Predicting CTCF-mediated chromatin interactions by integrating genomic and epigenomic features

Yan Kai[1,2,3], Jaclyn Andricovich[2,3], Zhouhao Zeng[1], Jun Zhu[4], Alexandros Tzatsos[2,3,*], and Weiqun Peng[1,*]

[1]Department of Physics, George Washington University (GWU), Washington DC 20052 and

[2]Cancer Epigenetics Laboratory, Department of Anatomy and Regenerative Biology, GWU, and

[3]GWU Cancer Center, GWU School of Medicine and Health Sciences, Washington DC 20052

[4]Systems Biology Center, National Heart Lung and Blood Institute, National Institute of Health, Bethesda, MD 20892

*Correspondence: atzatsos@gwu.edu, wpeng@gwu.edu

Abstract

The CCCTC-binding zinc finger protein (CTCF)-mediated network of long-range chromatin interactions is important for genome organization and function. Although this network has been considered largely invariant, we found that it exhibits extensive cell-type-specific interactions that contribute to cell identity. Here we present Lollipop—a machine-learning framework—which predicts CTCF-mediated long-range interactions using genomic and epigenomic features. Using ChIA-PET data as benchmark, we demonstrated that Lollipop accurately predicts CTCF-mediated chromatin interactions both within and across cell-types, and outperforms other methods based only on CTCF motif orientation. Predictions were confirmed computationally and experimentally by Chromatin Conformation Capture (3C). Moreover, our approach reveals novel determinants of CTCF-mediated chromatin wiring, such as gene expression within the loops. Our study contributes to a better understanding about the underlying principles of CTCF-mediated chromatin interactions and their impact on gene expression.

Introduction

Higher-order chromatin structure plays a critical role in gene expression and cellular homeostasis[1, 2, 3, 4, 5, 6, 7]. Genome-wide profiling of long-range interactions in multiple cell-types revealed that CCCTC-binding factor (CTCF) binds at loop anchors and delimits the boundaries of Topologically Associating Domains (TADs)[8, 9, 10, 11], suggesting that CTCF plays a central role in regulating the organization and function of the 3D genome[12, 13]. Depletion of CTCF revealed that it is required for chromatin looping between its binding sites and insulation of TADs[14, 15], and disruption of individual CTCF binding sites deregulated the expression of surrounding genes[16, 17, 18, 19]. Mechanistically, many of the CTCF-mediated loops define insulated neighborhoods that constrain promoter-enhancer interactions[13], and in some cases CTCF is directly involved in promoter-enhancer interactions[9, 10, 20].

The CTCF-mediated interaction network has been considered to be largely invariant across cell-types. However, in studies of individual loci, cell-type-specific CTCF-mediated interactions were found to be important in gene regulation[17, 21]. Furthermore, CTCF binding sites vary extensively across cell-types[22, 23]. These findings suggest that the repertoire of CTCF-mediated interactions can be cell-type specific, and it is necessary to understand the extent and functional role of cell-type-specific CTCF-mediated loops. If cell-type-specific interactions are prevalent and contribute to cellular function, it would be inappropriate to use the CTCF-mediated interactome derived from a different cell-type.

CTCF-mediated loops can be mapped through Chromatin Conformation Capture (3C)-based technologies[2]. Among them, Hi-C[9, 24] provides the most comprehensive coverage for identifying looping events. However, it requires billions of reads to achieve kilo-base resolution[9]. On the other hand, Chromatin Interaction Analysis using Paired End Tags (ChIA-PET) increases resolution by only targeting chromatin interactions associated with a protein of interest[10, 25, 26]. Recently developed protocols, including Hi-ChIP[27] and PLAC-seq[28], improved upon ChIA-PET in sensitivity and cost-effectiveness. Despite recent technical advances, experimental profiling of CTCF-mediated interactions remains difficult and costly, and few cell-types have been analyzed[9, 10, 24, 29]. Therefore, computational predictions that take advantage of the routinely available ChIP-seq and RNA-seq data is a desirable approach to guide the interrogation of the CTCF-mediated interactome for the cells of interest.

Here, we carried out comprehensive analysis of CTCF-mediated chromatin interactions using ChIA-PET data sets from multiple cell-types. We found that CTCF-mediated loops exhibit widespread plasticity and the cell-type-specific loops are biologically significant. Motivated by this observation, we developed Lollipop—a machine-learning framework based on random forests classifier—to predict the CTCF-mediated interactions using genomic and epigenomic features. Lollipop significantly outperforms methods based solely on convergent motif orientation when evaluated both within individual and across different cell-types. Our predictions were also experimentally confirmed by 3C. Moreover, our approach reveals novel determinants of CTCF-mediated chromatin wiring, such as gene expression within the loop.

Results

CTCF-mediated chromatin interactions exhibit cell-type specificity

We used the ChIA-PET2 pipeline[30] and analyzed published ChIA-PET data sets from three cell-lines (**Supplementary Table 1**): GM12878 (lympho-blastoid cells)[10], HeLa-S3 (cervical adenocarcinoma cells)[10], and K562 (chronic myelogenous leukemia cells)[29]. By using false discovery rate (FDR) ≤0.05 and paired-end tag (PET) number ≥2, we identified 51966, 16783, 13076 high-confidence chromatin loops for GM12878, HeLa, and K562, respectively (**Supplementary Table 2**). A significant fraction of loops was found to be cell-type-specific (67.9%, 26.2%, and 21.5% of loops in GM12878, HeLa, and K562, respectively (**Fig. 1a**). It is worth noting that the higher number of loops and cell-type-specific loops observed in GM12878 may be attributed to the higher sequencing depth of GM12878 ChIA-PET library (**Supplementary Table 2**).

To elucidate what contributes to this plasticity, we compared the CTCF binding sites identified in ChIA-PET data sets across the three cell-lines. We found that only 36% of CTCF binding sites are constitutive (i.e., "+++", **Fig. 1b**), consistent with previous reports[22, 23]. Besides cell-type-specific binding sites, rewiring of shared binding sites also contributes to the cell-type-specific loops (**Fig. 1c**).

Cell-type-specific CTCF-mediated loops contribute to gene regulation

Loops shared among different cell-types exhibit significantly higher interaction strength than the cell-type-specific loops (**Supplementary Fig. 1a**), questioning whether the latter are biologically relevant. To address this question, we asked whether these loops are involved in gene regulation. We found that cell-type-specific loops harbor a significantly higher ratio of tandem CTCF motif orientation compared to shared loops (**Supplementary Fig. 1b**), suggesting their involvement in gene regulation, considering that tandem loops exhibit more regulatory potential than convergent ones[10].

Super-enhancers (SEs) are defined as stretches of chromatin that cluster multiple enhancers decorated with H3K27ac. A recent study revealed that CTCF plays a critical role in the hierarchical organization of SEs[31]. Considering that SEs play critical roles in cell identity, development, and cancer[32, 33, 34], we examined whether they are enriched within cell-type-specific loops. Disease Ontology analysis using GREAT[35] confirmed that these SEs are linked with the corresponding disease origin of the three cell-types (**Supplementary Fig. 1c**). Comparison of SEs in HeLa and K562 identified three sets of SEs: HeLa-specific, common, and K562-specifc. HeLa-specific SEs are significantly enriched within HeLa-specific loops, compared to common SEs (**Fig. 1d** left panel). Similarly, K562-specific SEs are preferentially enriched within K562-specific loops compared to common SEs (**Fig. 1d** left panel). The same conclusion was reached when we compared GM12878 vs HeLa as well as GM12878 vs K562 (**Fig. 1d** central and right panels). Taken together, we found that cell-type-specific SEs are more likely to be associated with loops specific to that cell-type, suggesting the functional significance of cell-type-specific loops.

Consistently, differentially expressed genes (DEGs) between the three cell types are significantly associated with cell-type-specific loops (**Supplementary Fig. 1d**). Ingenuity Pathway Analysis (IPA)[36] revealed that DEGs between HeLa and K562 categorized based on loop association are enriched in distinct canonical pathways (**Fig. 1e**). Similar results were obtained in pair-wise comparisons between GM12878 and the other two cell lines

(**Supplementary Fig.1e-f**). For instance, **Fig. 1f** illustrates the loop architecture and epigenomic features of ROR2, a receptor involved in non-canonical Wnt signaling with a significant role in human carcinogenesis[37, 38]. ROR2 is highly expressed in K562 compared to HeLa, and these CTCF-mediated loops are present only in K562. The up-regulation of ROR2 expression is associated with a concomitant decrease of H3K27me3 and increase in H3K36me3 in the region, as well as the appearance of a K562-specific SE in the gene body.

Altogether, cell-type-specific CTCF-mediated loops are prevalent and may play a significant role in the transcriptional programs of cell-type-specific genes. Therefore, we sought to develop a computational approach to infer the CTCF-mediated loops.

An ensemble learning method to predict CTCF-mediated loops from genomic and epigenomic features

We employed a random forest classifier, a tree-based ensemble learning method, to predict CTCF-mediated loops. This classification method takes into consideration the complex interactions among features and is robust against overfitting[39, 40, 41]. The pipeline, named Lollipop, aims to find an optimized combination of genomic and epigenomic features to distinguish interacting from non-interacting pairs of CTCF sites. The schema of the pipeline is shown in **Fig. 2a**. The trained model can be used to predict CTCF-mediated loops in the same or a different cell-type.

For training purposes, the positive and negative loops were derived from ChIA-PET data sets[10, 29]. To ensure confident labeling of positive loops, we used stringent criteria (FDR <= 0.05 and at least 2 PETs connecting the two anchors). Negative loops were constructed by random pairing of CTCF binding sites and were 5 times as abundant as the positive loops. Additional rules to select negative loops included: (a) lack of PET in the ChIA-PET dataset; and (b) absence in the list of identified interactions from the Hi-C experiments (see methods for details).

A total of 77 features were derived from genomic and epigenomic data sets (**Fig. 2a**). Genomic features include loop length and features defined at the CTCF binding sites, including CTCF motif orientation, strength, and sequence conservation. We included loop length because it is an inherent determinant of contact frequency between two genomic regions[42], and motif orientation pattern because CTCF anchors preferentially adopt a convergent motif orientation[9]. Epigenomic features include chromatin accessibility, a variety of histone modifications, and architectural proteins CTCF and Cohesin (RAD21). For the use of DNase-seq and ChIP-seq data sets, three types of features were used: (a) local features defined at the anchors, (b) in-between features defined over the loop region, (c) and flanking features defined over the region from the loop anchor to the nearest CTCF binding event outside the loop (**Fig. 2b**). The use of the in-between features was motivated by a recent study[43] showing that signals over the loop regions were more important in predicting promoter-enhancer interactions than signals at anchors. In addition, given the insulator role of CTCF, we reasoned that the signals over the flanking regions might help to distinguish interacting from non-interacting CTCF binding sites. Finally, we also included gene expression within the looped region as a feature (see methods for details).

Assessment of Lollipop's performance within individual cell-types

We employed Receiver Operator Characteristic (ROC) and Precision-Recall (PR) curves with 10-fold cross-validation to assess the performance of Lollipop. To account for possible bias introduced by random partitioning of training data, we performed 5 iterations for cross-validation and reported the mean performance. For evaluation of Lollipop's performance, two methods

were used for comparison. Both methods are inspired by the finding that the CTCF motifs in anchors preferentially adopt convergent orientation[9, 10]: (a) The naïve method, which pairs a CTCF-bound motif that resides on the forward strand to the nearest downstream CTCF-bound motif that resides on the reverse strand (**Supplementary Fig. 2a**); (b) The Oti method[44], which iteratively applies the naïve method to CTCF binding sites selected by different signal intensity thresholds (see **Supplementary Fig. 2b** for illustration and methods for details). By doing so, the Oti method identifies more loops than the naïve method and partially recovers the nested structure of some CTCF-mediated loops.

Fig. 3a-b show that Lollipop achieved an area under ROC curve (AU-ROC) value of ≥0.97 and area under PR curve (AU-PR) value of ≥0.86 in all cell lines. Compared to other methods, Lollipop achieved similar or higher precision and superior recall. The latter can be partially attributed to the failure of naïve and Oti methods to capture tandem loops or loops without CTCF motif on anchors, which account for a significant fraction of CTCF-mediated loops (64% for GM12878, 61% for HeLa, 49% for K562). We then independently evaluated Lollipop's performance on convergent and non-convergent loops. Even on convergent loops, Lollipop achieved a superior recall score with a precision score comparable those of the naïve and Oti method (**Fig. 3c**). Furthermore, Lollipop also performed well in the prediction of non-convergent loops (**Fig. 3d**). In summary, Lollipop can account for the complexity of loop structures by integrating genomic and epigenomic features and outperforms methods that only consider the convergent CTCF motif orientation.

Feature analysis identified novel determinants of CTCF-mediated chromatin loops

Considering that convergent motif orientation does not suffice to identify CTCF-mediated loops, we ranked features that significantly improve the performance, by measuring the mean decrease impurity during training the random forests classifier[45]. We found that the average binding intensity of CTCF and Cohesin (RAD21) at the loop anchors are the most important features (**Fig. 4a** and **Supplementary Fig. 3a**), suggesting that sites with stronger CTCF and Cohesin binding are more likely to become anchors (**Supplementary Fig. 3b**), consistent with the observation that that these proteins are important for chromatin interactions[14, 15]. In addition, loop length and motif orientation pattern were amongst the top features, in agreement with previous results[9, 42]. The list also includes features defined within loop regions, among which gene expression was of particular interest. Regions inside positive loops exhibit significantly lower gene expression levels compared to negative loops (**Fig. 4b)**. This finding is supported by similar trends exhibited by histone marks for active gene bodies H3K79me2 and H3K36me3 (**Supplementary Fig. 3c**). Another interesting feature is the standard deviation of CTCF and Cohesin binding at the anchors (**Fig. 4a**). We therefore examined the relative fluctuation, defined as standard deviation divided by average intensity, of CTCF and Cohesin on anchor pairs of the positive and negative loops. As shown in **Fig. 4c** and **Supplementary Fig. 3d**, anchor-pair CTCF and RAD21 have significantly lower relative fluctuation in positive loops than in negative loops.

While CTCF binding at anchors is clearly critical for looping, formation of a loop requires wiring (i.e. physical interaction) between specific pair of anchors. We therefore asked what features contribute to the wiring. To this end, we changed negative loops to be random pairings of actual anchors, and then reanalyzed feature importance. As shown in **Supplementary Fig. 3e**, length, motif-orientation and expression are strongly contributing, whereas CTCF and Cohesin binding

at anchors become much less important. It is worth noting that more in-between features showed up in the list, compared to those in **Fig. 4a** and **Supplementary Fig. 3a**.

