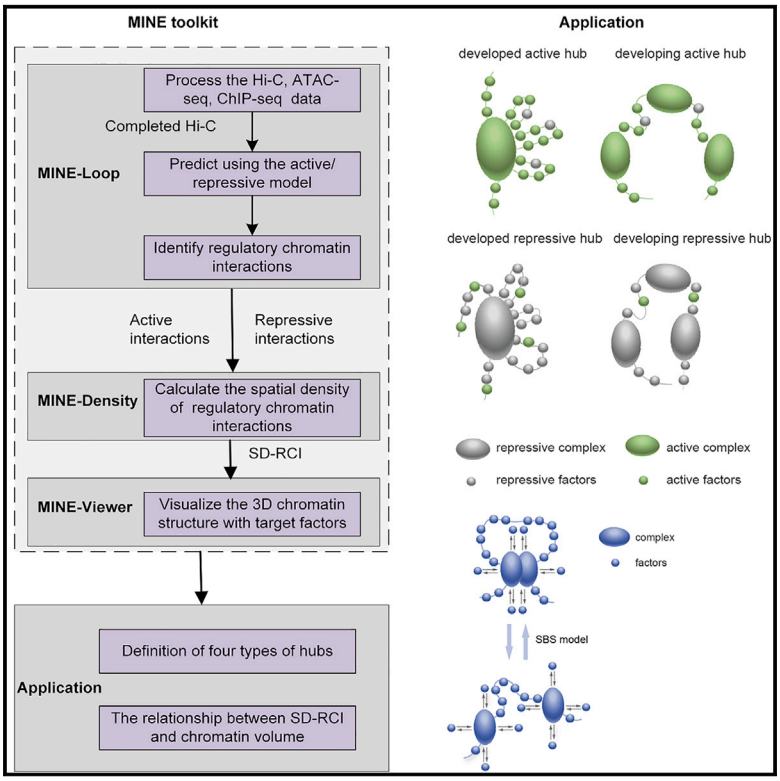
Cell Reports
Methods

Article

MINE is a method for detecting spatial density of regulatory chromatin interactions based on a multi-modal network

Graphical abstract



Highlights

- MINE is a tool for analyzing the spatial density of regulatory chromatin interactions
- The MINE package includes MINE-Loop, MINE-Density, and MINE-Viewer
- MINE suggests four types of chromatin hubs based on density of regulatory interactions
- MINE enables analysis of quantitative changes in chromatin structure

Authors

Haiyan Gong, Minghong Li, Mengdie Ji, ..., Yi Yang, Chun Li, Yang Chen

Correspondence

zxt@ies.ustb.edu.cn (X.Z.),
yc@ibms.pumc.edu.cn (Y.C.)

In brief

Gong et al. provide a toolkit, MINE, which includes MINE-Loop, MINE-Density, and MINE-Viewer to explore the relationship between spatial density of regulatory chromatin interactions, gene expression, and chromatin structure change.

Gong et al., 2023, Cell Reports Methods *3*, 100386
January 23, 2023
https://doi.org/10.1016/j.crmeth.2022.100386

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

MINE is a method for detecting spatial density of regulatory chromatin interactions based on a multi-modal network

Haiyan Gong,[1,5] Minghong Li,[1,5] Mengdie Ji,[2] Xiaotong Zhang,[1,3,*] Zan Yuan,[2] Sichen Zhang,[1] Yi Yang,[1] Chun Li,[4] and Yang Chen[2,6,*]

[1]Beijing Advanced Innovation Center for Materials Genome Engineering, School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
[2]State Key Laboratory of Medical Molecular Biology, Department of Biochemistry and Molecular Biology, Institute of Basic Medical Sciences, School of Basic Medicine, Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing 100005, China
[3]Shunde Innovation School, University of Science and Technology Beijing, Foshan 528399, China
[4]School of Mechanical Engineering, University of Science and Technology Beijing, Beijing 100083, China
[5]These authors contributed equally
[6]Lead contact
*Correspondence: zxt@ies.ustb.edu.cn (X.Z.), yc@ibms.pumc.edu.cn (Y.C.)
https://doi.org/10.1016/j.crmeth.2022.100386

MOTIVATION With the development of epigenomic technologies, we can obtain more and more Hi-C, ATAC-seq, ChIP-seq, and other epigenome data with high resolution; Although we can obtain some regulatory chromatin interactions that are anchored at some regulatory element, this would require us to calculate the loops that are anchoring to the location of the peaks. This process is challenging for the low ratio of regulatory chromatin interactions in Hi-C contact matrix. Therefore, we provide a toolkit, MINE, which includes MINE-Loop, MINE-Density, and MINE-Viewer, to enhance the ratio of regulatory chromatin interactions in Hi-C contact matrix and define the spatial density of regulatory chromatin interactions to explore the relationship between the spatial density of regulatory chromatin interactions, gene expression, and chromatin structure change.

SUMMARY

Chromatin interactions play essential roles in chromatin conformation and gene expression. However, few tools exist to analyze the spatial density of regulatory chromatin interactions (SD-RCI). Here, we present the multi-modal network (MINE) toolkit, including MINE-Loop, MINE-Density, and MINE-Viewer. The MINE-Loop network aims to enhance the detection of RCIs, MINE-Density quantifies the SD–RCI, and MINE-Viewer facilitates 3D visualization of the density of chromatin interactions and participating regulatory factors (e.g., transcription factors). We applied MINE to investigate the relationship between the SD-RCI and chromatin volume change in HeLa cells before and after liquid-liquid phase separation. Changes in SD-RCI before and after treating the HeLa cells with 1,6-hexanediol suggest that changes in chromatin organization was related to the degree of activation or repression of genes. Together, the MINE toolkit enables quantitative studies on different aspects of chromatin conformation and regulatory activity.

INTRODUCTION

With advances in 3D genome research, increasing evidence supports an essential role for chromatin interactions with nuclear regulatory factors in shaping chromatin conformation and the regulation of gene transcription. For example, the chromatin structure of A/B compartments and topologically associating domains (TADs) appear to be formed through distinct chromatin interactions (i.e., loops).[1] Previous research[2] has revealed that gene densities and GC content are correlated with the density of chromatin interactions (i.e., the number of interactions per Mb). Hou et al.[3] showed that gene density and transcription contribute to the partition of physical domains (i.e., regions with high gene density). Interactions between chromatin features associated with transcriptional activation or repression, such as Rad21, CCCTC-binding factor (CTCF), and H3K4me3, are correlated with gene expression.[4,5] Almassalha et al.[6] provided a "macrogenomic engineering" approach to regulate transcriptional activity in cancer cells by modulating chromatin density. Therefore, research on the spatial density of regulatory

chromatin interactions (SD-RCI) can further our understanding of the mechanisms responsible for chromatin folding and the relationship between chromatin conformation and gene expression. Technologies such as ChIA-PET[7] and HiChIP,[8] and tools like 3CPET[9] and ChIA-PET2,[10] can capture the RCIs for proteins of interest, such as DNA-binding regulatory proteins or RNA transcription factors. However, ChIA-PET and HiChIP data are only available for some cell lines, and their acquisition of different target proteins is costly, laborious, and time consuming. Hi-C[11] is a sequencing technology that quantifies the number of interactions between genome bins adjacent in 3D space but may be farther in a linear genome. Therefore, a method to detect RCIs by calling loops from Hi-C data can help reduce the cost. In this paper, we propose using MINE-Loop to identify special RCIs from high-resolution Hi-C data.

Among the numerous physical mechanisms of chromatin formation, one type of model[12–15] considers the formation of chromatin structures mediated by chromatin interactions with molecular factors, such as architectural proteins, histone marks, and non-coding RNAs.[16] Specifically, the strings and binders switch (SBS) model proposes that chromatin is a "self-avoiding polymer" surrounded by diffusive molecular factors (e.g., transcription factors) that anchor to cognate recognition sites on the chromosome to drive the chromatin folding process. The SBS model can be specifically applied to study the relationship between chromatin structural states, such as loops, TADs, or A/B compartments, and the density of regulatory factors (such as enhancer, promoter, and silencer). Studies investigating loops[17,18] have shown that CTCF mediates interactions with chromatin regions enriched with enhancer-regulated genes by altering chromatin domain structures. Similarly, Golkaram et al.[19] examined local chromatin density to quantify transcriptional regulatory components in a cell population and found that distinct TADs determined the distribution of gene expression. Jiang et al.[20] proposed the spatial density of open chromatin (SDOC) metric to characterize intra-TAD chromatin state and structure, where the SDOC refers to the ratio of the total number of accessible chromatin regions in a TAD to the total 3D space taken up by its physical structure. They found that TADs with decreased SDOC were enriched with repressed genes during T cell development in mice. While these studies investigated the relationship between chromatin structure and specific molecular factors (e.g., CTCF and transcription factors [TFs]), a method for simultaneous quantification of the spatial density of active or repressive RCIs (i.e., chromatin interactions that are anchoring regulatory elements to chromatin) is still lacking. Such a method could provide quantitative evidence supporting or refuting the SBS model of the relationship between chromatin structure and gene expression.

To address this gap, we introduce the multi-modal network (MINE) data analysis toolkit, including the MINE-Loop, MINE-Density, and MINE-Viewer tools, to explore the SD- RCI. MINE-Loop is a neural network model that integrates Hi-C, ChIP-seq,[21] and ATAC-seq[22] data to enhance the proportion of detectable RCIs. MINE-Density can be used to calculate the RCISD-RCI identified by MINE-Loop, and MINE-Viewer facilitates visualization of density and specific interactions with regulatory factors in 3D genomic structures. We explored the relationship between SD-RCI and the status of gene transcription in HepG2 cells. We also identify four distinct levels of interaction density related to the formation of transcriptionally active or repressive chromatin regions enriched for anchored regulatory factors (i.e., developed hubs) or regions with low density of anchored factors (developing hubs). Finally, we applied SD-RCI values to data obtained in a liquid-liquid phase separation (LLPS) experiment (i.e., the HeLa cell line treated or not with 1,6-hexanediol) to quantitatively describe changes in chromosome structure, which revealed that chromosome structure expands after treating with 1,6-hexanediol (Hex group). In summary, MINE provides a new method for quantitative analysis of chromatin conformations.

RESULTS

Overview of the MINE framework

MINE is a multi-modal method for detecting the SD-RCI (i.e., chromatin interactions that are anchoring regulatory elements to chromatin) that includes MINE-Loop, MINE-Density, and MINE-Viewer functions (Figure 1A). The pipeline schematics of using the MINE toolkit can be seen in Methods S1I–S1N).

Description of the MINE-Loop tool

Since raw Hi-C deep sequencing data contain a high degree of noise, the currently available loop (i.e., two chromatin regions that the interaction frequency is higher than that of the surrounding adjacent regions in the Hi-C contact matrix) callers commonly identify a relatively low proportion of RCIs. Hence, MINE-Loop was developed to obtain a larger proportion of RCIs from enhanced Hi-C data than from raw Hi-C data, where the RCIs are defined to be the chromatin loops that are anchoring with functional factors (e.g., CTCF, RAD21, SMC3, and POLR2A).

As shown in Figure 1B, the MINE-Loop tool first generates a high-resolution (i.e., 1 kb resolution) VC-normalized Hi-C contact matrix (Raw-hic) from raw Hi-C deep sequencing data obtained from GM12878, H1-hESC, and HepG2 cells. We then generated a masked Hi-C matrix (Masked-hic) from the Raw-hic file using ATAC-seq or histone ChIP-seq data obtained from the same cell line, resulting in a targeted Hi-C matrix that contains a large proportion of RCIs that are used to train the MINE-Loop network. The methodology for producing Masked-hic is described in "generation of Masked-hic."

Since the peaks called from the ATAC-seq and histone ChIP-seq data can indicate the localization of RCIs, we next generated a correlation matrix of ATAC-seq and ChIP-seq peaks by calculating the Pearson correlation coefficients between the peaks of these datasets. This correlation matrix serves as the first modal matrix for training the MINE-Loop model.

We then processed the Raw-hic matrix to generate the second modal matrix of training the MINE-Loop model. The Raw-hic matrix was downsampled at a ratio of $\frac{1}{s\times s}$ to build a smoothed Hi-C contact matrix at the same resolution as the Raw-hic, where the values of the $s\times s$ window were set as the average values of the $s\times s$ window. The essential reason to smooth Raw-hic is that the exact location of loops cannot be determined. We think this is beneficial for reducing noise (the noise refers to the chromatin interaction intensity values near the diagonal

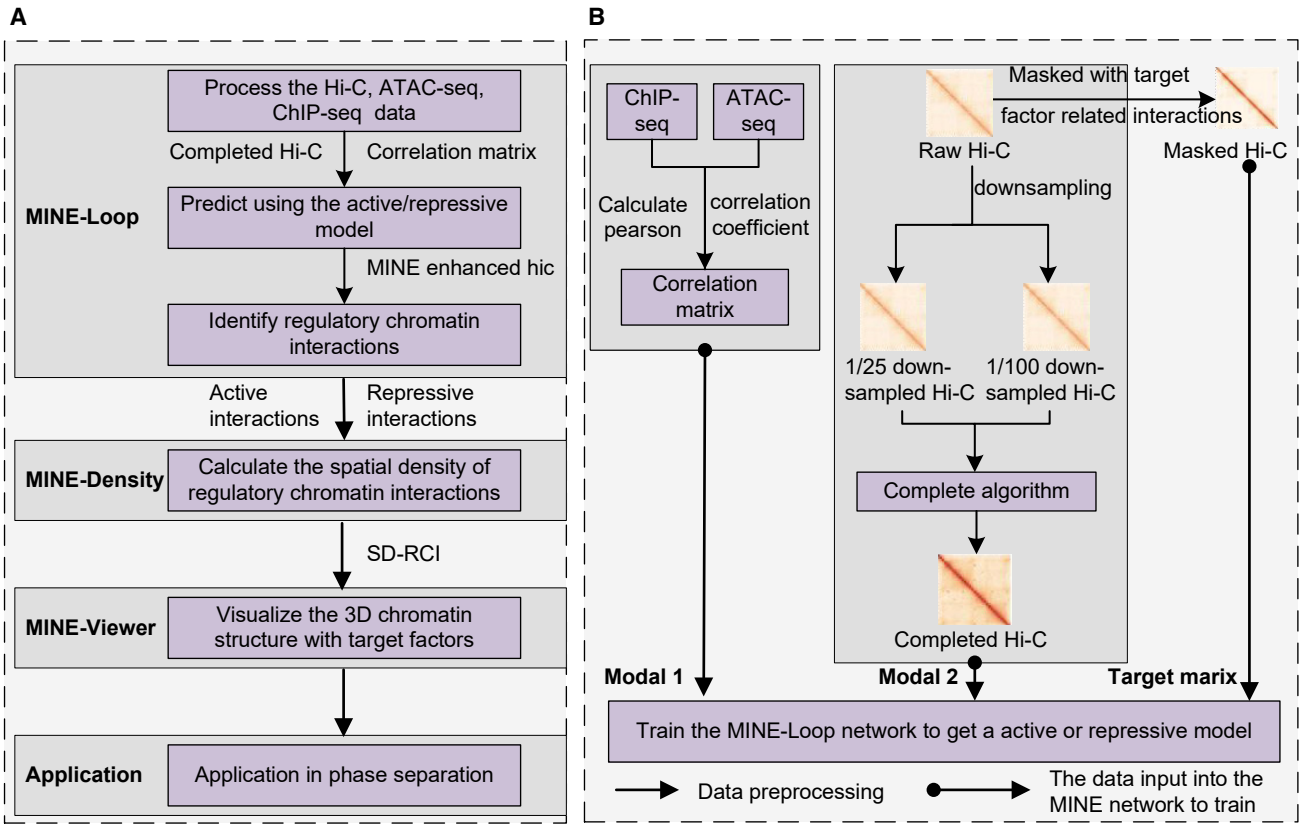


Figure 1. Overview of the MINE pipeline

(A) Workflow and analytical pipeline of the MINE method.

