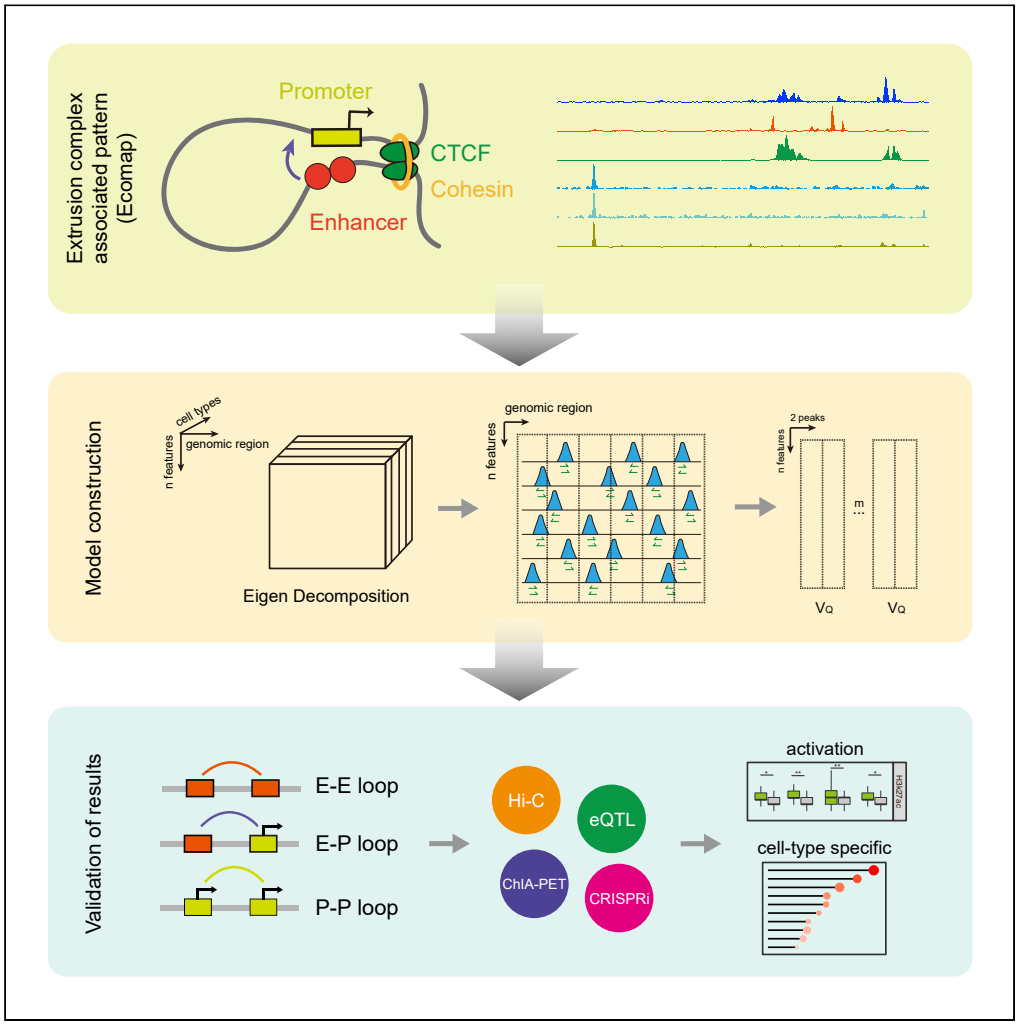
iScience

CellPress
OPEN ACCESS

Article

Integrating extrusion complex-associated pattern to predict cell type-specific long-range chromatin loops



Yajing Deng, Li Tang, Xiaolong Zhou, Wenkang Wang, Min Li

limin@mail.csu.edu.cn

Highlights

Extrusion complex-associated pattern facilitates the prediction of chromatin loops

Ecomap-loop outperforms the state-of-the-art unsupervised methods

Ecomap-loop can predict active functionality-related and cell type-specific loops

Deng et al., iScience 25, 105687
December 22, 2022
https://doi.org/10.1016/j.isci.2022.105687

iScience

CellPress
OPEN ACCESS

Article

Integrating extrusion complex-associated pattern to predict cell type-specific long-range chromatin loops

Yajing Deng,[1,2,3] Li Tang,[1,3] Xiaolong Zhou,[1] Wenkang Wang,[1] and Min Li[1,4,*]

SUMMARY

The chromatin loop plays a critical role in the study of gene expression and disease. Supervised learning-based algorithms to predict the chromatin loops require large priori information to satisfy the model construction, while the prediction sensitivity of unsupervised learning-based algorithms is still unsatisfactory. Therefore, we propose an unsupervised algorithm, Ecomap-loop. It takes advantage of extrusion complex-associated patterns, including CTCF, RAD21, and SMC enrichments, as well as the orientation distribution of CTCF motif of loops to build feature matrices; then the eigen decomposition model is employed to obtain the cell type-specific loops. We compare the performance of Ecomap-loop with the state-of-the-art unsupervised algorithm using Hi-C, ChIA-PET, expression quantitative trait locus (eQTL), and CRISPR interference (CRISPRi) screen data; the results show that Ecomap-loop achieves the best in four cell types. In addition, the functional analysis reveals the ability of Ecomap-loop to predict active functionality-related and cell type-specific loops.

INTRODUCTION

The three-dimensional architecture of chromatin plays an important role in maintaining normal gene expression levels,[1,2] marking cell specificity, and regulating cell growth and development.[3–5] In the past decade, the development of chromosome conformation capture (3C) techniques[6] facilitated the identification of chromatin interactions. To obtain higher resolution and throughput, the 3C-based assay has evolved from "one-to-one" (3C)[6] to "one-to-many" (circular chromosome conformation capture/chromosome conformation capture-on-chip, 4C),[7] "many-to-many" (chromosome conformation capture carbon copy, 5C),[8] and "high-throughput chromosome conformation capture" (Hi-C),[9,10] which enables high-throughput genome-wide remote chromatin interaction analysis. To detect the chromatin interactions mediated by specific proteins, chromatin interaction analysis using paired end tag sequencing (ChIA-PET)[11,12] and *in situ* Hi-C followed by chromatin immunoprecipitation (HiChIP) and proximity ligation-assisted ChIP-Seq (PLAC-seq)[13,14] have been proposed, which capture the effect of chromatin structure through the viewpoint of targeted proteins. The hierarchical chromatin organization can be studied mainly at four levels: chromosome territory, chromatin compartments, topologically associating domain (TAD), and chromatin loops.[3] The chromatin loops are the basic building blocks for the 3D architecture of chromatins, which establish regulatory networks between the distant elements through their physical proximity.[15] However, limited by the cost and technical issues of wet-lab experiments, it is still a great challenge to identify the chromosome loops of unrecognized cell types or species.

Recently, some machine learning-based algorithms have been emerged to solve the difficulties of identifying chromatin loops and investigating their regulatory function. These methods can be categorized into supervised and unsupervised, according to whether using 3C-identified loops for model training or not.[16,17] For the supervised algorithms, the multi-omics features or genomic sequences were usually used to construct the feature matrix, and the chromatin interactions from Hi-C, ChIA-PET, HiChIP, and so on were used to generate the positive and negative training sets.[18–21] However, these algorithms require large number of inputs to train the model, which is hard to apply to the uncharacterized cell types, and the prediction procedure usually takes a long time to complete. For the unsupervised algorithms, the distance and some other genomic characteristics were used as model features to infer the chromatin loops. Ernst et al.[22] and Thurman et al.[23] employed the correlations between enhancers and promoter

[1]Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, China

[2]School of Software, Xinjiang University, Urumqi 830091, China

[3]These authors contributed equally

[4]Lead contact

*Correspondence: limin@mail.csu.edu.cn

https://doi.org/10.1016/j.isci.2022.105687

DNaseI-hypersensitive sites (DHS) and the expression levels in specific regulatory regions to perform the prediction. PreSTIGE[24] utilized a linear domain model to link the enhancers to their target genes. EpiTensor[25] derived 3D interactions between distal genomic loci from 1D epigenomic data. Although these methods require less inputs and running time compared to the supervised methods, the accuracy of them is still unsatisfactory, which needs to be improved.[16]

With this in mind, we propose Ecomap-loop, which integrates extrusion complex-associated pattern to predict the cell type-specific chromatin loops. Ecomap-loop extracts the relevant patterns between three-dimensional structure of chromatin and epigenomics data and then generates the feature matrix with eigen decomposition, through which all the patterns are binned and summarized to the linear locus of genome. The predicted results are evaluated with 3C-based data, expression quantitative trait Loci (eQTLs), and clustered regularly interspaced short palindromic repeats interference or inhibition (CRISPRi) screen data, which showed that Ecomap-loop outperformed the other methods. Finally, the functional analysis indicated that Ecomap-loop can be used to predict the active functionality-related and cell type-specific loops.

