Connecting high-resolution 3D chromatin organization with epigenomics

Fan Feng[1], Yuan Yao[2], Xue Qing David Wang[3], Xiaotian Zhang[4], Jie Liu[1,2*]

[1] Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI, USA
[2] Department of Computer Science & Engineering, University of Michigan, Ann Arbor, MI, USA
[3] Center for Epigenetics, Van Andel Institute, Grand Rapids, MI, USA
[4] Department of Pathology, University of Michigan, Ann Arbor, MI, USA
contact: drjieliu@umich.edu

Abstract

The resolution of chromatin conformation capture technologies keeps increasing, and the recent nucleosome resolution chromatin contact maps allow us to explore how fine-scale 3D chromatin organization is related to epigenomic states in human cells. Using publicly available Micro-C datasets, we have developed a deep learning model, CAESAR, to learn a mapping function from epigenomic features to 3D chromatin organization. The model accurately predicts fine-scale structures, such as short-range chromatin loops and stripes, that Hi-C fails to detect. With existing epigenomic datasets from ENCODE and Roadmap Epigenomics Project, we successfully imputed high-resolution 3D chromatin contact maps for 91 human tissues and cell lines. In the imputed high-resolution contact maps, we identified the spatial interactions between genes and their experimentally validated regulatory elements, demonstrating CAESAR's potential in coupling transcriptional regulation with 3D chromatin organization at high resolution.

Introduction

Whereas 3D chromatin organization at the large scale of topologically associating domains (TADs) and compartments has been well characterized in many cell and tissue types by Hi-C technology [1], our understanding of fine-scale 3D chromatin organization at the nucleosome resolution has just begun [2, 3, 4]. With the increasing evidence that fine-scale chromatin organization at the nucleosome resolution is closely related to epigenomic state [5, 6], one intriguing question to ask is whether we can accurately extrapolate such high-resolution chromatin contact maps from epigenomic features such as chromatin accessibility, histone modifications, and transcription factor binding profiles. To explore this, we proposed CAESAR (Chromosomal structure And EpigenomicS AnalyzeR), a deep learning approach to predict nucleosome-resolution 3D chromatin contact maps from existing epigenomic features and lower-resolution Hi-C contact maps.

Our model leverages cutting-edge deep learning approaches to identify representations relevant to high-resolution chromatin organization. In particular, 1D convolutional and graph convolutional layers [7] identify epigenomic patterns over the linear chromatin fiber and over the 3D spatial chromatin organization that is relevant to impute high-resolution chromatin contact maps. With existing high-resolution Micro-C contact maps, Hi-C contact maps, and a number of cell-type matched epigenomic data on human H1-hESC (hESC), mouse ESC (mESC), and human foreskin fibroblasts (HFF), we systematically evaluated the model's performance across different chromosomes, across different cell types, and across different species. In the experiments, the model accurately imputes many fine-scale chromosomal structures that Hi-C sequencing fails to detect, including short-range chromatin loops and stripes. The model is more accurate at imputing evolutionarily conserved regions, active A compartment, and early-replicating regions, which indicates that the fine-scale 3D chromatin organization is strongly influenced by the nature of the epigenomic factors in these regions. The imputed chromatin contacts also recapitulate enhancer activities previously elucidated by CRISPRi experiments [8], and manifest expression quantitative trait loci (eQTLs) previously profiled by GTEx project [9]. CAESAR is also coupled with an attribution method which identifies epigenomic features explanatory to these fine-scale 3D chromatin structures. The explanatory features help to further subtype

fine-scale chromatin structures and elucidate the interplay between histone modifications and nucleosome level chromatin organization.

Our work is the first study connecting 3D genome organization with epigenomics at unprecedented resolution and scale. Our model produces high-resolution human chromatin contact maps for 57 tissue samples, 16 cell lines, 12 primary cells, and 6 *in vitro* differentiated cells. The imputed high-resolution contact maps are shared on a web server (https://nucleome.dcmb.med.umich.edu/), which allows users to easily navigate these fine-scale chromatin structures and the corresponding explanatory epigenomic features.

Results

A deep learning model imputing high-resolution chromatin contact maps

We proposed CAESAR, a supervised deep learning model to impute chromatin contact maps at nucleosome resolution. CAESAR's inputs include a lower-resolution Hi-C contact map and a number of histone modification features (e.g., H3K4me1, H3K4me3, H3K27ac, and H3K27me3), chromatin accessibility (e.g., ATAC-seq), and protein binding profiles (e.g., CTCF) (Supplementary Note 2). CAESAR captures the Hi-C contact map as a graph $\mathcal{G}$ with nodes representing genomic regions of 200 bp long, weighted edges representing chromatin contacts between the regions, and N epigenomic features modeled as N-dimensional node attributes. The architecture of CAESAR (Figures 1a and S1; Supplementary Note 3) includes ordinary 1D convolutional layers which extract local epigenomic patterns along the 1D chromatin fiber, and graph convolutional layers which extract spatial epigenomic patterns over the neighborhood specified by $\mathcal{G}$. The concatenated outputs from the convolutional layers capture all relevant features for one particular 200 bp bin, which are further fed into two parallel output layers — a fully-connected layer predicts the contact profile for each 200 bp bin, and an inner product layer predicts loops between bins. The outputs from the fully-connected layer and the inner product layer are summed up as CAESAR's final output. Using Micro-C contact maps from hESC, mESC, and HFF as the prediction target, the model was trained with backpropagation [10], in which the aforementioned convolutional features were learned adaptively. Other than leveraging a number of epigenomic features, our model architecture differs from HiCPlus [11] and DeepHiC [12] which treats Hi-C contact maps as images and performs grid-convolution to improve the resolution. With the graph convolutional networks and additional epigenomic features, CAESAR not only enhances the resolution of contact maps, but also predicts the structures which are not captured by Hi-C, including polycomb repressive regions, short-range loops and stripes (Figure 1b).