(B) An overview of the MINE-Loop architecture and workflow for model training.

region of a Hi-C contact matrix. Maass et al.[23] have proved that the chromatin interaction intensity near the diagonal region was much lower than that obtained by the Hi-C experiment.). To reduce the noise near the diagonal, we set s = 10 (a length of 10 kb region) to smooth the area around the diagonal, set s = 5 (the length of a 5 kb region, less than the average length of loops [5–200 kb][24]) to smooth the upper right corner area of the Hi-C contact matrix (the distal contact values in the Hi-C contact matrix). To extract more features of the Hi-C matrix, these down-sampled Hi-C matrices were then added by point-to-point, and the Hi-C matrix was then enhanced using the FAN method[25] to obtain a completed Hi-C matrix (Completed-hic).

Finally, the correlation matrix, Completed-hic, and Masked-hic are fed into the MINE-Loop network to map functions among the correlation matrix, Completed-hic, and Masked-hic. Once the model is trained, it is then applied to generate an enhanced Hi-C contact matrix (MINE-enhanced-hic) for any cell line using the Completed-hic and correlation matrix as inputs. To identify RCIs, MINE-enhanced-hic can then be fed into two available loop callers (i.e., FitHiC2[26] and mustache[27]). By surveying the input data requirements for different loop callers (Methods S1H), we found that only MUSTACHE and Fithic2 can identify loops with contact matrix (or some kind of data format that the contact matrix can be converted to) as input, while HiCCUPS[28] calls loops from .hic format file, cLoops[29] call loops require Mapped PETs info, HiC-ACT[30] calls loops from the output file from other methods (such as Fit-Hi-C/FitHiC2). Due to the output file predicted by MINE-Loops being only a contact matrix ($n \times n$ size), we only integrated MUSTACHE and Fithic2 in our MINE-Loop tool to call loops by transforming the Hi-C contact matrix ($n \times n$ size).

The MINE-Loop model can increase the proportion of different types of RCIs (active or repressive) by training with different histone modifications using ChIP-seq. For ChIP-seq data of different active-related histone modifications (i.e., H3K27ac or H3K4me3), an active MINE-Loop model (i.e., active model) can obtain a larger proportion of RCIs related to the control of DNA transcription machinery (i.e., active interactions), while ChIP-seq data of suppression-related histone modifications (i.e., H3K27me3 or H3k9me3) can be used to train a repressive MINE-Loop model (i.e., repressive model) to obtain a larger proportion of transcriptionally repressive chromatin interactions (i.e., repressive interactions).

Description of the MINE-Density and MINE-Viewer tools

Based on the RCIs identified by the MINE-Loop network, the ratio of the total number of active or repressive interactions in a TAD to the entire 3D space physically occupied by the TAD structure is defined as the SD-RCI. The 3D chromatin structure can be visualized with target factors (e.g., CTCF, genes, H3K4me3, and POLR2A) using the MINE-Viewer tool to

Table 1. Data used to train the active model or repressive model

Model	Epigenome data (input)	Data used to generate Masked-hic
Active	ATAC-seq, H3K27ac, H3K4me3 ChIP-seq	*cis*-regulatory element file
Active	ATAC-seq, H3K4me3 ChIP-seq	*cis*-regulatory element file
Active	ATAC-seq, H3K27ac, ChIP-seq	*cis*-regulatory element file
Repressive	H3K27me3, H3K9me3 ChIP-seq	H3K27me3, H3k9me3 ChIP-seq

investigate the density of target factor distribution in the 3D chromatin structure. Ultimately, MINE is applied to analyze changes in the SD-RCI before and after phase separation.

The three tools of the MINE method, including MINE-Loop, MINE-Density, and MINE-Viewer, enable exploration of the SD-RCI.

MINE-Loop facilitates detecting a high proportion of RCIs

MINE-Loop can detect RCIs in high-noise data

We next assessed whether the MINE-Loop analysis could increase the proportion of detectable RCIs compared with that obtained by current loop callers from raw Hi-C data using an active model as an example. We first generated a Completed-hic, correlation matrix, and Masked-hic of the GM12878 cell line that targeted factors (e.g., H3K4me3 and H3K27ac) specifically involved in DNA transcription to train and test the active model.

Hi-C data in the GM12878 cell line was downloaded from the 4DNucleome database (https://data.4dnucleome.org)[31] to generate the Completed-hic. The ATAC-seq and H3K27ac, H3K4me3 histone ChIP-seq data from the GM12878 cell line were downloaded from the ENCODE database and subsequently used to generate the correlation matrix. The annotation file of candidate *cis*-regulatory elements (CREs) in GM12878 cell line was downloaded from ENCODE to generate the Masked-hic.

Then, matrices generated for human chromosomes 1–17 in the GM12878 cell line dataset were used for training the active model, while the matrices generated for human chromosomes 18–22 in the GM12878 dataset were used to test the performance of the active model with the enhanced Hi-C data in the MINE-enhanced-hic output file. As Figure S1A shows, with the number of epochs in training, the accuracy of the training set and validation set increases. When the epoch equals about 20, the accuracy reaches saturation. To avoid overfitting, we choose to stop training when the epoch is about 20.

The MINE-Loop network includes MINE_Conv, maxPool_2D, and ConvTranspose_2D. To verify the influence of these modules on validation loss, as shown in Figure S3, we changed the network following three operations: (1) delete a layer of the network (“Remove one layer”), including “concat,” “ConvTranspose_2D,” and “MINE_Conv”; (2) remove half of the submodules of MINE_Conv module (“remove half of MINE_Conv”); (3) reduce the number of channels by half (“half channel for short”). Figure S1B shows that the verification loss value of MINE-Loop is lower than that of the other three networks, and “half channel for short” has the worst effect, followed by “Remove one layer” and “remove half of MINE_Conv.” We can conclude that the number of channels has the greatest impact on the model effect, followed by the upsampling layer, and the number of convolutional layers.

To evaluate the effect of FAN, we only use the 1/25 downsampling matrix to train MINE-Loop work and predict using 421.77 million, 601.74 million, and 4.01 billion Hi-C datasets (Figure S1C). The results show that the effect of a model trained using only the 1/25 downsampling matrix is much worse than using both the 1/25 and 1/100 downsampling matrices. Rao et al.[32] have shown that Hi-C contact matrices at different sequencing depth or resolutions can represent different features, such as A/B compartment, TAD, and loops. Forcato et al.[33] also showed that the reproducibility among replicates of the same dataset was low at all resolutions. Hence, we think that a Hi-C contact matrix with different sampling ratios can provide the model with different features. For example, in this experiment, 1/100 downsampling can compensate for some local information of the 1/25 downsampling matrix. When we only use a 1/25 downsampling matrix to train the MINE-Loop network, more features will be missed. When we use both the 1/25 and 1/100 downsampling matrix to train the model, the model will not rely on loop features of a certain resolution. A model trained in this way can help us find both proximal and distal loops. The validation results of MINE-Loop also evaluate the effect of FAN in this work.

The general applicability of the MINE-Loop model

To test the active model, FitHiC2[26] and mustache[27] were used to call intrachromosomal loops within a genomic distance of 2–100 kb in the MINE-enhanced-hic and Raw-hic matrices of human chromosomes 18–23 in GM12878 cells. The results showed a 12.5% and 9.7% overlap between loops called from MINE-enhanced-hic and Raw-hic data by FitHiC2 and mustache (Figure S2). However, the loops called from the MINE-enhanced-hic had 7–15 times more TFs related to the control of DNA transcription machinery (e.g., CTCF, RAD21, SMC3, and POLR2A) anchored at the same corresponding loops than that in raw Hi-C data (Figure S2A).

The same comparison was conducted for called loops ranging from 2 to 300 kb and 2 to 500 kb, which showed a greater number of anchored factors in the enhanced Hi-C data (Figure S2A). Taking the CTCF anchoring number as an example, “2-300kb” obtained more 2373 than “2-100kb,” “2-500kb” obtained more 1732 than “2-300kb.” These results suggested that MINE_loop analysis could also reveal long-range RCIs.

Having generated enhanced, integrative datasets related to RCIs as shown in Table 1 by training an active model, we then explored the effects of using different histone ChIP-seq data combinations to train active models. To this end, three combinations (combination (i): ATAC-seq, H3K27ac ChIP-seq, H3K4me3 ChIP-seq. Combination (ii): ATAC-seq, H3K27ac ChIP-seq. Combination (iii): ATAC-seq, H3K4me3 ChIP-seq) were used as inputs for active model training (Figures 2A–2D). Examination of anchor number for the CTCF, RAD21, SMC3, and POLR2A TFs in [2, 100 kb] loops (Figures 2A–2D) indicated that combination (ii) revealed fewer TF anchor points than other combinations

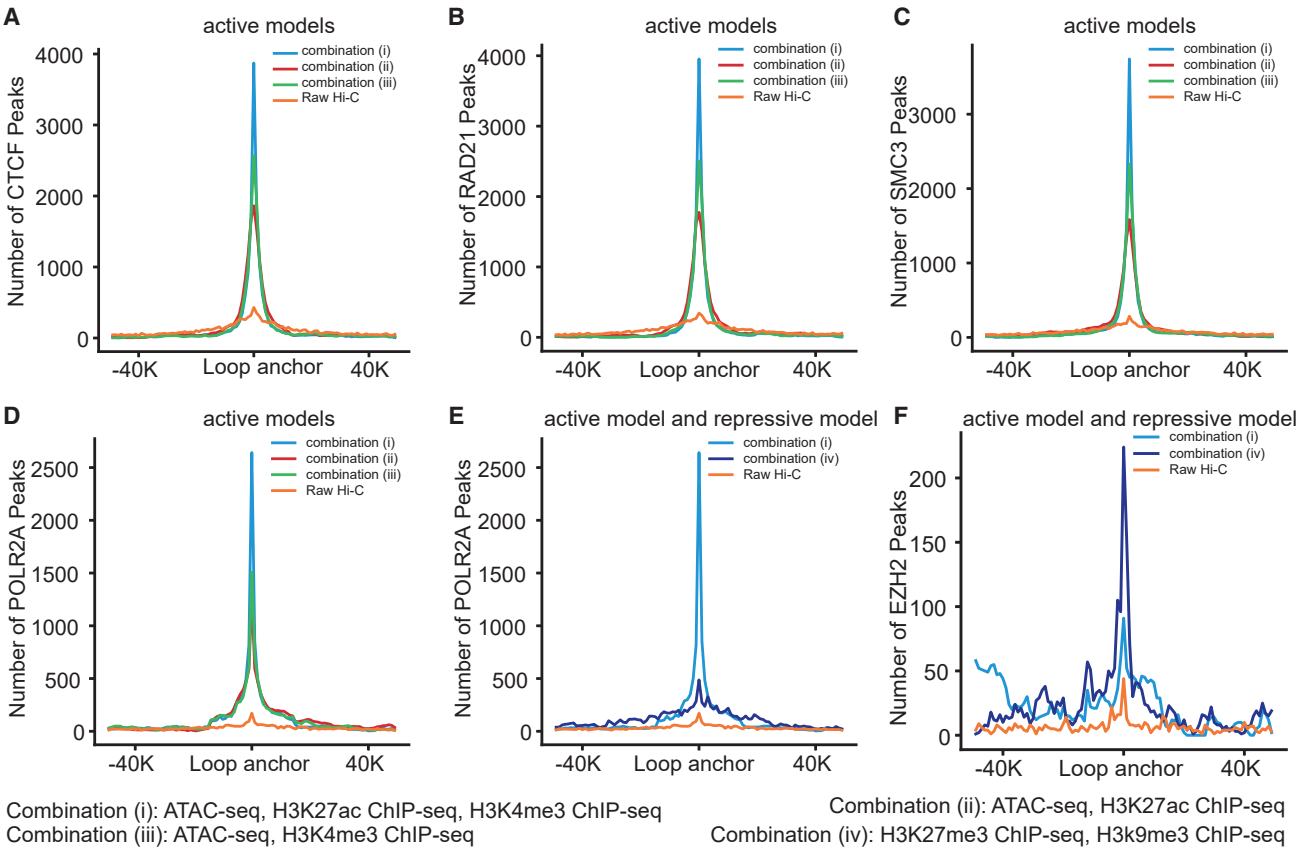


Figure 2. MINE-Loop can detect RCIs in high-noise data

The active and repressive models were trained in the GM12878 cell line using different epigenomic data combinations as inputs.