RESULT

Extrusion complex-associated pattern (Ecomap) facilitates the prediction of chromatin loops

Considering the loop extrusion model, in which a complex, including the proteins CCCTC-binding factor (CTCF) and cohesin, mediates the formation of loops by a process of extrusion.[26] The cohesin complex consists of structural maintenance of chromosomes protein (SMC), double-strand break repair protein (RAD21), and so on.[27] CTCF is a widely expressed class of transcription factors that are important for the local anchoring of loop structures.[10,12] In the extrusion model, when a loop is established and the extrusion complex stops sliding, the DNA located around the extrusion complex is maintained rigid (Figure 1A). Recently, some studies uncovered that the orientation of the CTCF motif is critical for the formation of the loop, which includes convergent, tandem (leftward and rightward), and divergent motif patterns (Figure 1B). And there is an orientation preference of convergent that with higher contact frequency than the other orientations.[10,12]

To investigate whether the extrusion complex-associated pattern facilitates the prediction of chromatin loops, we firstly collected H3K27ac ChIA-PET loops and characterized the type of loops into enhancer-enhancer (E-E), enhancer-promoter (E-P), and promoter-promoter (P-P) (see STAR Methods). The characterization showed E-P loops occupied the most percentage (43.35%), followed by E-E loops (32.1%) (Figure 1C), which was consistent with the finding that H3K27ac-mediated loops identified functional enhancer interactions.[28] Then ChIP-seq datasets of CTCF, SMC, and RAD21 in K562 cell line were collected and mapped to the different types of H3K27ac ChIA-PET loops, respectively (see STAR Methods); the results indicated the ChIP-seq peaks of three proteins enriched near the anchors of loop. As the promoter anchor positions of H3K27ac-mediated loops were high transcriptional activity related, which tend to have higher ChIP-seq signal, thus, the P-P loops showed the highest enrichment of ChIP-seq peaks, followed by E-P loops (Figure 1D). And we annotated all the ChIA-PET loops with CTCF motifs, of which 78% were associated with bound CTCF at both anchors; within these associated loops, 64% were convergent, 33% were tandem, and 3% were in the divergent orientation (Figure 1E). Our analysis result was consistent with the previous finding that the convergent orientation was required for the formation of loops.[10,26,29,30] Then we used a random forest classifier to compute the feature importance with mean decrease in impurity (MDI), which was defined as the mean and SD of accumulation of the impurity decrease within each classifier tree (Figure 1F). We observed that the ranking of importance was identical between three types of loops, suggesting the long-range chromatin loops were predictable through these extrusion complex-associated patterns. And the orientation of CTCF motifs showed higher importance than the others, which provided the basis for subsequent model design. The feature importance of CTCF motif orientation is similar for three types of anchors, indicating that the feature occupied similar weight in the prediction model for all the loops, not affected by the anchor types.

Ecomap-loop: Integrating extrusion complex-associated pattern to predict cell type-specific long-range chromatin loops

As extrusion complex (CTCF, RAD21, and SMC) plays important role in the formation of loops, the bounding pattern of which has been proved to benefit the prediction of chromatin interactions.[12] Here we propose an unsupervised algorithm Ecomap-loop (Figure 2), which integrates the bounding pattern of extrusion complex to predict the long-range chromatin loops in a cell type-specific manner. The model

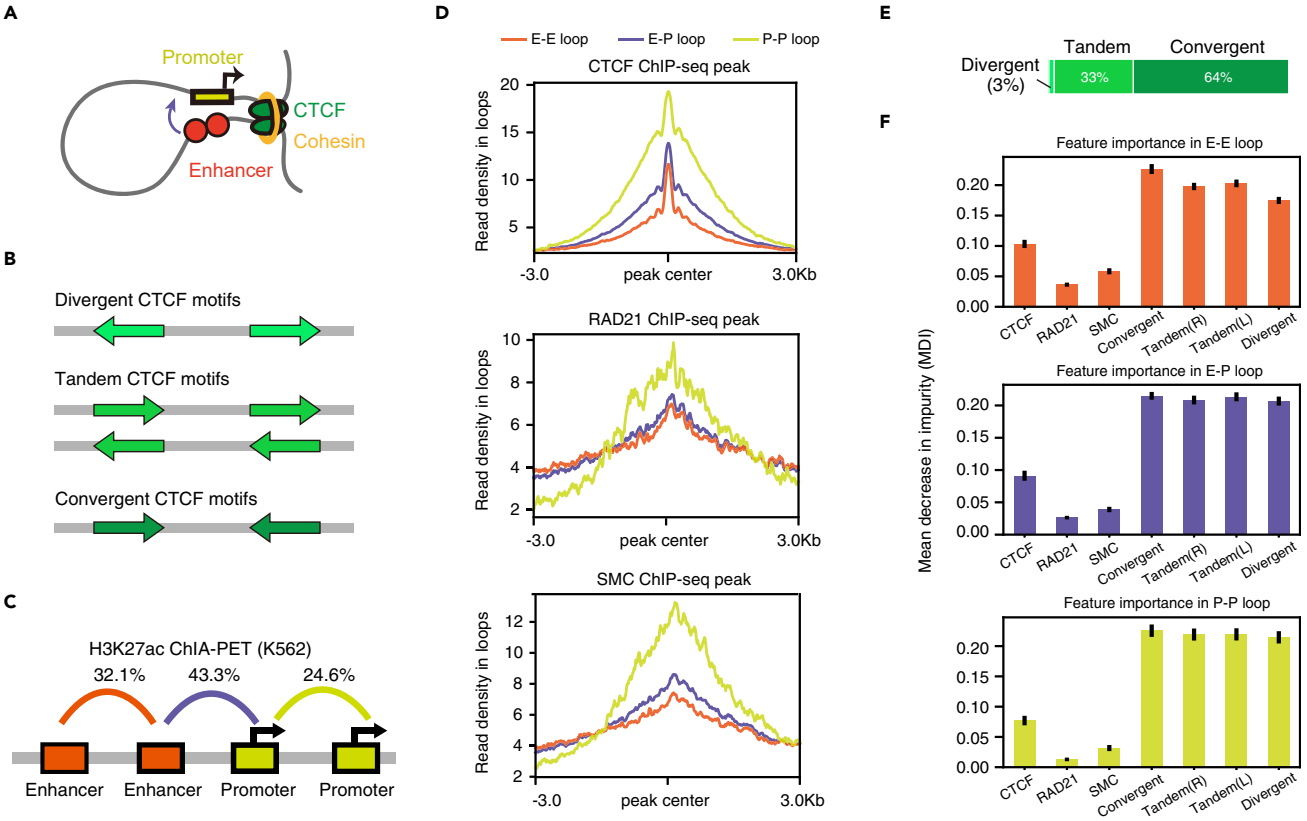


Figure 1. Extrusion complex-associated pattern (Ecomap) facilitates the prediction of chromatin loops

(A) Diagram of extrusion complex in loop formation.

(B) Diagram of three types of CTCF motif orientation (divergent, tandem, and convergent) bound to DNA.

(C) Percentage of E-E, E-P, and P-P loops in K562-H3K27ac ChIA-PET dataset.

(D) The ChIP-seq peak enrichment of CTCF, RAD21, and SMC in three types of loops of K562-H3K27ac ChIA-PET dataset.

(E) Percentage of divergent, tandem, and convergent CTCF motifs.

(F) MDI feature importance in three types of loops.

of Ecomap-loop can be divided into three parts: calculating the read coverage of CTCF, RAD21, and SMC on each fragment, the evaluation for CTCF motif orientations, and the eigen decomposition by using the assays including histone marks ChIP-seq and DNase-seq of different cell types (see STAR Methods).

Considering the final prediction effect and the changes of coverage rate led by the different lengths of different gene fragments, the number of base pairs is regarded as an evaluation indicator to present the coverage of CTCF, RAD21, and SMC on different fragments. Then the coverage value of three proteins is calculated on different fragments as V_c. Then we evaluate the matching probability and the orientations of CTCF motif for each fragment; a matching score V_F is calculated. As four orientations occur in different frequencies across all the loops, we assigned different weights to different orientations. Finally, the epigenomics data including histone mark ChIP-seq and DNase-seq data are used to build the eigen decomposition part, and we capture the peaks with covariation as chromatin interaction. We calculate an association score V_Q to measure the strength of interaction. The final score of Ecomap-loop is defined as the sum of the three parts.