As the features employed are correlated (**Fig. 4d** and **Supplementary Fig. 3f**), the feature importance scores might be skewed. To validate the ranking of feature importance, we applied the Recursive Feature Elimination method to evaluate the performance of the recursively reduced feature set. The results are consistent with the feature ranking from the mean decrease impurity (**Supplementary Table 3**). Last, performance evaluation under different feature sets suggests that near-optimal performance can be achieved by using ~16 features (**Fig. 4e**). These features include those derived from CTCF and RAD21 binding, loop length, CTCF motif orientation, gene expression, as well as epigenetic features (**Supplementary Table 3**).

Assessment of Lollipop's performance across cell-types

Having demonstrated Lollipop's superior performance within individual cell-types, we next used the model trained in one cell-type to make predictions and assessment in another cell-type (see methods for details). This is more realistic and challenging, as a large number of CTCF-mediated loops are cell-type-specific. In all three cell-types Lollipop achieved AU-ROC ≥ 0.93 and AU-PR ≥ 0.79 (**Fig. 5a-b**), only moderately lower than its performances within individual cell-types (**Fig. 3a-b**). It is worth noting that Lollipop outperforms motif-orientation based methods (**Fig. 5a-b**). Given that a loop consists of a pair of anchors and the wiring between them, we then dissected Lollipop's predictive power on anchors and wiring, respectively. For assessment of anchor prediction, we evaluated Lollipop by comparing the anchor usage of the predicted loops with that of loops identified from ChIA-PET in the target cell-type. For assessment of wiring prediction, we constructed negative loops by random pairing of actual anchors in the target cell-type (see methods for details). **Fig. 5c-d** show the PR curves demonstrating that Lollipop performed reasonably well in both, and better in predicting anchors than in predicting wiring. The results in terms of ROC (**Supplementary Fig. 4a-b**) are consistent with those in terms of PR.

Evaluation of de novo predictions of CTCF-mediated loops

After training Lollipop in individual cell-types, we then applied it to scan the genome of the same cell-type to make *de novo* genome-wide predictions. Lollipop predicted 67855, 38274, 32237 loops in GM12878, HeLa and K562, respectively. Notably, the number of predicted loops in GM12878 is much larger than those of the other two cell-types, due to the much larger number of loops identified by ChIA-PET in GM12878 (see last column of **Supplementary Table 2**). These loops were used in training the model and thus affect the number of predicted loops. Indeed, if we down-sample the GM12878 ChIA-PET library to 15% so that the number of called loops is on par with those in K562 and HeLa (see last column of **Supplementary Table 2**), the number of predicted loops is comparable to the number of predictions in K562 and HeLa.

As shown in **Supplementary Fig. 5a**, a large fraction of the predicted loops (48%, 73% and 77% for GM12878, HeLa and K562, respectively) was not supported by ChIA-PET under the stringent criterion of FDR<=0.05 and PET>=2 used for defining positive loops. However, if we relaxed the stringency to PET>=1 in ChIA-PET, the fraction of predicted loops not supported by ChIA-PET was significantly reduced, to 24%, 42% and 50% in GM12878, HeLa and K562, respectively. Similar result can be obtained with the down-sampled GM12878 library (**Supplementary Fig. 5b**). This observation raises the question of whether the predicted loops with less or no ChIA-PET support are indeed false positives. To address this question, we

carried out the following computational as well as experimental evaluations on those predicted loops without any ChIA-PET support.

First, we used the published Hi-C contact matrices for GM12878 and K562[9] (see methods for details) to evaluate these loops, and found that they have significantly higher contact frequencies than pairs of randomly-chosen genomic loci (**Fig. 6a**). For fair comparison, the control regions were sampled to have a length distribution matching those of the target loops. Second, we randomly selected two such cases and performed 3C experiments. **Fig. 6b** shows the sequence of the ligation junctions from the long-range interactions (PRKAG2-KMT2C and PDE6A-PDGFRB) in Hela. 3C-qPCR further confirmed the contact frequency of the PRKAG2-KMT2C loop in respect to neighboring HindIII fragments (**Supplementary Fig. 5d**).

Having shown that the predicted loops lacking ChIA-PET support could be real, we sought to understand why they were not observed in ChIA-PET. To this end, we performed scaling analysis in the ChIA-PET data of GM12878 cells, which received significantly higher sequencing coverage than those of K562 and HeLa (**Supplementary Table 2**). Specifically, we used the 15% down-sampled GM12878 ChIA-PET library to identify loops with the same approach employed for the full data set, and trained a classifier. We then applied this classifier to make genome-wide predictions. Of the 33463 predicted loops, 12047 are without any support from the down-sampled ChIA-PET data set. However, 46% of these loops find support in the full ChIA-PET library, and 20% of these loops even find significant support (**Fig. 6c**). This down-sampling process was repeated for 10 times and similar results were obtained (data not shown). Taken together, the scaling analysis suggests that insufficient sequencing depth contributes to the presence of predicted loops lacking support in ChIA-PET.

Topological properties of CTCF-mediated interaction network and associated biological functions

To gain a better understanding of these interactions, we took a systems approach to visualize and analyze the CTCF-mediated interactions. We constructed the CTCF-mediated interaction network by denoting the anchors as nodes and the long-range interactions as edges. As exemplified in **Fig. 7a**, where the interaction network on chromosome 1 (visualized using graph-tool V2.22, https://graph-tool.skewed.de) is shown, the CTCF-mediated interactions form a disconnected network encompassing many linear-polymer-like components. This is dramatically different from the RNA-PolII-mediated interaction network[46], which is dominated by one scale-free connected graph[46]. This dramatic difference in topological structure is also manifested in the degree distributions (**Supplementary Fig. 6**), where the distribution for RNA PolII exhibits a fatter tail.

It is worth noting that degrees of connections among the anchors vary. We therefore examined CTCF hubs, anchors involved in multiple interactions. Ranking anchors according to the degrees of connections, we defined hubs as those among the top 10% anchors and non-hubs as the bottom 10% (see methods for details), and identified 2914, 2111 and 1843 nodes for GM12878, HeLa and K562, respectively. Subsequent comparison between hubs and non-hub nodes revealed that hubs are (a) more conserved across cell-types than non-hubs, likely because they serve as the structural foci of genome organization in the nucleus, (b) characterized by significantly higher binding affinity for CTCF and Cohesin (**Fig. 7c**), and (c) associated with distinct biological functions. Gene ontology analysis[35] showed that the hubs are preferentially associated with immunology-related functions in GM12878 and K562 cells, but not in HeLa cells (**Fig. 7d**), consistent with the cellular origin of these cell-lines. For example, the

hubs in GM12878 and K562 cells were found to be significantly associated with antigen binding, and the GM12878 hubs were significantly associated with the MHC (major histocompatibility complex) protein complex. MHC is a set of cell surface proteins that are essential for immune system, while MHC class II (MHC-II) genes encode cell-surface glycoproteins that present antigens to CD4 T cells to initiate and control adaptive immune responses[47]. Our results were consistent with previous studies[47, 48] which found that CTCF plays an important role in controlling MHC-II gene expression.

Discussion

Here we showed that CTCF-mediated chromatin interactions exhibit extensive variations across cell-types. These cell-type-specific interactions are functionally important, as they are linked to differentially expressed genes and cell-type-specific SEs contributing to cell identity. However, genome-wide profiling of CTCF-mediated interactions is available in a very limited number of cell-types and conditions, as experimental approaches remain challenging and costly. Therefore, we developed Lollipop, a machine-learning framework, to make genome-wide predictions of CTCF-mediated loops using widely accessible genomic and epigenomic features. Using computational as well as experimental validations, we demonstrated that Lollipop performed well within and across cell-types. Analysis of the machine learning model revealed novel features associated with CTCF-mediated loops, and shed light on the rules underlying CTCF-mediated chromatin organization.

While previous studies focused on the significance of conserved CTCF binding at TAD boundaries or loop anchors, our study showed a significant proportion of CTCF-mediated interactions are cell-type-specific. Based on our analysis, both lineage-specific recruitment of architectural proteins and alternative wiring among available anchor sites contribute to the establishment of cell-type specificity. Although the process of establishing cell-type-specific is not well understood, it is conceivable that multiple factors combine to orchestrate a cell-type-specific chromatin context to promote the formation of a loop.

The convergent orientation of CTCF motifs at loop anchors is a prominent feature of CTCF-mediated interactions[9, 10], as it is also manifested by our model. However, model comparison demonstrated that motif orientation alone is limited in its predictive power, and inclusion of other features significantly improved the performance. Interestingly, we found that features for the loop regions, which are away from the anchors, contribute significantly to the predictive power, consistent with findings in enhancer-promoter interaction prediction[43]. Specifically, gene expression exhibits distinct distributions over positive loop regions compared to negative loops (**Fig. 4b**, and **Supplementary Fig. 4c**), which may be attributed to the enhancer-blocking role of CTCF loop anchors.

In evaluating our predictions, we showed that false positives could be due to mislabeling in the testing data. As advances in experimental protocols and continuous decreases in sequencing cost would result in better training data in reference cell-types, it is likely that the performance of Lollipop would further improve. Since CTCF plays a major role in defining regulatory domains, results obtained from our approach can potentially be used as constraints in predicting enhancer-promoter interactions, which remains a major challenge. Overall, CTCF-mediated chromatin interactions are critical for genome organization and function, and our study provides a computational tool for the exploration of the 3D organization of the genome.

Materials and Methods

Data availability

GM12878 and HeLa ChIA-PET data were downloaded from Gene Expression Omnibus (GEO) with accession number GSE72816[10]. K562 ChIA-PET data was downloaded from ENCODE[29] with accession number ENCLB559JAA. High-resolution genome-wide Hi-C contact matrices were obtained from GEO with accession number GEO63525[9]. DNase-Seq, ChIP-Seq and RNA-Seq data were downloaded from ENCODE and were aligned to hg19. The accession numbers for the data used in this study were summarized in **Supplementary Table 1**.

Lollipop is publically available in https://github.com/ykai16/Lollipop.

Identification of CTCF-mediated loops from CTCF ChIA-PET data

We employed ChIA-PET2 (v0.9.2)[30] to identify CTCF-mediated loops. Briefly, ChIA-PET2 involves linker filtering, PET mapping, PET classification, binding-site identification, and identification of long-range interactions. In the step of linker filtering, one mismatch was allowed in identifying reads with linkers. After linker removal, only reads with at least 15 bp in length were retained for further analysis for GM12878 and HeLa (read length = 150 bp). For K562, the read length was shorter (36 bp), therefore reads with at least 10 bp in length were retained for further analysis. In other steps, default values for parameters were used. Only uniquely mapped reads were kept, and PETs were de-duplicated. Significant loops were identified with a value of false discovery rate (FDR) <= 0.05. We further required that they are supported by at least two PETs (i.e., IAB >= 2).

We only considered long-range interactions whose length are less than 1 million bps (mb), for two reasons. First, vast majority of loops (93.2% for GM12878, 97.3% for HeLa, 98.1% for K562) are less than 1mb long. Similar observations were made in [10]. Second, insulated neighborhoods, the CTCF loops having higher potential in regulation of gene expression, were found to range from 25 kb to 940 kb[6, 16] (reviewed in [13]).

Comparison of CTCF-mediated loops among cell-types (Fig. 1a, Supplementary Fig. 1a-b)

An anchor is considered as shared by two cell-types if the respective genomic regions delineating this anchor overlap in the two cell-type. A loop is considered as shared by two cell-types if both anchors are shared by the two cell-types. A loop is considered cell-type specific if either of the two anchors are cell-type specific. The loops shared by all three cell-types were defined as GM12878 loops shared by both K562 and HeLa.

Analysis of CTCF binding sites in three cell-types (Fig. 1b)

CTCF peaks were determined by MACS2[49] in the ChIA-PET2 pipeline. A binding site was defined as peak summit +/- 500 bp. The binding sites in the three cell-types were classified into seven groups according to the overlapping pattern. Binding intensity for each site was represented by the log2 (RPKM) value over the summit +/- 2kb region. For each group, the binding sites were ordered in descending order according to binding intensity in a prioritized manner. Namely, CTCF binding sites present in GM12878 were ordered by their binding strengths in GM12878; CTCF binding sites not present in GM12878 were ordered by binding strengths in Hela and then in K562 accordingly. Seaborn (V 0.7.1, http://seaborn.pydata.org) was used to generate the heat map.

Super-enhancer analysis (Fig. 1d, Supplementary Fig. 1c)

Super-enhancers (SEs) were identified by the Ranking Ordering of Super-Enhancers algorithm (ROSE[33, 34]), using H3K27ac ChIP-Seq data as input and default parameters. Identified super-enhancers were then uploaded to Genomic Regions Enrichment of Annotations Tool (GREAT) V3.0.0[35] for GO analysis (**Supplementary Fig. 1c**). If a SE in one cell-type does not overlap with any SEs in a different cell-type, it is deemed as a SE specific to that cell-type. Otherwise, it is called a shared SE. We then counted the number of cell-type specific loops covering each type of SEs. The comparison between Hela and K562 is shown in **Fig. 1d**. For comparison between GM12878 and another cell-type, the GM12878 ChIA-PET data set is first randomly down-sampled to 15% of the original size so that the number of loops identified matched those from the ChIA-PET datasets of the other two cell-types (see **Supplementary Table 2**). Then analysis identical to that in **Fig. 1d** was carried out. The down sampling and follow-up analysis was repeated 10 times to ensure reproducibility, and standard-deviations were shown in the **Fig. 1d**.

Analysis of differentially expressed genes and their association with CTCF-mediated loops (Fig.1e, Supplementary Fig. 1d, e, f)

Each cell-line has two RNA-Seq replicates. Cufflinks V2.2.1[50] with default parameters (*q-value*=0.05) was used to identify the differentially expressed genes (DEG).