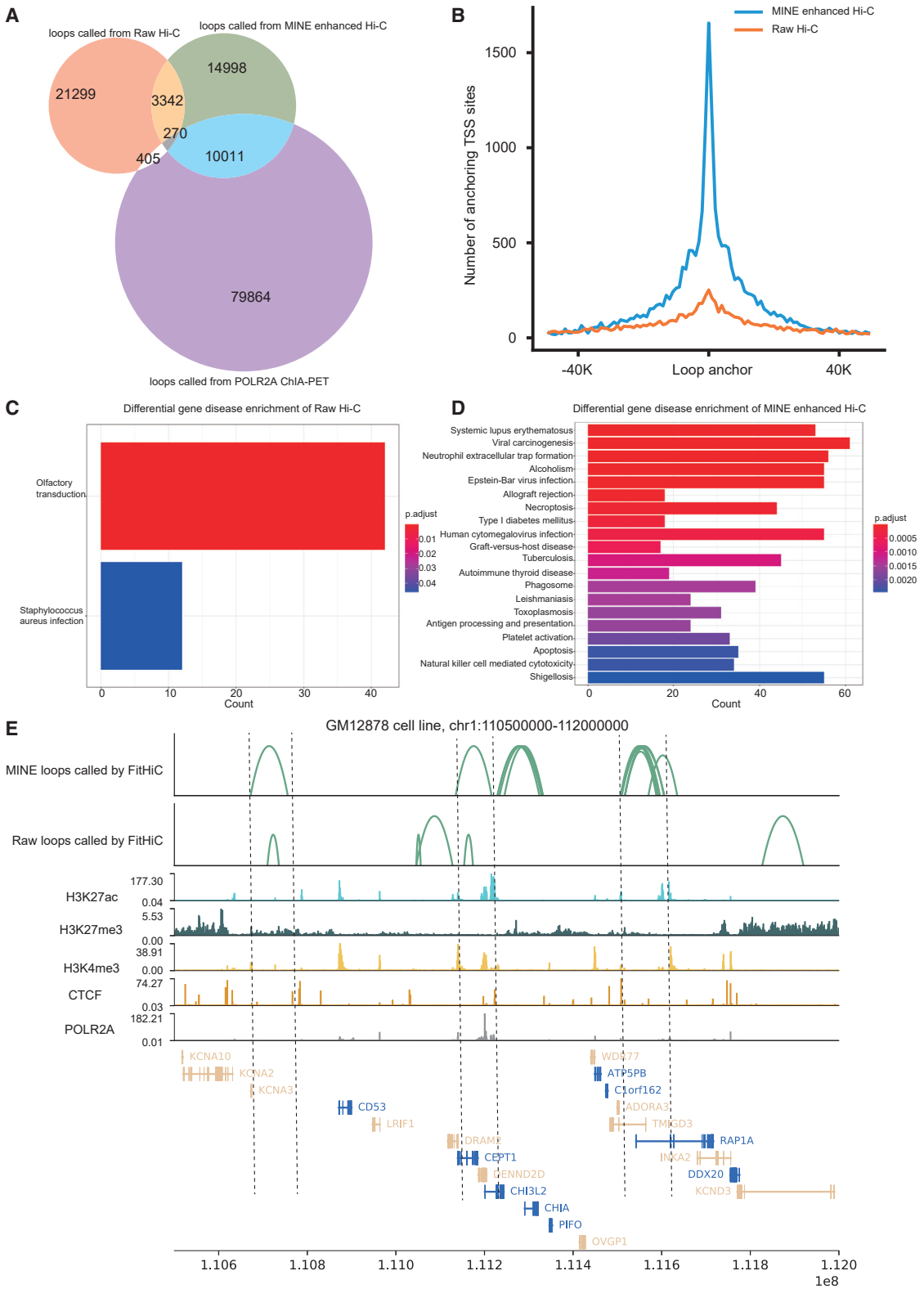
(A–D) Number of CTCF, RAD21, SMC3, and POLR2A transcription factor anchors in 2–100 kb loops called from active models trained with different epigenomic data combinations.

(E and F) Comparison of POLR2A and EZH2 anchors in loops called from active and repressive models.

in active interactions. This suggested that the more active-related histone ChIP-seq data used to train active models, the more active TF anchor chromatin interactions will be detectable.

To determine the MINE-Loop model's general applicability, we applied the active model trained using the combination (i) dataset in GM12878 cell line to predict using the combination (i, ii, iii) datasets in the IMR90 cell line as input data. The results show that the three combinations all perform better than raw Hi-C data in the number of anchoring TFs (Figure S3A), suggesting that MINE-Loop model does not require all epigenome data used in the training process. Besides the IMR90 cell line, the data from K562, H1-hESC, and HepG2 cell lines is also conducted using the active model trained in the GM12878 cell line. The results show that, within the genomic distance of 2–100, loops called from MINE-enhanced-hic of IMR90 (Figure S9) and K562 (Figure S3B) cell lines both can anchor more TFs than from Raw-hic, but less than from GM12878 (Figures 2A–2D), H1-hESC (Figure S3C), and HepG2 (Figure S3D) cell lines. The results suggest that the sequencing depth of the raw Hi-C data has an impact on the model's effect, where the sequencing depth of K562 and IMR90 cell lines (~1 billion) is much lower than that of GM12878 (~4.01 billion), H1-hESC (~3.22 billion), and HepG2 (2.02 billion) cell lines. To further validate the influence of sequencing depth of Hi-C data on the prediction effect, we downloaded Hi-C data from 4dnucleome.org under accession numbers 4DNFI9ZWZ5BS (421.77 million), 4DNFI7J8BQ4P (601.74 million), and 4DNFI1UEG1HD (4.01 billion), and enhanced these Hi-C data using the active model trained using the combination (i) dataset in the GM12878 cell line. Figure S4 shows that the deeper the sequencing depth is, the better the prediction effect of the model is.

Following the identification of loops containing transcriptional machinery genes that are actively transcribed, we then used a repressive MINE-Loop model (repressive model) trained with suppression-related histone marks (i.e., H3K27me3 and H3K9me3) target ChIP-seq data downloaded from ENCODE[34] (accession numbers ENCSR000DRX and ENCSR000AOX) to assess whether MINE-Loop could improve detection of transcriptionally repressive chromatin interactions. Comparison of POLR2A and EZH2 (a transcript factor related to long-term transcriptional inhibition) factors between active and repressive models trained with GM12878 cell line data showed that 2–100 kb loops in the repressive model obtained more EZH2 anchors than those of POLR2A (Figures 2E and 2F), indicating that



(legend on next page)

active models successfully identified more transcriptional activation-related interactions while the repressive model identified more transcriptional inhibition-related interactions. In agreement with our experimental evaluation of the active model, the repressive model trained with GM12878 cell line data was also used for prediction in other cell lines. In the K562 (Figure S3E) and HepG2 (Figure S3F) cell lines, loops called from MINE-enhanced-hic revealed a greater number of EZH2 anchors than that obtained from Raw-hic. Collectively, these data demonstrated that the MINE-Loop tool can facilitate the detection of a high proportion of active and repressive RCIs.

MINE-Loop facilitates the detection of functional chromatin loops

We next sought to verify whether the loops with anchored TFs called from the MINE-enhanced-hic overlapped with the ChIA-PET region. We found that 42.43%, 21.64% of the CTCF and POLR2A ChIA-PET region, overlapped with the active loops anchored CTCF and POLR2A in the HepG2 cell line, suggesting that loops called from MINE-enhanced-hic include many transcription factor binding interactions (Figure S5A). The Venn graph of Raw-hic, MINE-enhanced-hic, and POLR2A ChIA-PET in the GM12878 cell line showed that MINE-enhanced-hic could overlap with more POLR2A ChIA-PET region (about 9,606) than Raw-hic (Figure 3A). The number of anchoring TSS sites around the active loops (Figure 3B) further proved that MINE-enhanced-hic could detect more functional chromatin loops. By doing disease ontology (Do) enrichment analysis of the genes around these ~405 loops detected by Raw-hic but not by MINE-enhanced-hic (Figure S5C), we found these genes are enriched in terms related to "adenocarcinoma," "myeloma," where "adenocarcinoma," "myeloma" are more likely to be associated with cancer cells (e.g., Rpmi-8226 and U266) than with a normal B cell line (GM12878). Therefore, MINE-enhanced-hic cannot detect these 405 loops.

To validate whether the loops called from MINE-enhanced-hic could anchor more cell-specific genes than from Raw-hic. We next performed Do enrichment analysis to investigate the predicted functions of genes close to the anchor of active loops that were differentially detected between Raw-hic and MINE-enhanced-hic datasets generated with GM12878 or HepG2 cell lines. The differential genes enriched in active loops called from MINE-enhanced-hic GM12878 data were involved in "immune-related" processes (Figures 3C and 3D). In the HepG2 cell line, the differential genes from MINE-enhanced-hic were enriched in terms related to liver disease, which was consistent with the characteristics of HepG2 cells, while differential genes from Raw-hic were enriched in terms related to sensory perception of smell (Figure S5B). This shows that the differential loops of MINE-enhanced-hic compared with Raw-hic are enriched for genes functionally related to characteristics of that cell line.

To further investigate the enrichment pattern in a locus that loops called from Raw-hic and MINE-enhanced-hic are quite different, we chose the chr1: 110500000–112000000 region in chromosome 1. Upon close inspection of the region (Figure 3E), we found that the differential loops of MINE-enhanced-hic were enriched with H3K27ac, H3K27me3, H3K9me3, and POLR2A signals. These results showed that MINE-Loop could enhance the detection of functional RCIs.

Spatial density of RCIs and gene transcription

To investigate how the active and repressive models affect the SD-RCI, we refer to the definition of the SDOC metric[20] used to quantitatively measure the intra-TAD chromatin state and structure, and propose the SD-RCI (Figure 4A). The SD-RCI is defined as the ratio of the total number of active or repressive interactions in a TAD to the entire 3D space taken up by the physical structure of the TAD. (The calculation steps of SD-RCI are described in "spatial density of regulatory chromatin interactions.") Based on the definition of SD-RCI, the MINE-density tool is developed to calculate the SD-RCI.

To explore the relationship between the SD-RCI and the status of gene transcription in the HepG2 cell line, we first generated active and repressive loops within a genomic distance of 2–100 kb (Figure 4) and 2–300 kb (Figures S6D–S6F) using the active model and repress model of MINE-Loop. Based on the active and repressive loops, we calculated SD-RCI with a unit of TAD and divided the genome structure into four levels (i.e., ultra_high, high, middle, and low) based on the value of δ in the Gaussian distribution of SD-RCI (Figures 4B and 4E). To further explore the relationship between the number of RCIs and the volume of chromatin structure, we analyzed the relationship between the number of active loops or repressive loops in a TAD and an estimation of the volume of the TAD (i.e., an estimation of the physical space the TADs are occupying), respectively (Figures 4C and 4F). Details on our calculations of TAD volume are described in "calculation of TAD volume" in STAR Methods. The results show that a structure with higher SD-RCI (i.e., high, middle level) can cover more RCIs with a smaller volume of TADs, consistent with the definition of SD-RCI.

To investigate the gene expression changes within the TADs, we calculated the gene expression value log (RPKM) for different levels of TADs from the active model, where the gene expression data were downloaded from NCBI under accession number GSE184697. The result shows that the gene expression value is positively correlated with the SD-RCI level (Figure 4D), whereas the log (RPKM) calculated from the repress model shows an opposite change (Figure 4G). The results further proved that the repress model could detect more chromatin interactions related to inhibition of transcription, and the active model can detect more chromatin interactions related to the promotion of transcription. We also calculate the difference (p value)

Figure 3. MINE-Loop facilitates the detection of functional chromatin loops

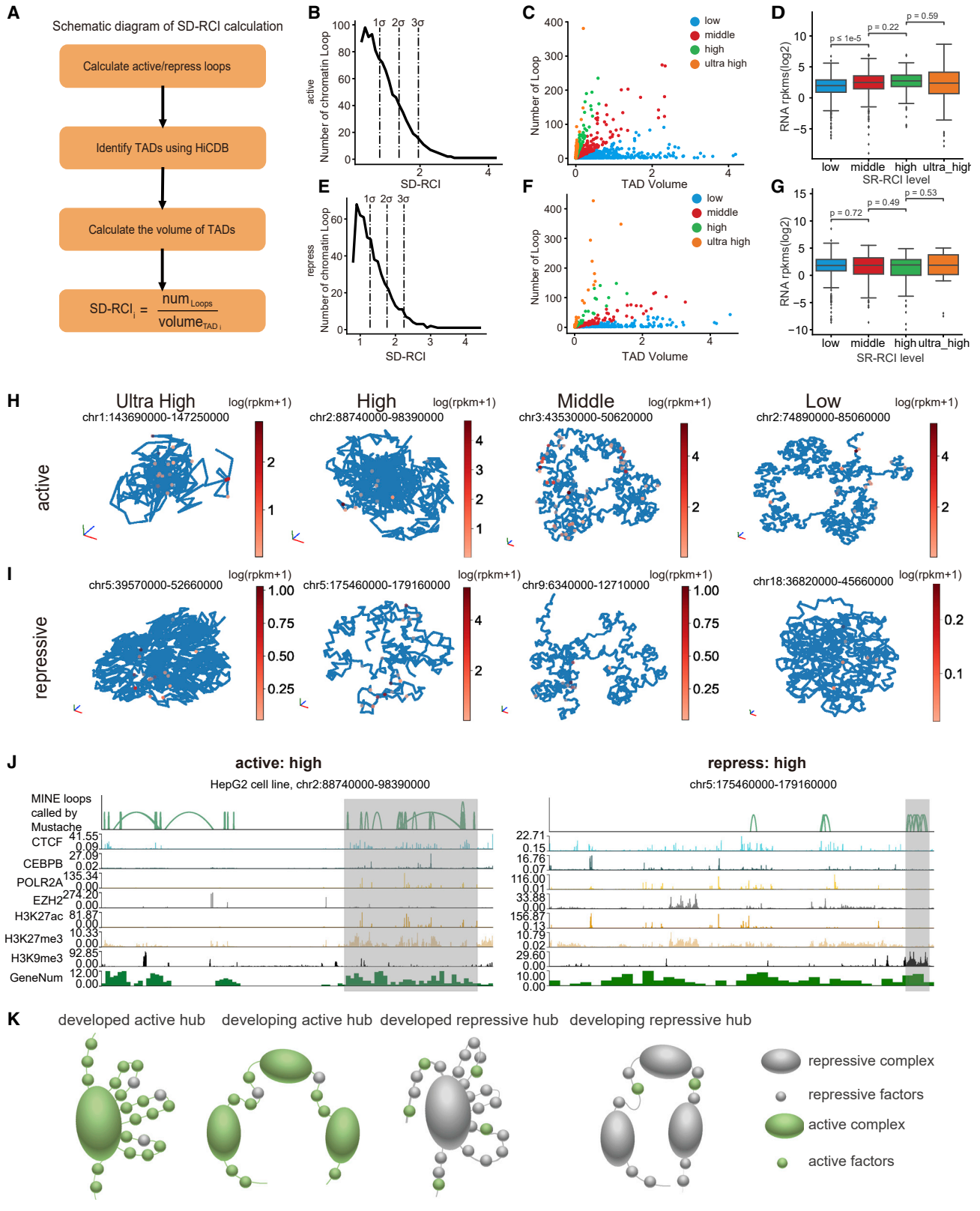
The active model was trained by GM12878 cell line with data of combination (i).