Evaluation of predicted chromatin loops with 3C-based experimental datasets

To evaluate the prediction results of Ecomap-loop, we downloaded the Hi-C experiment data of K562, GM12878, IMR90, and HepG2 with 5–10 kb resolution[10] (see STAR Methods), which were regarded as positive samples. As most of the earlier unsupervised methods did not provide source code, here we used the state-of-the-art method EpiTensor for comparison. The predicted loops from EpiTensor were ranked in terms of their association scores (AS), and the predicted loops from Ecomap-loop were ranked in terms

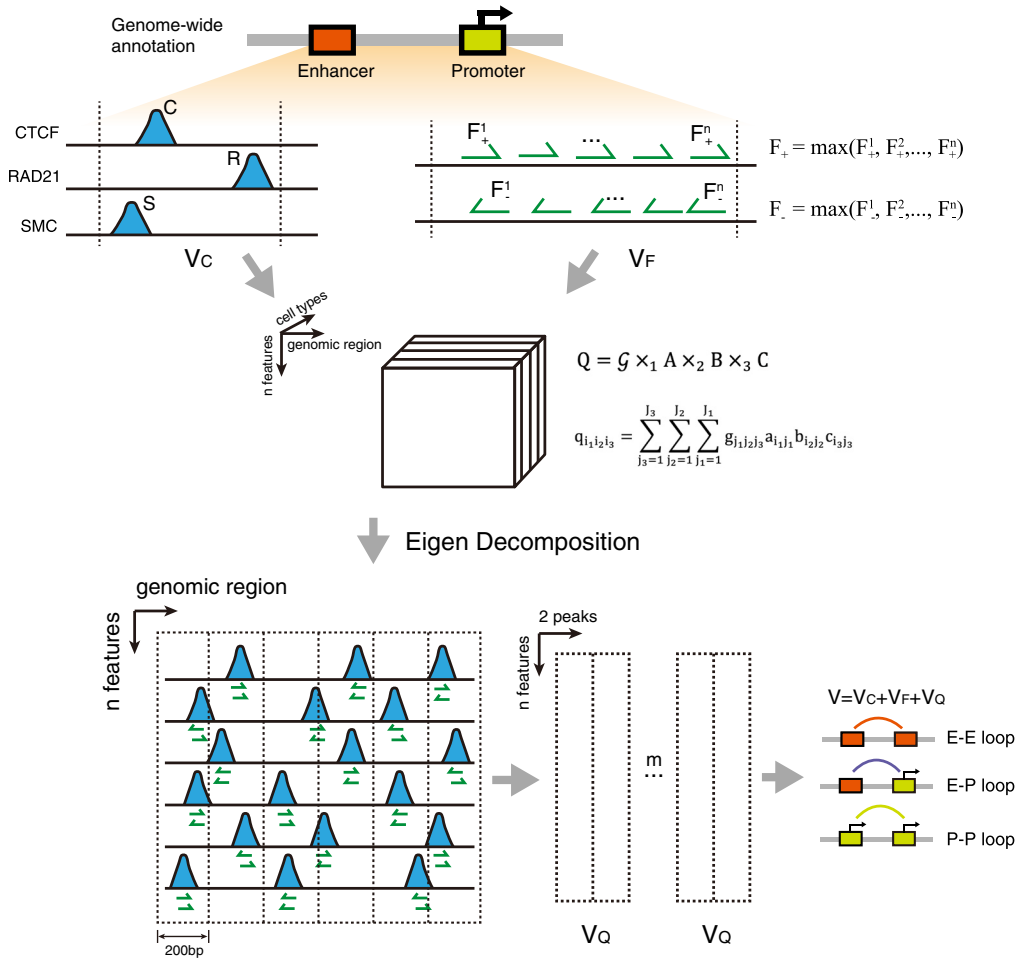


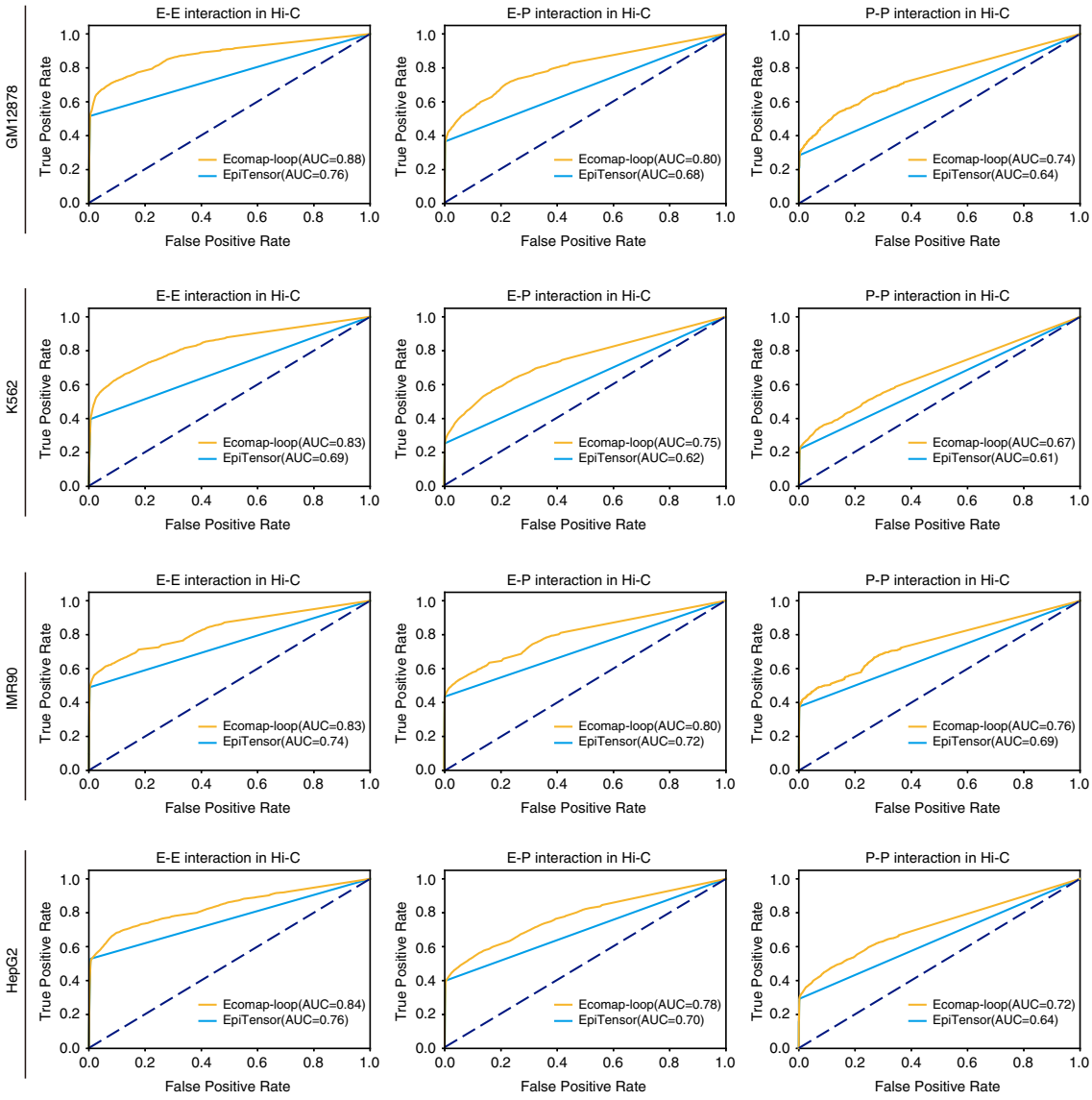
Figure 2. The schema of Ecomap-loop

The model of Ecomap-loop can be divided into three parts: calculating the read coverage of CTCF, RAD21, and SMC on each fragment as V_c, calculating the evaluation score for CTCF motif orientations as V_F, and the eigen decomposition by using the assays including histone marks ChIP-seq and DNase-seq of different cell types as V_Q. The final E-E, E-P, and P-P loops are measured with score *V*. The blue peaks indicate the ChIP-seq peaks across the genome. The green arrows indicate the CTCF motif orientation.

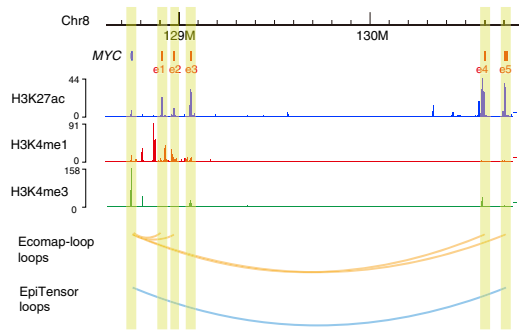
of the final evaluation scores (ES). The predicted loops validated by the Hi-C contacts were defined as true positives; the predicted loops not validated by the Hi-C contacts were defined as false positives; the loops not predicted while validated by the Hi-C contacts were defined as false negatives; and the loops not predicted and not validated by the Hi-C contacts were defined as true negatives. To generate the receiver operating characteristic (ROC) curve, we changed the threshold of AS and ES gradually to calculate a series of sensitivity and specificity values. Then the area under the curve (AUC) for different cell lines were calculated. The ROC curves of EpiTensor are simple fold lines, which may be due to the low number of positive samples predicted by EpiTensor. And the AUC values of Ecomap-loop were higher than those of EpiTensor in three loop types across the cell lines of K562, GM12878, IMR90, and HepG2 in which Ecomap-loop achieved the highest AUC increasing of 20.9% in the E-P loop (K562) dataset (Figure 3A).