For comparison between HeLa and K562, a DEG was deemed to be associated with HeLa-specific loops if it is within one or more HeLa-specific loops but not within any K562-specific loops. If a DEG is covered only by one or more shared loops, this DEG is deemed to be associated with shared loops. Following the criteria described above, we obtained three sets of DEGs respectively associating with HeLa-specific loops, shared loops, K562-specific loops. These three sets of DEGs were then subject to GO analysis using 'Ingenuity Pathway Analysis' [36]. The GO terms whose P-value are no less than 1e-3 in all three gene sets were then removed. The result is shown in **Fig. 1e**. Color key represents the -log10 (P-value). For comparison between GM12878 and another cell-type (**Supplementary Fig. 1 e, f**), the GM12878 ChIA-PET library is first randomly down-sampled to 15% of the original size so that the number of loops identified matched those of the ChIA-PET libraries from the other two cell-types.

For **Supplementary Fig. 1d**, non-DEG genes were those with the least significant expression changes as ranked by P-value, with group size matching to that of the corresponding DEG group.

Identification of CTCF motif occurrences

The position frequency matrix of CTCF for human was downloaded from Jaspar 2016 (http://jaspar.genereg.net)[51] . CTCF motif occurrences were identified by the FIMO package (V4.11.1[52]) with the P-value < 1e-5. In total, 110879 motif occurrences were identified.

Preparation of training data

Positive loops were identified using ChIA-PET2 pipeline with FDR<=0.05 and IAB >=2, with loop length restricted to be in the range of 10 kb to 1mb. The choice of the lower limit of 10 kb is because the ChIA-PET-identified loops with length below 10 kb are likely caused by self-ligation in library preparation[25]. The reason for the upper limit of 1mb is given above. Negative loops were constructed by random pairing of CTCF binding sites, with loop length ranging from 10 kb to 1mb. The number of negative interactions was chosen to be 5 times that of the positive interactions. To ensure accurate labeling, we further required that the negative loops (**1**) do not

receive any ChIA-PET support; and (**2**) are not present in the CTCF-mediated interactions identified from the Hi-C experiments [9].

Feature calculation (Fig. 2a, b)

Genomic features include motif strength, motif orientation, conservation score and loop length. Motif strength represents how similar the underlying sequence is to the CTCF consensus motif. The motif strength score was provided by FIMO[52]. The motif strength score of a CTCF binding site (summit +/- 1000bp) was represented by the strength of the motif occurrence within the site. If a CTCF binding site have more than one motif occurrences, the highest score was used. If there is no motif occurrence, 0 would be assigned. The feature of motif orientation was represented by the following rule: If neither anchor has CTCF motif, we assign a value of 0; If one anchor has no motif and the other has one or more than one motifs, we assign a value of 1; If both anchors have one or more motif occurrences, the orientation of each anchor is determined by the orientation of its strongest motif occurrence. Divergent orientation would be assigned a value of 2, tandem orientation would be assigned a value of 3, and convergent orientation would be assigned a value of 4. For conservation, we used the 100 way phastCons score downloaded from UCSC (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/phastCons100way)[53]. The conservation score of a CTCF binding site was defined as the mean value of the conservation score of each nucleotide in the summit +/- 20 bp region.

Functional genomic features include chromosome accessibility profiled by DNase-Seq, histone modifications, CTCF and Cohesin binding profiles profiled by ChIP-Seq, and gene expression profiled by RNA-Seq. DNase-Seq and ChIP-Seq data were de-duplicated and then subject to pre-processing to remove noise as follows. For DNase-Seq data, peaks were downloaded from ENCODE[29]. For ChIP-Seq data, SICER (V1.1)[54] were used to identify enriched regions with FDR 1e-5. For histone modifications with diffused signal (H3K27me3, H3K36me3, H3K9me3, H3K79me2), window size = 200 bp, gap size = 600 bp were used. For other ChIP-Seq libraries, window size = gap size = 200 bp were used. For both DNase-Seq and ChIP-Seq, only reads located on signal-enriched regions were used for feature calculation. For RNA-Seq data, gene expressions were calculated using Cufflinks[50] with default parameters. Each dataset was characterized by three types of features: local features, in-between features and flanking features, as illustrated in **Fig. 2b**. Local features were defined around anchors, represented by the signal intensity (RPKM value) over the CTCF summit position +/- 2kb region. In-between feature is represented by the average signal intensity (RPKM value) over a presumed loop region. The value of the expression feature is defined as the average FPKM value of the genes whose promoters are located inside the presumed loop. The flanking features are represented by the RPKM value over the region from the loop anchor to the nearest CTCF binding event identified in the CTCF ChIP-Seq.

Implementation of the naïve method and the Oti method (Supplementary Fig. 2)

The naïve method is implemented by pairing a CTCF-bound motif that resides on the forward strand to the nearest downstream CTCF-bound motif that resides on the reverse strand (**Supplementary Fig. 2a**). The Oti method was introduced in [44]. It ranked all the active motif sites in terms of CTCF peak strength in descending order. First all active motif sites were used to construct loops by the naïve method. Then, the same procedure was repeated for the top 80%, top 60%, top 40% and top 20% active motif sites. The loops constructed in different rounds were then pooled together. The Oti method is illustrated in **Supplementary Fig. 2b**.

Performance evaluation within individual cell-types (Fig. 3)

In **Fig. 3c, d**, the performance was evaluated at the looping probability cut-off of 0.5.

Evaluation of feature importance (Fig. 4a, d, e and Supplementary Fig. 3a, e, f)

Predictive importance scores of features were obtained from the "feature_importances " attribute of the trained random forest classifier[55]. The ranking of the top 20 features was visualized in **Fig. 4a** and **Supplementary Fig. 3a**. Pearson correlations of the in-between features calculated in positive interactions were used to generate the correlation matrix. The correlation matrix was subject to hierarchical clustering, as shown in **Fig. 4d and Supplementary Fig. 3f**. Recursive Feature Elimination (RFE) method was used to validate the analysis of the feature importance. After each iteration, model performance was evaluated in terms of Area Under Receiver Operating Characteristics (AU-ROC) curve and Area Under Precision Recall (AU-PR) curve. The performance vs. feature number was plotted in **Fig. 4e**.

For feature importance analysis of wiring prediction (**Supplementary Fig. 3e**). Negative data was prepared as follows: the anchors of positive loops were used to construct negative loops by random pairing. The number of negative loops were set to be 3 times that of positive loops. Other procedures on construction of negative loops were the same as described in the section of 'Preparation of training data'. Positive data remained unchanged.

Performance evaluation across cell-types (Fig. 5 and Supplementary Fig. 4)

In the across-cell-type performance evaluation, the model trained in cell-type A was applied to the cell-type B, using training data prepared in B for evaluation of performance.

For evaluation of anchor prediction, the anchors of positive loops in cell-type B were labeled positive, while the anchors belonging only to negative loops in cell-type B were labeled negative. The anchors of predicted loop were compared with positive and negative labels for evaluation of anchor prediction. This evaluation was repeated under different thresholds of looping probability to generate the PR and ROC curves (**Fig. 5c and Supplementary Fig. 4a**).

For evaluation of wiring prediction, the anchors of positive loops in cell-type B were used to construct negative loops by random pairing. The model trained in cell-type A was then applied to the training data of cell-type B for evaluation.

Computational evaluation of predicted CTCF-mediated loops (Fig. 6a, c and Supplementary Fig. 5a, b)

Models trained in a cell-type was used to predict loops genome-widely in the same cell-type. Predicted loops were then compared with loops identified from ChIA-PET datasets and categorized into three groups. 'Significant' loops denote those supported by ChIA-PET under the stringent criterion of FDR<=0.05 and PET>=2. 'With evidence' loops denote those supported by ChIA-PET reads but do not meet the stringent criterion mentioned above. 'No support' loops denote those without any support from ChIA-PET. The numbers of loops in each group were shown **in Supplementary Fig. 5a**.

Down sampling of ChIA-PET library in GM12878 cells: The ChIA-PET library was first randomly down-sampled to 15% of the original size, followed by loop identification using ChIA-PET2 and preparation of training data. Trained model was used to make genome-wide predictions. The predicted loops were categorized into three groups by comparing with loop calls using the down-sampled library, as described above. The result was shown in **Supplementary Fig. 5b**.

Evaluation of predicted loops without any ChIA-PET support using Hi-C data (**Fig. 6a**). 10 kb resolution Hi-C contact matrices for GM12878 and K562[9] were used for validation. The contact matrices were normalized by Knight and Ruiz (KR) normalization vector[9]. For each cell-type, we collected contact frequencies from the contact matrix for those predicted loops without any ChIA-PET support. As a control, we chose a matching set of random pairs of genomic locations as anchors with matching length-distribution. We then collected the contact frequencies of this control set. The two contrasting distributions of contact frequencies are shown. HeLa cell was not included in this analysis because the Hi-C library and Hi-C derived contact matrix are not available.

Scaling analysis in GM12878 cells. Predicted loops belonging to the 'No support' group in the down-sampled ChIA-PET library (yellow slice in **Supplementary Fig. 5b**) were compared with the loops identified using the full GM12878 ChIA-PET library and categorized into three groups, as shown in **Fig. 6b**.

Experimental validation using Chromosome Conformation Capture (3C) (Fig. 6b, Supplementary Fig. 5c-d)

The loops used for experimental validation were randomly selected from the loops predicted by Lollipop but not observed in ChIA-PET, as described above. For the 3C assay, cells were fixed and nuclei were prepared as in ChIP experiments. Nuclei were resuspended in 500 µl 1.2X CutSmart buffer (NEB) with 14 µl 10% SDS, and incubated at 37°C for 1 hour. SDS was sequestered by the addition of 50 µl 20% Triton X-100, and incubated at 37°C for 1 hour. Next, 5-20 µl "undigested" was reserved, and 400 U of HindIII was added to the remaining sample and digested overnight at 37°C with end-over-end rotation. The second day, 5-20 µl of "digested" material was reserved, and 40 µl of 20% SDS was added to remaining sample to inactivate HindIII by incubating at 65°C for 25 minutes. The samples were transferred to 15 mL conical tubes and diluted with the following 1.15X ligation buffer recipe: 352 µl 10X T4 ligase buffer (NEB), 2.71 ml water, and 187.5 µl 20% Triton X-100. Samples were incubated at 37°C for 1 hour. Next, 5000 U T4 ligase was added, and ligation took place with gentle end-over-end rotation at 16°C for 4 hours, and then 45 minutes at room temperature. Reverse crosslinking took place by the addition of 300 µg (30 µl) Proteinase K at 65°C, overnight. On day three, 300 µg RNase-A was added, and samples were placed at 37°C for one hour. To begin DNA extraction, 4 ml of phenol-chloroform was added, samples were vortexed for a full minute, and centrifuged at 2,200 x *g* for 15 minutes. The aqueous phase was collected in a new 50 ml tube and diluted with an equal volume of water (4 ml) and with 800 µl of 2 M sodium acetate pH 5.6; next 20 ml of ethanol was added, samples were inverted 10 times, and placed at -80°C for 1-4 hours to precipitate the DNA. The samples were centrifuged at 2,200 x *g* for 45 minutes at 4°C and washed with 70% ethanol. The 3C libraries were then allowed to dry briefly, without letting the pellet become dull. The libraries were re-suspended in 100-600 µl of 10 mM Tris. The digestion efficiency, as well as the quality and quantity of 3C libraries, were assessed before downstream analyses. The Q5 Taq polymerase (NEB) was used for PCR reactions using the following protocol: 98°C 30 sec, 35 cycles [98°C 10 sec, 70°C 15 sec, 72°C 10 sec], 72°C 2 min. Reactions were run on 2% agarose gels and analyzed using the ImageLab software (BioRad). Bands were extracted and sequenced (Eurofins) to confirm specificity of primers and loop identity. Data points plotted in the contact matrix are the averages of duplicates ± StDev from two independent library preparations. Primers were designed using a uni-directional strategy [56] and used are provided in **Supplementary Table 4**.

Analysis of CTCF-mediated interaction network (Fig. 7)

Construction of CTCF-mediated interaction network. We used nodes to represent anchors and edges to represent loops. Graph-tool (V2.22, https://graph-tool.skewed.de) was used for visualization of networks (**Fig. 7a**). In identification of hubs, anchors were ranked according to the degree of connection in descending order. Anchors with the same degree of connection were further ranked according to CTCF binding intensity in descending order. The top 10% anchors were defined as hubs, while the bottom 10% as non-hubs.

Functional enrichment analysis of hubs (**Fig. 7d**). Hubs were uploaded to GREAT (V3.0.0) [35] for functional enrichment analysis. The whole set of CTCF anchors was used as background. The GO terms in 'Molecular Functions' with P-value<1e-4 in each cell-type were shown.

References

1. Bickmore WA. The spatial organization of the human genome. *Annu Rev Genomics Hum Genet* **14**, 67-84 (2013).

2. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet* **14**, 390-403 (2013).

3. Dixon JR, Gorkin DU, Ren B. Chromatin Domains: The Unit of Chromosome Organization. *Mol Cell* **62**, 668-680 (2016).

4. Dixon JR*, et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331-336 (2015).

5. Gorkin DU, Leung D, Ren B. The 3D genome in transcriptional regulation and pluripotency. *Cell Stem Cell* **14**, 762-775 (2014).

6. Ji X*, et al.* 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell Stem Cell* **18**, 262-275 (2016).

7. Bonev B, Cavalli G. Organization and function of the 3D genome. *Nat Rev Genet* **17**, 772 (2016).

8. Dixon JR*, et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380 (2012).

9. Rao SS*, et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680 (2014).

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

11. Huang J, Marco E, Pinello L, Yuan GC. Predicting chromatin organization using histone marks. *Genome Biol* **16**, 162 (2015).

12. Ghirlando R, Felsenfeld G. CTCF: making the right connections. *Genes Dev* **30**, 881-891 (2016).

13. Hnisz D, Day DS, Young RA. Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell* **167**, 1188-1200 (2016).

14. Nora EP*, et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944 e922 (2017).

15. Zuin J*, et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc Natl Acad Sci U S A* **111**, 996-1001 (2014).

16. Dowen JM*, et al.* Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* **159**, 374-387 (2014).

17. Hanssen LLP*, et al.* Tissue-specific CTCF-cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. *Nat Cell Biol* **19**, 952-961 (2017).

18. Hnisz D*, et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454-1458 (2016).

19. Narendra V*, et al.* CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* **347**, 1017-1021 (2015).

20. Ren G*, et al.* CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. *Mol Cell* **67**, 1049-1058 e1046 (2017).

21. Hou C, Dale R, Dean A. Cell type specificity of chromatin organization mediated by CTCF and cohesin. *Proc Natl Acad Sci U S A* **107**, 3651-3656 (2010).

22. Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K. Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res* **19**, 24-32 (2009).

