much attention has thus been drawn to understand the interactions of enhancer and promoters. Several 3C-based experimental approaches have been modified to target at enhancer-promoter (EP) interactions, such as promoter capture Hi-C (PC-HiC), Pol ll ChIA-PET and H3K27ac Hi-ChIP. However, these methods usually have low power and high false positive rate due to various technical issues. On the other hand, genetic perturbation approaches like CRISPR interference are able to identify a subset of EP interactions with high confidence, but may fail to detect more transient or dynamic interactions genome-wide. The development of computational methods for predicting EP interactions is therefore limited by a lack of good ground truth datasets. We envision that this challenge will be solved by continuous evolution of proximity-based 3D genome mapping techniques that detect chromatin interactions at higher resolution, or massive parallel CRISPRi systems that characterize regulatory effects for larger amounts of enhancers.

The ultimate goal of understanding EP interactions is to be able to predict transcriptional changes in specific contexts. Currently, very few studies have focused on the quantitative relationship between chromatin contact frequency and gene expression. A recent work has revealed a nonlinear relationship at a specific loci with only one

pair of enhancer and promoter [137]. This work reported that gene expression can be sensitive to small interaction frequency changes. Conceptually, this system can be made more complicated by adding additional elements like competing enhancers and CTCF binding sites to closely simulate the genomic environment of disease risk loci. Such studies can provide necessary resources to refine mechanistic models that predict gene expression from epigenetic features.

References

1. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* **7,** 29–59 (2006).
2. Ong, C.-T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics* **12,** 283–293 (2011).
3. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. en. *Science* **326.** Publisher: American Association for the Advancement of Science Section: Report, 289–293 (Oct. 2009).
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. en. *Genome Research* **19.** Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab, 521–532 (Apr. 2009).
5. Xi, W. & Beer, M. A. Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy. *PLoS Computational Biology* **14** (Dec. 2018).
6. Xi, W. & Beer, M. A. Loop competition and extrusion model predicts CTCF interaction specificity. *Nature communications* **12,** 1–15 (2021).
7. Rowley, M. J. & Corces, V. G. Organizational principles of 3D genome architecture. en. *Nature Reviews Genetics* **19.** Number: 12 Publisher: Nature Publishing Group, 789–800 (Dec. 2018).
8. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. en. *Nature* **485.** Number: 7398 Publisher: Nature Publishing Group, 376–380 (May 2012).
9. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. en. *Cell* **159,** 1665–1680 (Dec. 2014).
10. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. en. *Cell Reports* **15,** 2038–2049 (May 2016).
11. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. en. *Proceedings of the National Academy of Sciences* **112,** E6456–E6465 (Nov. 2015).
12. Davidson, I. F. *et al.* DNA loop extrusion by human cohesin. en. *Science* **366.** Publisher: American Association for the Advancement of Science Section: Research Article, 1338–1345 (Dec. 2019).

13. Kim, Y., Shi, Z., Zhang, H., Finkelstein, I. J. & Yu, H. Human cohesin compacts DNA by loop extrusion. en. *Science* **366.** Publisher: American Association for the Advancement of Science Section: Report, 1345–1349 (Dec. 2019).

14. Deng, W. *et al.* Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* **149,** 1233–1244 (2012).

15. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human molecular genetics* **12,** 1725–1735 (2003).

16. Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nature Reviews Genetics* **20,** 437–455 (2019).

17. Amano, T. *et al.* Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. *Developmental cell* **16,** 47–57 (2009).

18. Gröschel, S. *et al.* A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia. *Cell* **157,** 369–381 (2014).

19. Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538,** 265–269 (2016).

20. Lupiáñez, D. G. *et al.* Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. en. *Cell* **161,** 1012–1025 (May 2015).

21. Redin, C. *et al.* The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. *Nature genetics* **49,** 36–45 (2017).

22. Xing, M. *et al.* Genomic and epigenomic EBF1 alterations modulate TERT expression in gastric cancer. *The Journal of clinical investigation* **130,** 3005–3020 (2020).