(A) The Venn graph of loops called from Raw-hic, MINE-enhanced-hic, and POLR2A ChIA-PET data.

(B) Comparison of anchoring transcription start site (TSS) number between MINE-enhanced-hic and Raw-hic.

(C and D) The differential gene Do (disease) enrichment of Raw-hic and MINE-enhanced-hic.

(E) Visualization of loops, H3K27ac, H3K27me3, H3K4me3, and POLR2A target ChIP-seq tracks.



(legend on next page)

between the RPKM distribution of genes corresponding to different SD-RCI levels, where the p values were calculated by the two-sided Mann-Whitney-Wilcoxon test with Bonferroni correction. The result shows a large difference ($p <= 1e-5$) between the low SD-RCI level and other SD-RCI levels in the active model. This means that there is a large differential gene expression when the SD-RCI levels are from low to middle. To explore the regulatory element distribution in different SD-RCI levels, we downloaded the super-enhancer (SE) and typical-enhancer (TE) datasets in the HepG2 cell line from the Sedb database (http://www.licpathway.net/sedb/) and statistically analyzed the proportion of SE and TE number and density (Figure S7). The results show that the TE accounts for a higher proportion at the low SD-RCI level, the SE accounts for a higher proportion at the middle SD-RCI level. Then, we analyzed the gene expression ([FPKM] fragments per kilobase of exon model per million mapped fragments) changes with the culture time of HepG2 cells, where the genes anchored at SE and TE at low and middle SD-RCI. The gene expression data of HepG2 cultured in 0, 1, 3, and 5 days were downloaded from NCBI with accession number GSE128763. Figure S7C shows that the average FPKM of SE at middle SD-RCI level is much higher than SE in low SD-RCI level. This means that SE can work at the middle SD-RCI level in the process of cell culture. Therefore, an analysis of the chromatin regions at middle SD-RCI level can effectively help researchers to explore the relevant genes or regulatory elements that play important roles in cell differentiation.

To explore the reason for the differences in gene expression between the active model and the repressive model at different SD-RCI levels, we developed a MINE-viewer tool to do 3D structure visualization with the gene expression for the four levels of TADs using the Pastis-PM2[35] algorithm (Figures 4H and 4I). The results show that the spatial density of TAD in the high level is higher than in middle and low levels. Then, we visualized the loops, CTCF, H3K27ac, H3K27me3, and H3K9me3 histone ChIP-seq tracks in the active and repress high-level TADs. We found that the active high-level TAD regions are enriched with CTCF and H3K27ac, the repressive high-level TAD regions are enriched with H3K9me3, which represses the transcriptional activity of genes (Figure 4J). The visualization of the other three SD-RCI levels in active or repressive regions can be seen in Figures S8 and S9. We quantified the enrichments by calculating the average signal p value (where the p value is extracted from the bigwig file of CTCF, POLR2A, or EZH2 target ChIP-seq) at different SD-RCI levels. Tables S1 and S2 show that, for the active model, CTCF and POLR2A in a higher SD-RCI level obtained a larger average p value than in a low SD-RCI level for the active model, and EZH2 in a higher SD-RCI level obtained a larger average p value than in a low SD-RCI level for the repressive model. The above results show that RCIs called from an active model or repressive model are enriched with active-related TFs or repressive-related TFs.

Based on the active or repressive loops identified by active model or repressive model, the genome regions can be spatially divided into active and repressive hubs, with active hubs in regions enriched with active TFs (e.g., CTCF, POLR2A, and SMC3) and repressive hubs in regions enriched with transcriptional repressors (e.g., EZH2). Using SD-RCI values, active and repressive hubs can be further defined into developed hubs, which have high SD-RCI (middle, high, and ultra-high level) or developing hubs, which have lower SD-RCI (SD-RCI levels low) (Figure 4K). By observing the TAD volume and loop number distribution of the different hubs (Figure S33), we find that the developing hubs own high TAD volume and low loop number. This result is consistent with the definition of these hubs. Hubs form in chromatin regions through active or repressive loops with TFs. Active or repressive hubs form where corresponding regulatory elements anchor to ensure the transcription or repression of genes required or not, respectively, for organismal function. When the proportion of RCIs in the chromatin space increases (i.e., forms a developed hub), then there is a higher frequency of anchoring by the corresponding regulatory element at a gene requiring its regulation. When the proportion of regulatory elements in the chromatin space is low (i.e., in developing hubs), the frequency of the required regulatory interaction is lower. So, genes in these developing hubs are upregulated more slowly, or require additional recruitment factors, while repression from an active state is also slower. This conclusion is consistent with the results that the average gene expression is higher in a higher SD-RCI level in the active model, and the average gene expression is lower in a higher SD-RCI level in the repressive model, as shown in Figures 4D and 4G.

In conclusion, the MINE-density and MINE-Viewer tools provide us with a view of the 3D visualization of spatial density of RCIs, and allow us to explore the active and repressive genome by calculating the SD-RCI of active and repressive interactions, respectively.

Spatial RCI density reflects changes in chromosome structure

MINE was next applied to data obtained from an LLPS experiment[36] reported in 2021 to verify whether MINE could identify the effects of 1,6-hexanediol-mediated LLPS disruption on

Figure 4. SD-RCI is correlated with gene expression

(A) Schematic diagram of SD-RCI calculation.

(B and E) The distribution of the number of active loops changes with the SD-RCI value calculated by the SD-RCI method.

(C and F) The dot plot of the number of RCIs and volume of TADs.

(D and G) The boxplot of RPKM and SD-RCI degree, where p values were calculated by two-sided Mann-Whitney-Wilcoxon test with Bonferroni correction, the SD-RCI is divided into 4° according to the value of δ in the Gaussian distribution of (A).

(H and I) The 3D genome TAD structure visualization with the gene expression strength in four levels from the active and repress model. The 3D structure of this region corresponds to the active (repress) high level in (B) and (C).

(J) Visualization of loops, CTCF, and histone mark ChIP-seq tracks. Loops are identified from MINE-enhanced-hic by using MUSTACHE.

(K) Four types of hubs are defined by the active or repressive SD-RCI.

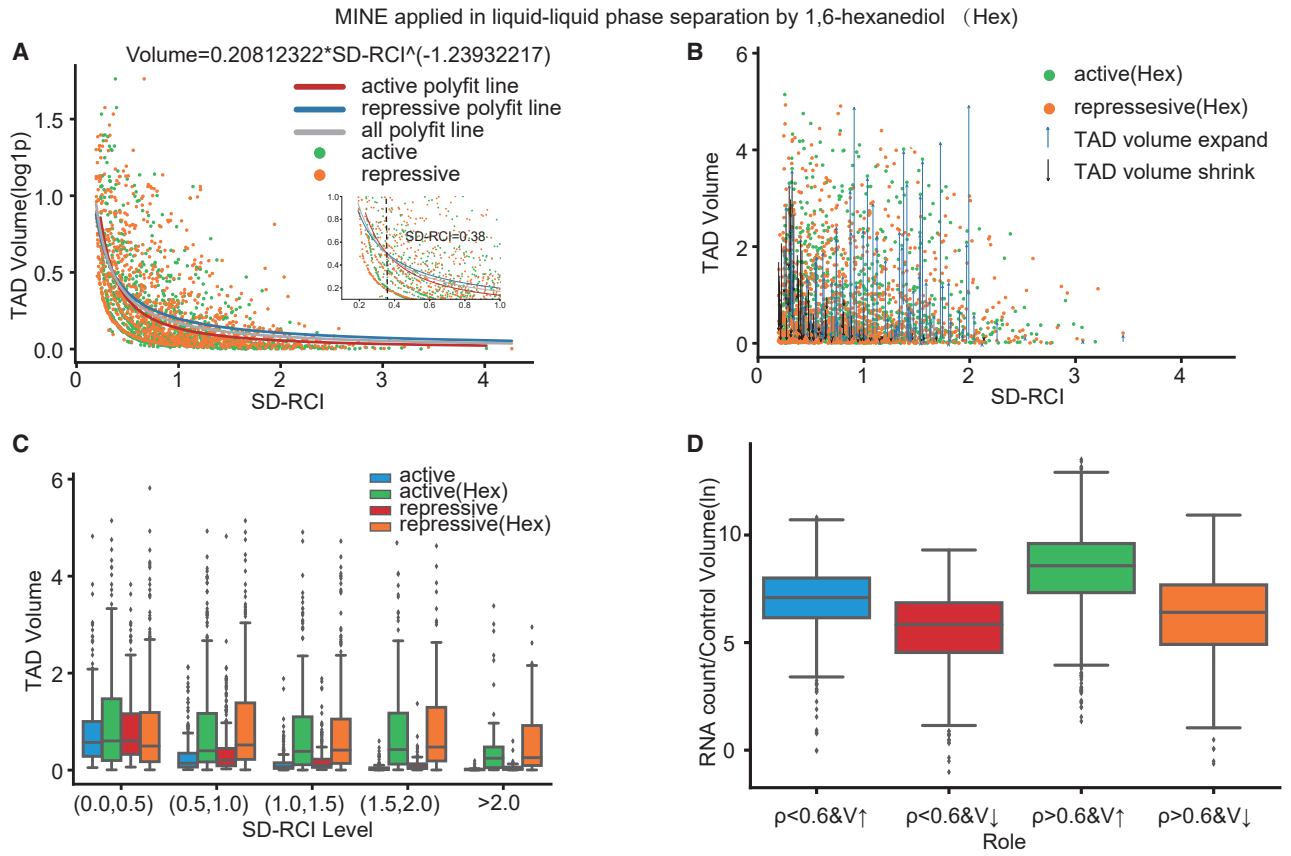


Figure 5. SD-RCI is correlated with chromosome structure before and after the effects of 1,6-hexanediol-mediated LLPS disruption

(A and B) The distribution of the volume of TADs changes with the SD-RCI value before and after liquid-liquid phase separation.

(C) The boxplot of the volume of TADs and the range of SD-RCI.

(D) The boxplot of the gene counts ratio calculated from RNA-seq data (control group) and the four types, where p represents the value of SD-RCI, and V represents the volume of TADs.

RCIs (Figures 5 and 6). For this analysis, Hi-C and epigenomic data obtained from the HeLa cell line treated (Hex group) or not (control group) with 1,6-hexanediol were downloaded and processed according to the MINE workflow described above.

First, we examined the relationship between these TAD volumes and SD-RCI values calculated from the active and repressive models before and after phase separation, and fitted the corresponding power function curve. We found that the active volume was larger than the repressive volume under the same SD-RCI condition when the SD-RCI of the HeLa cell line was <0.38 (i.e., the intersection value in Figure 5A), whereas the active volume was smaller than the repressive volume when the SD-RCI was >0.38 (Figure 5A). By comparing the volume of TADs at active and repressive states under the same SD-RCI condition, we could determine that, if the volume of active TADs is larger than repressive TADs with the same SD-RCI value, then the number of active loops is less than the repressive loops (Figure S33). As we know, active regulatory interactions promote higher gene expression and repressive regulatory interactions repress gene expression. Therefore, when SD-RCI is lower than the intersection value (0.38) in Figure 5A, the active or repressive TADs both tend to have low gene expression. When SD-RCI is higher than the intersection value (0.38), the active or repressive TADs both tend to have high gene expression.

Enhanced Hi-C data revealed that the TAD volume was larger after drug treatment than before treatment under the same SD-RCI conditions (Figures 5B and 5C). TADs were then categorized into four types ((1) SD-RCI <0.6 and control volume < Hex volume, (2) SD-RCI $\geq$ 0.6 and control volume > Hex volume, (3) SD-RCI <0.6 and control volume > Hex volume, (4) SD-RCI $\geq$ 0.6 and control volume < Hex volume, where 0.6 was the SD-RCI inflection point when count became positive calculated from the count-SD-RCI curve, as shown in Figure S11A) according to the change in TAD volume (whether increased or decreased) and the size of SD-RCI before and after drug treatment.