To further validate the prediction results, we collected the ChIA-PET experimental datasets of four cell types from ENCODE[31] as positive samples (see STAR Methods). Similar as the validation of Hi-C experimental data, we change the AS and ES gradually to generate the ROC curve and the definition of true positives, false positives, false negatives, and true negatives depending on the consistency between predicted loops and ChIA-PET loops. The comparison results indicated that Ecomap-loop outperformed EpiTensor

A


GM12878
E-E interaction in Hi-C
Ecomap-loop(AUC=0.88)
EpiTensor(AUC=0.76)
E-P interaction in Hi-C
Ecomap-loop(AUC=0.80)
EpiTensor(AUC=0.68)
P-P interaction in Hi-C
Ecomap-loop(AUC=0.74)
EpiTensor(AUC=0.64)
K562
E-E interaction in Hi-C
Ecomap-loop(AUC=0.83)
EpiTensor(AUC=0.69)
E-P interaction in Hi-C
Ecomap-loop(AUC=0.75)
EpiTensor(AUC=0.62)
P-P interaction in Hi-C
Ecomap-loop(AUC=0.67)
EpiTensor(AUC=0.61)
IMR90
E-E interaction in Hi-C
Ecomap-loop(AUC=0.83)
EpiTensor(AUC=0.74)
E-P interaction in Hi-C
Ecomap-loop(AUC=0.80)
EpiTensor(AUC=0.72)
P-P interaction in Hi-C
Ecomap-loop(AUC=0.76)
EpiTensor(AUC=0.69)
HepG2
E-E interaction in Hi-C
Ecomap-loop(AUC=0.84)
EpiTensor(AUC=0.76)
E-P interaction in Hi-C
Ecomap-loop(AUC=0.78)
EpiTensor(AUC=0.70)
P-P interaction in Hi-C
Ecomap-loop(AUC=0.72)
EpiTensor(AUC=0.64)
True Positive Rate
False Positive Rate

B


Chr8
129M
130M
MYC
e1 e2 e3
e4 e5
H3K27ac
H3K4me1
H3K4me3
Ecomap-loop loops
EpiTensor loops

C

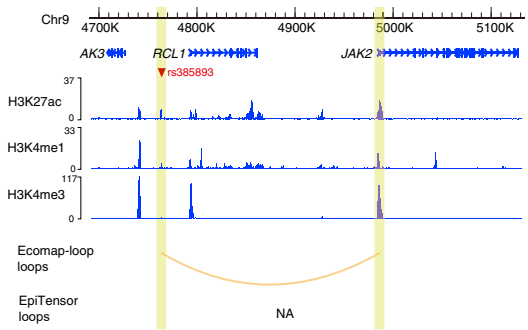

Chr9
4700K
4800K
4900K
5000K
5100K
AK3
RCL1
JAK2
rs385893
H3K27ac
H3K4me1
H3K4me3
Ecomap-loop loops
EpiTensor loops
NA

Figure 3. Evaluation of predicted chromatin loops with Hi-C experimental datasets

(A) The receiver operating characteristic (ROC) curves are generated by changing the threshold of AS and ES gradually; the area under the curve (AUC) for different cell types were calculated.

(B) Example of prediction results near *MYC* gene locus.

(C) Example of prediction results near *JAK2* gene locus.

in all the cell types and loop types, in which Ecomap-loop achieved the highest AUC increasing of 15.0% in the P-P loop (GM12878) dataset (Figure S1).

In the study of Fulco,[32] the interactions between *MYC* and 7 enhancers were identified in K562 cell line. Here we checked the prediction results of Ecomap-loop and EpiTensor near the *MYC* locus with K562 datasets, which showed Ecomap-loop can predict the loops of *MYC*-e1, *MYC*-e2, *MYC*-e4, and *MYC*-e5, while EpiTensor only predicted the loop of *MYC*-e5 (Figure 3B). In the study of Mattews,[33] *JAK2* gene promoter interacts with an enhancer 222 kb away and relates to the myeloproliferative disorder. The enhancer has an H3K27ac peak and harbor the SNP rs385893. Presence of SNP within the H3K27ac peak alters the transcription factor binding property of that region and thus causes reduced interaction with *JAK2* promoter. The prediction results showed Ecomap-loop can predict the loop of *JAK2*-rs385893, while no loop was detected by EpiTensor in the region (Figure 3C).

Overall, the validation comparison between Ecomap-loop and EpiTensor in Hi-C and ChIA-PET experimental datasets across four cell types revealed the high sensitivity of Ecomap-loop to predict E-E, E-P, and P-P chromatin loops.

Validation of predicted chromatin loops with eQTL and CRISPRi datasets

eQTLs are genetic loci that control the expression level of genes for quantitative traits, which paralleled the adoption of genome-wide association studies (GWAS) to analyze the complex traits and disease in humans. It has become common to interpret noncoding variant-gene associations using eQTL data.[34,35] Here we obtained the reliable chromatin associations from eQTL datasets of each cell line to validate the predicted loops (see STAR Methods). Because the number of eQTL loops was relatively less than the number of loops detected by sequence-based techniques (such as Hi-C and ChIA-PET), we used the overlapping percentage between predicted loops and eQTL loops to measure the precision of Ecomap-loop and EpiTensor. Before calculation, the predicted loops from Ecomap-loop and EpiTensor were ranked by AS and ES, respectively, and the top 20% chromatin loops were retained for the calculation. The comparison results revealed that the precision of Ecomap-loop outperformed EpiTensor in four cell lines and three loop types. In addition, it was expected that the overlapping percentage of E-E loops was observed the highest across three loop types for both Ecomap-loop and EpiTensor as eQTL detected the chromatin associations of noncoding variants and most of them linked to enhancers (Figure 4A).

We next validated Ecomap-loop-predicted loops by comparing them to functionally validated enhancer-promoter pairs identified via systematic CRISPRi screen.[36] In the study of Klann et al., several candidate regulatory elements were perturbed and the expression changes of gene *HBE1* were detected in K562 cell line.[36] We collected the ChIP-seq signal tracks of H3K27ac, H3K4me1, H3K4me3, CTCF, RAD21, and SMC3 in K562 and mapped the Virtual 4C profile,[37] CRISPRi screen interaction, and predicted loops from Ecomap-loop to these tracks. The mapping results showed the CRISPRi screen interactions were predicted by Ecomap-loop. Besides, the furthest upstream loops predicted by Ecomap-loop have been validated by 4C profile (Figure 4B).

Functional analysis revealed the cell type-specific prediction ability of Ecomap-loop

To analyze the regulatory functionality of predicted loops, we firstly annotated the loop anchors with active histone marks and filtered the loops with both anchors active (see STAR Methods). We calculated the percentage of active loops for each cell type and loop type, which showed that the prediction of GM12878 cell type had the most active loops (>60%) in three loop types. And E-E loop was observed with the highest percentage of active loops across four cell types (Figure 5A). Then we extracted the ChIP-seq peak signals of H3K27ac, H3K4me1, and H3K4me3 at the locus of anchors; the active loops were observed with higher histone marks binding signal than the common loops in four cell types. The p value was calculated by the Wilcoxon test, indicating that the active loops had higher chromatin activity than the common loops (Figure 5B).