Accurately predicting high resolution chromatin contact maps

With existing Micro-C data on mESC, hESC, and HFF, we evaluated CAESAR in three different sets of experiments, including a *cross-chromosome* experiment, a *cross-cell type* experiment, and a *cross-species* experiment, so as to evaluate the model's generalizability in different scenarios. In the cross-validation experiment on hESC, we divided the human chromosomes into a train set, a test set, and a tune set of similar sizes (Supplementary Notes 4 and 5). CAESAR and two baseline models, including HiCPlus [11] which only used low-resolution chromatin contact maps, and HiC-Reg [13] which only used epigenomic features, were trained with the train set and evaluated with the test set (Supplementary Note 6). We used the tune set to tune hyperparameters. For CAESAR and HiC-Reg, 6 epigenomic features were used, including ATAC-seq, CTCF, H3K4me1, H3K4me3, H3K27ac, and H3K27me3. CAESAR outperformed HiCPlus and HiC-Reg in terms of the stratum-adjusted correlation coefficient (SCC) with the observed Micro-C contact map (Figure 2a). The results demonstrated that it is necessary to leverage both the contact maps and epigenomic features in the prediction of high-resolution contact maps. In the cross-cell type experiment, we used the same train set of chromosomes to build a model on HFF, and then tested it on hESC with the same test set of chromosomes as in the cross-chromosome experiments. The HFF-trained model imputed almost as well as the hESC-trained model for chromatin contacts within 100 kb and 200 kb range (Figure 2b). In the

cross-species experiment, we trained the model on mESC and tested the performance on hESC. In order to stay consistent with cross-chromosome and cross-cell-type evaluation, we also divided mouse chromosomes into train, tune, and test sets of similar sizes. We trained the model with mESC's train set and then tested its performance on the same aforementioned test set of hESC. It was observed that the model trained on mESC also moderately generalized to hESC, and the generalization deteriorates as the contact distance increases.

In addition, we tested CAESAR's performance in predicting fine-scale structures including loops and stripes. In the test set of HFF, CAESAR captured 50% of the loops and 61% of the stripes from Micro-C contact maps at 1 kb resolution, whereas only less than 1% were captured from the input Hi-C contact maps (Figures 2c and 2e; Supplementary Notes 7 and 8). Since loops called from two Hi-C replicates only agree $\sim$60% [14], we believe that our imputed contact map recovers a good portion of these fine-scale structures. By piling up all the loop and stripe regions called from the Micro-C contact maps, we observed comparable enrichment from our predicted high-resolution contact maps and the observed Micro-C contact maps, but the pile-up results from the input Hi-C contact maps showed little enrichment (Figures 2d and 2f).

Factors influencing CAESAR's performance

In order to optimize CAESAR's efficiency, we next explored the factors influencing its performance. As CAESAR's principle inputs are epigenomic and Hi-C data, we began by evaluating the minimum required number of datasets to achieve good imputed results. Four sets of epigenomic features were chosen based on common availability (Figure 3a), and we observed comparable performance among the 13-epi, 7-epi, 6-epi, and 3-epi models (Figure 3b). Although the SCC of the 3-epi model (including ATAC-seq, CTCF, and H3K27ac) did not drop significantly, it over-predicted fine-scale structures (Supplementary Note 8). Therefore, we recommend using the commonly profiled 6 epigenomic features in CAESAR. We also asked what is the requirement for input Hi-C contact maps. Using Hi-C data from Rao *et. al.* [1] and Krietenstein *et. al.* [3], we tested four contact maps, including the original Hi-C contact maps with around 1 billion contacts, two down-sampled Hi-C contact maps with 100 million and 25 million contacts, and a surrogate Hi-C contact map with 1 billion contacts aggregated from four unmatched cell lines. The surrogate contact map acts as a replacement when no chromatin contact map is available for a particular cell type. Although the SCC curve does not drop significantly with the down-sampled contact maps, surrogate Hi-C performs better (Figure 3c). There, if the matched Hi-C contact map is unavailable to complement the epigenomic data in a particular analysis, a surrogate contact map can be used in CAESAR.

We further investigated the relationship between CAESAR's performance, measured with Spearman's correlation between the imputed and the observed Micro-C contact maps, and evolutionary conservation, measured with phastCons scores. It was observed that the model imputed more accurately in the regions with higher evolutionary conservation (Figure 3d). In addition, we also discovered that the model imputes more accurately in A compartment than B compartment, and in early-replicating regions than late-replicating regions (Figures 3e and 3f). The results indicate that fine-scale chromatin organization is more closely related to the 6 epigenomic factors at evolutionarily conserved regions, A compartment, and early-replicating regions.

Recapitulating CRISPRi-validated enhancer activities

With publicly available epigenomic data, we imputed high-resolution chromatin contact maps for 15 human cancer cell lines (Supplementary Table 3b). In some cancer cell lines, noncoding regions with their regulating genes have been interrogated by CRISPR interference (CRISPRi) technology [8]. The profiled CRISPRi score indicates genomic loci's capability to regulate an essential gene, and the peaks (both positive and negative) often correspond to enhancers and promoters.

We used the CRISPRi scores profiled near two essential genes - *MYC* and *GATA1*, to validate our imputed contact maps. On the imputed contact maps for the chronic myelogenous leukemia cell K562, *MYC*

gene strongly interacts with *PVT1*, which matches with the peaks of CRISPRi scores at *PVT1* locus (Figure 4a). The imputed contact map also showed a significant interaction between *GATA1* and *HDAC6*, which matches the CRISPRi score peak at *HDAC6* locus (Figure 4b). The matching of chromatin contacts and CRISPRi score peaks demonstrates our model recapitulates gene-enhancer interactions in cancer cell lines.

Recovering eQTL-gene interactions

With the large-scale epigenomic data available from ENCODE and Roadmap Epigenomics Project, we imputed the high-resolution contact maps for 57 human tissue samples and 2 cell lines – IMR-90 and GM12878 (Supplementary Tables 3a and 3b). With eQTLs profiled by GTEx [9], we asked whether our imputed chromatin contacts are enriched between genes and their eQTLs in the corresponding tissue or cell line. Previous works [15] have shown eQTLs are enriched in tissue-specific frequently interacting regions on Hi-C contact maps at 40 kb resolution, but a large portion of eQTLs reside too close to their gene transcriptional start sites (TSS) to be seen on a low-resolution contact map (Figure S3a). For example, three eQTLs that are specific in heart left ventricle (HLV) are associated with the *NIFK* gene, with distances to the TSS at 5 kb, 7 kb, and 16 kb, respectively. The interactions between the three eQTLs and their TSS cannot be observed on the low-resolution Hi-C contact maps, whereas they appear on the CAESAR-imputed contact maps (Figure 5a). Among the three loops between the TSS and eQTLs, the one anchored at eQTL i appears exclusively on the imputed contact map of HLV, whereas the ones anchored at eQTLs ii and iii are also found on the imputed contact map of lung and the HFF Micro-C contact map respectively (Figure 5a). In another example region where six eQTLs of the *TTC7A* gene shared between pancreas and stomach reside 15 - 31 kb downstream the TSS, both loops and stripes are observed on the imputed contact maps of the two tissues, but not on the imputed contact map of lung tissues or the low-resolution Hi-C contact map (Figure 5b).