To further determine whether changes in volume from pre- to post-phase separation were related to the intensity of gene transcription, we calculated the gene counts ratio (i.e., $count/|V_{before} - V_{after}|$, where count is gene count, V_{before} is the volume before LLPS, V_{after} is the volume after LLPS) for the four types of changes in TAD volume. The results showed that TADs with high gene counts ratio were more likely to increase in volume, while TADs with low gene counts would likely decrease in

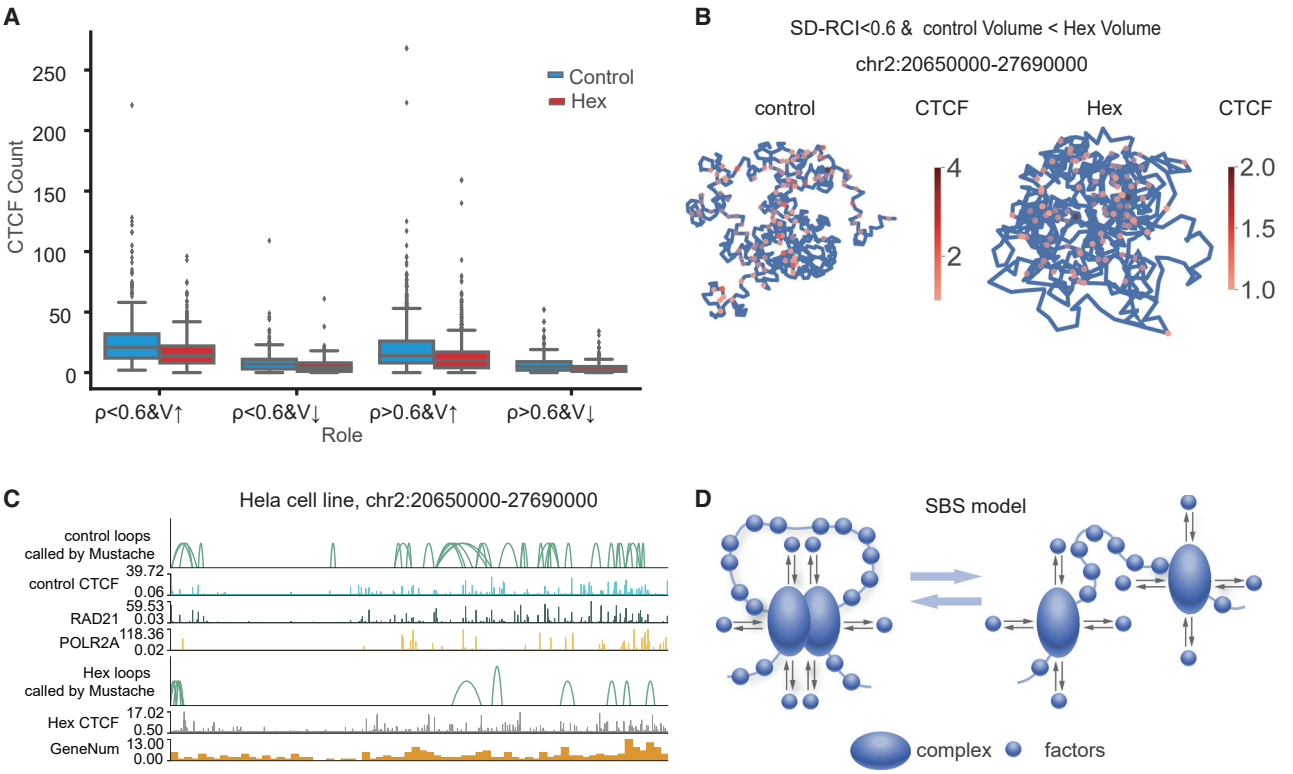


Figure 6. Exploration of mechanisms to explain the reason for changes in TAD volume

(A) The boxplot of the CTCF counts calculated from control group and Hex group.

(B) The 3D structure marked with CTCF anchor intensity of TAD typed with "SD-RCI < 0.6 & control Volume "Hex Volume."

(C) The visualization of loops and tracks from control group and Hex group.

(D) The mechanism of chromatin interaction formation and unwinding.

volume (Figure 5D). Previous studies[37–39] have established that regions with a high density of genes are mostly in open chromatin regions (A compartment), while regions with low gene density are generally in closed chromatin regions (B compartment). We therefore concluded that, in the A compartment, the volume of TADs was more likely to increase following LLPS, while the volume of TADs was likely to decrease in the B compartment after LLPS. This conclusion is consistent with the established findings.[36,40]

To explore the reason for the volume change after LLPS, we calculated the CTCF count distribution of the four types of TADs and found TADs with high CTCF counts were more likely to increase in volume (Figure 6A). The four types of TADs were then visualized as three-dimensional structures marked with CTCF anchor intensity before and after LLPS (CTCF intensity is shown in Figures 6B and S11B). Set the type of "SD-RCI<0.6 & control Volume ≪ Hex Volume" as an example, the visualization of 3D structure and loops showed the control group obtained more loops than the Hex group (Figure 6C). The visualization of other three type can be found in Figures S12A–S12C. This means that the 1,6-hexanediol (Hex) breaks or forms some loops, which makes the volume of TAD increase or decrease. In the discussion, we use the SBS (Figure 6D)[15,41] model to describe the reason for an increase in chromosome volume.

DISCUSSION

In consideration of these RCIs identified through MINE-Loop, SD-RCI can serve as a metric for quantitative exploration of the relationship between active or repressive chromatin interactions and the gene transcriptional status for a given TAD region. In this work, we define four levels of SD-RCI (ultra_high, high, middle, and low) to assess the relationship between SD-RCI and gene transcription. By comparing the expression strength of genes at different SD-RCI levels in the HepG2 cell line, we found that a higher SD-RCI in the active model is more conducive for gene transcription, and conversely in repressive models, higher SD-RCI is associated with greater transcriptional inhibition.

In analyses investigating the relationship between SD-RCI and chromosomal structure, SD-RCI values were used to compare HeLa cell volumes before and after LLPS (i.e., the HeLa cell line treated or not with 1,6-hexanediol). We found that the overall volume population cells increased following drug treatment. We propose that the SBS[15,41] model can explain this change in volume through the formation of loops and domains resulting from chromatin contacts between distant loci mediated by molecular factors, such as TFs. Before treating the HeLa cell with 1,6-hexanediol, the density of CTCF in 3D structure is higher in the control group (Figures 6A and 6C), a stable chromatin loop is formed through entropic force[42] exerted by other small molecular factors

in the nucleus localized in complexes attached to the chromatin. After treating the HeLa cell with 1,6-hexanediol, the density of CTCF in the 3D structure is lower (Figures 6A and 6C), the complex is no longer attached to the same chromatin location, and the small molecular factors in the nucleus exert entropic force on each complex individually, resulting in disruption of the previously stable chromatin loops, loosening the chromatin, and resulting in increased volume. Previous study[1] has shown that TAD structures and the A/B compartments are primarily formed by chromatin loop extrusion, ultimately resulting in a higher overall cell volume after treatment with 1,6-hexanediol.

In summary, we established a deep-learning-based framework by integrating multiple omics datasets (i.e., ATAC-seq and ChIP-seq) to reduce noise and increase the proportion of detectable RCIs. Compared with the methods of simple overlapping between histone modification and raw loops manually, MINE-Loop can detect a high proportion of RCIs by just inputting a few types of histone modification ChIP-seq and Hi-C contact matrix data. Applying the MINE pipeline to explore the relationship between SD-RCI and gene transcriptional status led to the discovery of four levels of spatial density in chromatin interactions that reflect the relationship between SD-RCI and gene regulation. We then applied MINE to data obtained from a LLPS experiment (i.e., treating the HeLa cell with 1,6-hexanediol), which showed that the 3D conformation of active and repressive models are consistent with the results[36] that the 1,6-hexanediol treatment caused the enlargement of nucleosome clutches and their more uniform distribution in the nuclear space. Finally, the mechanism underlying structural changes in TADs before and after LLPS was explained by SBS model.[15,41]

Limitations of the study

(1) This paper lacks enough physical characteristics of the four types of chromatin hubs and their epigenetic makeup; although examples of analyzing the changes of RNA-seq expression during cell differentiation, and chromosome formation change. There is still no proper characterization of those regions together with a mechanistic explanation of their formation.

(2) The Hi-C datasets with more than 400 million filtered reads as the inputs of MINE-Loop model are suggested to get a better prediction performance.

(3) The MINE toolkit is limited for two loop callers (FitHiC2 and mustache). In the future, we may improve it to accommodate more loop callers.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2022.100386.

ACKNOWLEDGMENTS

This work was funded by grants from the National Natural Science Foundation of China (81890994, 31871343, and 61971031), the Foshan Higher Education Foundation (BKBS202203), the Scientific and Technological Innovation Foundation of Shunde Graduate School, USTB (BK20BF009), the National Key R&D Program of China (2018YFA0801402), and the CAMS Innovation Fund for Medical Sciences (2020-RC310-009, 2021-RC310-007, 2021-I2M-1-020, and 2022-I2M-JB-003). The authors wish to thank Chen Fengling in Tsinghua University, Kailong Li in Peking University, and Xudong Wu in Tianjin Medical University for their suggestions about data analysis. We also thank Xu Miaomiao in University of Science and Technology Beijing for her assistance on the color and layout of figures. All authors thank the reviewers for their critical suggestions.

AUTHOR CONTRIBUTIONS

Y.C., X.T.Z., H.Y.G., and M.H.L. conceived and designed the project. H.Y.G. and M.H.L. performed the experiments. M.D.J., Z.Y., S.C.Z., and Y.Y. contributed to the implementation of the research. C.L. contributed to the design of figures. H.Y.G. and M.H.L. completed the figures and writing of the paper with the guidance of Y.C. and X.T.Z.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 30, 2022
Revised: September 15, 2022
Accepted: December 16, 2022
Published: January 12, 2023

REFERENCES

1. de Wit, E. (2020). TADs as the caller calls them. J. Mol. Biol. *432*, 638–642. https://doi.org/10.1016/j.jmb.2019.09.026.
2. Sandhu, K.S., Li, G., Poh, H.M., Quek, Y.L.K., Sia, Y.Y., Peh, S.Q., Mulawadi, F.H., Lim, J., Sikic, M., Menghi, F., et al. (2012). Large-scale functional organization of long-range chromatin interaction networks. Cell Rep. *2*, 1207–1219. https://doi.org/10.1016/j.celrep.2012.09.022.
3. Hou, C., Li, L., Qin, Z.S., and Corces, V.G. (2012). Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. Mol. Cell *48*, 471–484. https://doi.org/10.1016/j.molcel.2012.08.031.

4. Seitan, V.C., Faure, A.J., Zhan, Y., McCord, R.P., Lajoie, B.R., Ing-Simmons, E., Lenhard, B., Giorgetti, L., Heard, E., Fisher, A.G., et al. (2013). Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments. Genome Res. *23*, 2066–2077. https://doi.org/10.1101/gr.161620.113.

5. Li, Y., He, Y., Liang, Z., Wang, Y., Chen, F., Djekidel, M.N., Li, G., Zhang, X., Xiang, S., Wang, Z., et al. (2018). Alterations of specific chromatin conformation affect ATRA-induced leukemia cell differentiation. Cell Death Dis. *9*, 200. https://doi.org/10.1038/s41419-017-0173-6.

6. Almassalha, L.M., Bauer, G.M., Wu, W., Cherkezyan, L., Zhang, D., Kendra, A., Gladstein, S., Chandler, J.E., VanDerway, D., Seagle, B.-L.L., et al. (2017). Macrogenomic engineering via modulation of the scaling of chromatin packing density. Nat. Biomed. Eng. *1*, 902–913. https://doi.org/10.1038/s41551-017-0153-2.

7. Fullwood, M.J., and Ruan, Y. (2009). ChIP-based methods for the identification of long-range chromatin interactions. J. Cell. Biochem. *107*, 30–39. https://doi.org/10.1002/jcb.22116.

8. Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J., and Chang, H.Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat. Methods *13*, 919–922. https://doi.org/10.1038/nmeth.3999.

9. Djekidel, M.N., Liang, Z., Wang, Q., Hu, Z., Li, G., Chen, Y., and Zhang, M.Q. (2015). 3CPET: finding co-factor complexes from ChIA-PET data using a hierarchical Dirichlet process. Genome Biol. *16*, 288. https://doi.org/10.1186/s13059-015-0851-6.

10. Li, G., Chen, Y., Snyder, M.P., and Zhang, M.Q. (2017). ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. Nucleic Acids Res. *45*, e4. https://doi.org/10.1093/nar/gkw809.

11. Van Berkum, N.L., Lieberman-Aiden, E., Williams, L., Imakaev, M., Gnirke, A., Mirny, L.A., Dekker, J., and Lander, E.S. (2010). Hi-C: a method to study the three-dimensional architecture of genomes. JoVE, 1869. https://doi.org/10.3791/1869.

12. Di Stefano, M., Paulsen, J., Lien, T.G., Hovig, E., and Micheletti, C. (2016). Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. Sci. Rep. *6*, 35985. https://doi.org/10.1038/srep35985.

13. Di Pierro, M., Zhang, B., Aiden, E.L., Wolynes, P.G., and Onuchic, J.N. (2016). Transferable model for chromosome architecture. Proc. Natl. Acad. Sci. USA *113*, 12168–12173. https://doi.org/10.1073/pnas.1613607113.

14. Brackley, C.A., Brown, J.M., Waithe, D., Babbs, C., Davies, J., Hughes, J.R., Buckle, V.J., and Marenduzzo, D. (2016). Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. Genome Biol. *17*, 59. https://doi.org/10.1186/s13059-016-0909-0.

15. Fiorillo, L., Bianco, S., Esposito, A., Conte, M., Sciarretta, R., Musella, F., and Chiariello, A.M. (2020). A modern challenge of polymer physics: novel ways to study, interpret, and reconstruct chromatin structure. WIREs Comput. Mol. Sci. *10*, e1454. https://doi.org/10.1002/wcms.1454.

16. Jung, N., and Kim, T.K. (2021). Advances in higher-order chromatin architecture: the move towards 4D genome. BMB Rep. *54*, 233–245. https://doi.org/10.5483/BMBRep.2021.54.5.035.

17. Oti, M., Falck, J., Huynen, M.A., and Zhou, H. (2016). CTCF-mediated chromatin loops enclose inducible gene regulatory domains. BMC Genom. *17*, 252. https://doi.org/10.1186/s12864-016-2516-6.

18. Ren, G., Jin, W., Cui, K., Rodrigez, J., Hu, G., Zhang, Z., Larson, D.R., and Zhao, K. (2017). CTCF-mediated enhancer-promoter interaction is a critical regulator of cell-to-cell variation of gene expression. Mol. Cell *67*, 1049–1058.e6. https://doi.org/10.1016/j.molcel.2017.08.026.