23. Sheng, T. *et al.* Integrative epigenomic and high-throughput functional enhancer profiling reveals determinants of enhancer heterogeneity in gastric cancer. *Genome medicine* **13,** 1–25 (2021).

24. Ho, S. W. T. *et al.* Regulatory enhancer profiling of mesenchymal-type gastric cancer reveals subtype-specific epigenomic landscapes and targetable vulnerabilities. *Gut* (2022).

25. Krijger, P. H. L. & De Laat, W. Regulation of disease-associated gene expression in the 3D genome. *Nature reviews Molecular cell biology* **17,** 771–782 (2016).

26. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. en. *Nature Reviews Genetics* **19.** Number: 7 Publisher: Nature Publishing Group, 453–467 (July 2018).

27. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *science* **295,** 1306–1311 (2002).

28. Schmitt, A. D., Hu, M. & Ren, B. Genome-wide mapping and analysis of chromosome architecture. *Nature reviews Molecular cell biology* **17,** 743–755 (2016).

29. Simonis, M. *et al.* Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). *Nature genetics* **38,** 1348–1354 (2006).

30. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome research* **16,** 1299–1309 (2006).

31. Fullwood, M. J. *et al.* An oestrogen-receptor--bound human chromatin interactome. en. *Nature* **462.** Number: 7269 Publisher: Nature Publishing Group, 58–64 (Nov. 2009).

32. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature methods* **13,** 919–922 (2016).

33. Yu, M., Juric, I., Abnousi, A., Hu, M. & Ren, B. in *Enhancers and Promoters* 181–199 (Springer, 2021).

34. Gall, J. G. & Pardue, M. L. Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *Proceedings of the National Academy of Sciences* **63,** 378–383 (1969).

35. Barutcu, A. R., Maass, P. G., Lewandowski, J. P., Weiner, C. L. & Rinn, J. L. A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. *Nature communications* **9,** 1–11 (2018).

36. Luperchio, T. R. *et al.* Chromosome conformation paints reveal the role of lamina association in genome organization and regulation. *BioRxiv,* 122226 (2017).

37. Beagrie, R. A. *et al.* Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543,** 519–524 (2017).

38. Quinodoz, S. A. *et al.* Higher-order inter-chromosomal hubs shape 3D genome organization in the nucleus. *Cell* **174,** 744–757 (2018).

39. Zheng, M. *et al.* Multiplex chromatin interactions with single-molecule precision. *Nature* **566,** 558–562 (2019).

40. Kempfer, R. & Pombo, A. Methods for mapping 3D chromosome architecture. *Nature Reviews Genetics* **21,** 207–226 (2020).

41. De Gennes, P.-G. & Gennes, P.-G. *Scaling concepts in polymer physics* (Cornell university press, 1979).

42. Bates, F. S. & Fredrickson, G. H. Block copolymer thermodynamics: theory and experiment. *Annual review of physical chemistry* **41,** 525–557 (1990).

43. Mirny, L. A. The fractal globule as a model of chromatin architecture in the cell. *Chromosome research* **19,** 37–51 (2011).

44. Barbieri, M. *et al.* Complexity of chromatin folding is captured by the strings and binders switch model. en. *Proceedings of the National Academy of Sciences* **109.** Publisher: National Academy of Sciences Section: Biological Sciences, 16173–16178 (Oct. 2012).

45. Bianco, S. *et al.* Polymer physics predicts the effects of structural variants on chromatin architecture. en. *Nature Genetics* **50.** Number: 5 Publisher: Nature Publishing Group, 662–667 (May 2018).

46. Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. & Mirny, L. A. Chromatin organization by an interplay of loop extrusion and compartmental segregation. en. *Proceedings of the National Academy of Sciences* **115.** Publisher: National Academy of Sciences Section: PNAS Plus, E6697–E6706 (July 2018).

47. Di Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G. & Onuchic, J. N. Transferable model for chromosome architecture. en. *Proceedings of the National Academy of Sciences* **113,** 12168–12173 (Oct. 2016).