23. Wang H*, et al.* Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res* **22**, 1680-1688 (2012).

24. Lieberman-Aiden E*, et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293 (2009).

25. Li G*, et al.* Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. *BMC Genomics* **15 Suppl 12**, S11 (2014).

26. Fullwood MJ*, et al.* An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58-64 (2009).

27. Mumbach MR*, et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13**, 919-922 (2016).

28. Fang R*, et al.* Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res* **26**, 1345-1348 (2016).

29. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

30. Li G, Chen Y, Snyder MP, Zhang MQ. ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. *Nucleic Acids Res* **45**, e4 (2017).

31. Jialiang Huang KL, Wenqing Cai, Xin Liu, Yuannyu Zhang, Stuart H.Orkin, Jian Xu, Guo-Cheng Yuan. Dissecting super-enhancer hierarchy based on chromatin interactions. In: *bioRxiv(doi:10.1101/149583)* (ed^(eds) (2017).

32. Hnisz D*, et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-947 (2013).

33. Loven J*, et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320-334 (2013).

34. Whyte WA*, et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319 (2013).

35. McLean CY*, et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28**, 495-501 (2010).

36. Kramer A, Green J, Pollard J, Jr., Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30**, 523-530 (2014).

671 37. Bolzoni M*, et al.* Myeloma cells inhibit non-canonical wnt co-receptor ror2 expression in human
672 bone marrow osteoprogenitor cells: effect of wnt5a/ror2 pathway activation on the osteogenic
673 differentiation impairment induced by myeloma cells. *Leukemia* **27**, 451-463 (2013).

674
675 38. Yuan Y*, et al.* The Wnt5a/Ror2 noncanonical signaling pathway inhibits canonical Wnt signaling
676 in K562 cells. *Int J Mol Med* **27**, 63-69 (2011).

677
678 39. Ho TK. The random subspace method for constructing decision forests. *Ieee T Pattern Anal* **20**,
679 832-844 (1998).

680
681 40. Wang HH. Pattern classification with random decision forest. *2012 International Conference on*
682 *Industrial Control and Electronics Engineering (Icicee)*, 128-130 (2012).

683
684 41. Xue J, Zhao YX. Random-Forests-based phonetic decision trees for conversational speech
685 recognition. *Int Conf Acoust Spee*, 4169-4172 (2008).

686
687 42. Ay F, Noble WS. Analysis methods for studying the 3D architecture of the genome. *Genome Biol*
688 **16**, 183 (2015).

689
690 43. Whalen S, Truty RM, Pollard KS. Enhancer-promoter interactions are encoded by complex
691 genomic signatures on looping chromatin. *Nat Genet* **48**, 488-496 (2016).

692
693 44. Oti M, Falck J, Huynen MA, Zhou H. CTCF-mediated chromatin loops enclose inducible gene
694 regulatory domains. *BMC Genomics* **17**, 252 (2016).

695
696 45. Kramer O. Scikit-Learn. *Stud Big Data* **20**, 45-53 (2016).

697
698 46. Sandhu KS*, et al.* Large-scale functional organization of long-range chromatin interaction
699 networks. *Cell Rep* **2**, 1207-1219 (2012).

700
701 47. Majumder P, Gomez JA, Chadwick BP, Boss JM. The insulator factor CTCF controls MHC class II
702 gene expression and is required for the formation of long-distance chromatin interactions. *J Exp*
703 *Med* **205**, 785-798 (2008).

704
705 48. Majumder P, Boss JM. CTCF controls expression and chromatin architecture of the human major
706 histocompatibility complex class II locus. *Mol Cell Biol* **30**, 4211-4223 (2010).

707
708 49. Zhang Y*, et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

709

50. Trapnell C*, et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562-578 (2012).

51. Mathelier A*, et al.* JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **44**, D110-115 (2016).

52. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017-1018 (2011).

53. Rosenbloom KR*, et al.* The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res* **43**, D670-681 (2015).

54. Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics* **25**, 1952-1958 (2009).

55. Pedregosa F*, et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**, 2825-2830 (2011).

56. Naumova N, Smith EM, Zhan Y, Dekker J. Analysis of long-range chromatin interactions using Chromosome Conformation Capture. *Methods* **58**, 192-203 (2012).

Acknowledgement

We thank Dr. Michael Beer for valuable discussions and suggestions. We thank Nick Waring, Stephanie Perkail and Coen Lap for proof-reading and editing. This research was supported by National Institute of Health (Division of Intramural Research of National Lung and Blood Institute) to Zhu, a Lemonade Stand Foundation Young Investigator Award, and National Cancer Institute grants (R00CA158582, R21CA182662, and R03CA212068) to Tzatsos, a George Washington University Cross-Disciplinary Research Fund to Tzatsos and Peng, and National Institute of Allergy and Infection Diseases grants (R21 AI113806, R01 AI121080) to Peng.

Author Contributions

Y.K., A.T. and W.P. conceived the project. A.T. and W.P. supervised this study. Y.K. and W.P. developed the method and analyzed the results. Y.K. wrote the software. J.A. performed the 3C experiments. Z.Z. and J.Z. contributed to methodology design. Y.K., A.T. and W.P. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Figure 1

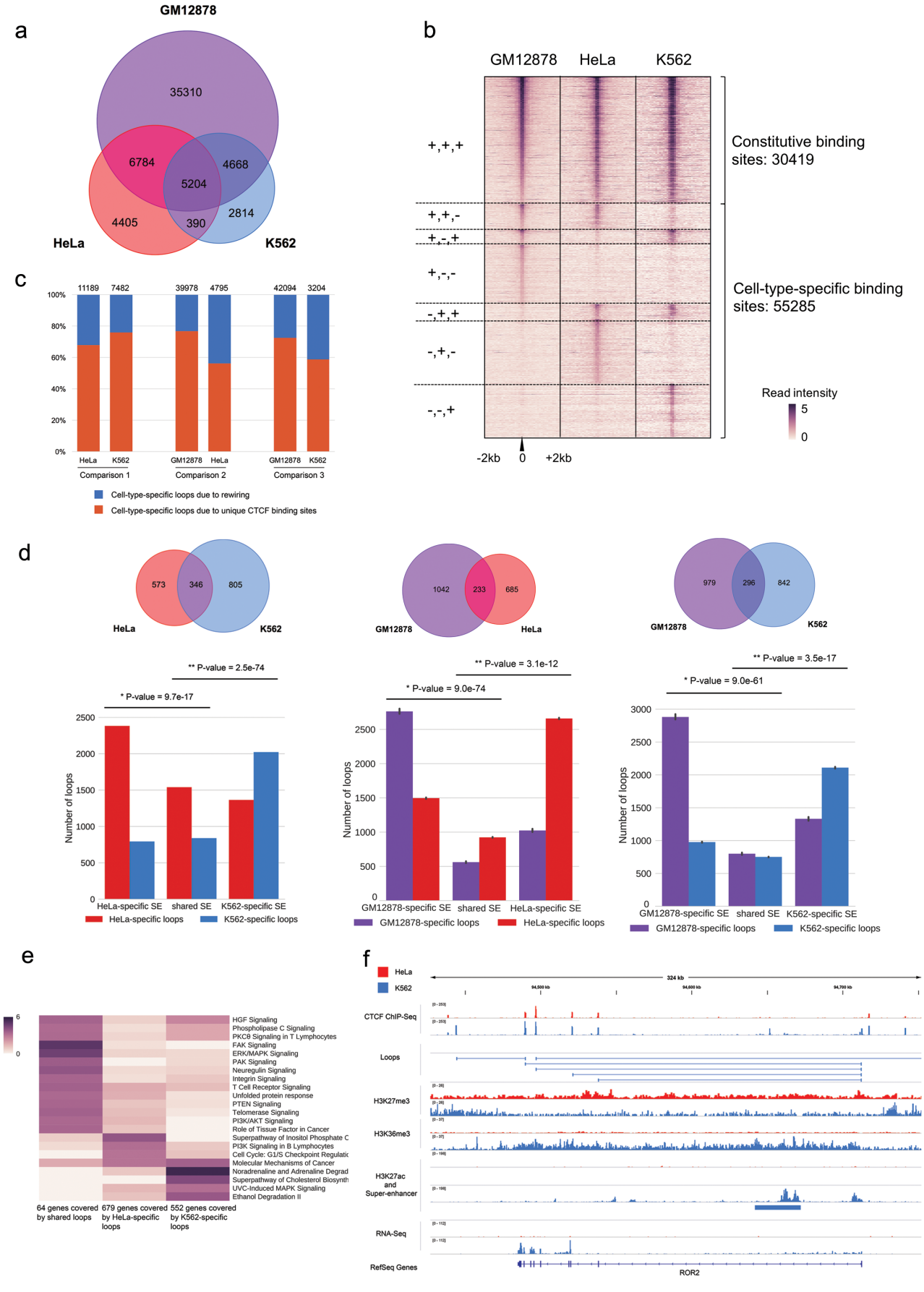


Figure 1. CTCF-mediated loops exhibit cell-type-specificity.

(**a**) Venn diagram of CTCF-mediated loops identified from ChIA-PET experiments in GM12878, HeLa and K562.

(**b**) Heat map of CTCF binding sites in GM12878, HeLa and K562. Each row represents a CTCF binding event identified in ChIA-PET in at least one cell-type. The binding sites are divided into seven groups based on the presence (+) or absence (-) of CTCF binding. Color key shows the log2-transformed value of reads per kilobase per million reads (RPKM).

(**c**) Cell-type-specific CTCF binding and rewiring between common CTCF binding sites contribute to cell-type-specific loops.

(**d**) Cell-type-specific SEs are enriched with cell-type-specific loops. Top: Venn diagram of SEs in pairwise comparison of cell types. Bottom: Number of cell-type-specific loops covering cell-type-specific and shared SEs. P-values were calculated by Chi-square test. The GM12878 ChIA-PET dataset was down-sampled to 15% of the original size so that the number of identified loops matched those of the other ChIA-PET datasets. The down sampling and further analysis was repeated 10 times and the standard-deviations were shown.

(**e**) Canonical pathway enrichment analysis of differentially expressed genes associated with K562-specific, HeLa-specific and shared CTCF-mediated loops, respectively.
Color represents the -log10 (P-value).

(**f**) Genome browser snapshot of ROR2 locus. ROR2 is expressed and associated with CTCF-mediated loops in K562 but not in HeLa. Expression of ROR2 in K562 is associated with a concomitant decrease of H3K27me3 and increase of H3K36me3 within the gene body, as well as the appearance of a K562-specific SE. The ChIP-Seq and RNA-seq signals are represented in RPKM values.

Figure 2

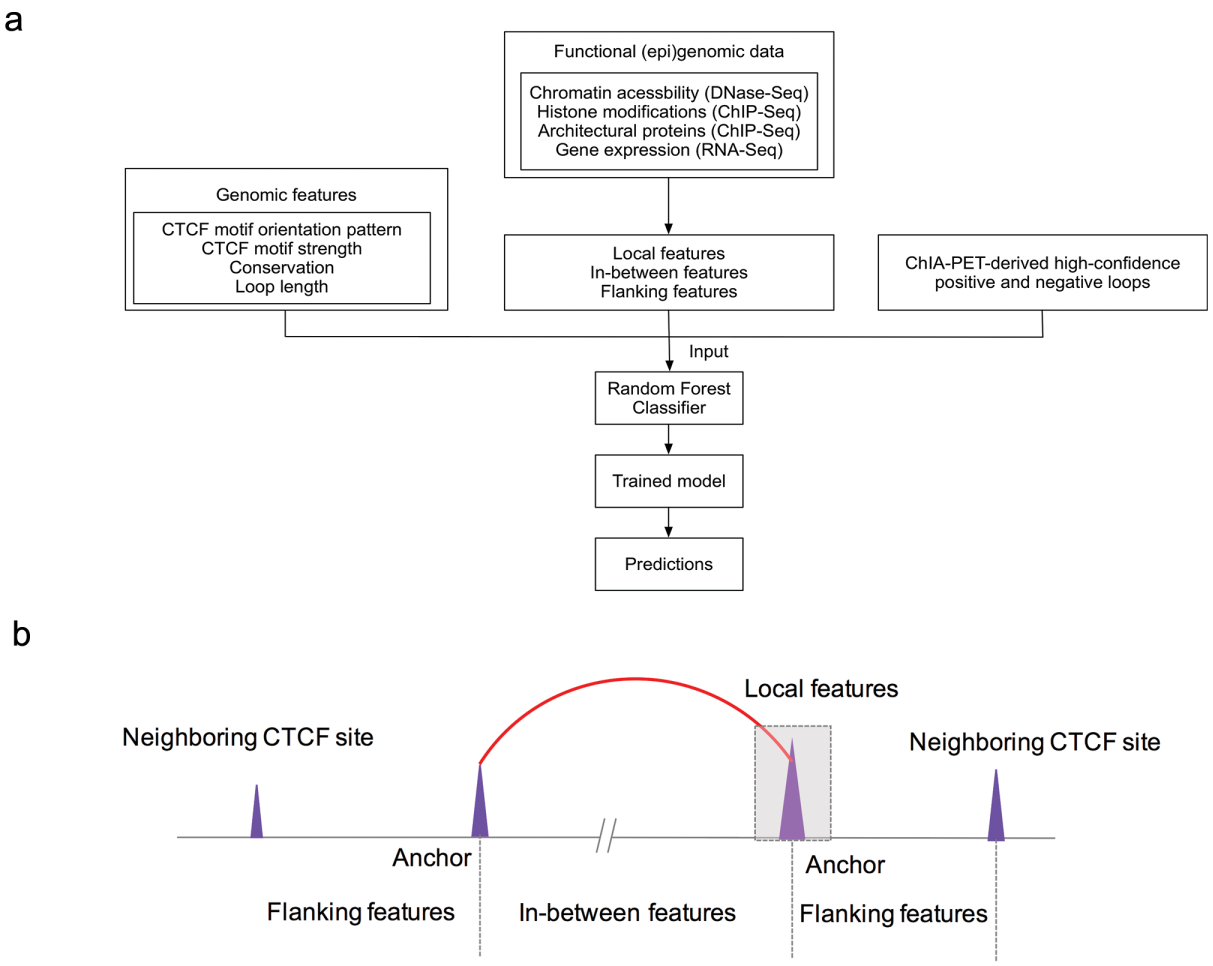


Figure 2. Illustration of the Lollipop pipeline and types of features.