19. Golkaram, M., Jang, J., Hellander, S., Kosik, K.S., and Petzold, L.R. (2017). The role of chromatin density in cell population heterogeneity during stem cell differentiation. Sci. Rep. *7*, 13307. https://doi.org/10.1038/s41598-017-13731-3.

20. Jiang, S., Li, H., Hong, H., Du, G., Huang, X., Sun, Y., Wang, J., Tao, H., Xu, K., Li, C., et al. (2021). Spatial density of open chromatin: an effective metric for the functional characterization of topologically associated domains. Brief. Bioinform. *22*, bbaa210. https://doi.org/10.1093/bib/bbaa210.

21. Schmidt, D., Wilson, M.D., Spyrou, C., Brown, G.D., Hadfield, J., and Odom, D.T. (2009). ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. Methods *48*, 240–248. https://doi.org/10.1016/j.ymeth.2009.03.001.

22. Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: a method for assaying chromatin accessibility genome-wide. Curr. Protoc. Mol. Biol. *109*, 21.29.1–21.21.29. https://doi.org/10.1002/0471142727.mb2129s109.

23. Maass, P.G., Barutcu, A.R., Weiner, C.L., and Rinn, J.L. (2018). Inter-chromosomal contact properties in live-cell imaging and in Hi-C. Mol. Cell *70*, 188–189.

24. Jackson, D.A., Dickinson, P., and Cook, P.R. (1990). The size of chromatin loops in HeLa cells. EMBO J. *9*, 567–571. https://doi.org/10.1016/j.molcel.2018.02.007.

25. Achanta, R., Arvanitopoulos, N., and Süsstrunk, S. (2017). Extreme Image Completion (Ieee), pp. 1333–1337. https://doi.org/10.1109/ICASSP.2017.7952373.

26. Kaul, A., Bhattacharyya, S., and Ay, F. (2020). Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. Nat. Protoc. *15*, 991–1012. https://doi.org/10.1038/s41596-019-0273-0.

27. Roayaei Ardakany, A., Gezer, H.T., Lonardi, S., and Ay, F. (2020). Mustache: multi-scale detection of chromatin loops from Hi-C and micro-C maps using scale-space representation. Genome Biol. *21*, 256. https://doi.org/10.1186/s13059-020-02167-0.

28. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021.

29. Cao, Y., Chen, Z., Chen, X., Ai, D., Chen, G., McDermott, J., Huang, Y., Guo, X., and Han, J.D.J. (2020). Accurate loop calling for 3D genomic data with cLoops. Bioinformatics *36*, 666–675. https://doi.org/10.1093/bioinformatics/btz651.

30. Lagler, T.M., Abnousi, A., Hu, M., Yang, Y., and Li, Y. (2021). HiC-ACT: improved detection of chromatin interactions from Hi-C data via aggregated Cauchy test. Am. J. Hum. Genet. *108*, 257–268. https://doi.org/10.1016/j.ajhg.2021.01.009.

31. Dekker, J., Belmont, A.S., Guttman, M., Leshyk, V.O., Lis, J.T., Lomvardas, S., Mirny, L.A., O'shea, C.C., Park, P.J., Ren, B., et al. (2017). The 4D nucleome project. Nature *549*, 219–226. https://doi.org/10.1038/nature23884.

32. Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021.

33. Forcato, M., Nicoletti, C., Pal, K., Livi, C.M., Ferrari, F., and Bicciato, S. (2017). Comparison of computational methods for Hi-C data analysis. Nat. Methods *14*, 679–685. https://doi.org/10.1038/nmeth.4325.

34. Zhang, J., Lee, D., Dhiman, V., Jiang, P., Xu, J., McGillivray, P., Yang, H., Liu, J., Meyerson, W., Clarke, D., et al. (2020). An integrative ENCODE resource for cancer genomics. Nat. Commun. *11*, 3696. https://doi.org/10.1038/s41467-020-14743-w.

35. Varoquaux, N., Ay, F., Noble, W.S., and Vert, J.-P. (2014). A statistical approach for inferring the 3D structure of the genome. Bioinformatics *30*, i26–i33. https://doi.org/10.1093/bioinformatics/btu268.

36. Ulianov, S.V., Velichko, A.K., Magnitov, M.D., Luzhin, A.V., Golov, A.K., Ovsyannikova, N., Kireev, I.I., Gavrikov, A.S., Mishin, A.S., Garaev, A.K.,

et al. (2021). Suppression of liquid–liquid phase separation by 1, 6-hexanediol partially compromises the 3D genome organization in living cells. Nucleic Acids Res. *49*, 10524–10541. https://doi.org/10.1093/nar/gkab249.

37. Stadhouders, R., Filion, G.J., and Graf, T. (2019). Transcription factors and 3D genome conformation in cell-fate decisions. Nature *569*, 345–354. https://doi.org/10.1038/s41586-019-1182-7.
38. Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T.C., Robins, A.J., Dalton, S., and Gilbert, D.M. (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res. *20*, 761–770. https://doi.org/10.1101/gr.099655.109.
39. Barutcu, A.R., Lajoie, B.R., McCord, R.P., Tye, C.E., Hong, D., Messier, T.L., Browne, G., van Wijnen, A.J., Lian, J.B., Stein, J.L., et al. (2015). Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. Genome Biol. *16*, 214. https://doi.org/10.1186/s13059-015-0768-0.
40. Liu, X., Jiang, S., Ma, L., Qu, J., Zhao, L., Zhu, X., and Ding, J. (2021). Time-dependent effect of 1, 6-hexanediol on biomolecular condensates and 3D chromatin organization. Genome Biol. *22*, 230. https://doi.org/10.1186/s13059-021-02455-3.
41. Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.-M., Dostie, J., Pombo, A., and Nicodemi, M. (2012). Complexity of chromatin folding is captured by the strings and binders switch model. Proc. Natl. Acad. Sci. USA *109*, 16173–16178. https://doi.org/10.1073/pnas.1204799109.
42. Cook, P.R., and Marenduzzo, D. (2018). Transcription-driven genome organization: a model for chromosome structure and the regulation of gene expression tested through simulations. Nucleic Acids Res. *46*, 9895–9906. https://doi.org/10.1093/nar/gky763.
43. Liu, S., Su, Y., Yin, H., Zhang, D., He, J., Huang, H., Jiang, X., Wang, X., Gong, H., Li, Z., et al. (2021). An infrastructure with user-centered presentation data model for integrated management of materials data and services. npj Comput. Mater. *7*, 88. https://doi.org/10.1038/s41524-021-00557-x.
44. Krietenstein, N., Abraham, S., Venev, S.V., Abdennur, N., Gibcus, J., Hsieh, T.-H.S., Parsi, K.M., Yang, L., Maehr, R., Mirny, L.A., et al. (2020). Ultrastructural details of mammalian chromosome architecture. Mol. Cell *78*, 554–565.e7. https://doi.org/10.1016/j.molcel.2020.03.003.
45. ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia of DNA elements) project. Science *306*, 636–640. https://doi.org/10.1126/science.1105136.
46. Robinson, J.T., Turner, D., Durand, N.C., Thorvaldsdóttir, H., Mesirov, J.P., and Aiden, E.L. (2018). Juicebox. js provides a cloud-based visualization system for Hi-C data. Cell Syst. *6*, 256–258.e1. https://doi.org/10.1016/j.cels.2018.01.001.
47. Rocha, P.P., Raviram, R., Bonneau, R., and Skok, J.A. (2015). Breaking TADs: insights into hierarchical genome organization. Epigenomics *7*, 523–526. https://doi.org/10.2217/epi.15.25.
48. Zhang, Y., An, L., Xu, J., Zhang, B., Zheng, W.J., Hu, M., Tang, J., and Yue, F. (2018). Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. Nat. Commun. *9*, 750. https://doi.org/10.1038/s41467-018-03113-2.
49. Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual Losses for Real-Time Style Transfer and Super-resolution (Springer), pp. 694–711. https://doi.org/10.1007/978-3-319-46475-6_43.
50. Chen, F., Li, G., Zhang, M.Q., and Chen, Y. (2018). HiCDB: a sensitive and robust method for detecting contact domain boundaries. Nucleic Acids Res. *46*, 11239–11250. https://doi.org/10.1093/nar/gky789.
51. Hunter, J.D. (2007). Matplotlib: a 2D graphics environment. Comput. Sci. Eng. *9*, 90–95.
52. Carlson, M., Falcon, S., Pages, H., and Li, N. (2019). Org. Hs. Eg. Db: Genome Wide Annotation for Human. R package version *3*.
53. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. Innovation *2*, 100141. https://doi.org/10.1016/j.xinn.2021.100141.
54. Yarberry, W. (2021). Dplyr. In CRAN Recipes (Springer), pp. 1–58.
55. Wickham, H., Chang, W., and Henry, L. (2012). ggplot2. Computer Software. Retrieved from. http://ggplot2.org.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
GM12878, Hi-C	https://data.4dnucleome.org/files-processed/4DNFI1UEG1HD/	4dnucleome: 4DNFI1UEG1HD
GM12878,ATAC-seq	https://www.encodeproject.org/experiments/ENCSR637XSC/	ENCODE: ENCSR637XSC
GM12878, H3K27ac ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000AKC/	ENCODE: ENCSR000AKC
GM12878, H3K4me3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR057BWO/	ENCODE: ENCSR057BWO
GM12878, Cis-Regulatory Elements	https://www.encodeproject.org/annotations/ENCSR820WFY/	ENCODE: ENCSR820WFY
GM12878, H3K27me3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000DRX/	ENCODE: ENCSR000DRX
GM12878, H3K9me3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000AOX/	ENCODE: ENCSR000AOX
GM12878,CTCF ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000DKV/	ENCODE: ENCSR000DKV
GM12878,RAD21 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000BMY/	ENCODE: ENCSR000BMY
GM12878,SMC3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000DZP/	ENCODE: ENCSR000DZP
GM12878,POLR2A ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000EAD/	ENCODE: ENCSR000EAD
GM12878,EZH2 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000ARD/	ENCODE: ENCSR000ARD
GM12878,Annotation file	ftp://ftp.ensembl.org/pub/release-104/gff3/homo_sapiens/Homo_sapiens.GRCh38.104.chr.gff3.gz	Ensembl: Homo_sapiens.GRCh38.104.chr.gff3.gz
H1-hESC, Hi-C	https://data.4dnucleome.org/files-processed/4DNFI2TK7L2F/	4dnucleome: 4DNFI2TK7L2F
H1-hESC, ATAC-seq	https://data.4dnucleome.org/files-processed/4DNFICPNO4M5	4dnucleome: 4DNFICPNO4M5
H1-hESC, H3K27ac ChIP-seq	https://www.encodeproject.org/experiments/ENCSR880SUY/	ENCODE: ENCSR880SUY
H1-hESC, H3K4me3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR443YAS/	ENCODE: ENCSR443YAS
H1-hESC, Cis-Regulatory Elements	https://www.encodeproject.org/annotations/ENCSR597SZL/	ENCODE: ENCSR597SZL
H1-hESC, CTCF ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000BNH/	ENCODE: ENCSR000BNH
H1-hESC, RAD21 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000BLD/	ENCODE: ENCSR000BLD
H1-hESC, POLR2A ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000BHN/	ENCODE: ENCSR000BHN
K562,Hi-C	https://data.4dnucleome.org/files-processed/4DNFITUOMFUQ/	4dnucleome: 4DNFITUOMFUQ
K562,ATAC-seq	https://www.encodeproject.org/experiments/ENCSR483RKN/	ENCODE: ENCSR483RKN

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
K562,H3K27ac ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000AKP/	ENCODE: ENCSR000AKP
K562,H3K4me3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000AKU/	ENCODE: ENCSR000AKU
K562,H3K9me3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000APE/	ENCODE: ENCSR000APE
K562,H3K27me3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000EWB/	ENCODE: ENCSR000EWB
K562,Cis-Regulatory Elements	https://www.encodeproject.org/annotations/ENCSR301FDP/	ENCODE: ENCSR301FDP
K562,CTCF ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000DMA/	ENCODE: ENCSR000DMA
K562,RAD21 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000BKV/	ENCODE: ENCSR000BKV
K562,SMC3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000EGW/	ENCODE: ENCSR000EGW
K562,POLR2A ChIP-seq	https://www.encodeproject.org/experiments/ENCSR388QZF/	ENCODE: ENCSR388QZF
K562,EZH2 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000AQE/	ENCODE: ENCSR000AQE
IMR90,Hi-C	https://data.4dnucleome.org/files-processed/4DNFIH7TH4MF/	4dnucleome: 4DNFIH7TH4MF
IMR90,ATAC-seq	https://www.encodeproject.org/experiments/ENCSR200OML/	ENCODE: ENCSR200OML
IMR90,H3K27ac ChIP-seq	https://www.encodeproject.org/experiments/ENCSR002YRE/	ENCODE: ENCSR002YRE
IMR90,H3K4me3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR087PFU/	ENCODE: ENCSR087PFU
IMR90,H3K9me3 ChIP-seq	https://www.encodeprojct.org/experiments/ENCSR055ZZY/	ENCODE: ENCSR055ZZY
IMR90,Cis-Regulatory Elements	https://www.encodeproject.org/annotations/ENCSR599FOY/	ENCODE: ENCSR599FOY
IMR90,CTCF ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000EFI/	ENCODE: ENCSR000EFI
IMR90,RAD21 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000EFJ/	ENCODE: ENCSR000EFJ
IMR90,SMC3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000HPG/	ENCODE: ENCSR000HPG
IMR90,POLR2A ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000EFK/	ENCODE: ENCSR000EFK
HepG2,Hi-C	https://data.4dnucleome.org/experiment-set-replicates/4DNESC2DEQIJ/	4dnucleome: 4DNESC2DEQIJ
HepG2,ATAC-seq	https://www.encodeproject.org/experiments/ENCSR042AWH/	ENCODE: ENCSR042AWH
HepG2,H3K27ac ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000AMO/	ENCODE: ENCSR000AMO
HepG2,H3K4me3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR575RRX/	ENCODE: ENCSR575RRX
HepG2,H3K9me3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000ATD/	ENCODE: ENCSR000ATD
HepG2,H3K27me3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000AOL/	ENCODE: ENCSR000AOL