48. Strom, A. R. *et al.* Phase separation drives heterochromatin domain formation. *Nature* **547,** 241–245 (2017).

49. Cao, F. & Fullwood, M. J. Inflated performance measures in enhancer–promoter interaction-prediction methods. en. *Nature Genetics* **51.** Number: 8 Publisher: Nature Publishing Group, 1196–1198 (Aug. 2019).

50. Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature genetics* **48,** 488–496 (2016).

51. Yang, Y., Zhang, R., Singh, S. & Ma, J. Exploiting sequence-based features for predicting enhancer–promoter interactions. *Bioinformatics* **33,** i252–i260 (2017).

52. Li, W., Wong, W. H. & Jiang, R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic acids research* **47,** e60–e60 (2019).

53. Zhang, R., Wang, Y., Yang, Y., Zhang, Y. & Ma, J. Predicting CTCF-mediated chromatin loops using CTCF-MP. *Bioinformatics* **34,** i133–i141 (2018).

54. Kai, Y. *et al.* Predicting CTCF-mediated chromatin interactions by integrating genomic and epigenomic features. en. *Nature Communications* **9.** Number: 1 Publisher: Nature Publishing Group, 4221 (Oct. 2018).

55. Belokopytova, P. S., Nuriddinov, M. A., Mozheiko, E. A., Fishman, D. & Fishman, V. Quantitative prediction of enhancer–promoter interactions. *Genome research* **30,** 72–84 (2020).

56. Zhang, S., Chasman, D., Knaack, S. & Roy, S. In silico prediction of high-resolution Hi-C interaction matrices. *Nature communications* **10,** 1–18 (2019).

57. Fudenberg, G., Kelley, D. R. & Pollard, K. S. Predicting 3D genome folding from DNA sequence with Akita. *Nature methods* **17,** 1111–1117 (2020).

58. Schwessinger, R. *et al.* DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nature methods* **17,** 1118–1124 (2020).

59. Cao, Q. *et al.* Reconstruction of enhancer–target networks in 935 samples of human primary cells, tissues and cell lines. *Nature genetics* **49,** 1428–1436 (2017).

60. Singh, S., Yang, Y., Póczos, B. & Ma, J. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quantitative Biology* **7,** 122–137 (2019).

61. Schreiber, J., Libbrecht, M., Bilmes, J. & Noble, W. S. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. *BioRxiv,* 103614 (2018).

62. Hsieh, T.-H. S. *et al.* Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. en. *Cell* **162,** 108–119 (July 2015).

63. Narendra, V. *et al.* CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. en. *Science* **347.** Publisher: American Association for the Advancement of Science Section: Report, 1017–1021 (Feb. 2015).

64. Phillips, J. E. & Corces, V. G. CTCF: Master Weaver of the Genome. en. *Cell* **137,** 1194–1211 (June 2009).

65. Stigler, J., Çamdere, G. Ö., Koshland, D. E. & Greene, E. C. Single-Molecule Imaging Reveals a Collapsed Conformational State for DNA-Bound Cohesin. en. *Cell Reports* **15,** 988–998 (May 2016).

66. De Wit, E. *et al.* CTCF Binding Polarity Determines Chromatin Looping. en. *Molecular Cell* **60,** 676–684 (Nov. 2015).

67. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. en. *Cell* **171,** 305–320.e24 (Oct. 2017).

68. Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. en. *Cell* **162,** 900–910 (Aug. 2015).

69. Conte, M. *et al.* Polymer physics indicates chromatin folding variability across single-cells results from state degeneracy in phase separation. en. *Nature Communications* **11.** Number: 1 Publisher: Nature Publishing Group, 3289 (July 2020).

70. Brackey, C. A., Marenduzzo, D. & Gilbert, N. Mechanistic modeling of chromatin folding to understand function. en. *Nature Methods* **17.** Number: 8 Publisher: Nature Publishing Group, 767–775 (Aug. 2020).