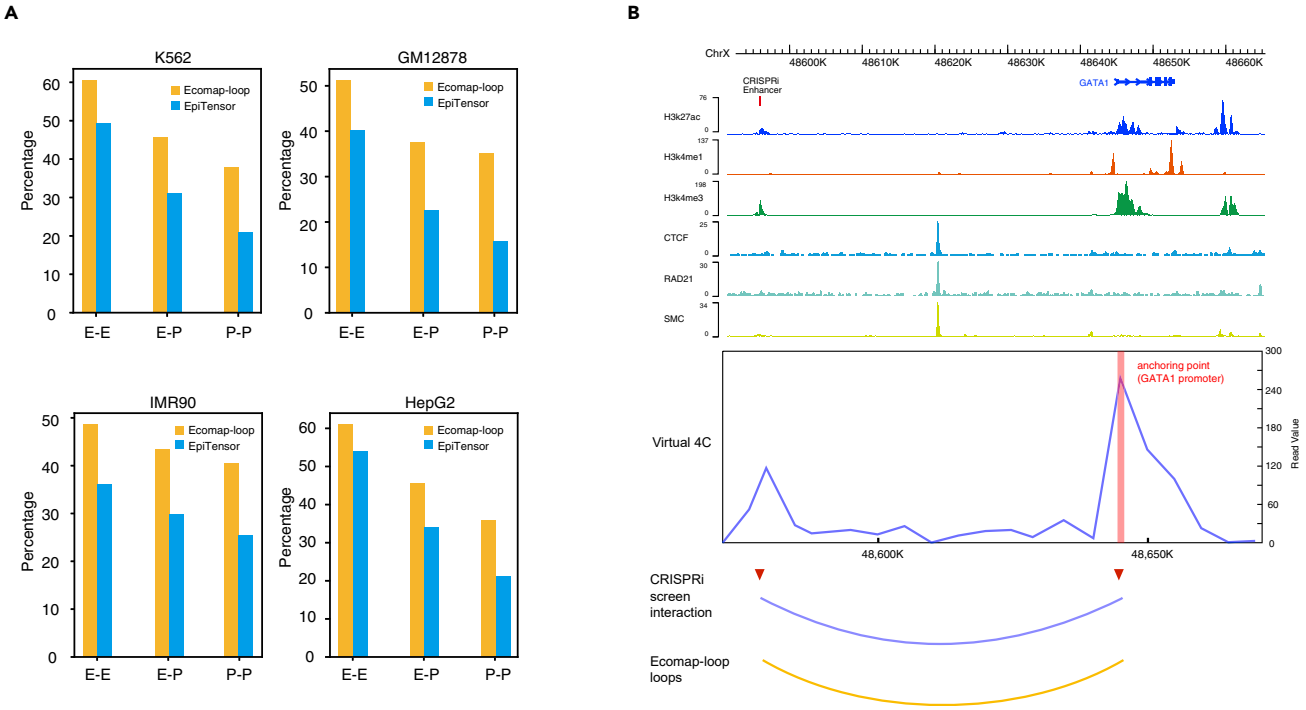


Figure 4. Validation of predicted chromatin loops with eQTL and CRISPRi datasets

(A) Overlapping percentage between eQTL loops and Ecomap-loop/EpiTensor-predicted loops in four cell types and three loop types.

(B) Genomic tracks of virtual 4C profile, CRISPRi screen interactions, and Ecomap-loop predictions near the *GATA1* gene locus. The red triangles indicate the CRISPRi testing positions.

We next extracted the genomic positions of active anchors for each cell type individually, then annotated the positions with the nearest genes, and the active gene sets were used to detect the gene ontology (GO) enrichment by Metascape[38] with the p value cutoff of 0.01, minimum overlap of 3, and minimum enrichment of 1.5.[38] The significant enriched terms related to the cell type identity were selected and showed. For K562, the GO terms enriched in leukemia and immunity. For GM12878, the GO terms enriched in lymphocyte activation and related regulation process. For IMR90, the GO terms enriched in lung cancer and morphogenesis. For HepG2, the GO terms enriched in liver development and disease. These results indicated that the loop anchors contributed to the corresponding cell identity, suggesting the loops predicted by Ecomap-loop were cell type-specific (Figure 5C).

DISCUSSION

With the rapid development of 3C-based techniques and high-throughput sequencing, we have known that human interphase chromosomes are folded into multiple layers of hierarchical structures, including chromatin territory, compartment, topologically associated domain (TAD), and chromatin loop. Among them, the chromatin loop by definition is two genomic loci that are physically closer in the nucleus than their intervening sequences, which play an important role in gene expression and disease-associated studies. Recently, some supervised-learning algorithms have been developed to eliminate the obstacles of wet-lab experiments, while these algorithms require large data input and long running time. Thus, a fast and easy-to-use algorithm with high sensitivity is required in this area. In this study, we develop an unsupervised-based algorithm, Ecomap-loop, to predict the cell type-specific long-range chromatin loops.

The contribution of CTCF and cohesin to the formation of E-P loops is still an open question. Some studies have revealed the extrusion complexes do not contribute to the E-P interactions significantly,[39] while other cases found that CTCF is directly involved in E-P interactions.[10,12,40] Although the opposing statement exists, the consensus view is that CTCF and cohesin are the important mediator of chromatin loops.[9,10] In our study, we did not use the extrusion complexes to distinguish the anchor type of enhancer/promoter. For the classification of promoter/enhancer, we used GENCODE data and the EnhancerAtlas to annotate the loops (see STAR Methods). Through these steps, we get the prediction of E-P loops on the basis of all the chromatin loops.

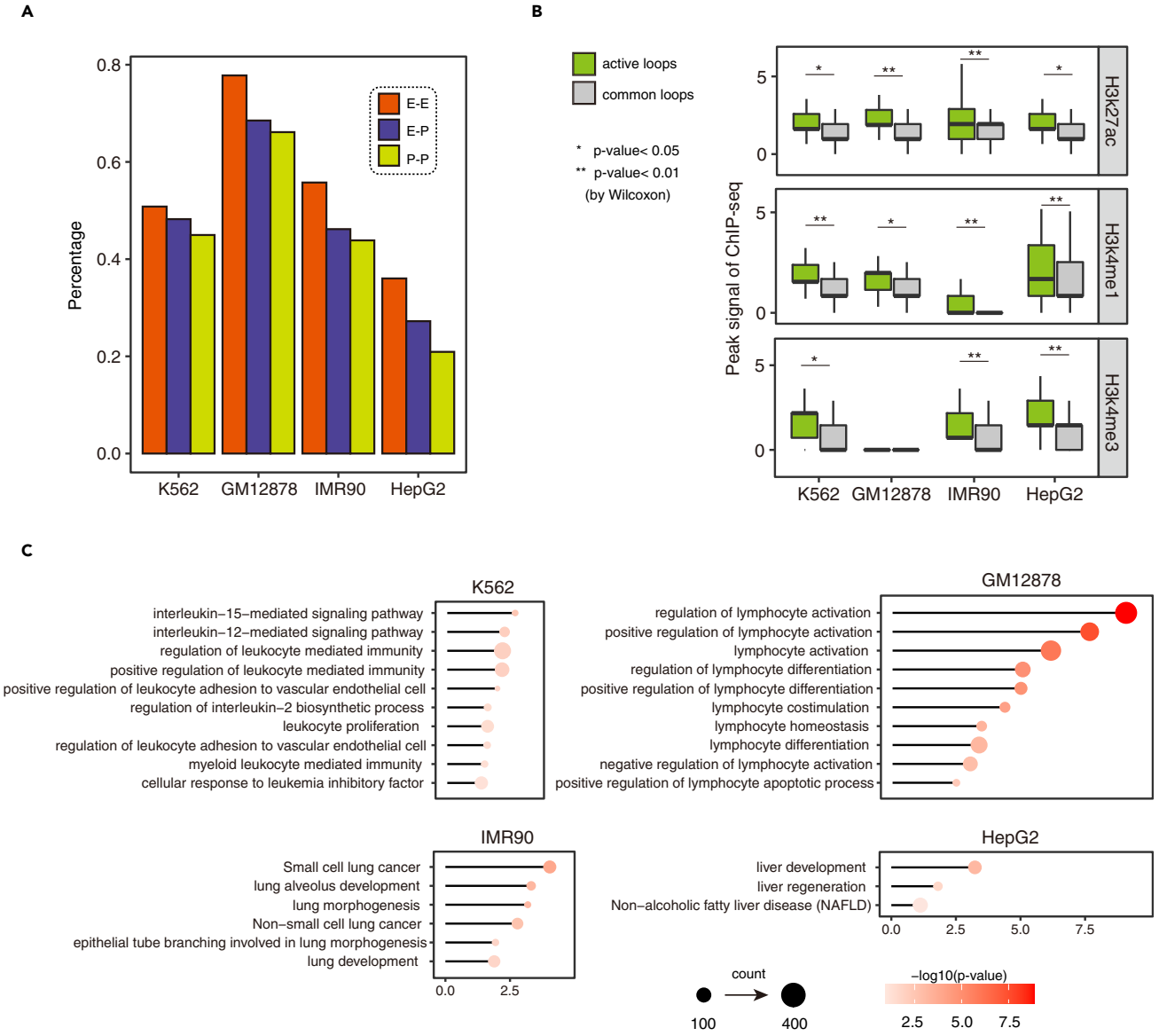


Figure 5. Functional analysis revealed the cell type-specific prediction ability of Ecomap-loop

(A) Percentage of active loops in four cell types and three loop types predicted by Ecomap-loop.

(B) Histone mark ChIP-seq peak signal of active loop anchors in four cell types.

(C) Cell identity GO enrichment of active loop anchors in four cell types.

Recently, some studies have incorporated CTCF and cohesin binding information to predict 3D loops in silico. In the studies of Oti et al.[41] and Matthews et al.,[42] the features of CTCF and cohesin were considered individually, and only the interactions anchored by cohesin and CTCF were predicted. In our study, we employed the extrusion complex-associated pattern, including CTCF orientation and CTCF/cohesin bounding, to construct an unsupervised-learning model to perform the prediction. Ecomap-loop can predict all the possible loops across the genome, including the CTCF/cohesin-mediated ones. Besides, the prediction results of Ecomap-loop were classified into E-E, E-P, and P-P.