To evaluate the overall contact enrichment between eQTLs-TSS pairs, we piled up the regions between tissue-specific eQTLs and their TSS. The enrichment of eQTL-TSS contacts, which does not appear on low-resolution Hi-C contact maps, is the most significant on the imputed contact maps of the corresponding tissue or cell line. The moderate enrichment on the Micro-C contact map from an unmatched cell line HFF demonstrates the eQTL-TSS interactions are not necessarily exclusive even if the eQTL is tissue or cell line-specific (Figure 5c). This suggests that some fine structural interactions are conserved across tissues or cell types but the regulatory functions remain specific.

Identifying epigenomic features relevant to fine-scale 3D chromatin organization

Although deep learning models are often referred to as "black boxes", their outputs can be traced back and interpreted. In our model, we used *integrated gradient* [16] to attribute the predicted chromatin contacts to each genomic locus of each input epigenomic feature. The attribution results illustrate which parts of the epigenomic features are the most determinative for the model's predictions. By attributing the entire contact map to all epigenomic features, we evaluated the overall contribution for each feature, and low attribution is another reason for leaving H3K4me2 out from the 7-epi model besides limited availability (Figure S4a).

This method can be applied to arbitrary regions on the contact map, which allows us to connect fine-scale structures with the most explanatory epigenomic features. Surprisingly, many of the peaks in the input epigenomic features do not necessarily help the model to predict fine-scale structures. For example, the H3K27ac peaks showed negative attribution in predicting the stripe in Figure 6a and the loop in Figure 6b. With attribution calculated by *integrated gradient*, the predicted chromatin structures can be further analyzed and subtyped (Supplementary Note 9).

Discussion

Our study is the first effort to connect nucleosome-resolution chromatin structures with epigenomic features. Leveraging the currently available Micro-C contact maps for hESC, mESC, and HFF from the 4DN consortium and the corresponding epigenomic profiles from ENCODE and Roadmap Epigenomics Project, we systematically mapped 1D epigenomic profiles to fine-scale 3D chromatin structures with CAESAR. The mapping was validated by high SCCs with observed Micro-C contact maps and the accurate capture of fine-scale loops and stripes. CAESAR can be applied to generate high-resolution contact maps for any cell line or tissue as long as their common epigenomic features are profiled. Our model further connects transcriptome with fine-scale structures and epigenomics by identifying the spatial interactions between genes and regulatory elements. Therefore, the imputed high-resolution contact maps will be useful for target finding, hypotheses generating, and other downstream analyses. All imputed human chromatin contact maps across 57 tissues, 16 cell lines, 12 primary cells, and 6 *in vitro* differentiated cells have been made publicly available on our web server (http://nucleome.dcmb.med.umich.edu/) for ease of access by biomedical researchers to perform further analyses (Supplementary Table 1; Supplementary Note 10).

While CAESAR presents a novel way to investigate fine details of 3D chromatin structure, we note that it is an evolving methodology with certain shortcomings that can be improved. First, since Micro-C data mostly outperforms Hi-C in the detection of short-range interactions, CAESAR also performs best at genomic distances of less than 200 kb. As a result of this, CAESAR-imputed contact maps are not well suited for analyses of large 3D chromatin structures such as TADs or compartments. Second, because Micro-C and Hi-C generate short-read sequences, our study is still limited to pairwise chromatin contacts, and therefore higher-order interactions are insufficiently studied. Third, our analyses showed that CAESAR performed well according to multiple evaluation metrics, yet there was clear bias towards A compartment, evolutionarily conserved regions, and early-replicating regions. This is likely a reflection that the epigenomic features in the study are generally more enriched in these regions. As such, it is possible that including additional epigenomic features may shift this bias effect accordingly. Fourth, though CAESAR demonstrated clear relationships between epigenomic features and 3D fine-scale chromatin organization, we did not observe significant improvement in imputed contact maps with increasing number of epigenomic datasets. This suggests that epigenomic data may not explain all the features observed in 3D chromatin organization. There may be unexplored layers of genetic and/or epigenetic information that play a role in the organization of chromatin inside the nucleus. So far, CAESAR demonstrated a framework for jointly analyzing 3D chromatin structures and 1D epigenomic features at a matched resolution, and further integration of 1D DNA sequences is possible. For example, our model can potentially include DNA sequences as features and elucidate 3D QTLs [17] in the context of high-resolution chromatin organization.

Online Methods

Model training

CAESAR takes both epigenomic features and Hi-C contact maps as inputs. Based on the availability of epigenomic features, we trained four models with different epigenomic features — one model with 13 epi-features including ATAC-seq, CTCF, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3, H3K79me2, Nanog, and Rad21; one model with 7 epi-features including ATAC-seq, CTCF, H3K4me1, H3K4me2, H3K4me3, H3K27ac, and H3K27me3; one model with 6 epi-features including ATAC-seq, CTCF, H3K4me1, H3K4me3, H3K27ac, and H3K27me3; and one model with 3 epi-features including ATAC-seq, CTCF, and H3K27me3. Due to high computational burden, it is impossible to feed the entire contact map into the memory, and therefore we used a 250 kb sliding window with 50 kb step length along the diagonal (e.g., 0-250,000; 50,000-300,000; 100,000-350,000; ...) to select the regions and fed them one by one into the model.