(**a**) Schematic of the Lollipop pipeline. In training data, positive loops were generated from high-confidence interactions identified from ChIA-PET, and negative loops were random pairs of CTCF binding sites without interactions in ChIA-PET or significant contact in Hi-C dataset. A diverse set of features, generated from genomic and epigenomic data, was used to characterize the interactions. A random forests classifier distinguished interacting CTCF binding sites from non-interacting ones. The performance of resulting classifier was then evaluated. Trained model can be used to scan the genome and predict de novo CTCF-mediated loops in the same or a different cell-type.

(**b**) Illustration of local, in-between, and flanking features.

Figure 3

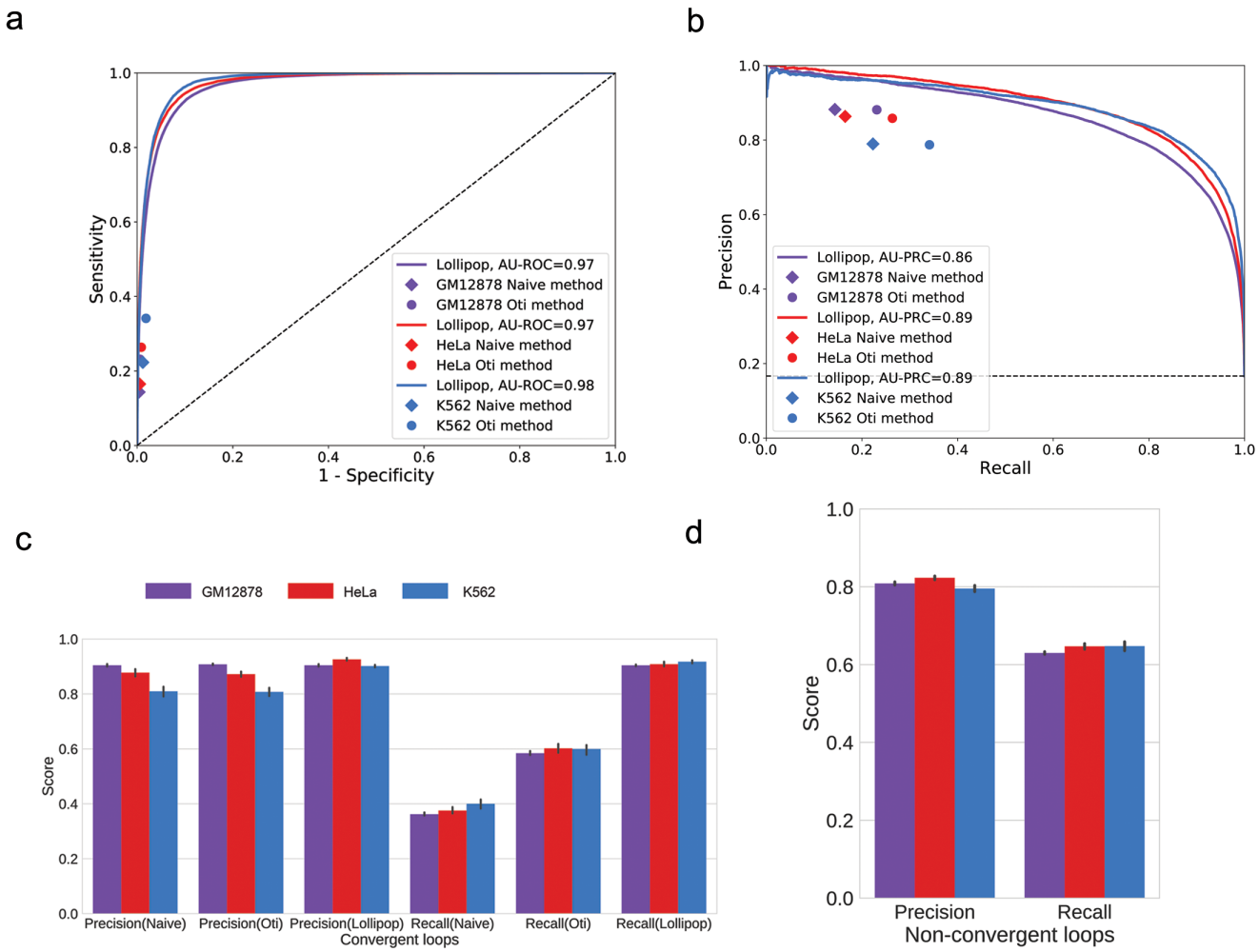


Figure 3. Performance evaluation within individual cell types

(**a,b**) Performance evaluation using (A) Receiver Operating Characteristic (ROC) and (B) Precision-Recall (PR) curve. Performance of the naïve and Oti methods are represented by diamonds and circles, respectively. Results in GM12878, HeLa and K562 are shown in purple, red and light blue, respectively.

(**c**) Comparison of the precision and the recall of the three methods in predicting convergent loops.

(**d**) Evaluation of Lollipop's performance on non-convergent loops, which include tandem loops, divergent loops and loops without CTCF motifs in the anchors.

Figure 4

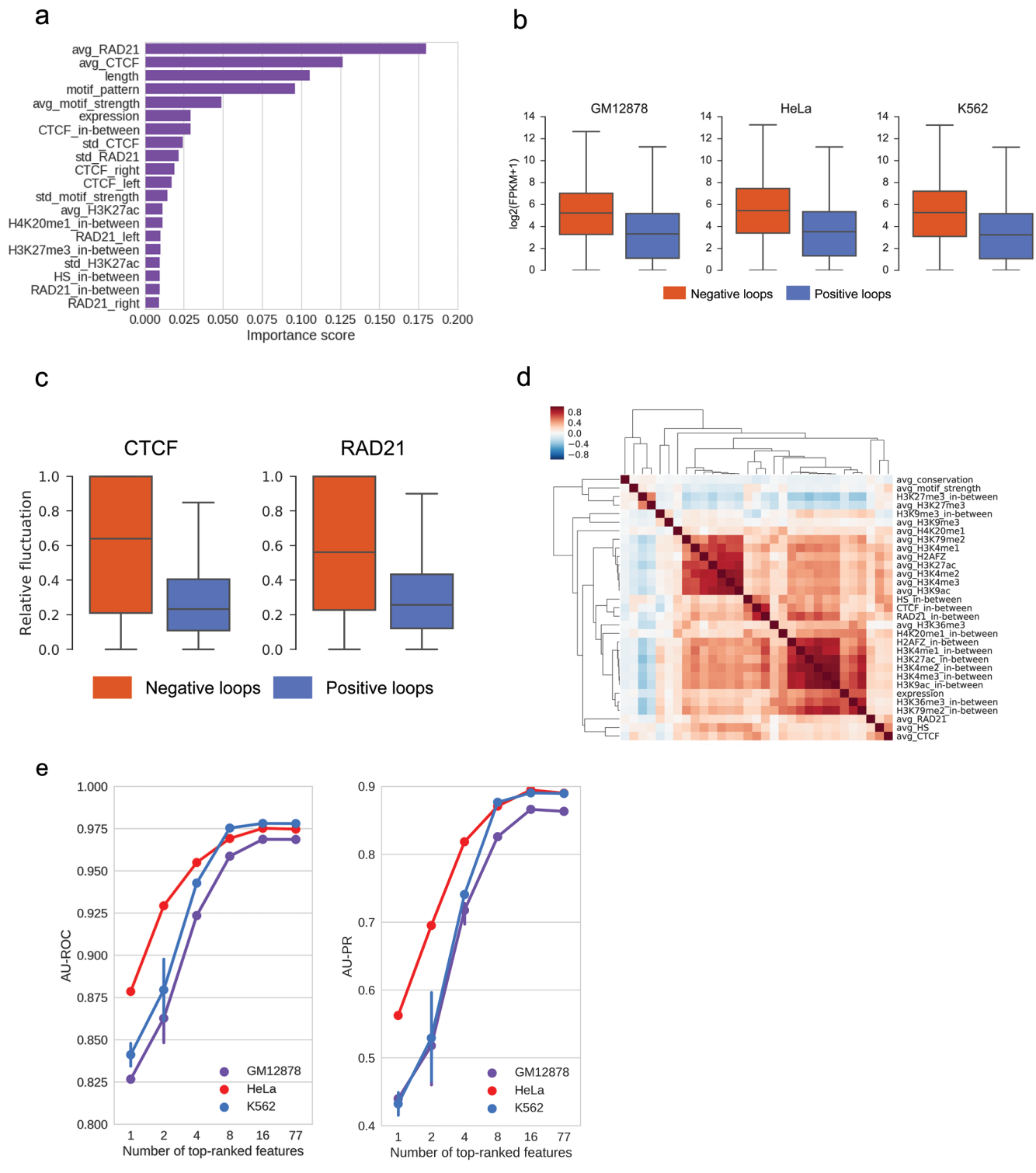


Figure 4. Feature analysis identified novel determinants of CTCF-mediated chromatin loops.

(**a**) Ranking of predictive importance of the top 20 features in the model trained in GM12878 cells. Predictive importance is measured by mean decrease impurity in the training process. 'avg' and 'std' represent the mean and standard deviation of the signal intensity on both anchors. '_left' and '_right' represent the flanking features while '_in-between' is the signal intensity within the loop.

(**b**) Distributions of average gene expression levels within negative and positive loops. The positive and negative loops were defined in the training data, with those without any promoters inside the loops excluded in this analysis. In all three cases, P-value was < 1e-300 using Mann-Whitney U test.

(**c**) Distribution of the relative fluctuations of CTCF and RAD21 binding intensities on paired anchors of negative and positive loops in GM12878 cells. Relative fluctuation was defined as the ratio of standard deviation to mean intensity of anchor pairs. In both cases, P-value was < 1e-300 using Mann-Whitney U test.

(**d**) Heatmap of feature correlations in GM12878. On anchors, active histone marks are highly correlated. Along the loop regions, active histone marks and expression exhibit strong correlation. In addition, RAD21, CTCF and DNase hypersensitive sites are strongly correlated. Spearman's rank correlation and hierarchical clustering were used.

(**e**) Recursive Feature Elimination analysis on feature reduction. Left: AU-ROC; Right: AU-PR.

Figure 5

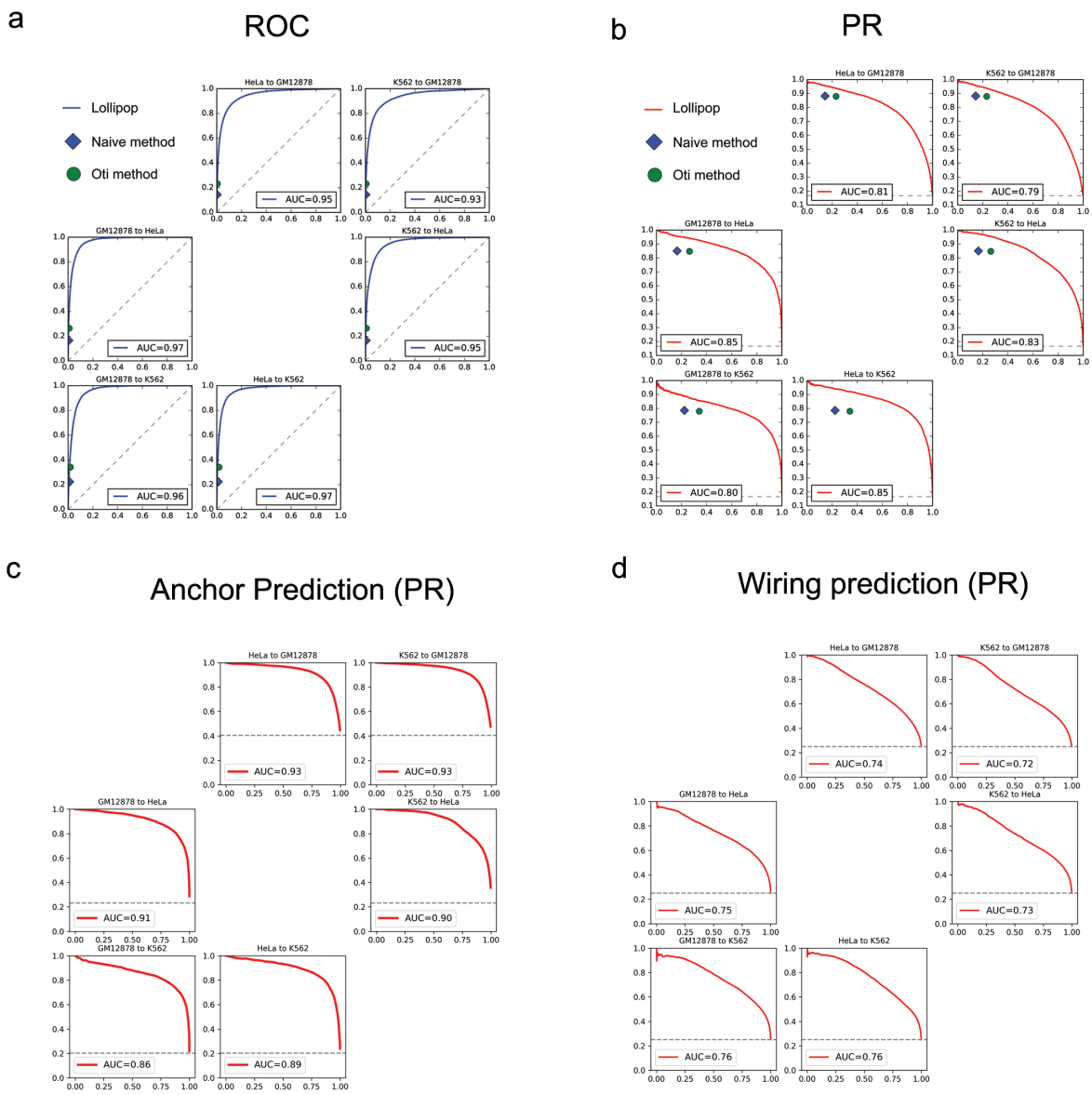


Figure 5. Assessment of Lollipop's performance across cell-types.

(**a, b**) Across-cell-type performance evaluation using (**A**) ROC and (**B**) PR curves. In each subplot, 'cell A to cell B', applies the model trained from cell-type A to the data of cell-type B. For comparison, the performance of the naïve and Oti methods in each cell-type were represented by diamonds and circles, respectively.

(**c**) Performance evaluation of anchor prediction using PR curve.

(**d**) Performance evaluation of wiring prediction using PR curve.

The dash lines in (a-d) represent baseline performance.