The prediction results of Ecomap-loop have been validated by Hi-C, ChIA-PET, eQTL, and CRISPRi data. The benchmarking results show that Ecomap-loop outperforms the state-of-the-art unsupervised algorithm EpiTensor. For further comparison, we evaluated the number of loops predicted, median loop length, and the number of genes covered for the two methods (Table S1). The inputs for Ecomap-loop

and EpiTensor are different. The data matrices used for EpiTensor are generated from histone ChIP-seq data. The input for Ecomap-loop included the CTCF, SMC, and RAD21 ChIP-seq data and the CTCF motif matrix with orientation information. To make the comparison as fair as possible, we processed Hi-C and ChIA-PET datasets with the same steps for Ecomap-loop and EpiTensor to generate the positive samples. Then we defined the true positive, false negative, true negative, and false negative as described by the EpiTensor paper. To generate the ROC curve, we changed the threshold of loop score gradually to calculate the sensitivity and specificity.

To check the ability of predicting inactive loops, we extracted the ChIP-seq signal of H3K27me3 at the locus of anchors; the inactive loops were observed with higher H3K27me3 histone mark binding signal than the common loops (Figure S2). Overall, Ecomap-loop can predict not only active loops but also inactive loops, which facilitates the further mining of gene regulation mechanism under the context of 3D architecture.

Limitations of the study

Compared with other unsupervised learning-based algorithms to predict the chromatin interactions, Ecomap-loop makes use of the orientation distribution of CTCF motif of loops and the enrichments of CTCF, RAD21, and SMC on loops to improve the prediction accuracy. However, there are still some limitations. Firstly, Ecomap-loop requires a variety of input files, including CTCF, RAD21, and SMC ChIP-seq data, CTCF motif matrix, and gene annotation files. For different cell lines, the publicly accessible files may not completely satisfy the requirements. Secondly, Ecomap-loop takes a substantial amount of computational resources and time to process the data for cell lines in the whole genome, which needs to be improved in the future. Thirdly, Ecomap-loop divides the whole genome into different segments, including promoter, enhancer, and other regions, while the regions overlapped with both promoter and enhancer should be treated properly.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Annotation of loop types
  - Enrichment of ChIP-seq peaks in loops
  - Implementation of Ecomap-loop
  - Process of 3C-based experimental datasets
  - Process of eQTL datasets
  - Detection of active loops
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105687.

ACKNOWLEDGMENTS

This work was supported by grants from the National Natural Science Foundation of China under Grants (No. 62225209) [M.L.], the science and technology innovation program of Hunan Province (2021RC4008) [M.L.], and the Fundamental Research Funds for the Central Universities of Central South University (2021zzts0203) [L.T.]. We are grateful to the High-Performance Computing Center of Central South University for partial support of this work.

AUTHOR CONTRIBUTIONS

L.T. and M.L. conceived the presented idea. Y.D. and L.T. collected the data and designed the model. Y.D. wrote the source code. X.Z. helped improve the bioinformatics analysis. W.W. helped organize the code. L.T. and M.L. aided in interpreting the results and provided input on the data presentation. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 11, 2022
Revised: November 10, 2022
Accepted: November 25, 2022
Published: December 22, 2022

REFERENCES

1. Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell *161*, 1012–1025. https://doi.org/10.1016/j.cell.2015.04.004.

2. Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.-L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., et al. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454–1458. https://doi.org/10.1126/science.aad9024.

3. Zheng, H., and Xie, W. (2019). The role of 3D genome organization in development and cell differentiation. Nat. Rev. Mol. Cell Biol. *20*, 535–550. https://doi.org/10.1038/s41580-019-0132-4.

4. Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B.R., Mirny, L.A., and Dekker, J. (2013). Organization of the mitotic chromosome. Science *342*, 948–953. https://doi.org/10.1126/science.1236083.

5. Zhang, H., Lam, J., Zhang, D., Lan, Y., Vermunt, M.W., Keller, C.A., Giardine, B., Hardison, R.C., and Blobel, G.A. (2021). CTCF and transcription influence chromatin structure re-configuration after mitosis. Nat. Commun. *12*, 5157. https://doi.org/10.1038/s41467-021-25418-5.

6. Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311. https://doi.org/10.1126/science.1067799.

7. Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., de Laat, W., and van Steensel, B. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nat. Genet. *38*, 1348–1354. https://doi.org/10.1038/ng1896.

8. Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309. https://doi.org/10.1101/gr.5571506.

9. Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293. https://doi.org/10.1126/science.1181369.

10. 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.

11. Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-α-bound human chromatin interactome. Nature *462*, 58–64. https://doi.org/10.1038/nature08497.

12. Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., et al. (2015). CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell *163*, 1611–1627. https://doi.org/10.1016/j.cell.2015.11.024.

13. 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.

14. Fang, R., Yu, M., Li, G., Chee, S., Liu, T., Schmitt, A.D., and Ren, B. (2016). Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. Cell Res. *26*, 1345–1348. https://doi.org/10.1038/cr.2016.137.

15. Kumar, S., Kaur, S., Seem, K., Kumar, S., and Mohapatra, T. (2021). Understanding 3D genome organization and its effect on transcriptional gene regulation under environmental stress in plant: a chromatin perspective. Front. Cell Dev. Biol. *9*, 774719. https://doi.org/10.3389/fcell.2021.774719.

16. Tang, L., Zhong, Z., Lin, Y., Yang, Y., Wang, J., Martin, J.F., and Li, M. (2022). EPIXplorer: a web server for prediction, analysis and visualization of enhancer-promoter interactions. Nucleic Acids Res. *50*, W290–W297. https://doi.org/10.1093/nar/gkac397.

17. Tao, H., Li, H., Xu, K., Hong, H., Jiang, S., Du, G., Wang, J., Sun, Y., Huang, X., Ding, Y., et al. (2021). Computational methods for the prediction of chromatin interaction and organization using sequence and epigenomic profiles. Brief Bioinform *22*, bbaa405. https://doi.org/10.1093/bib/bbaa405.

18. Whalen, S., Truty, R.M., and Pollard, K.S. (2016). Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. Nat. Genet. *48*, 488–496. https://doi.org/10.1038/ng.3539.

19. Cao, Q., Anyansi, C., Hu, X., Xu, L., Xiong, L., Tang, W., Mok, M.T.S., Cheng, C., Fan, X., Gerstein, M., et al. (2017). Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. Nat. Genet. *49*, 1428–1436. https://doi.org/10.1038/ng.3950.

20. Belokopytova, P.S., Nuriddinov, M.A., Mozheiko, E.A., Fishman, D., and Fishman, V. (2020). Quantitative prediction of enhancer–promoter interactions. Genome Res. *30*, 72–84. https://doi.org/10.1101/gr.249367.119.

21. Tang, L., Hill, M.C., Wang, J., Wang, J., Martin, J.F., and Li, M. (2020). Predicting unrecognized enhancer-mediated genome topology by an ensemble machine learning model. Genome Res. *30*, 1835–1845. https://doi.org/10.1101/gr.264606.120.

22. Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49. https://doi.org/10.1038/nature09906.

23. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82. https://doi.org/10.1038/nature11232.

24. Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal·lari, R., Lupien, M., Markowitz, S., and Scacheri, P.C. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. *24*, 1–13. https://doi.org/10.1101/gr.164079.113.

25. Zhu, Y., Chen, Z., Zhang, K., Wang, M., Medovoy, D., Whitaker, J.W., Ding, B., Li, N., Zheng, L., and Wang, W. (2016). Constructing 3D interaction maps from 1D epigenomes. Nat. Commun. *7*, 10812. https://doi.org/10.1038/ncomms10812.

26. Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion

explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. USA *112*, E6456–E6465. https://doi.org/10.1073/pnas.1518552112.

27. Peters, J.-M., Tedeschi, A., and Schmitz, J. (2008). The cohesin complex and its roles in chromosome biology. Genes Dev. *22*, 3089–3114. https://doi.org/10.1101/gad.1724308.

28. Mumbach, M.R., Satpathy, A.T., Boyle, E.A., Dai, C., Gowen, B.G., Cho, S.W., Nguyen, M.L., Rubin, A.J., Granja, J.M., Kazane, K.R., et al. (2017). Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. Nat. Genet. *49*, 1602–1612. https://doi.org/10.1038/ng.3963.

29. de Wit, E., Vos, E.S.M., Holwerda, S.J.B., Valdes-Quezada, C., Verstegen, M.J., Teunissen, H., Splinter, E., Wijchers, P.J., Krijger, P.H.L., and de Laat, W. (2015). CTCF binding polarity determines chromatin looping. Mol. Cell *60*, 676–684. https://doi.org/10.1016/j.molcel.2015.09.023.

30. Ghirlando, R., and Felsenfeld, G. (2016). CTCF: making the right connections. Genes Dev. *30*, 881–891. https://doi.org/10.1101/gad.277863.116.

31. ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. https://doi.org/10.1038/nature11247.

32. Fulco, C.P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S.R., Perez, E.M., Kane, M., Cleary, B., Lander, E.S., and Engreitz, J.M. (2016). Systematic mapping of functional enhancer–promoter connections with CRISPR interference. Science *354*, 769–773. https://doi.org/10.1126/science.aag2445.

33. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190–1195. https://doi.org/10.1126/science.1222794.

34. Gilad, Y., Rifkin, S.A., and Pritchard, J.K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. Trends Genet. *24*, 408–415. https://doi.org/10.1016/j.tig.2008.06.001.

35. Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S., and Cheung, V.G. (2004). Genetic analysis of genome-wide variation in human gene expression. Nature *430*, 743–747. https://doi.org/10.1038/nature02797.

36. Klann, T.S., Black, J.B., Chellappan, M., Safi, A., Song, L., Hilton, I.B., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2017). CRISPR–Cas9 epigenome editing enables high-throughput screening for functional regulatory elements in the human genome. Nat. Biotechnol. *35*, 561–568. https://doi.org/10.1038/nbt.3853.

37. Wang, Y., Song, F., Zhang, B., Zhang, L., Xu, J., Kuang, D., Li, D., Choudhary, M.N.K., Li, Y., Hu, M., et al. (2018). The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol. *19*, 151. https://doi.org/10.1186/s13059-018-1519-9.

38. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. *10*, 1523. https://doi.org/10.1038/s41467-019-09234-6.

39. Hnisz, D., Day, D.S., and Young, R.A. (2016). Insulated neighborhoods: structural and functional units of mammalian gene control. Cell *167*, 1188–1200. https://doi.org/10.1016/j.cell.2016.10.024.

40. 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.

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

42. Matthews, B.J., and Waxman, D.J. (2018). Computational prediction of CTCF/cohesin-based intra-TAD loops that insulate chromatin contacts and gene expression in mouse liver. Elife *7*, e34077. https://doi.org/10.7554/elife.34077.

43. Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J.E., Mudge, J.M., Sisu, C., Wright, J.C., Armstrong, J., Barnes, I., et al. (2021). GENCODE 2021. Nucleic Acids Res. *49*, D916–D923. https://doi.org/10.1093/nar/gkaa1087.

44. Gao, T., and Qian, J. (2020). EnhancerAtlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species. Nucleic Acids Res. *48*, D58–D64. https://doi.org/10.1093/nar/gkz980.

45. Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330. https://doi.org/10.1038/nature14248.

46. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. https://doi.org/10.1038/35057062.

47. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

48. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165. https://doi.org/10.1093/nar/gkw257.

49. Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics *27*, 1017–1018. https://doi.org/10.1093/bioinformatics/btr064.

50. Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Lemma, R.B., Turchi, L., Blanc-Mathieu, R., Lucas, J., Boddie, P., Khan, A., Manosalva Pérez, N., et al. (2022). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res. *50*, D165–D173. https://doi.org/10.1093/nar/gkab1113.

51. Reiff, S.B., Schroeder, A.J., Kırlı, K., Cosolo, A., Bakker, C., Mercado, L., Lee, S., Veit, A.D., Balashov, A.K., Vitzthum, C., et al. (2022). The 4D Nucleome Data Portal as a resource for searching and visualizing curated nucleomics data. Nat. Commun. *13*, 2365. https://doi.org/10.1038/s41467-022-29697-4.

52. Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., and Aiden, E.L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. *3*, 95–98. https://doi.org/10.1016/j.cels.2016.07.002.

53. Lee, M.G., Villa, R., Trojer, P., Norman, J., Yan, K.-P., Reinberg, D., Di Croce, L., and Shiekhattar, R. (2007). Demethylation of H3K27 regulates polycomb recruitment and H2A ubiquitination. Science *318*, 447–450. https://doi.org/10.1126/science.1149042.

54. Herz, H.-M., Mohan, M., Garruss, A.S., Liang, K., Takahashi, Y.H., Mickey, K., Voets, O., Verrijzer, C.P., and Shilatifard, A. (2012). Enhancer-associated H3K4 monomethylation by Trithorax-related, the Drosophila homolog of mammalian Mll3/Mll4. Genes Dev. *26*, 2604–2620. https://doi.org/10.1101/gad.201327.112.

55. Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. USA *107*, 21931–21936. https://doi.org/10.1073/pnas.1016071107.

56. Allis, C.D., and Jenuwein, T. (2016). The molecular hallmarks of epigenetic control. Nat. Rev. Genet. *17*, 487–500. https://doi.org/10.1038/nrg.2016.59.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
GENCODE	Frankish et al., 2021[43]	https://www.gencodegenes.org/
EnhancerAtlas	Gao and Qian, 2020[44]	http://www.enhanceratlas.org/downloadv2.php
ROADMAP	Roadmap Epigenomics Consortium et al., 2015[45]	http://www.roadmapepigenomics.org/
ENCODE	ENCODE Project Consortium, 2012[31]	https://www.encodeproject.org/(See Table S2)
GEO	Rao et al., 2014[10]	http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525
UCSC	Lander et al., 2001[46]	https://genome.ucsc.edu/
Software and algorithms		
Ecomap-loop	This study	https://github.com/CSUBioGroup/Ecomap-loop
EpiTensor	Zhu et al., 2016[25]	http://wanglab.ucsd.edu/star/EpiTensor/
BEDTools(v2.30)	Quinlan and Hall, 2010[47]	https://bedtools.readthedocs.io/en/latest/
Deeptools	Ramírez et al., 2016[48]	https://deeptools.readthedocs.io
FIMO(v5.4)	Grant et al., 2011[49]	https://meme-suite.org/meme/doc/fimo.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Min Li (limin@mail.csu.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 URLs for the datasets are listed in the key resources table. The accession numbers of publicly Epigenomics datasets used in this study are shown in Table S2.

Source code and tutorials are publicly available online at https://github.com/CSUBioGroup/Ecomap-loop.

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

METHOD DETAILS

Annotation of loop types

The entire genome is divided into three parts: promoter, enhancer and other. Other regions are defined as the remaining portion of the genome not overlapping with annotated promoters and enhancers. We download the GFF3 file of human GRCh37 (v19) in GENCODE (https://www.gencodegenes.org/).[43] The promoter regions are defined as the region 1000 bp upstream and 1000 bp downstream of each gene transcription start site. The enhancer regions are accessible in EnhancerAtlas (http://www.enhanceratlas.org/downloadv2.php)[44] which provides enhancers predicted by combining the results of different analyses of high-throughput data in humans (hg19). All the predicted loop anchors overlap with the promoters or enhancers with at least 1 bp were retained, here BEDTools[47] is used to count the overlapping length. Then we extract the promoter parts and enhancer parts from the predicted loops, and concentrate on the promoter-enhancer pairs, promoter-promoter pairs, and enhancer-enhancer pairs.

Enrichment of ChIP-seq peaks in loops

We firstly download the raw fastq sequencing files of K562-H3K27ac ChIA-PET, then trim the linkers and align the reads to reference hg19, the alignment results are converted to bigwig format with bamCoverage of deepTools.[48] And the ChIP-seq peaks of CTCF, RAD21, and SMC are collected in bed format, then the regions of E-E, E-P and P-P loops were used to extract the corresponding alignments from ChIA-PET bigwig file. We next calculate the reads density with the bin length of 10bp for three loop types at the ChIP-seq peak center with 3 kb upstream and 3 kb downstream. Finally, the matrices of read density are used to generate the enrichment plot.

Implementation of Ecomap-loop

The implementation of Ecomap-loop can be divided into three parts, calculating the read coverage of CTCF, RAD21 and SMC on each fragment, the evaluation for CTCF motif orientations and the eigen decomposition by using the assays including histone marks ChIP-seq and DNase-seq of different cell types.

As CTCF, RAD21 and SMC are reported as the extrusion complex and play important role in loop formation, the ChIP-seq peaks of CTCF, RAD21 and SMC are downloaded from ENCODE (https://www.encodeproject.org/) for each cell line, then coverageBed function of BEDTools[47] is used to obtain the coverage rate and the number of base pairs of CTCF, RAD21 and SMC peaks covering with on each fragment. Considering the final prediction effect and the changes of coverage rate led by the different lengths of different fragments, here we use the number of base pairs as an evaluation indicator to present the coverage of CTCF, RAD21 and SMC on different fragments. To balance the impact of such a high coverage threshold, we have experimentally assigned different coefficients to the coverage and determined its final coefficient to be 0.1. Then we combine all the promoters and enhancers genome-wide as E-E, E-P, and P-P pairs. After matching all the CTCF data, RAD21 data and SMC data to the pairs, we sort them by the genomic locus, through which the fragment with smaller genomic coordinate is placed in front, named fragment-1, and the fragment with larger genomic coordinate is placed behind, named fragment-2. The coverage of CTCF, RAD21 and SMC are defined as $V_c = 0.1 \times (C_1 + C_2 + R_1 + R_2 + S_1 + S_2)$, where C_1 represents the CTCF coverage score of fragment-1, and C_2 represents the CTCF coverage score of fragment-2. Similarly, R_1, R_2, S_1 and S_2 are the RAD21 coverage score of fragment-1, the RAD21 coverage score of fragment-2, the SMC coverage score of fragment-1, and the SMC coverage score of fragment-2, respectively.