We split all chromosomes into train, tune, and test sets of similar sizes (Supplementary Note 4). We

used the train set to train the parameters and the tune set to choose hyperparameters (Supplementary Note 5). During training, the parameters were optimized by minimizing the mean squared error (MSE) with Adam algorithm [18]. Because the model has two parts, one for predicting contact profiles and one for predicting loops (Figure S1 and Supplementary Note 3), we employed a sequential training strategy as follows. First, the loop predicting part was trained, in which the model was optimized targeting only the observed Micro-C contacts in loop regions (i.e, 10 kb$\times$10 kb squares centered at Micro-C loops) instead of the entire contact map. Second, we trained the contact profile part with the residual contact map (i.e., the observed Micro-C contact map minus the outputs of the loop predicting part). The outputs from the two parts were summed up to generate the predicted contact maps.

Evaluation experiments

Three sets of cross-validation experiments were performed. First, the *cross-chromosome* model was trained with the train set of hESC, and tested on the test set of hESC. Second, the *cross-cell type* model was trained with the train set of HFF, and tested on the test set of hESC. Third, the *cross-species* model was trained with the train set of mESC, and tested on the test set of hESC.

To compare CAESAR with baselines and evaluate how much they improve original Hi-C contact maps, we calculated the stratum-adjusted correlation coefficient (SCC) [19] between the observed Micro-C contact map and 1) the CAESAR-imputed contact map, 2) the contact maps imputed by other baseline methods (Supplementary Note 6), and 3) the interpolated Hi-C contact map. Other than evaluating SCC, we also called and compared the loops and stripes from the CAESAR-imputed contact maps, the Micro-C contact maps, and the Hi-C contact maps. We implemented a fast loop calling approach and a stripe calling approach to call loops and stripes at 1 kb resolution (Supplementary Notes 7 and 8; Figure S2). We compared the loops and stripes called from 1) the CAESAR-imputed contact map, 2) the observed Micro-C contact map, and 3) the interpolated Hi-C contact map to generate a Venn diagram. We piled up all stripe and loop regions called from Micro-C contact maps in 1) the CAESAR-imputed contact map, 2) the observed Micro-C contact map, and 3) the interpolated Hi-C contact map.

Correlating model performance with evolutionary conservation, A/B compartments, and replication timing

We tested whether the model performance is correlated with evolutionary conservation, A/B compartments, and replication timing. The genome was split into 250 kb mutually exclusive fragments. For each fragment, we imputed the OE-normalized contact map at 200 bp resolution and smoothed it with a 5$\times$5 uniform kernel. We calculated the Spearman's correlation coefficient between the imputed and the observed Micro-C contact maps to evaluate the model's performance at this fragment.

The 100-way hg38 phastCons scores [20] were used to quantify evolutionary conservation. We processed the hg38 phastCons scores into 250 kb resolution and performed a correlation test between the model performance (i.e., the Spearman's correlation coefficients) and the phastCons scores. Then, the fragments were clustered into three groups, top 10%, top 10-50%, and the others, according to their phastCons score ranking. A box plot of spearman's correlation coefficients was plotted for each group.

The A/B compartments were called at 250 kb resolution. By checking the sign of the first eigenvector of the normalized contact map [21], we separated all 250 kb bins into two groups. The one with more enriched H3K27ac was labeled as A compartment, while the other B compartment. The two-sided student's *t*-test was applied to identify whether the two groups have significantly different Spearman's coefficient.

Similarly, early-late replication timing is defined by the sign of the two-stage repli-seq signal[22]. We processed the repli-seq signal at 250 kb resolution and separated the fragments into two groups, early-replicating regions and late-replicating regions. The two-sided student's *t*-test was applied to identify whether the two groups have significantly different Spearman's coefficient.

Attribution by integrated gradient

We used integrated gradient to identify each input dimension's contribution to the output. Let $\mathbf{X}$ denote the input epigenomic signals

$$\mathbf{X} = \begin{bmatrix} X_1^{(s_1)} & \dots & X_n^{(s_1)} \\ \dots & \dots & \dots \\ X_1^{(s_m)} & \dots & X_n^{(s_m)} \end{bmatrix} \in \mathcal{R}^{m \times n},$$

in which $s_1, \dots, s_m$ are m epigenomic signals (e.g., ATAC-seq, CTCF, etc.) and $1, 2, \dots, n$ are the indices of 200 bp bins. CAESAR takes $\mathbf{X}$ as input and learns a mapping function F: $\mathcal{R}^{m \times n} \rightarrow \mathcal{R}^{n \times n}$ to predict $n \times n$ chromatin contacts between n bins (denoted as $\mathbf{Y}$). Integrated gradient [16] attributes the output to each input dimension of $\mathbf{X}$ by calculating a path integral of the gradient $\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$. Gradient $\frac{\partial \mathbf{Y}}{\partial \mathbf{X}}$ is a measure to quantify how much each dimension of $\mathbf{X}$ influences $\mathbf{Y}$, which reveals the contribution from each input dimension. The path integral starts from a pre-defined "background" $\mathbf{X_0}$ and ends at $\mathbf{X}$, and thus it accumulates the contributions of each input dimension from the background to real input $\mathbf{X}$ [23]. Here we used a matrix of all zeros as the epigenomic background. As demonstrated in [16], a straight-line path is efficient at disentangling the input features. Formally, the attribution of the t-th epigenomic signal s_t at bin i is:

$$A(X_i^{(s_t)}) = \int_{\alpha=0}^{1} \frac{\partial y}{\partial \gamma_i^{(s_t)}(\alpha)} \frac{\partial \gamma_i^{(s_t)}(\alpha)}{\partial \alpha} d\alpha$$

in which y can be $\mathbf{Y}$ or a part of $\mathbf{Y}$, $\frac{\partial y}{\partial \gamma(\alpha)}$ is the gradient, γ is the path, and $\gamma_i^{(s_t)}(\alpha)$ is the dimension corresponding to $X_i^{(s_t)}$ in the path.

By calculating the attribution towards the entire output, we obtained an overall attribution from each epigenomic feature, in which the scale of its absolute value indicates the magnitude of its importance (Figure S4a). Alternatively, the attribution can be calculated for an arbitrary region on the contact map, e.g., a chromatin loop or a chromatin stripe, and used for further subtyping of these loops and stripes (Supplementary Note 9; Figure S4c).