Figure 6

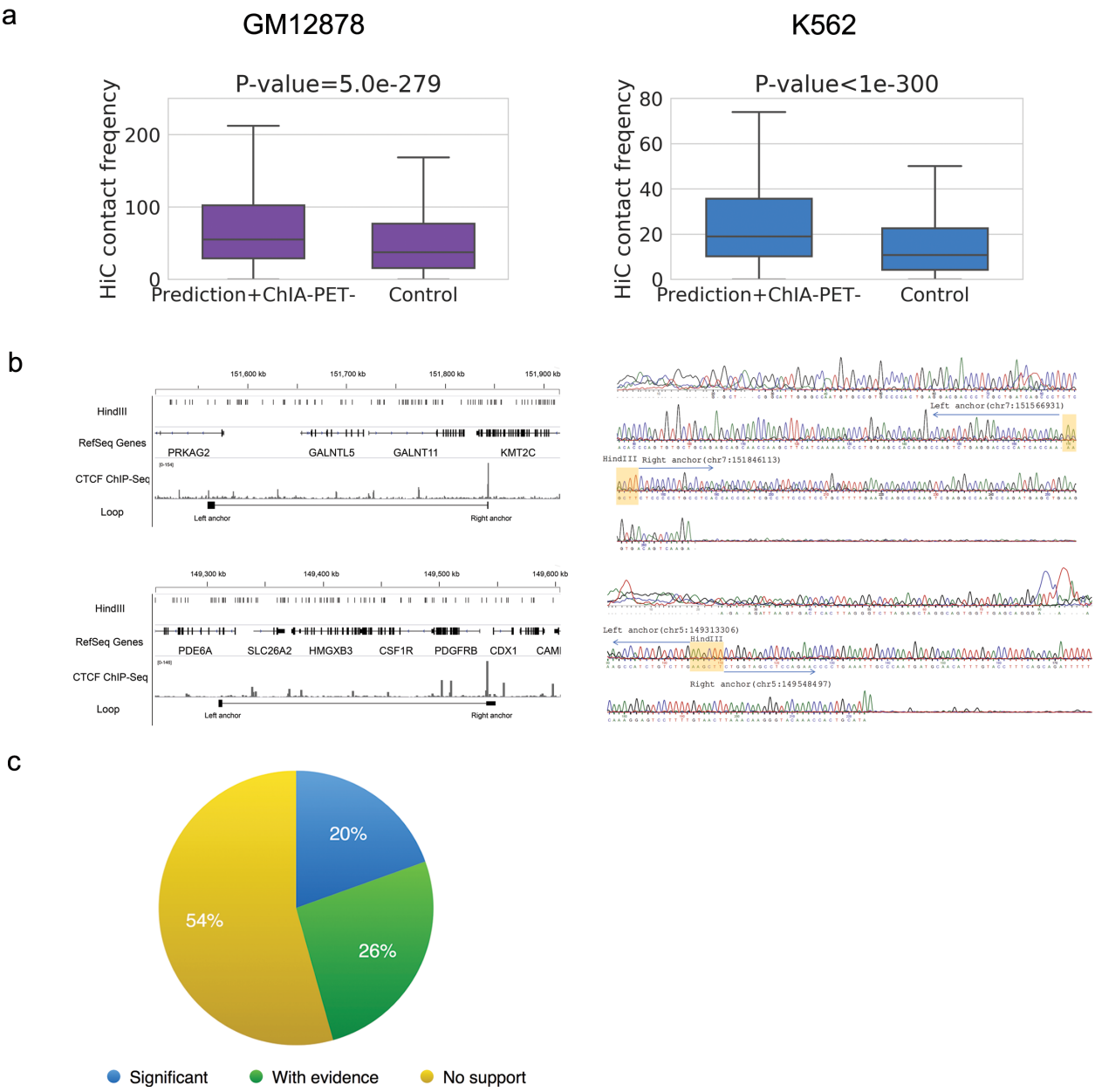


Figure 6. Validation of predicted CTCF-mediated interactions.

(**a**) CTCF-mediated loops predicted by Lollipop but lacking ChIA-PET support exhibit significantly higher contact frequency than background in Hi-C experiments. P-values were calculated using Mann-Whitney U test.

(**b**) Validation of two loops predicted by Lollipop, but not present in the HeLa ChIA-PET data set. Left: schematic of PRKAG2-KMT2C (chr7:151560677-151843260; top) and PDE6A-PDGFRB loop (chr5:149312517-149547724; bottom). Right: Sanger sequencing confirmation of the ligation junctions. Shaded areas in the right panels indicate the HindIII ligation junctions.

(**c**) Scaling analysis of loop prediction. Loops predicted using a model trained on the down-sampled (to 15%) GM12878 library, but lacks support in the down-sampled library (i.e., the yellow slice in **Supplementary Fig. 5b**) are evaluated by the full ChIA-PET data. 46% of these loops find support.

Figure 7

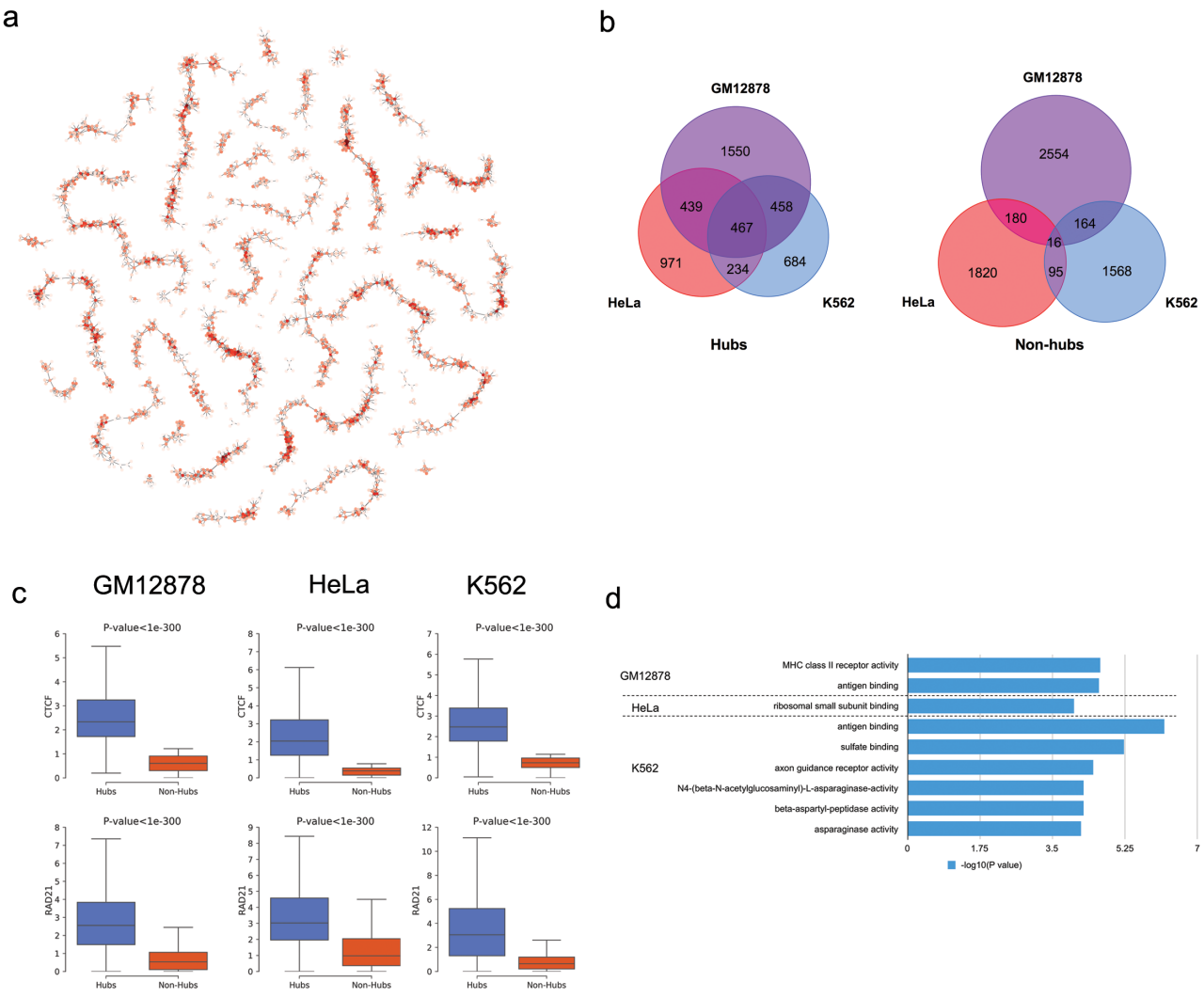


Figure 7. Topological properties of the CTCF-mediated interaction network and their association with biological functions.

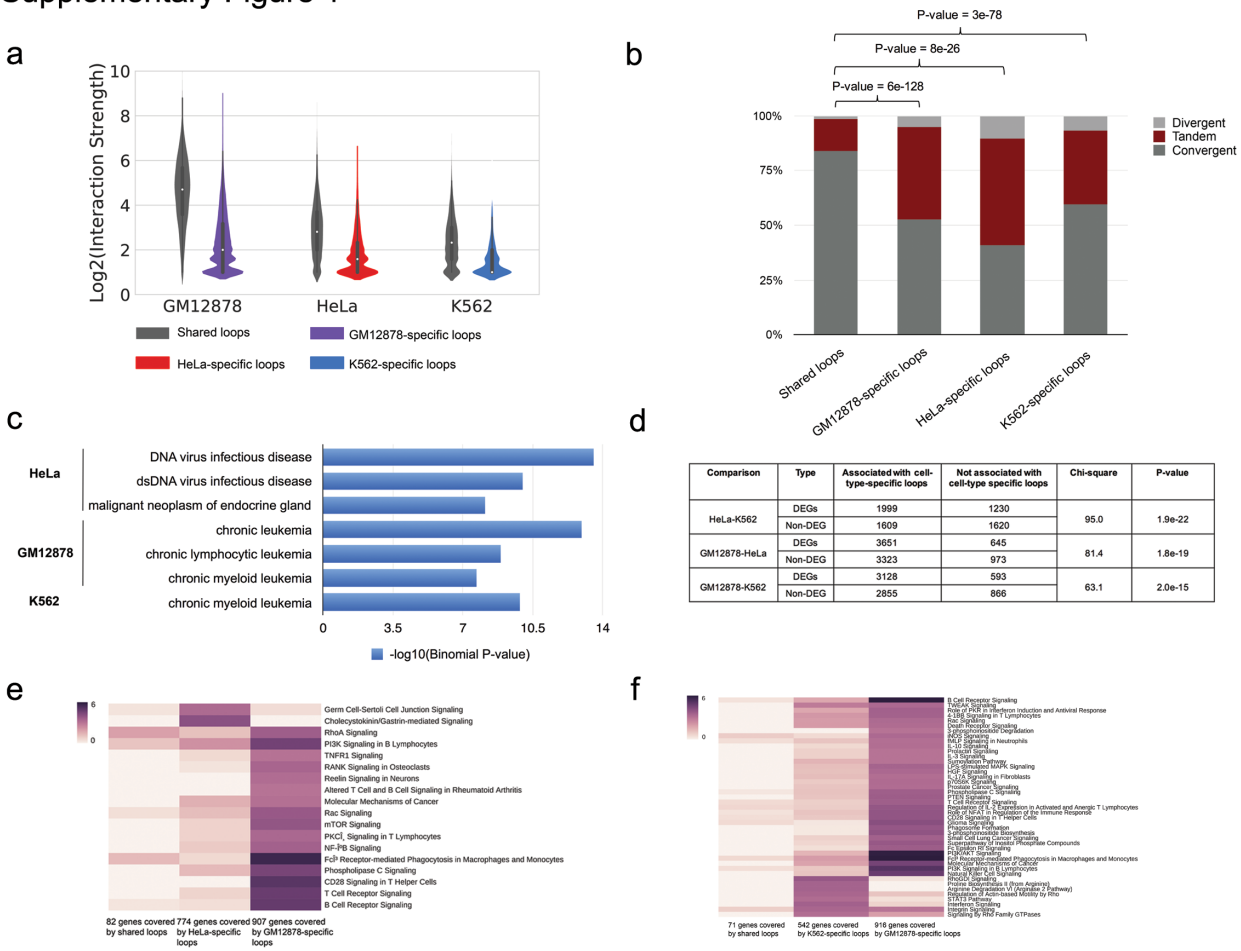
(**a**) Visualization of the CTCF-mediated interaction network of chromosome 1 in GM12878 cells. Each node represents an anchor, with color representing degree of connection. Each edge represents an interaction.

(**b**) Overlap of predicted hubs and non-hubs among each cell-type. Hubs are more conserved than non-hubs.

(c) Distribution of the binding affinity of architectural proteins, CTCF (top) and RAD21 (bottom), on predicted hubs and non-hubs.

(d) Functional enrichment analysis of hubs using GREAT. The x-axis represents the binomial P-values.

Supplementary Figure 1



Comparison	Type	Associated with cell-type-specific loops	Not associated with cell-type specific loops	Chi-square	P-value
HeLa-K562	DEGs	1999	1230	95.0	1.9e-22
	Non-DEG	1609	1620		
GM12878-HeLa	DEGs	3651	645	81.4	1.8e-19
	Non-DEG	3323	973		
GM12878-K562	DEGs	3128	593	63.1	2.0e-15
	Non-DEG	2855	866		

Supplementary Figure 1. CTCF-mediated loops exhibit cell-type-specificity.

(**a**) Violin plots show that shared CTCF-mediated loops are stronger than cell-type-specific loops. Interaction strength is defined as the number of Paired-End Tags (PETs) connecting the anchors.