71. Kornberg, R. D. Chromatin Structure: A Repeating Unit of Histones and DNA: Chromatin structure is based on a repeating unit of eight histone molecules and about 200 DNA base pairs. *Science* **184,** 868–871 (1974).

72. Hsieh, T.-H. S. *et al.* Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. en. *Molecular Cell* **78,** 539–553.e8 (May 2020).

73. Wood, C. & Tonegawa, S. Diversity and joining segments of mouse immunoglobulin heavy chain genes are closely linked and in the same orientation: implications for the joining mechanism. *Proceedings of the National Academy of Sciences* **80,** 3030–3034 (1983).

74. Riggs, A. DNA methylation and late replication probably aid cell memory, and type I DNA reeling could aid chromosome folding and enhancer function. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **326,** 285–297 (1990).

75. Nasmyth, K. Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. *Annual review of genetics* **35,** 673 (2001).

76. Strunnikov, A. V., Larionov, V. L. & Koshland, D. SMC1: an essential yeast gene encoding a putative head-rod-tail protein is required for nuclear division and defines a new ubiquitous protein family. *The Journal of cell biology* **123,** 1635–1648 (1993).

77. Wendt, K. S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451,** 796–801 (2008).

78. Kurukuti, S. *et al.* CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proceedings of the national academy of sciences* **103,** 10684–10689 (2006).

79. Katz, D. J., Beer, M. A., Levorse, J. M. & Tilghman, S. M. Functional characterization of a novel Ku70/80 pause site at the H19/Igf2 imprinting control region. *Molecular and cellular biology* **25,** 3855–3863 (2005).

80. Nichols, M. H. & Corces, V. G. A CTCF code for 3D genome architecture. *Cell* **162,** 703–705 (2015).

81. Lengronne, A. *et al.* Cohesin relocation from sites of chromosomal loading to places of convergent transcription. *Nature* **430,** 573–578 (2004).

82. Busslinger, G. A. *et al.* Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. *Nature* **544,** 503–507 (2017).

83. Davidson, I. F. & Peters, J.-M. Genome folding through loop extrusion by SMC complexes. *Nature reviews Molecular cell biology* **22,** 445–464 (2021).

84. Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced Regulatory Sequence Prediction Using Gapped k-mer Features. en. *PLOS Computational Biology* **10.** Publisher: Public Library of Science, e1003711 (July 2014).

85. Granek, J. A. & Clarke, N. D. Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biology* **6,** R87 (Sept. 2005).

86. Kim, E., Kerssemakers, J., Shaltiel, I. A., Haering, C. H. & Dekker, C. DNA-loop extruding condensin complexes can traverse one another. en. *Nature* **579.** Number: 7799 Publisher: Nature Publishing Group, 438–442 (Mar. 2020).

87. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. en. *Cell* **163,** 1611–1627 (Dec. 2015).

88. Li, G., Chen, Y., Snyder, M. P. & Zhang, M. Q. ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. en. *Nucleic Acids Research* **45.** Publisher: Oxford Academic, e4–e4 (Jan. 2017).

89. Schones, D. E., Smith, A. D. & Zhang, M. Q. Statistical significance of cis-regulatory modules. en. *BMC Bioinformatics* **8,** 19 (Jan. 2007).

90. Hashimoto, H. *et al.* Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. en. *Molecular Cell* **66,** 711–720.e3 (June 2017).

91. Holzmann, J. *et al.* Absolute quantification of cohesin, CTCF and their regulators in human cells. *eLife* **8** (eds Sherratt, D. J., Struhl, K. & Sherratt, D. J.) Publisher: eLife Sciences Publications, Ltd, e46269 (June 2019).

92. Hansen, A. S., Pustova, I., Cattoglio, C., Tjian, R. & Darzacq, X. CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *eLife* **6** (ed Sherratt, D.) Publisher: eLife Sciences Publications, Ltd, e25776 (May 2017).

93. Cattoglio, C. *et al.* Determining cellular CTCF and cohesin abundances to constrain 3D genome models. *eLife* **8** (eds Sherratt, D. J., Struhl, K. & Sherratt, D. J.) Publisher: eLife Sciences Publications, Ltd, e40164 (June 2019).