High-resolution contact map imputation for 91 human tissues and cell lines

As the cross-cell type model is validated, we used the trained model to impute high-resolution chromatin contact maps for other human tissues and cell lines. We collected the epigenomic signals from a total number of 57 tissue samples, 16 cell lines, 12 primary cells, and 6 *in vitro* differentiated cells (Supplementary Note 2). If the ATAC-seq signal was unavailable, DNase-seq was collected as an alternative. The 6-epi CAESAR model trained with both hESC and HFF's train set was used. For IMR-90, GM12878, and K562, we used their deeply sequenced (above 1B contacts) Hi-C contact maps as input. For cell lines or tissues without Hi-C or with only shallowly sequenced Hi-C, we used the surrogate Hi-C as input (Supplementary Note 2).

Validation of imputed contact maps with CRISPRi in cancer cell lines

The profiled CRISPRi score indicates the strength a genomic locus regulates a gene, and the peaks (both positive and negative) correspond to enhancers and promoters. We binned the CRISPRi scores at 200-bp resolution. On the imputed high-resolution contact maps, we selected the region near *MYC* gene (chr8: 12,765,000-12,785,000) and *GATA1* gene (chrX: 48,725,000-48,825,000) for K562. The contacts in these regions were jointly analyzed with CRISPRi scores.

Validation of imputed contact maps with eQTLs in human tissues

To process the raw eQTL data, we identified the 200 bp bin where each variant and its corresponding TSS locates and the contacts between the variant bin and TSS bin. We only kept the eQTL-TSS "bin pairs" which are 1) less than 150 kb apart, and 2) specific in only one tissue or cell line. The piled-up analysis was applied to the eQTL-TSS interactions in 1) the CAESAR-imputed contact map, 2) the Micro-C contact map of hESC and HFF, and 3) the interpolated Hi-C contact map. For each eQTL-TSS pair, a square region (51 pixels$\times$51 pixels) centered at their contact was collected. The regions from each contact map were piled up, averaged and further visualized.

Code availability

The source code is publicly available in the GitHub repository https://github.com/liu-bioinfo-lab/caesar.

References

[1] S. S. P. Rao, M. H. Huntley, N. Durand, C. Neva, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 59(7):1665–1680, 2014.

[2] Tsung-Han S Hsieh, Claudia Cattoglio, Elena Slobodyanyuk, Anders S Hansen, Oliver J Rando, Robert Tjian, and Xavier Darzacq. Resolving the 3D landscape of transcription-linked mammalian chromatin folding. *Molecular Cell*, 2020.

[3] Nils Krietenstein, Sameer Abraham, Sergey V Venev, Nezar Abdennur, Johan Gibcus, Tsung-Han S Hsieh, Krishna Mohan Parsi, Liyan Yang, René Maehr, Leonid A Mirny, et al. Ultrastructural details of mammalian chromosome architecture. *Molecular Cell*, 2020.

[4] Masae Ohno, Tadashi Ando, David G Priest, Vipin Kumar, Yamato Yoshida, and Yuichi Taniguchi. Sub-nucleosomal genome structure reveals distinct nucleosome folding motifs. *Cell*, 176(3):520–534, 2019.

[5] Andrew B Stergachis, Brian M Debo, Eric Haugen, L Stirling Churchman, and John A Stamatoyannopoulos. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science*, 368(6498):1449–1454, 2020.

[6] Sandy L Klemm, Zohar Shipony, and William J Greenleaf. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, 20(4):207–220, 2019.

[7] T.N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. arXiv:1609.02907, 2016.

[8] Charles P Fulco, Mathias Munschauer, Rockwell Anyoha, Glen Munson, Sharon R Grossman, Elizabeth M Perez, Michael Kane, Brian Cleary, Eric S Lander, and Jesse M Engreitz. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. *Science*, 354(6313):769–773, 2016.

[9] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, Fernando Garcia, Nancy Young, et al. The genotype-tissue expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 2013.

[10] Paul J Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[11] Yan Zhang, Lin An, Jie Xu, Bo Zhang, W Jim Zheng, Ming Hu, Jijun Tang, and Feng Yue. Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus. *Nature Communications*, 9(1):750, 2018.

[12] Hao Hong, Shuai Jiang, Hao Li, Guifang Du, Yu Sun, Huan Tao, Cheng Quan, Chenghui Zhao, Ruijiang Li, Wanying Li, et al. DeepHiC: A generative adversarial network for enhancing Hi-C data resolution. *PLoS Computational Biology*, 16(2):e1007287, 2020.

[13] Shilu Zhang, Deborah Chasman, Sara Knaack, and Sushmita Roy. In silico prediction of high-resolution Hi-C interaction matrices. *Nature Communications*, 10(1):1–18, 2019.

[14] Abbas Roayaei Ardakany, Halil Tuvan Gezer, Stefano Lonardi, and Ferhat Ay. Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. *Genome Biology*, 21, 2020.

[15] Jingting Yu, Ming Hu, and Chun Li. Joint analyses of multi-tissue Hi-C and eQTL data demonstrate close spatial proximity between eQTLs and their target genes. *BMC Genetics*, 20(1):43, 2019.

[16] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017.

[17] David U Gorkin, Yunjiang Qiu, Ming Hu, Kipper Fletez-Brant, Tristin Liu, Anthony D Schmitt, Amina Noor, Joshua Chiou, Kyle J Gaulton, Jonathan Sebat, et al. Common DNA sequence variation influences 3-dimensional conformation of the human genome. *Genome Biology*, 20(1):1–25, 2019.

[18] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

[19] T. Yang, F. Zhang, G. G. Yardımcı, F. Song, R. C. Hardison, W. S. Noble, F. Yue, and Q. Li. HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Research*, 27(11):1939–1949, 2017.

[20] A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005.

[21] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

[22] Claire Marchal, Takayo Sasaki, Daniel Vera, Korey Wilson, Jiao Sima, Juan Carlos Rivera-Mulia, Claudia Trevilla-García, Coralin Nogues, Ebtesam Nafie, and David M Gilbert. Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq. *Nature Protocols*, 13(5):819, 2018.

[23] Sergio Albeverio, Rafael Høegh-Krohn, and Sonia Mazzucchi. *Mathematical theory of Feynman path integrals: an introduction*, volume 523. Springer Science & Business Media, 2008.

Acknowledgements

The research was supported by NIH R35 HG011279. The authors deeply appreciate the feedback from the 4DN Joint Analysis Working Group, Drs. George Zhang and Russell Ryan from the University of Michigan.