There are four orientations of CTCF motif on interactions, including convergent, tandem leftward, tandem rightward and divergent. To evaluate the orientations of CTCF motif on both ends of loops, we employ FIMO,[49] which is a motif scanning tool in the MEME suit to scan each promoter and enhancer fragment. FIMO needs a motif file containing MEME formatted motifs and a sequence file in FASTA format as input, then reports all possible positions in each sequence that match a motif with the corresponding stand, matched sequence, log likelihood ratio score, p value, and q-value. We next process the orientations of CTCF motifs on each fragment, firstly, we download the hg19 genome sequence in UCSC (https://hgdownload.soe.ucsc.edu/) and the meme motif format data of CTCF in JASPER (https://jaspar.genereg.net/)[50] to get the position-dependent letter-probability matrices that describe the probability of each possible letter at each position in the pattern. Then we extract the sequences in.fasta format for each promoter and enhancer by using getfasta function in BEDTools package. Next, we use the FIMO to identify the candidate CTCF binding sites and their corresponding chains. Finally, we filter for maximum value in both forward and reverse strands of each promoter and enhancer from the output files of FIMO and preserve the DNA strand information for each fragment.

We use F_{1+} to represent the CTCF motif score of fragment-1 on the forward strand, F_{2-} represents the CTCF score of fragment-2 on the reverse strand, F_{1-} represents the CTCF score of fragment-2 on the reverse strand, and F_{2+} represents the CTCF score of fragment-2 on the forward strand. Thus, the convergent orientation of CTCF on both ends of interactions can be described as $F_c = \sqrt{(F_{1+} \times F_{2-})}$. Similarly, the tandem rightward orientation and the tandem leftward orientation are described as $F_{tr} = \sqrt{(F_{1+} \times F_{2+})}$ and $F_{tl} = \sqrt{(F_{1-} \times F_{2-})}$, respectively, and the $F_d = \sqrt{(F_{1-} \times F_{2+})}$ stands for the divergent orientation. However, these orientations have different frequencies across all the loops. The convergent orientation has been proved be the majority part (around 64.5–92%) in four orientations, while the divergent orientation is rare because of its structural instability. Therefore, we assign different weights

for these orientations in different experiments and validate the corresponding results in four cell lines to choose the optimal weights for different orientations. The equation to evaluate the orientations of CTCF motifs on both ends of interactions is defined as following,

$$V_F = 0.7 \times F_c + 0.15 \times F_{tr} + 0.15 \times F_{tl}$$

Finally, we use epigenomics data including histone ChIP-seq and DNase-seq data to construct the eigen Q, which can be divided into three feature matrices, as shown below,

$$Q = \mathcal{G} \times_1 A \times_2 B \times_3 C$$

Where A, B, C represent the feature matrix of the cell type, the genomic locus of epigenomic data and the epigenomics data such as DNase-seq data and different histone mark ChIP-seq data by eigen decomposition. $\mathcal{G}$ is the Core third order matrix among three feature matrices. The definitions of the 1-mode product $\mathcal{G} \times_1 A$, the 2-mode product $\mathcal{G} \times_2 B$ and the 3-mode product $\mathcal{G} \times_3 C$ are shown as below,

$$(\mathcal{G} \times_1 A)_{i_1 j_2 j_3} = \sum_{j_1=1}^{J_1} g_{j_1 j_2 j_3} a_{i_1 j_1}$$

$$(\mathcal{G} \times_2 B)_{j_1 i_2 j_3} = \sum_{j_2=1}^{J_2} g_{j_1 j_2 j_3} b_{i_2 j_2}$$

$$(\mathcal{G} \times_3 C)_{j_1 j_2 i_3} = \sum_{j_3=1}^{J_3} g_{j_1 j_2 j_3} c_{i_3 j_3}$$

Thus, we can get another equation as following,

$$q_{i_1 i_2 i_3} = \sum_{j_3=1}^{J_3} \sum_{j_2=1}^{J_2} \sum_{j_1=1}^{J_1} g_{j_1 j_2 j_3} a_{i_1 j_1} b_{i_2 j_2} c_{i_3 j_3}$$

where $q_{i_1 i_2 i_3}$ is the specific value in (i_1, i_2, i_3) of Q, which is the same as $g_{j_1 j_2 j_3}$, $a_{i_1 j_1}$, $b_{i_2 j_2}$, $c_{i_3 j_3}$.

Here we focus on the feature matrix of the genomic locus. Then we capture the peaks with co-variation across different cell types and epigenomic datasets by dimensionality reduction, inferring that there is a physical association between them, and determining the type of association based on the gene region in which the peak loci are located. We use $V_Q = \sqrt{h_1 \times h_2}$ to define the association between two peaks, where h_1 and h_2 are the strength of two peaks, respectively. And the final score of Ecomap-loop is defined as $V = V_c + V_F + V_Q$.

Process of 3C-based experimental datasets

To evaluate the predicted loops, we used the public Hi-C and ChIA-PET datasets of K562, GM12878, IMR90 and HepG2 to generate the positive loops. For Hi-C datasets, the Hi-C matrix are downloaded from 4DN data portal (https://data.4dnucleome.org/)[51] with resolution of 5 kb, then we call the Hi-C interactions by HiCCUPS[52] with default parameters. Since we concentrate on the E-E, E-P, and P-P loops, here we narrow down the Hi-C interactions with the promoters from ENCODE and the enhancers from EnhancerAtlas in a cell type-specific manner. For the ChIA-PET datasets, we obtain the interactions from ENCODE with bedpe format, like Hi-C data, we use promoters and enhancers to narrow down the ChIA-PET interactions.

We divided the whole genome into different segments, including promoter, enhancer, and others. All the enhancer and promoter segments are combined to obtain the possible promoter-promoter, enhancer-enhancer and promoter-enhancer pairs. For each pair, we got a final score *V*, which is the sum of V_C, V_F, V_Q. We arranged the possible pairs according to the score *V*. The pairs with *V* greater than the pre-setting threshold are regarded as loops predicted by Ecomap-loop, which are positive samples, and the others are regarded as loops not predicted by Ecomap-loop. Then, we regarded the pairs whose both-end overlapped with the loops in Hi-C data as loops validated by Hi-C experiments, and the other pairs are defined as

loops not validated by Hi-C experiments. The process of ChIA-PET data for the verification is the same as Hi-C.

Process of eQTL datasets

We curate the published eQTL datasets from eQTL Catalog (https://www.ebi.ac.uk/eqtl/Data_access/), for K562, lymphoblastoid cell line (LCL) eQTL data are used; for GM12878, blood tissue eQTL data are used; for IMR90, lung tissue eQTL data are used; for HepG2, liver tissue eQTL data are used. The variant-gene pairs with the highest PIP within each credible set are extracted as the candidate chromatin loops, and we extend 1000bp length on both ends for each variant, the regions of extended variants are regarded as left anchors, and the paired target genes are regarded as right anchors.

Detection of active loops

Active and inactive promoters/enhancers are connective with the bounding pattern of histone marks. The histone mark H3K27me3 is regarded as the sign of inactive promoters and inactive enhancers,[53,54] while active enhancers have deposition of H3K27ac,[55] and H3K4me3 localizes at the active promoter regions.[56] In this study, active promoters are defined as the promoter regions overlapped with the H3K4me3 peaks and not overlapped with the H3K27me3 peaks, while active enhancers are defined as the enhancer regions enriched with H3K27ac. These active promoter-active enhancer pairs, active promoter-active promoter pairs and active enhancer-active enhancer pairs (validated by Hi-C experimental data, ChIA-PET experimental data etc.) are assumed as active loops. Here we use the intersect function of BEDTools to obtain the active enhancers by extracting the intersections of enhancers and the H3K27ac peaks for corresponding cell types. The active promoters are accessible by extracting the intersections of promoters and subtracting the H3K4me3 peaks to extract the differences of the intersections and H3K27me3 peaks.

QUANTIFICATION AND STATISTICAL ANALYSIS

Data were analyzed using Python. Details of specific statistical analyses are included in the main text. The AUC curves were generated using the matplotlib in python. The bar graph, box graph, and arc graph were generated with the R package ggplot2. For differences between the peak signals of histone marks, we used the Wilcoxon test to calculate the p value. Statistical significance was defined as $p < 0.05$.