94. Li, G. *et al.* ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. en. *Genome Biology* **11,** R22 (Feb. 2010).

95. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997 [q-bio].* arXiv: 1303.3997 (May 2013).

96. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9,** R137 (Sept. 2008).

97. Mathelier, A. *et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. en. *Nucleic Acids Research* **44.** Publisher: Oxford Academic, D110–D115 (Jan. 2016).

98. Plimpton, S. *Fast parallel algorithms for short-range molecular dynamics* English. Tech. rep. SAND-91-1144 (Sandia National Labs., Albuquerque, NM (United States), May 1993).

99. Gorkin, D. U. *et al.* Common DNA sequence variation influences 3-dimensional conformation of the human genome. *Genome Biology* **20,** 255 (Nov. 2019).

100. Hu, M. *et al.* HiCNorm: removing biases in Hi-C data via Poisson regression. en. *Bioinformatics* **28.** Publisher: Oxford Academic, 3131–3133 (Dec. 2012).

101. Greenwald, W. W. *et al.* Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. en. *Nature Communications* **10.** Number: 1 Publisher: Nature Publishing Group, 1054 (Mar. 2019).

102. Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. en. *Cell* **169,** 930–944.e22 (May 2017).

103. Guo, Y. *et al.* CRISPR-mediated deletion of prostate cancer risk-associated CTCF loop anchors identifies repressive chromatin loops. en. *Genome Biology* **19,** 160 (Oct. 2018).

104. Haarhuis, J. H. I. *et al.* The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. en. *Cell* **169,** 693–707.e14 (May 2017).

105. 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.** Publisher: John Wiley & Sons, Ltd, 3573–3599 (Dec. 2017).

106. Fulco, C. P. *et al.* Systematic mapping of functional enhancer–promoter connections with CRISPR interference. en. *Science* **354.** Publisher: American Association for the Advancement of Science Section: Report, 769–773 (Nov. 2016).

107. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. en. *Nature Genetics* **51.** Number: 12 Publisher: Nature Publishing Group, 1664–1669 (Dec. 2019).

108. Thakore, P. I. *et al.* Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. en. *Nature Methods* **12.** Number: 12 Publisher: Nature Publishing Group, 1143–1149 (Dec. 2015).

109. Klann, T. S. *et al.* CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. en. *Nature Biotechnology* **35.** Number: 6 Publisher: Nature Publishing Group, 561–568 (June 2017).

110. Liu, S. J. *et al.* CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. en. *Science* **355.** Publisher: American Association for the Advancement of Science Section: Research Article (Jan. 2017).

111. Qi, Z. *et al.* Tissue-specific Gene Expression Prediction Associates Vitiligo with SUOX through an Active Enhancer. en. *bioRxiv.* Publisher: Cold Spring Harbor Laboratory Section: New Results, 337196 (June 2018).

112. Ulirsch, J. C. *et al.* Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. en. *Cell* **165,** 1530–1545 (June 2016).

113. Wakabayashi, A. *et al.* Insight into GATA1 transcriptional activity through interrogation of cis elements disrupted in human erythroid disorders. en. *Proceedings of the National Academy of Sciences* **113.** Publisher: National Academy of Sciences Section: Biological Sciences, 4434–4439 (Apr. 2016).

114. Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. en. *Molecular Cell* **66,** 285–299.e5 (Apr. 2017).

115. Xu, B. *et al.* Selective inhibition of EZH2 and EZH1 enzymatic activity by a small molecule suppresses MLL-rearranged leukemia. en. *Blood* **125.** Publisher: American Society of Hematology, 346–357 (Jan. 2015).

116. Huang, J. *et al.* Dissecting super-enhancer hierarchy based on chromatin interactions. en. *Nature Communications* **9.** Number: 1 Publisher: Nature Publishing Group, 943 (Mar. 2018).

117. Hnisz, D., Day, D. S. & Young, R. A. Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. en. *Cell* **167,** 1188–1200 (Nov. 2016).

118. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. en. *Nature* **583.** Number: 7818 Publisher: Nature Publishing Group, 699–710 (July 2020).

119. Beer, M. A., Shigaki, D. & Huangfu, D. Enhancer Predictions and Genome-Wide Regulatory Circuits. *Annual Review of Genomics and Human Genetics* **21.** _eprint: https://doi.org/10.1146/annurev-genom-121719-010946, 37–54 (2020).

120. Li, Q. V. *et al.* Genome-scale screens identify JNK–JUN signaling as a barrier for pluripotency exit and endoderm differentiation. en. *Nature Genetics* **51.** Number: 6 Publisher: Nature Publishing Group, 999–1010 (June 2019).

121. Cao, F., Zhang, Y., Loh, Y. P., Cai, Y. & Fullwood, M. J. Predicting chromatin interactions between open chromatin regions from DNA sequences. en. *bioRxiv.* Publisher: Cold Spring Harbor Laboratory Section: New Results, 720748 (July 2019).

122. Walker, D. M., Freddolino, P. L. & Harshey, R. M. A Well-Mixed E. coli Genome: Widespread Contacts Revealed by Tracking Mu Transposition. en. *Cell* **180,** 703–716.e18 (Feb. 2020).

123. Schwarzer, W. *et al.* Two independent modes of chromatin organization revealed by cohesin removal. en. *Nature* **551.** Number: 7678 Publisher: Nature Publishing Group, 51–56 (Nov. 2017).

124. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. en. *Nature Reviews Genetics* **19.** Number: 7 Publisher: Nature Publishing Group, 453–467 (July 2018).

125. Giorgetti, L. *et al.* Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription. en. *Cell* **157,** 950–963 (May 2014).

126. Andreoletti, G., Pal, L. R., Moult, J. & Brenner, S. E. Reports from the fifth edition of CAGI: The Critical Assessment of Genome Interpretation. en. *Human Mutation* **40.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.23876, 1197–1201 (2019).

127. Shigaki, D. *et al.* Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. en. *Human Mutation* **40.** _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/humu.23797, 1280–1291 (2019).

128. Kreimer, A. *et al.* Predicting gene expression in massively parallel reporter assays: A comparative study. *Human mutation* **38,** 1240–1250 (2017).

129. Beer, M. A. Predicting enhancer activity and variant impact using gkm-SVM. *Human Mutation* **38,** 1251–1258 (2017).

130. Carter, D., Chakalova, L., Osborne, C. S., Dai, Y.-f. & Fraser, P. Long-range chromatin regulatory interactions in vivo. *Nature genetics* **32,** 623–626 (2002).

131. Symmons, O. *et al.* Functional and topological characteristics of mammalian regulatory domains. *Genome research* **24,** 390–400 (2014).

132. Symmons, O. *et al.* The Shh topological domain facilitates the action of remote enhancers by reducing the effects of genomic distances. *Developmental cell* **39,** 529–543 (2016).

133. Anania, C. *et al.* In vivo dissection of a clustered-CTCF domain boundary reveals developmental principles of regulatory insulation. en. *Nature Genetics* **54.** Number: 7 Publisher: Nature Publishing Group, 1026–1036 (July 2022).

134. Chakraborty, S. *et al.* High affinity enhancer-promoter interactions can bypass CTCF/cohesin-mediated insulation and contribute to phenotypic robustness. *BioRxiv,* 2021–12 (2022).

135. Reilly, S. K. *et al.* Direct characterization of cis-regulatory elements and functional dissection of complex genetic associations using HCR–FlowFISH. *Nature genetics* **53,** 1166–1176 (2021).

136. Gabriele, M. *et al.* Dynamics of CTCF-and cohesin-mediated chromatin looping revealed by live-cell imaging. *Science* **376,** 496–501 (2022).

137. Zuin, J. *et al.* Nonlinear control of transcription through enhancer–promoter interactions. *Nature* **604,** 571–577 (2022).