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
HepG2,CTCF ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000AMA	ENCODE: ENCSR000AMA
HepG2,RAD21 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000EEG/	ENCODE: ENCSR000EEG
HepG2,SMC3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000EDW/	ENCODE: ENCSR000EDW
HepG2,POLR2A ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000EEM/	ENCODE: ENCSR000EEM
HepG2,EZH2 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000ARI/	ENCODE: ENCSR000ARI
HepG2,CEBPB ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000BQI/	ENCODE: ENCSR000BQI
Hela,Hi-C	https://data.4dnucleome.org/experiment-set-replicates/4DNESCMX7L58/	4dnucleome: 4DNESCMX7L58
Hela,Hi-C	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138543	NCBI: GSE138543
Hela,H3K27ac ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000AOC/	ENCODE: ENCSR000AOC
Hela,H3K4me3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000AOF/	ENCODE: ENCSR000AOF
Hela,H3K27me3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000APB/	ENCODE: ENCSR000APB
Hela,H3K9me3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000AQO/	ENCODE: ENCSR000AQO
Hela, CTCF ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000AOA/	ENCODE: ENCSR000AOA
Hela, CTCF ChIP-seq (Hex5)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE138543	NCBI: GSE138543
Hela, POLR2A ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000BGO/	ENCODE: ENCSR000BGO
Hela,RAD21 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000EDE/	ENCODE: ENCSR000EDE
Hela, SMC3 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000ECS/	ENCODE: ENCSR000ECS
Hela, EZH2 ChIP-seq	https://www.encodeproject.org/experiments/ENCSR000ATC/	ENCODE: ENCSR000ATC
Prediction of model trained using Hi-C, ATAC-seq and H3K4me3 ChIP-seq in GM12878 cell line	http://mged.nmdms.ustb.edu.cn/storage/data/28463526	NMDMS: https://doi.org/10.12110/mater10.121.NKRDP.20221209.ds.63930883e571e2448aaed532
Prediction of model trained using Hi-C, ATAc-seq and H3K27ac ChIP-seq in GM12878 cell line	http://mged.nmdms.ustb.edu.cn/storage/data/28463525	NMDMS: https://doi.org/10.12110/mater10.121.NKRDP.20221209.ds.63930883e571e2448aaed532
Prediction of model trained using Hi-C, ATAC-seq H3K4me3 and H3K27ac ChIP-seq in GM12878 cell line	http://mged.nmdms.ustb.edu.cn/storage/data/28463524	NMDMS: https://doi.org/10.12110/mater10.121.NKRDP.20221209.ds.63930883e571e2448aaed532
Prediction of model trained using Hi-C, H3K9me3 and H3K27me3 ChIP-seq in GM12878 cell line	http://mged.nmdms.ustb.edu.cn/storage/data/28463523	NMDMS: https://doi.org/10.12110/mater10.121.NKRDP.20221209.ds.63930883e571e2448aaed532
Software and algorithms		
fithic2	https://github.com/ay-lab/fithic	github: fithic
mustache	https://github.com/ay-lab/mustache	github: mustache

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
HiCDB	https://github.com/ChenFengling/HiCDB	github: HiCDB
juicer	https://github.com/aidenlab/juicer	github: juicer
SDOC	https://github.com/birmjiangs/Code- for-main-results	github: SDOC
MINE toolkit	https://github.com/MICL-biolab/MINE	zenodo: https://doi.org/10.5281/zenodo.7388730

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Yang Chen (yc@ibms.pumc.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table. The full processed data (ie, the model predictions when trained on different datasets) has been deposited at a general use repository NMDMS (http://nmdms.ustb.edu.cn/),[43] and the accession number of datasets is https://doi.org/10.12110/mater10.121.NKRDP.20221209.ds.63930883e571e2448aaed532. In addition, this data will be shared directly by the lead contact upon request.

The analysis code is available in the GitHub repository (https://github.com/MICL-biolab/MINE). The pipeline of using MINE toolkit is provided in Methods S1I–S1N).

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data description

The description of Hi-C data. Hi-C data for six human cell lines (GM12878, IMR90, K562, H1-hESC, HepG2, and Hela) are downloaded from https://data.4dnucleome.org/ [31] under accession number 4DNFI1UEG1HD, 4DNFIH7TH4MF, 4DNFITUOMFUQ,[32] 4DNFI2TK7L2F,[44] 4DNFICSTCJQZ, and 4DNESCMX7L, respectively.

The description of epigenome data that are used for training. For the active model, ATAC-seq data and ChIP-seq data of H3K27ac and H3K4me3 mainly target transcription-related factors are selected, for the repressive model, ChIP-seq data of H3K9me3 and H3K27me3 that are associated with gene repression are selected.

To verify that the MINE-Loop method can help identify more regulatory chromatin interactions, the CTCF, RAD21, SMC3, POLR2A ChIP-seq data are selected to verify the gene transcription-related interaction, and POLR2A, EZH2 ChIP-seq data are selected to verify the gene repression-related interactions.

To verify whether the loops called from the MINE-enhanced-hic can overlap the loops called from ChIA-PET data, CTCF and POLR2A ChIA-PET data in HepG2 cell line were obtained from ENCODE[45] under accession number ENCSR411IVB and ENCSR857MYZ, respectively.

Down sampling Hi-C data

Downsampling Hi-C matrix were simulated by downsampling the VC-normalized Hi-C matrix obtain from.hic format file using juicer-box[46] (In order to ensure that the contact matrix of all chromosome can be extracted from the hic format file (hic is a popular used format), we choose VC-normalization method. Because KR-normalization method may not support for some chromosomes at 1kb resolution due to high sparsity (Knight and Ruiz, 2013), ICE is provided by hiclib with the input format of hdf5, hm, bychr (HDF5), or is provided by HiC-Pro the input format of SAM, validpair, but is not supporting for hic format file. In the future, we will consider train the KR-normalized and ICE-normalized contact matrix provided by other data format. But, up to now, in order to ensure that most of the obtained HI-C data can be used for mine-loop prediction, we only train MINE-Loop model using VC-normalized Contact matrix.). The downsampling strategy is described as Methods S1C: *downsampling ratio* $= \frac{1}{s \times s}$ is defined, then, values of the $s \times s$ window were all set as the average value of the $s \times s$ window.

Complete downsampling Hi-C matrices

To obtain more Hi-C matrix features, we first synthesize these downsampling Hi-C matrices by point-to-point addition, and then complete the Hi-C matrix after addition through the FAN method,[25] a matrix completion algorithm suited for the sparse input matrix whose 99% pixels are randomly missing. For the FAN method, we first define the row or column of the matrix after superimposing as f, define the signal as g containing ones and zeros with similar length of f, as described in Equation 1. Then, f and g are convolved with the same kernel h. Finally, we can obtain the completed matrix by performing an element-wise division of the two convolved signals, as Equation 2 shows.

$$g[n] = \begin{cases} 1 & if\ \exists f[n] \\ 0 & otherwise \end{cases} \quad \text{(Equation 1)}$$

$$HR_c[n] = \frac{(f * h)[n]}{(g * h)[n]} \quad \text{(Equation 2)}$$

Where n represents the position of the signal, g[n], f[n], and $HR_c[n]$ represent the signal g, f, and the finally obtained complete Hi-C data at position n, respectively.

Preproces epigenome data (Generation of Correlation matrix)

Same as Hi-C matrix, we divided the chromosome into n fragments with a length of len (in this paper, len = 1000 base), and calculated the signal p value f_i of the *ith* fragment using ChIP-seq or ATAC-seq data as the following formula to get a $n \times 1$ matrix.

$$f_i = \frac{\sum_{j=1}^{s_i} (location(end_j) - location(start_j)) \times p\_value_j}{len} \quad \text{(Equation 3)}$$

Where s_i is the number of fragments divided by ChIP-seq data within the fragment i, $location(start_j)$ and $location(end_j)$ is the start and end location of the jth fragment in ChIP-seq data, p_value_j is the signal p value of the jth fragment in ChIP-seq data.

Then we combine multiple matrixes in columns to obtain a feature matrix $F = [F_1 \quad F_2 \quad ... \quad F_m]^T$ and normalize it to be 0∼1 with $n \times m$ size, where m is the sample number of the epigenomics data. We define the i-th row of feature matrix F as $F_i = [f_{i,1} \quad f_{i,2} \quad ... \quad f_{i,n}]$, the average of matrix F as $\overline{F}$, $\tilde{F} = F - \overline{F} = [F_1 - \overline{F}_1, \quad F_2 - \overline{F}_2, \quad ..., \quad F_n - \overline{F}_n]^T = [\tilde{F}_1, \quad \tilde{F}, \quad ..., \quad \tilde{F}_n]^T$, the transpose of $\tilde{F}$ as $\tilde{F}'$, $dot = [\tilde{F}_1^{\,2}, \quad \tilde{F}_2^{\,2}, \quad ..., \quad \tilde{F}_n^{\,2}]^T$, then the Pearson correlation coefficient between pairwise of F can be calculated as $PCC_F = \frac{\tilde{F} \cdot \tilde{F}'}{\sqrt{dot \cdot dot'}}$, where dot' is the transpose of dot.

In order to explore the biological patterns represented by the relationship matrix, we count the number of loops identified by FitHiC from Hi-C matrix, ATAC-Seq, H3K27ac, CTCF, and H3K27me3 Chain-seq peaks in GM12878 cell line. Methods S1E shows that the number of loops identified by the Hi-C data is much smaller than the number of peaks identified by the epigenomic data, suggesting that a high proportion of regulatory chromatin interactions cannot be identified solely by the method of identifying loops from Hi-C data. Since the formation of chromatin loops is related to the two boundaries of chromatin topological domains (TADs) in a certain proportion (Anania et al., 2022; Crane et al., 2015; Tang et al., 2015), and a large number of epigenomic signals, such as CTCF, histone marker H3K4me3 and other factors, are enriched at the boundaries of TADs (Dixon et al., 2012) (Yu et al., 2017). Therefore, we add the loops (TADs boundaries) features by calculating the epigenomic correlation matrix. If the signals have same trends, then the correlation value is positive, else the correlation value is negative. For example, we visualize the ChIP-seq/ATAC-seq correlation matrix (lower triangle in Methods S1F) and the corresponding Hi-C matrix (upper triangle in Methods S1F) in the NBN genomic region (chr8: 89,920,000–90000000). We choose the genomic regions of A, B, C, D, E, where B, C, E have peaks in the ChIP-seq track, A, D do not have signals in the ChIP-seq track. Methods S1F shows that the correlation values between A and B, A and C, A and D, B and D are −1, the correlation values between A and D, C and E are +1. The results show that the peaks in the correlation matrix can anchor the locations of loops in the Hi-C matrix. It can be seen that the loops in the Hi-C matrix can be well increased by adding epigenomic relationship matrix for model training.

Matrix normalization

For the data used in the training, **the completed matrix** obtained from section "Complete downsampling Hi-C matrices", **Pearson correlation coefficient matrix** obtained from section "Preproces epigenome data (Generation of Correlation matrix)" and **VC-normalized Hi-C matrix** obtain from.hic format file using juicerbox[46] all need to be normalized as Equations 4, 5, 6, and 7 shows. We define Val_{ij} as the value of row i and column j, num pairs as the number of pairs whose value is not zero, nums as the 1/1000 of the sum of num pairs as Equation 1, max_*num* as the minimum value of the largest nums value in the matrix. We first set the values greater than max_*num* to be max_*num* for these matrices, then do normalization as Equations 6 and 7 to limit the value to be between 0 and 255.

$$nums = \left(\sum num\_pairs\right) \Big/ 1000 \quad \text{(Equation 4)}$$

$$Val_{ij} = \max\_num \quad if \quad Val_{ij} \geq \max\_num \tag{Equation 5}$$

$$Val_{ij} = \begin{cases} (Val_{ij}/10) \times \lg(10) & if \quad Val_{ij} \leq 10 \\ \lg(Val_{ij}) & if \quad Val_{ij} > 10 \end{cases} \tag{Equation 6}$$

$$Val = (Val / \max(Val)) \times 255 \tag{Equation 7}$$

Generation of Masked-hic

In order to ensure that the MINE-Loop method can obtain more regulatory chromatin interactions, the VC-normalized Hi-C matrix after normalization as Equations 4, 5, 6, and 7 was masked with the Candidate Cis-Regulatory Elements or other ChIP-seq data to be the target high-resolution Hi-C used in training **(Masked-hic).** The mask operation obeys the following principle: for ChIP-seq data, only these interaction values with peaks in the ChIP-seq data at both positions will be retained; for the *cis*-regulatory element, we first set values of all locations in the *cis*-regulatory element file to be 1, then retain the interaction values of Hi-C matrix when values of both ends in the *cis*-regulatory element file are 1.

With the above operations, we can obtain all data for training: completed Hi-C (**Completed-hic**), Pearson correlation coefficient matrix (M_c) and the target high-resolution Hi-C (**Masked-hic**).

In order to compare with the raw high-resolution Hi-C matrix at the same scale, the raw high-resolution Hi-C matrix was also normalized to be 0–255 as Equations 4, 5, 6, and 7. In the remaining manuscript, we name **the Hi-C used for comparison with the enhanced Hi-C as Raw-hic**.

Matrix-based sample division

Divide the Masked-hic, Completed-hic and M_c into $n \times K \times K$ sub-matrices, where n represents the number of sub-matrices, K is the dimension of sub-matrix (in this paper, we define K = 400), each sub-matrix is treated as a sample. The reason for choosing a400×400 size is that the average size of TADs is 0.4Mb∼2Mb,[47] according to the previous data enhancement algorithm HiCPlus,[48] the interaction within the genomic distance of $0.4Mb \times 0.4Mb$ can retain more local information. As shown in Methods S1D, since the Hi-C matrix is a symmetric matrix, only the upper right part of the diagonal is reserved. Along the upper right corner of the Hi-C matrix, sub-matrices of400×400 are taken as samples from top to bottom in a step of 400 kb. Each small square in the figure represents a400×400 sub-matrix, and 5 sub-matrices are taken (that is, the two interaction sites are within the 2Mb genome range, because the average size of TADs is within the 1Mb genome distance range. Outside of TADs, few significant interactions are existing).