Author Contributions

F.F. and J.L. conceived the idea. F.F. and J.L. designed the model and algorithms. F.F. implemented the model and performed the experiments. F.F. and Y.Y. collected the relevant datasets. Y.Y. implemented the web server. F.F. and J.L. wrote the manuscript. X.Q.D.W. and X.Z. provided feedback regarding experiments and the manuscript.

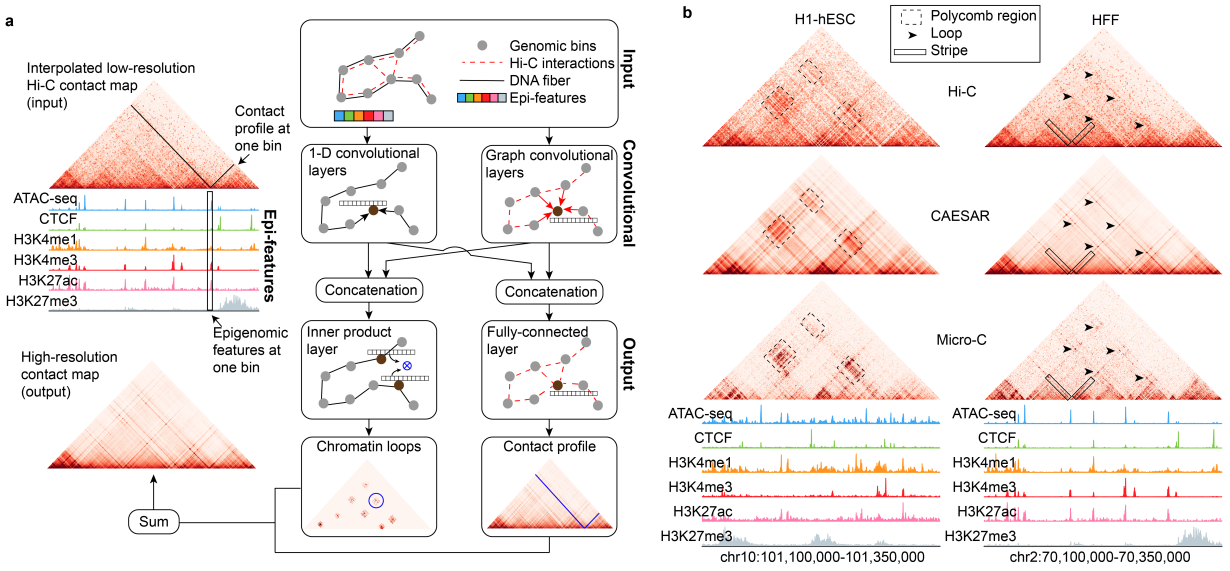

Figure 1: Overview of the model.
a, Model architecture. The model inputs are a Hi-C contact map and a number of epigenomic features including histone modifications, chromatin accessibility, and protein binding profiles. The lower-resolution Hi-C contact map is first interpolated into a 200 bp-resolution contact map, and then transformed into a graph $\mathcal{G}$ in which the nodes represent 200 bp genomic bins and the edges represent the interpolated contacts between the nodes. The epigenomic features are assigned to the corresponding nodes as node attributes. The inputs are fed into 1D convolutional and graph convolutional layers to generate hidden representations, which extract features from both nearby genomic regions along the 1D DNA sequence and spatially-contacting regions specified by $\mathcal{G}$. The output layers take input the hidden representations and predict the contact profile at each 200 bp bin as well as the chromatin contacts between bins. **b**, In an example region, the polycomb interactions are accurately predicted by CAESAR. In another example region, loops and stripes undetected by Hi-C are accurately predicted by CAESAR.