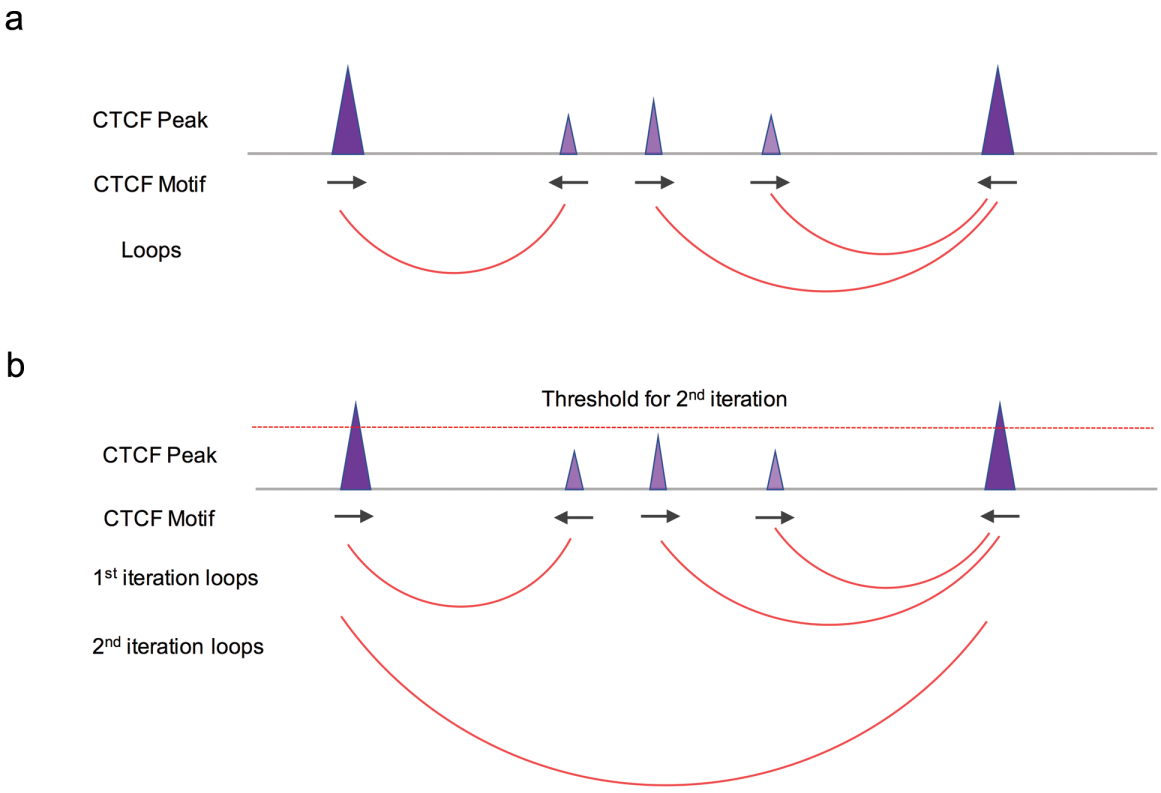
(**b**) Stacked bar plot comparing the pattern of motif orientation between cell-type-specific and shared loops. The P-values were calculated using Chi-square test.

(**c**) Disease Ontology analysis of SEs using GREAT reveals the disease origin of the three cell-types.

(**d**) Contingency table for the number of loops associated with DEGs and Non-DEGs among the three cell lines. Pair-wise comparison was shown.

(**e-f**) Canonical pathway enrichment analysis of DEGs associated with cell-type-specific and shared loops in (**e**) HeLa-GM12878 and (**f**) K562-GM12878 comparison. Color represents the -log10 (P-value).

Supplementary Figure 2

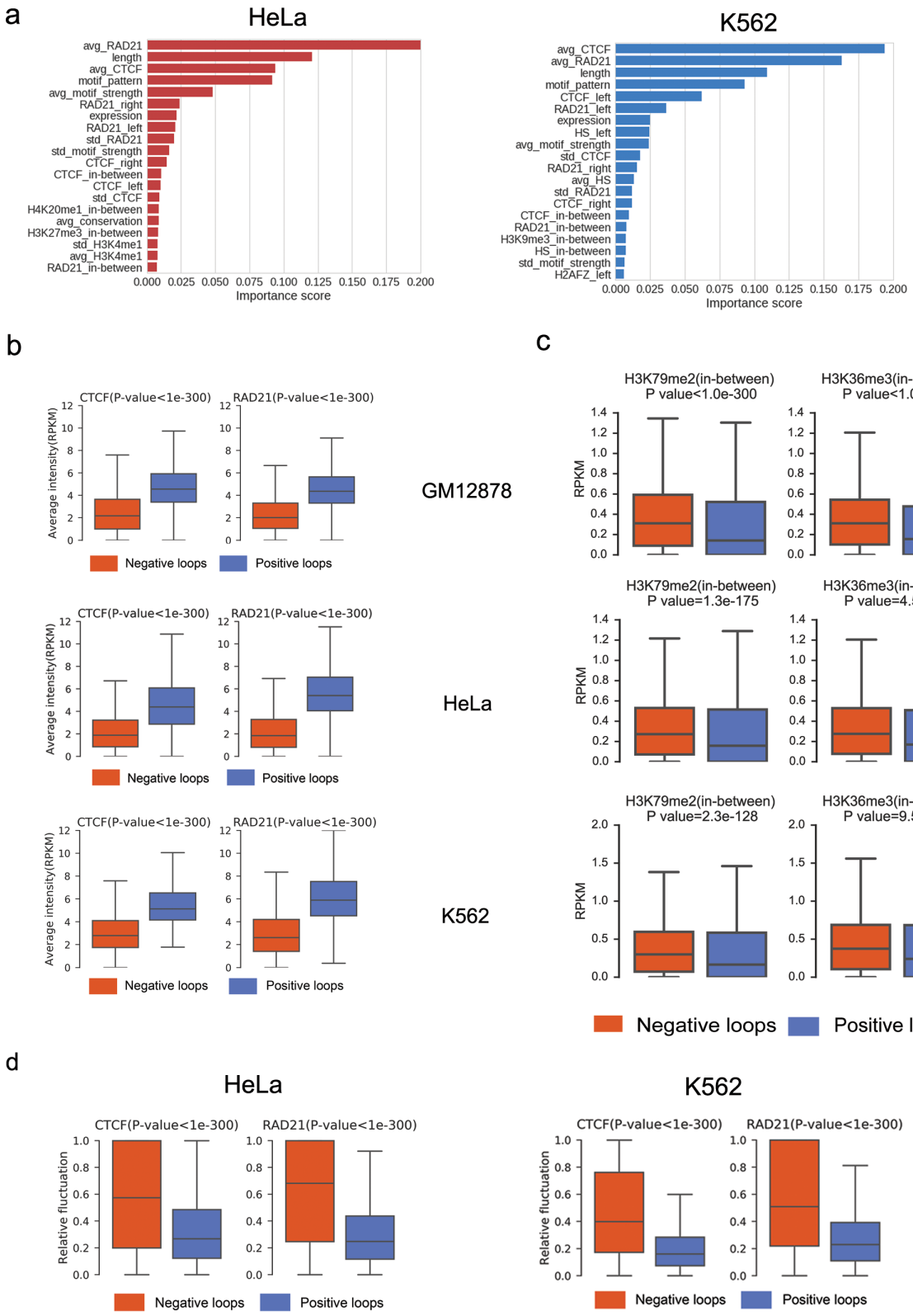


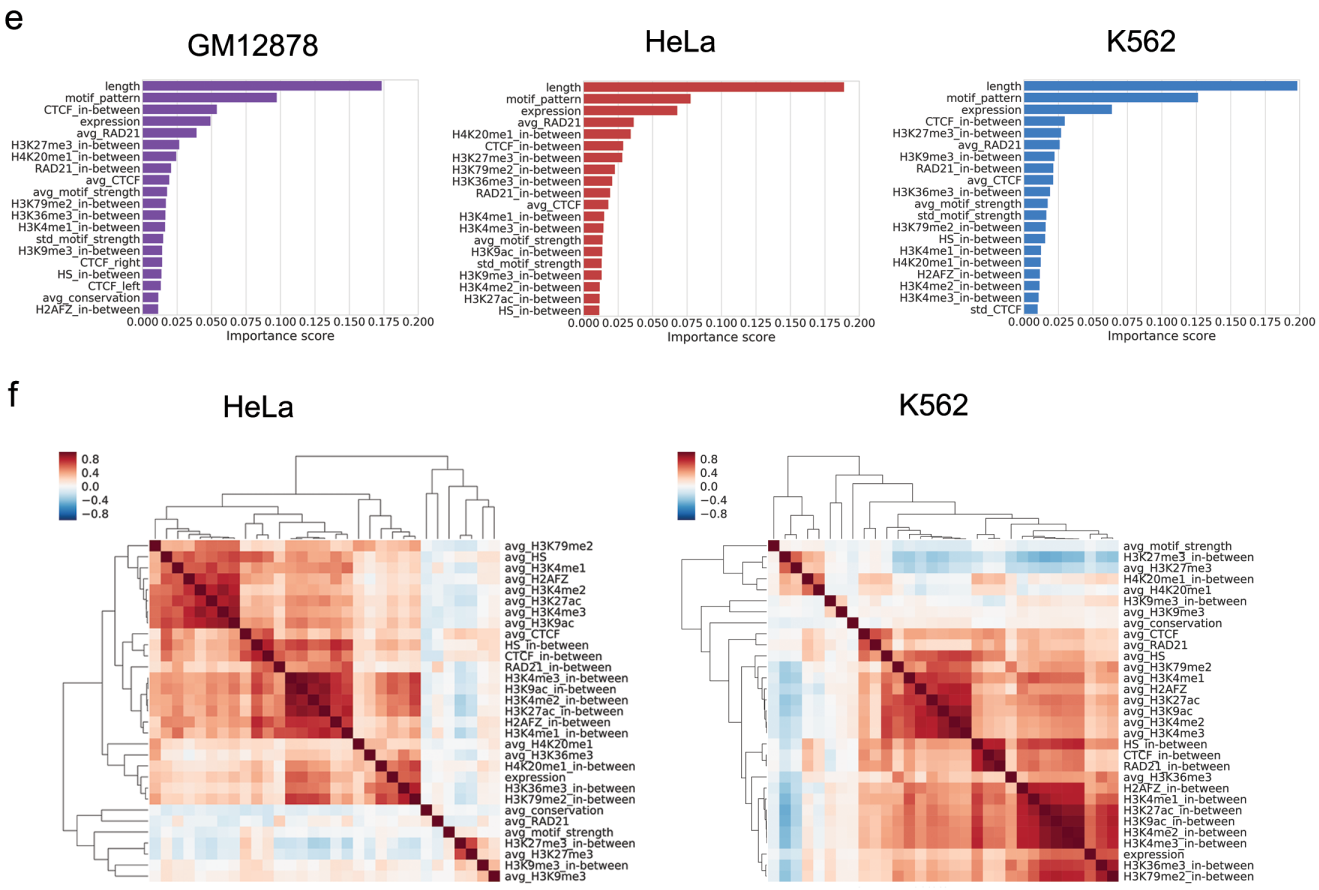
Supplementary Figure 2. Illustration of the naïve and Oti method.

(**a**) Illustration of the naïve method. This method pairs a CTCF-bound motif that resides on the forward strand to the nearest downstream CTCF-bound motif that resides on the reverse strand.

(**b**) Illustration of the Oti method. It constructed loops in iterations by increasing the threshold of CTCF binding intensity. In each iteration, CTCF-bound motifs whose binding intensity are above the threshold were chosen, and naïve method was applied to construct loops. The loops constructed in different iterations were pooled together for the eventual result.

Supplementary Figure 3





Supplementary Figure 3. Results of feature analysis in K562 and Hela cells are consistent with those in GM12878.

(**a**) Ranking the predictive importance of the top 20 features in the model trained in HeLa and K562.

(**b**) Distributions of the average binding intensity of CTCF and RAD21 on anchors in negative and positive loops in the three cell lines. P-values were calculated using Mann-Whitney U test.

(**c**) Distributions of the intensity of the indicated histone marks within negative and positive loops.

(**d**) Distributions of relative fluctuations of CTCF and RAD21's binding intensities on paired anchors of negative and positive loops in HeLa and K562 cells. Relative fluctuation was defined as the ratio of standard deviation to average value.

(**e**) Ranking the predictive importance of the top 20 features in wiring prediction. The model was trained in the rewiring data (see methods for details) of GM12878, HeLa and K562 cells, respectively.

(**f**) Heatmaps of feature correlation in HeLa and K562 cells.

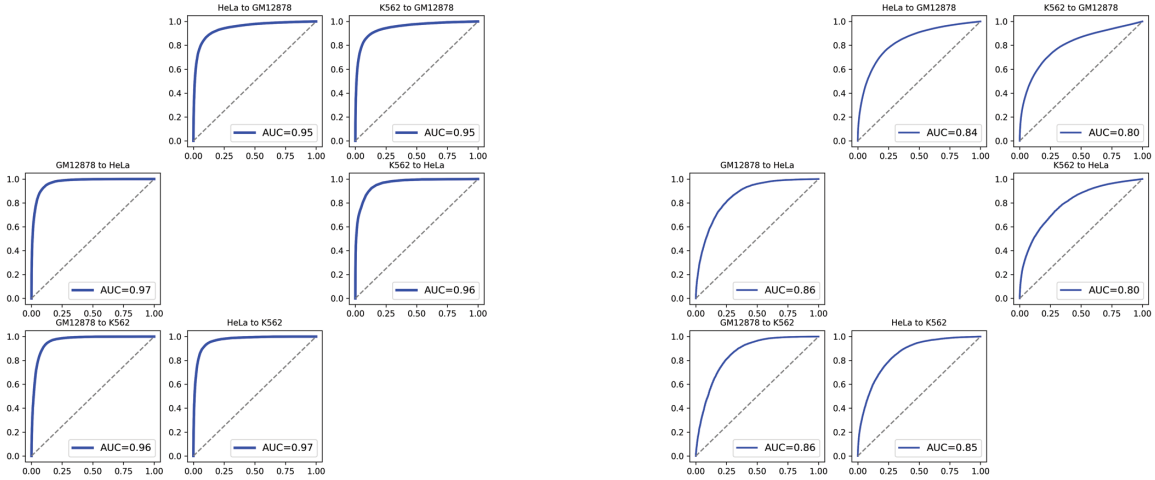
Supplementary Figure 4

a

Anchor prediction (ROC)

b

Wiring prediction (ROC)