Structure of the MINE-loop network

The implementation of the network is shown in Methods S1A. The network structure is divided into three type layers: MINE_Conv, maxPool2D, and ConvTranspose_2D as Equations 8, 9, 10, and 11 show. The first type (*i.e.*, MINE_Conv in Methods S1A, containing Conv2d, BatchNorm2d, ReLU, Conv2d, BatchNorm2d, and ReLU, is proposed to extract and present the Hi-C pattern. The second (*i.e.*, MaxPool2d in Methods S1A), is designed to perform dimensionality reduction operations. The third (*i.e.*, ConvTranspose2d in Methods S1A), is designed to perform an upsampling operation. As Methods S1A shows, we divided the network's input into two parts: the completed Hi-C data HR_c and the correlation matrix of the epigenomics data M_c. For HR_c, we use MEMR_Conv Module following with a max pool layer to do downsampling, and ConvTranspose 2D layer to do upsampling times to get a result (c5). For M_c, we use MINE_Conv Module three times to get a result (e3) and merge c5 and e3, and put them into the network (MINE_Conv Module two times) to get the final enhanced Hi-C matrix.

$$f(X) = \max(0, w_1 * X + b_1) \tag{Equation 8}$$

$$w_1 = \frac{\gamma w_0}{\sqrt{Var(X) + \varepsilon}} \tag{Equation 9}$$

$$b_1 = \frac{\gamma(X - mean(X))}{\sqrt{Var(X) + \varepsilon}} + \beta \tag{Equation 10}$$

$$F_1 = f(f(X)) \tag{Equation 11}$$

where $eps = 1e-5$, Var represents the variance, γ, β represent the learned coefficient matrix gamma and beta, F_1 represents the output of MINE_Conv.

Model training and testing for MINE-loop

Data (**Completed-hic, correlation matrix and Masked-hic**) of chromosome 1–17 in GM12878 cell line are used for training the model in this paper, data of chromosome 18–22 in GM12878 are used for testing, and the ChIP-seq data of transcription factors (RAD21, SMC3, POLR2A) from GM12878, IMR90, K562, H1-hESC and HepG2 cell lines are used for verification.

MINE-enhanced-hic matrix and Masked-hic matrix are used as the input of L1 Loss (Equation 12) and Perceptual Loss[49] (Equation 13) for comparison and scoring. Since MINE-enhanced-hic may be very sparse in some subgraphs, this will affect the judgment of the Perceptual Loss on the result and indirectly affect the training result. Therefore, we remove these training data, if the number of value points of the attention interaction sub-matrix is less than 10% of the sub-matrix scale.

$$L_1 = |y' - y| \quad \text{(Equation 12)}$$

$$\ell(y', y) = \frac{1}{C_j H_j W_j} \|\varphi_j(y') - \varphi_j(y)\|_2^2 \quad \text{(Equation 13)}$$

Where $\varphi_j(y)$ is a feature map of shape $C_j \times H_j \times W_j$, the loss is a loss using the squared, normalized Euclidean distance between features.

As Table 1 shows, MINE-loop can be divided into active model and repress model with different epigenome data as the input of model training and Masked-hic. For the active model, ATAC-seq, H3K27ac, H3K4me3 ChIP-seq are chose to be epigenome data to train model; For the repressive model, H3K27me3, H3k9me3 ChIP-seq are chose to be epigenome data to train model. For the training of active model, we chose the *cis*-regulatory element file as the data source of the attention matrix for training, since there are many attention interactions and the model can be fully trained, we choose the best result when the loss score of the test set is the lowest. For the training of repressive model or when the number of attention points in the matrix is too small, we believe that the model is prone to functional overfitting due to too few comparison points. Therefore, we use the iterative version model of the test dataset with the lowest validation loss score as the final trained model.

Spatial density of regulatory chromatin interactions

The calculation steps of SD-RCI are as follows: (1) the Pastis-PM2[35] algorithm was used to get the 3D coordinates of these bins for the TADs; (2) calculate the volume of TAD_i ($volume_{TAD_i}$) constructed by the 3D coordinates (X, Y, Z) of these bins as the raw volume of TAD_i; (3) calculate the raw SD-RCI of TAD_i: $SD - RCI_i = \frac{num_{Loops}}{volume_{TAD_i}}$, where num_{Loops} is the total number of loops in each TAD region. (4) the number of total loops may increase or decrease as sequencing depth or loop-calling algorithms changes. Same as SDOC,[20] we used quantile normalization to normalize the raw SD-RCI value to Gaussian distribution (mean = 0, SD= 1). (5) the SD-RCI of the ith TAD in a chromosome can be formulated as follows. Where the TADs are identified by predicted by HiCDB[50] using the following parameters: HiCDB({'/home/data/sample'},10,000,'hg38','ref','hg38'), where resolution = 10,000, chrsizes = hg38; ref = hg38.

$$SD - RCI_i = \sum_{j=0}^{n, j \neq i} SD - RCI_j F_{ijnorm} \quad \text{(Equation 14)}$$

Where n is the number of TADs predicted by HiCDB,[50] $SD - RCI_i$ is the ith TAD on a same chromosome, F_{ijnorm} is the TAD-TAD normalized contact frequency between TAD_i and TAD_j can be formulated as follows:

$$F_{i,j} = \frac{m}{L_i \times L_j} \quad \text{(Equation 15)}$$

$$F_{ijnorm} = \frac{F_{i,j} - \mu_d}{\delta_d} \quad \text{(Equation 16)}$$

Where L_i, L_j represent the length of TAD_i, TAD_j in a chromosome, m is the sum of contact frequency between TAD_i and TAD_j, μ_d is the loess regressed pairwise contact frequency, δ_d is the loess regressed SD calculated by the lowess function of statsmodels Python library. We let $F_{ijnorm} = 0$ if genomic distance$dist_genomic(TAD_i, TAD_j) < 2Mb$ or $F_{ijnorm} < 1$.

Detail of Pastis-PM2

Pastis-PM2[35] algorithm assumes that the counts between two loci in Hi-C contact matrix follow a Poisson distribution whose intensity decreases with the physical distances between the loci. Pastis-PM2 can automatically adjust the transfer function relating the spatial distance to the Poisson intensity and infer a genome structure that best explains the observed data. Pastis-PM2 treats each loci as a point, therefore, the output of Pastis-PM2 is the 3D coordinates of all loci.

Calculation of TAD volume

Given a TAD, the 3D coordinates of all loci in the TAD can be calculated by Pastis-PM2[35] algorithm. Then, the 3D coordinates of all loci in the TAD are used to calculate a convex hull, where the volume of the convex hull is defined as the TAD volume.

Expression changes associated with level changes in HepG2 cell line

Genes were classified based on the levels (ultra_high, high, middle, low) divided by SD-RCI in HepG2 cell line. The counts file of RNA-seq data in the HepG2 cell line was obtained from GEO under accession number GSE117815, RPKM value was calculated as the following formula.

$$RPKM_i = \frac{count_i \Big/ \left(\frac{\sum_i^n count_i}{1000000}\right)}{len_i} \quad \text{(Equation 17)}$$

Where $RPKM_i$ is the ith gene's RPKM value, $count_i$ is the ith gene's count number, len_i is the length of the ith gene.

Important definitions in MINE work

1) Hi-C: High-through chromosome conformation capture.

2) RCI: regulatory chromatin interaction.

3) Raw-hic: the raw high-resolution Hi-C matrix calculated from deeply sequenced Hi-C data using VC-normalized method used to do comparative analysis.

4) Masked-hic: the 1 kb high-resolution Hi-C matrix from raw high-resolution Hi-C masked by location of interest like candidate Cis-Regulatory Elements.

5) MINE-enhanced-hic: the 1 kb resolution Hi-C matrix predicted by MINE method.

6) Completed-hic: Hi-C matrix completed from many downsampling Hi-C matrix with different downsampling ratio (the visualization of Raw-hic, MINE-enhanced-hic, and Completed-hic can be seen from Methods S1G**)**.

7) active model: The MINE-Loop model trained using ATAC-seq data and ChIP-seq data for different targeted factors (e.g., H3K4me3 and H3K27ac) specifically involved in DNA transcription.

8) repressive model: The MINE-Loop model trained using ChIP-seq data target with suppression-related epigenomic marks (i.e., H3K27me3, H3k9me3).

9) active loops: loops identified from the active model.

10) repressive loops: loops identified from the repressive model.

11) raw-Loop-fithic: loops called from Raw-hic using FitHiC2.

12) active-Loop-fithic: loops called from MINE-enhanced-hic of active model using FitHiC2.

13) raw-Loop-mustache: loops called from Raw-hic using mustache.

14) active-Loop-mustache: loops called from MINE-enhanced-hic of active model using mustache.

15) SD-RCI: the spatial density of regulatory chromatin interactions, i.e., the ratio of the total number of active or repressive interactions in a TAD to the entire 3D space taken up by the physical structure of the TAD.

16) Four levels of TADs' SD-RCI value: ultra-high, high, middle and low based on the value of δ in the Gaussian distribution of SD-RCI.

17) The TADs were categorized into four types (①SD-RCI <0.6 and control volume < Hex volume, ②SD-RCI $\geq$ 0.6 and control volume > Hex volume, ③SD-RCI <0.6 and control volume > Hex volume, ④SD-RCI $\geq$ 0.6 and control volume < Hex volume, where 0.6 was the SD-RCI inflection point when count became positive calculated from the gene count - SD-RCI curve as shown in Figure S11) according to the change in TAD volume (whether increased or decreased) and the size of SD-RCI before and after drug treatment.

18) active hubs: TAD regions that are enriched with active factors.

19) developed active hubs: TAD regions that are enriched with active factors, and the SD-RCI level is middle, high and ultra-high.

20) developing active hubs: TAD regions that are enriched with active factors, and the SD-RCI level is low.

21) repressive hubs: TAD regions that are enriched with repressive factors.

22) developed repressive hubs: TAD regions that are enriched with repressive factors, and the SD-RCI level is middle, high and ultra-high.

23) developing repressive hubs: TAD regions that are enriched with repressive factors, and the SD-RCI level is low.

QUANTIFICATION AND STATISTICAL ANALYSIS

In this paper, we mainly verify whether the MINE-enhanced-hic matrix can be used to detect more regulatory chromatin interactions than from the Raw-hic matri from the following aspects:

Model verification

As shown in Methods S1B, we move some import network model and train network again.

Verify Biologically

1) for the active model aimed at enhancing the interactions related to the promotion of transcription, we choose to explore the overlap number of loops identified by MINE-enhanced-hic and Raw-hic within different genomic distance range (2-100kb, 2-300kb and 2-500kb genomic distance), the number of CTCF, RAD2, SMC, POLR2A TFs, and transcription start site (TSS) anchoring around loops called from MINE-enhanced-hic and Raw-hic, to verify the MINE-Loop method can help to identify more loops related to the promotion of transcription.
2) Verify the influence of different types of ChIP-seq data combinations on the effect of MINE-Loop: study the influence of the model obtained by combining different epigenome data used for MINE-Loop training input on the data enhancement effect.
3) Verify whether MINE-enhanced-hic can enrich more functional genes: Go enrichment analysis was performed on immune-activated cells (human B lymphocyte line GM12878) and human liver cancer cell line (HepG2) to verify whether MINE-enhanced-hic can enrich more functional genes.

Verify the general applicability of MINE-Loop model

1) the ability to do prediction in other cell lines data by using the active model trained by the GM12878 cell line.
2) the ability to do prediction with different combinations of histone target ChIP-seq data as the prediction input.

Call loops using FitHiC2 and mustache

By surveying the input data requirements for different loop callers (Methods S1H), we found only MUSTACHE and Fithic2 can identify loops with contact matrix (or some kind of data format that the contact matrix can be converted to) as input, while HiCCUPS[28] call loops from.hic format file, cLoops[29] call loops require Mapped PETs info, HiC-ACT[30] call loops from the output file from other methods (such as Fit-Hi-C/FitHiC2). Due to the output file predicted by MINE-Loops being only a contact matrix ($n \times n$ size), we only integrated MUSTACHE and Fithic2 in our MINE-Loop tool to call loops by transforming the Hi-C contact matrix ($n \times n$ size).

Loops from Hi-C data are called by FitHiC2[26] with parameters as the following: resolution = 1kb, distLowThres = 2kb, and distUpThres = 100, 300 or 500kb, p value<0.015, q-value (FDR obtained by applying Benjamini-Hochberg correction to the p values) <0.015. We renamed loops called from Raw-hic using FitHiC2 as raw-Loop-fithic, loops called from MINE-enhanced-hic of active model using FitHiC2 as active-Loop-fithic.

For the mustache[27] tool, loops from raw high-resolution Hi-C data were called from the.hic format file with default parameters. loops from MINE-enhanced-hic matrix were called from text format with parameters as the following: resolution = 1kb, pt (p-Value threshold) = 0.5. We renamed loops called from Raw-hic using mustache as raw-Loop-mustache, loops called from MINE-enhanced-hic of active model using mustache as active-Loop-mustache.

Overlap of raw-hic and MINE-enhanced-hic

Both ends of loops called separately from Raw-hic and MINE-enhanced-hic are anchored within 2kb genomic distance are defined as overlapping.

Call loops using ChIA-PET2

Loops from ChIA-PET data are called by ChIA-PET2 tool[10] with parameters and limitations as the following: -m 1 -A ACGCGATATCTTATC -B AGTCAGATAAGATAT.

Analysis of loops anchoring transcription factor

Calculate counts of CTCF, RAD21, SMC3, POLR2A, EZH2 CTCF, RAD21, SMC3 peak around loops (distance to the loop anchor point: −40kb∼+40kb) called from Raw-hic and MINE-enhanced-hic. curves of factors peak count with distance to loop anchor point were plotted using python library matplotlib.[51]

Go enrichment analysis in differential loop anchors

The differential loops called from MINE-enhanced-hic and Raw-hic refer to the un-overlapped loops regions (the pink and green regions) as shown in Figure S2A. Gene Go enrichment analysis were performed using R package org.Hs.eg.db[52](3.14.0), clusterProfiler[53] (4.0), dplyr[54] (1.0.7) and ggplot2[55] (3.3.5).