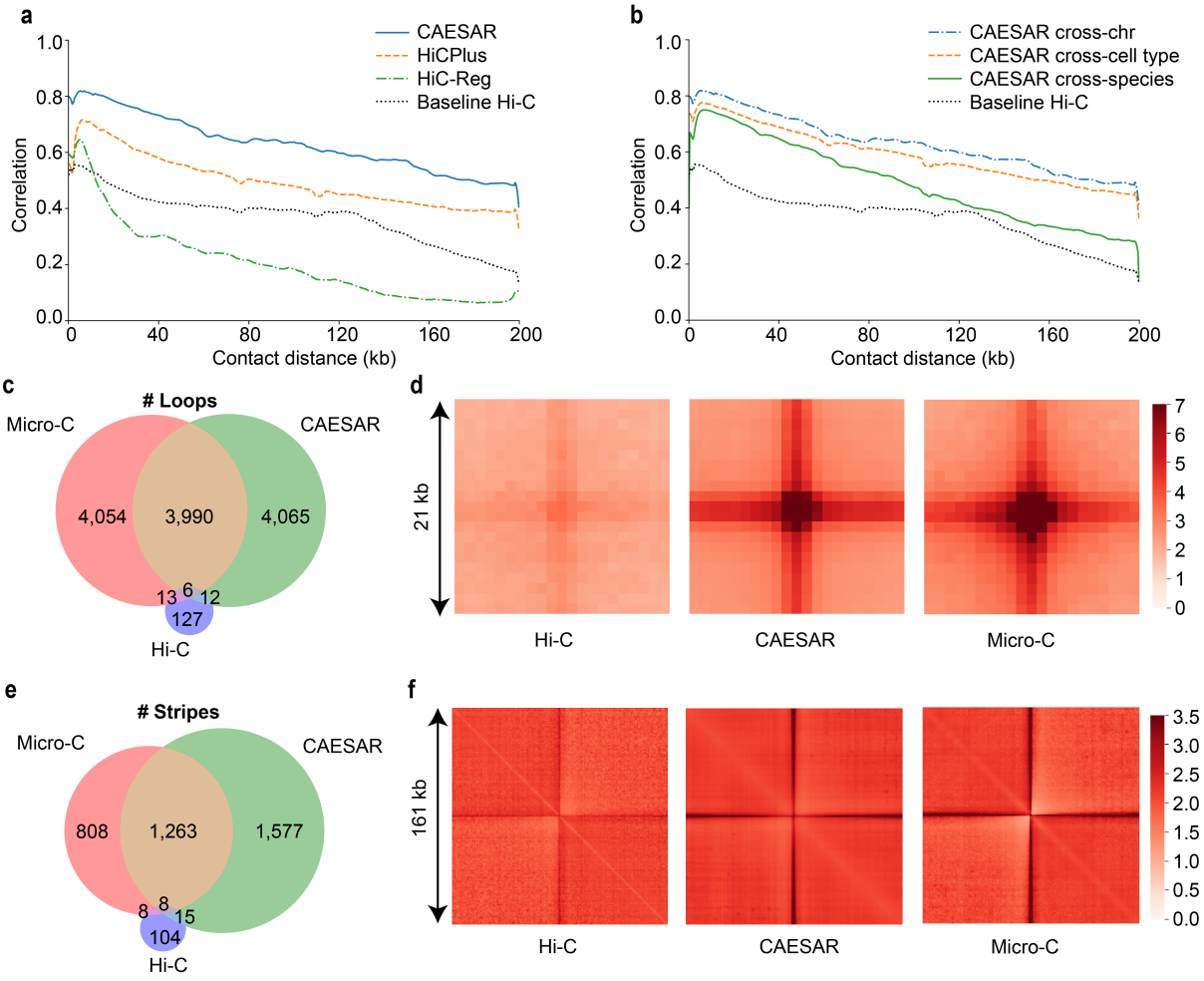

Figure 2: Evaluating CAESAR's performance in multiple tasks.
a, The distance-stratified Pearson's correlation with the observed Micro-C contact map from CAESAR and two baselines, HiC-Reg and HiCPlus, in a cross-chromosome experiment. The black dotted lines in **a** and **b** are the correlation between the input Hi-C contact map and the observed Micro-C contact map. **b**, The distance-stratified Pearson's correlation with the observed Micro-C contact map from CAESAR in 1) a cross-chromosome experiment (train on hESC train set and test on hESC test set), 2) a cross-cell type experiment (train on HFF train set and test on hESC test set), and 3) a cross-species experiment (train on mESC train set and test on hESC test set). **c**, The Venn diagram of the loops called from 1) the input Hi-C contact map, 2) the CAESAR-imputed contact map, and 3) the observed Micro-C contact map. **d**, The pile-up visualization of the loops called from 1) the input Hi-C contact map, 2) the CAESAR-imputed contact map, and 3) the observed Micro-C contact map. **e**, The Venn diagram of the stripes called from 1) the input Hi-C contact map, 2) the CAESAR-imputed contact map, and 3) the observed Micro-C contact map. **f**, The pile-up visualization of the stripes called from 1) the input Hi-C contact map, 2) the CAESAR-imputed contact map, and 3) the observed Micro-C contact map.