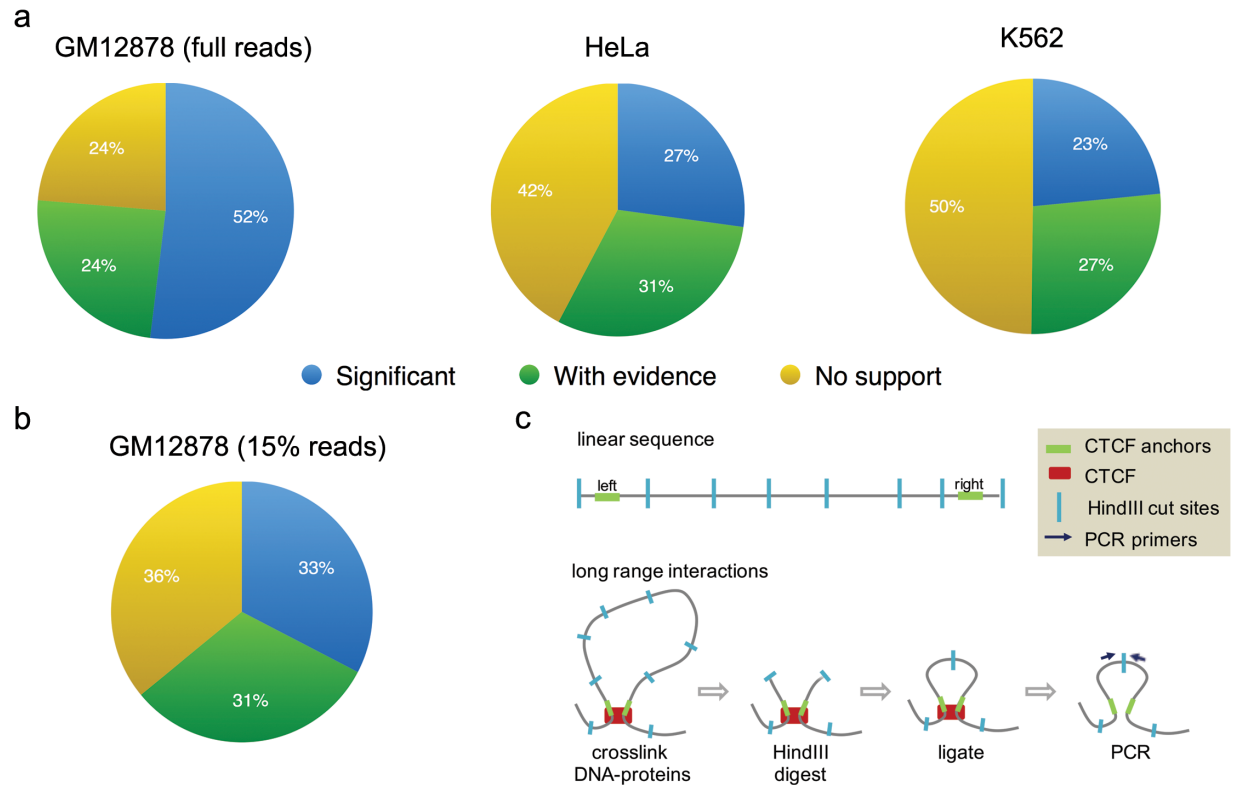


Supplementary Figure 4. Assessment of Lollipop's performance across cell-types.

(**a**) Performance evaluation of anchor prediction using ROC curve.

(**b**) Performance evaluation of wiring prediction using ROC curve.

Supplementary Figure 5



Supplementary Figure 5. Validation of predicted CTCF-mediated interactions

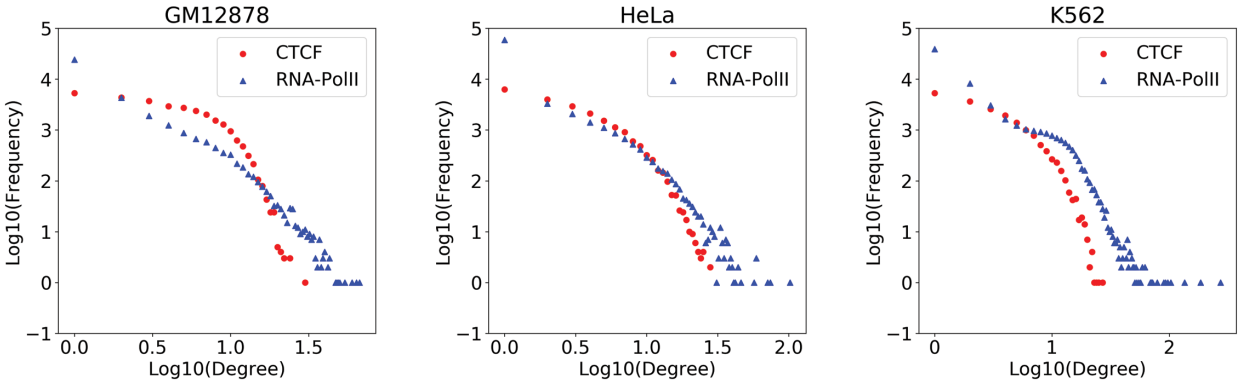
(**a**) The distribution of de novo predicted loops, as compared to original ChIA-PET data. 'Significant' denotes loops with FDR <=0.05 and PET number >=2 in ChIA-PET. 'With evidence' denotes predicted loops with less-significant ChIA-PET support (i.e., FDR > 0.05 or PET =1). 'No support' denotes predicted loops without any ChIA-PET support.

(**b**) The distribution of predicted loops using a down-sampled (to 15%) GM12878 library for model building, followed by genome-wide prediction and comparison with ChIA-PET data. The number of loops observed in the downscaled GM12878 library is similar to those of K562 and HeLa (see **Supplementary Table 2**).

(**c**) Illustration demonstrating the major steps of 3C experiments.

(**d**) 3C-qPCR analysis shows the relative abundance of PRKAG2 anchor to KMT2C anchor and adjacent HindIII fragments (**Fig. 6b** top panel). Tracks from top to bottom: HindIII cut sites, designed primer for testing interaction, CTCF ChIP-Seq, motif occurrences, and relative quantification of the 3C interaction.

Supplementary Figure 6



Supplementary Figure 6. The connection degree distribution for the CTCF- and RNA-PolII-mediated interaction network. De novo predictions from Lollipop were used for the CTCF network, whereas loops identified from RNA-PolII ChIA-PET were used for the RNA-PolII network.

Supplementary Table 1: Used data sets

Data	GM12878	HeLa	K562
ChIA-PET	GSE72816[4]	GSE72816[4]	ENCLB559JAA[1,2]
Hi-C	GSE63525[3]		GSE63525[3]
DNase-Seq	ENCFF000SKV[1,2]	ENCFF000SPJ[1,2]	ENCFF000SVI[1,2]
RNA-Seq	ENCFF000FBU[1,2] ENCFF000FBV[1,2]	ENCFF158RCK[1,2] ENCFF169ZTB[1,2]	GSM765393[1,2]
ChIP-Seq (CTCF)	ENCFF000ARG[1,2]	ENCFF000BAJ[1,2]	ENCFF000YLT[1,2]
ChIP-Seq (RAD21)	ENCFF000OBV[1,2]	ENCFF000XKH[1,2]	ENCFF084HTD[1,2]
ChIP-Seq (H2AZ)	ENCFF001SUD[1,2]	ENCFF000BAX[1,2]	ENCFF000BWO[1,2]
ChIP-Seq (H3K4me1)	ENCFF000ARY[1,2]	ENCFF000BBA[1,2]	ENCFF000BXK[1,2]
ChIP-Seq (H3K4me2)	ENCFF000ATG[1,2]	ENCFF000BCH[1,2]	ENCFF000BXT[1,2]
ChIP-Seq (H3K4me3)	ENCFF000ATS[1,2]	ENCFF000BCO[1,2]	ENCFF000BXW[1,2]
ChIP-Seq (H3K9ac)	ENCFF000ATY[1,2]	ENCFF000BCW[1,2]	ENCFF000BYK[1,2]
ChIP-Seq (H3K9me3)	ENCFF000AUH[1,2]	ENCFF000BBG[1,2]	ENCFF000BYT[1,2]
ChIP-Seq (H3K27ac)	ENCFF000ASI[1,2]	ENCFF000BBN[1,2]	ENCFF000BWZ[1,2]
ChIP-Seq (H3K27me3)	ENCFF000ASK[1,2]	ENCFF000BBS[1,2]	ENCFF000BXA[1,2]
ChIP-Seq (H3K36me3)	ENCFF000ASX[1,2]	ENCFF000BCC[1,2]	ENCFF000BXE[1,2]
ChIP-Seq (H3K79me2)	ENCFF000ATT[1,2]	ENCFF000BCQ[1,2]	ENCFF000BYC[1,2]
ChIP-Seq (H4K20me1)	ENCFF000AUT[1,2]	ENCFF000BDC[1,2]	ENCFF001QWY[1,2]
ChIP-Seq Input	ENCFF000AQZ[1,2]	ENCFF000BAI[1,2]	ENCFF000BVZ[1,2]
ChIP-Seq Input	ENCFF651WEV[1,2]	ENCFF469INX[1,2]	ENCFF000QEK[1,2]

[1]Consortium, E. P. (2012). "An integrated encyclopedia of DNA elements in the human genome." Nature **489**(7414): 57-74.

[2]Sloan, C. A., E. T. Chan, J. M. Davidson, V. S. Malladi, J. S. Strattan, B. C. Hitz, I. Gabdank, A. K. Narayanan, M. Ho, B. T. Lee, L. D. Rowe, T. R. Dreszer, G. Roe, N. R. Podduturi, F. Tanaka, E. L. Hong and J. M. Cherry (2016). "ENCODE data at the ENCODE portal." Nucleic Acids Res **44**(D1): D726-732.

[3]Rao, S. S., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander and E. L. Aiden (2014). "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." Cell **159**(7): 1665-1680.

[4]Tang, Z., O. J. Luo, X. Li, M. Zheng, J. J. Zhu, P. Szalaj, P. Trzaskoma, A. Magalska, J. Wlodarczyk, B. Ruszczycki, P. Michalski, E. Piecuch, P. Wang, D. Wang, S. Z. Tian, M. Penrad-Mobayed, L. M. Sachs, X. Ruan, C. L. Wei, E. T. Liu, G. M. Wilczynski, D. Plewczynski, G. Li and Y. Ruan (2015). "CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription." Cell **163**(7): 1611-1627.

Supplementary Table 2: Analysis results of ChIA-PET data sets

Cell-type	Raw reads (in million)	Unique PETs (in million)	IAB >= 2 loops	FDR <= 0.05 loops	IAB >=2 and FDR <= 0.05 loops
GM12878 (full reads)	680	39.8	93914	73511	51966
GM12878 (15% reads)	102	13.1	37125	22248	15569
HeLa	531	21.1	42430	25047	16783
K562	195	6.6	23884	23377	13076

Supplementary Table 3: Top-ranked features from the Recursive Feature Elimination analysis

*Numbers inside the parentheses indicate the times of top-ranked feature set appears.
*'avg' and 'std' represent the mean and standard deviation of the signal intensity on both anchors. '_left' and '_right' represent the flanking features while '_in-between' means the signal intensity in the loop region.

	GM12878
Top 1 feature	avg_RAD21 (5)
Top 2 features	avg_CTCF, avg_RAD21 (3) motif_pattern, avg_RAD21 (2)
Top 4 features	length, motif_pattern, avg_CTCF, avg_RAD21 (4) length, avg_CTCF, CTCF_in-between, avg_RAD21 (1)
Top 8 features	length, motif_pattern, avg_motif_strength, avg_CTCF, std_CTCF, CTCF_in-between, avg_RAD21, expression (5)
Top 16 features	length, motif_pattern, avg_motif_strength, HS_in-between, avg_H3K4me1, avg_H3K27ac, avg_CTCF, std_CTCF, CTCF_in-between, CTCF_left, CTCF_right, avg_RAD21, std_RAD21, RAD21_in-between, RAD21_left, expression (2) length, motif_pattern, avg_motif_strength, HS_in-between, HS_left, avg_H3K4me1, avg_H3K27ac, avg_CTCF, std_CTCF, CTCF_in-between, CTCF_left, CTCF_right, avg_RAD21, std_RAD21, RAD21_in-between, expression (3)
	HeLa
Top 1 feature	avg_RAD21 (5)
Top 2 features	length, avg_RAD21 (5)
Top 4 features	length, motif_pattern, avg_CTCF, avg_RAD21 (5)
Top 8 features	length, motif_pattern, avg_motif_strength, avg_CTCF, CTCF_in-between, avg_RAD21, std_RAD21, expression (3) length, motif_pattern, avg_motif_strength, avg_H3K4me1, avg_CTCF, avg_RAD21, std_RAD21, expression (2)
Top 16 features	length, motif_pattern, avg_motif_strength, avg_HS, avg_H3K4me1, avg_CTCF, std_CTCF, CTCF_in-between, CTCF_left, CTCF_right, avg_RAD21, std_RAD21, RAD21_in-between, RAD21_left, RAD21_right, expression (5)
	K562
Top 1 feature	avg_CTCF (3) avg_RAD21 (2)
Top 2 features	length, avg_CTCF (1) avg_CTCF, avg_RAD21 (3) length, avg_RAD21 (1)
Top 4 features	length, avg_CTCF, CTCF_left, avg_RAD21 (5)
Top 8 features	length, motif_pattern, avg_CTCF, std_CTCF, CTCF_in-between, CTCF_left, avg_RAD21, expression (3) length, motif_pattern, avg_CTCF, std_CTCF, CTCF_in-between, CTCF_left, avg_RAD21, RAD21_left (2) length, motif_pattern, HS_left, avg_CTCF, std_CTCF, CTCF_in-between, CTCF_left, avg_RAD21 (1)
Top 16 features	length, motif_pattern, avg_motif_strength, HS_in-between, HS_left, avg_CTCF, std_CTCF, CTCF_in-between, CTCF_left, CTCF_right, avg_RAD21, std_RAD21, RAD21_in-between, RAD21_left, RAD21_right, expression (2) length, motif_pattern, avg_motif_strength, std_HS, HS_in-between, HS_left, avg_CTCF, std_CTCF, CTCF_in-between, CTCF_left, CTCF_right, avg_RAD21, std_RAD21, RAD21_in-between, RAD21_left, expression; (3)

Supplementary Table 4: Designed primers for 3C validation

Primer Name	Sequence (5' to 3')
KMT2C_U2	FGGAGAGGATGATGGTGCTGTGTAT
KMT2C_U1	CTTGATCGTTTCTCACTCCTTTCA
KMT2C_L	CTTGACTGTCACCTTCAGCTCATC
KMT2C_D1	GACATACCAGAGCAATAACCTGGA
KMT2C_D3	AGCAGCAAATGAATCAGCTCAG
KMT2C_D4	AGTGGTGTCAATGCTGGTTTTC
KMT2C_R	ATCACTGTCTAGCTGCCCGTTC
PDGFRB_L	TATGCAGTGGTTTGTACCCTTG
PDGFRB_R	GTGGCACCATAATCATCCCTAT