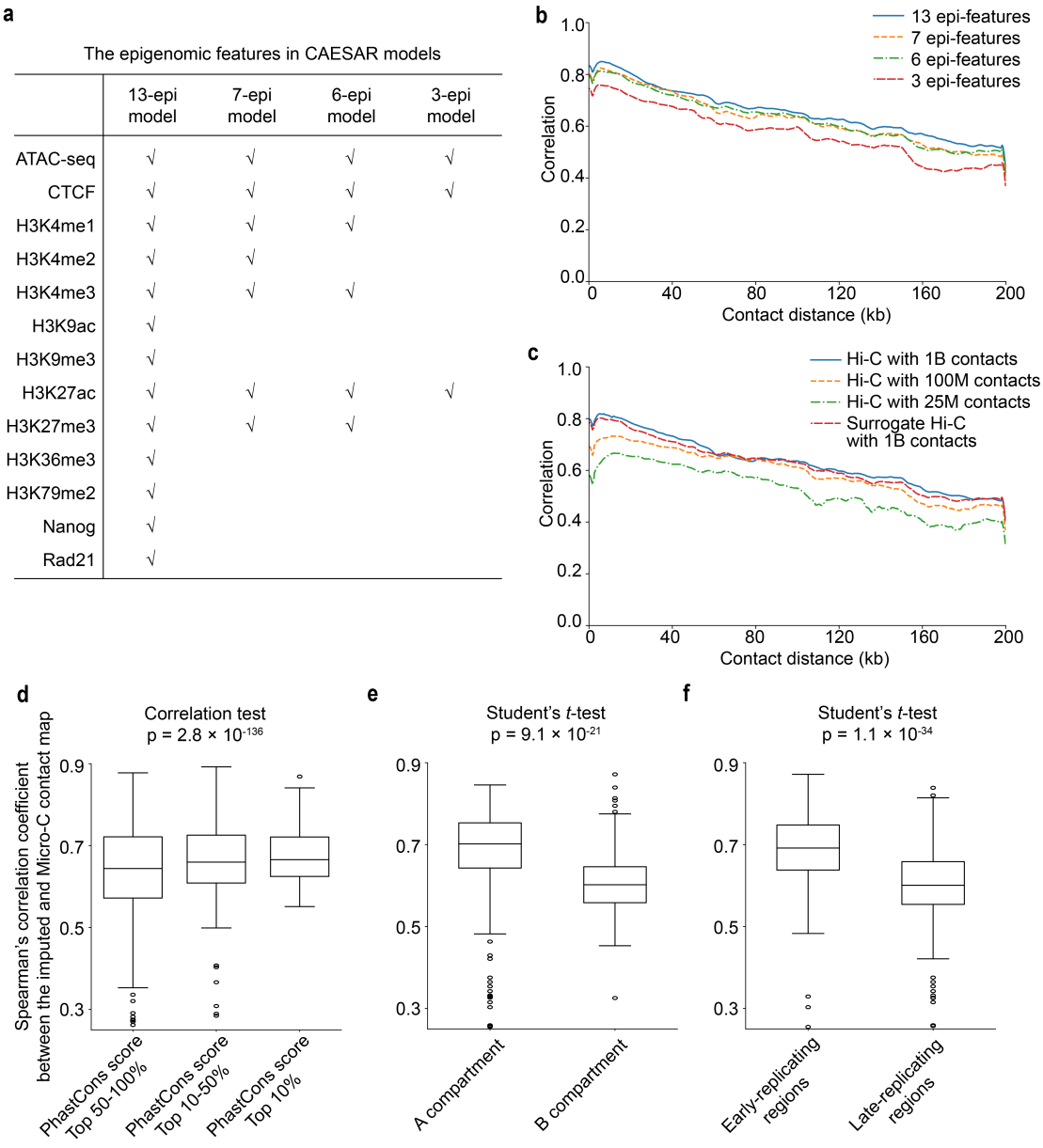

Figure 3: The relationships between CAESAR's performance with Hi-C quality, the number of epigenomic features, evolutionary conservation, A/B compartments, and early/late replication timing.
a, The epigenomic features in 13-epi, 7-epi, 6-epi, and 3-epi CAESAR models are listed in the table, which are chosen based on common availability. **b**, The distance-stratified Pearson's correlation with the observed Micro-C contact map from CAESAR in a cross-cell type experiment with different numbers of epigenomic features (i.e., 13, 7, 6, and 3). **c**, The distance-stratified Pearson's correlation with the observed Micro-C contact map from CAESAR in a cross-cell type experiment when 1) using the original Hi-C contact map with about 1 billion contacts, 2) randomly down-sampling the Hi-C contact map at different down-sampling rates (resulting in 100 million and 25 million chromatin contacts), and 3) using a surrogate Hi-C contact map with 1 billion contacts aggregated from HFF, GM12878, IMR-90, and K562 with equal proportions. **d**, The model performance in a specific region is quantified by the Spearman's correlation coefficient between the CAESAR-imputed and the Micro-C contact map. In cross-chromosome and cross-cell-type experiments, the model performance (i.e., Spearman's correlation coefficient) is significantly correlated with evolutionary conservation evaluated by sequence alignment scores. In the boxplots, the center line indicates median; the box limits are upper and lower quartiles; the whiskers are 1.5×interquartile range; the points are outliers. **e**, In cross-chromosome and cross-cell-type experiments, the correlation coefficient is significantly larger in A compartment than in B compartment. **f**, In cross-chromosome and cross-cell-type experiments, the correlation coefficient is significantly larger in early-replicating regions than in late-replicating regions.

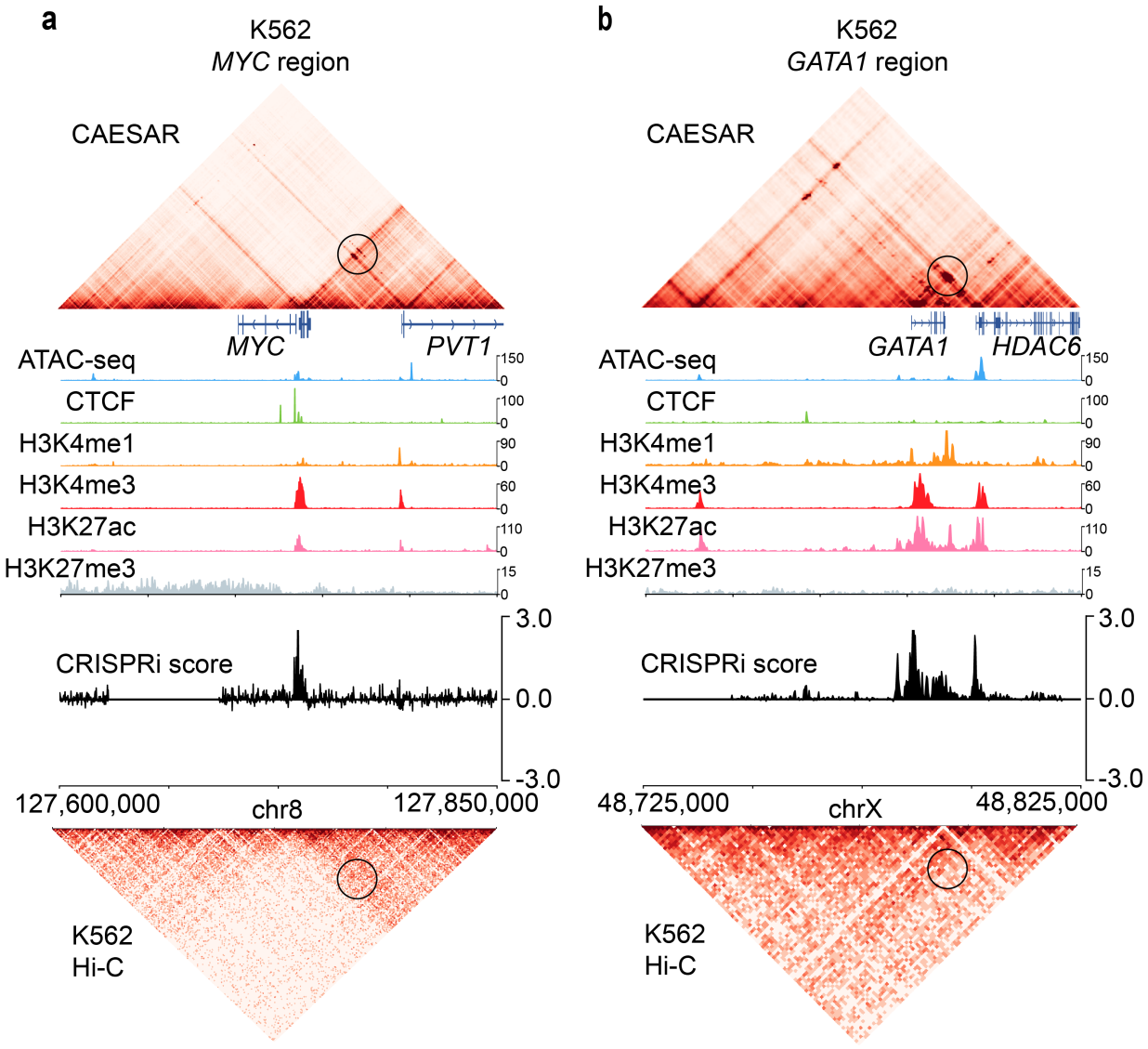

Figure 4: The interactions between genes and their CRISPRi-validated enhancers in CAESAR-imputed contact maps.
a, The CAESAR-imputed contact map of K562 at *MYC* region (chr8: 127,600,000-127,850,000) demonstrates significant contacts between *MYC* and *PVT1*, which agree with with CRISPRi score peaks, but are not shown on the original input Hi-C contact map. The magnitude of the epigenomic features is the observed value divided by the genome-wide average. **b**, The CAESAR-imputed contact map of K562 at *GATA1* region (chrX: 48,725,000-48,825,000) demonstrates significant contacts between *GATA1* and *HDAC6*, which agree with with CRISPRi score peaks, but are not shown on the original input Hi-C contact map.

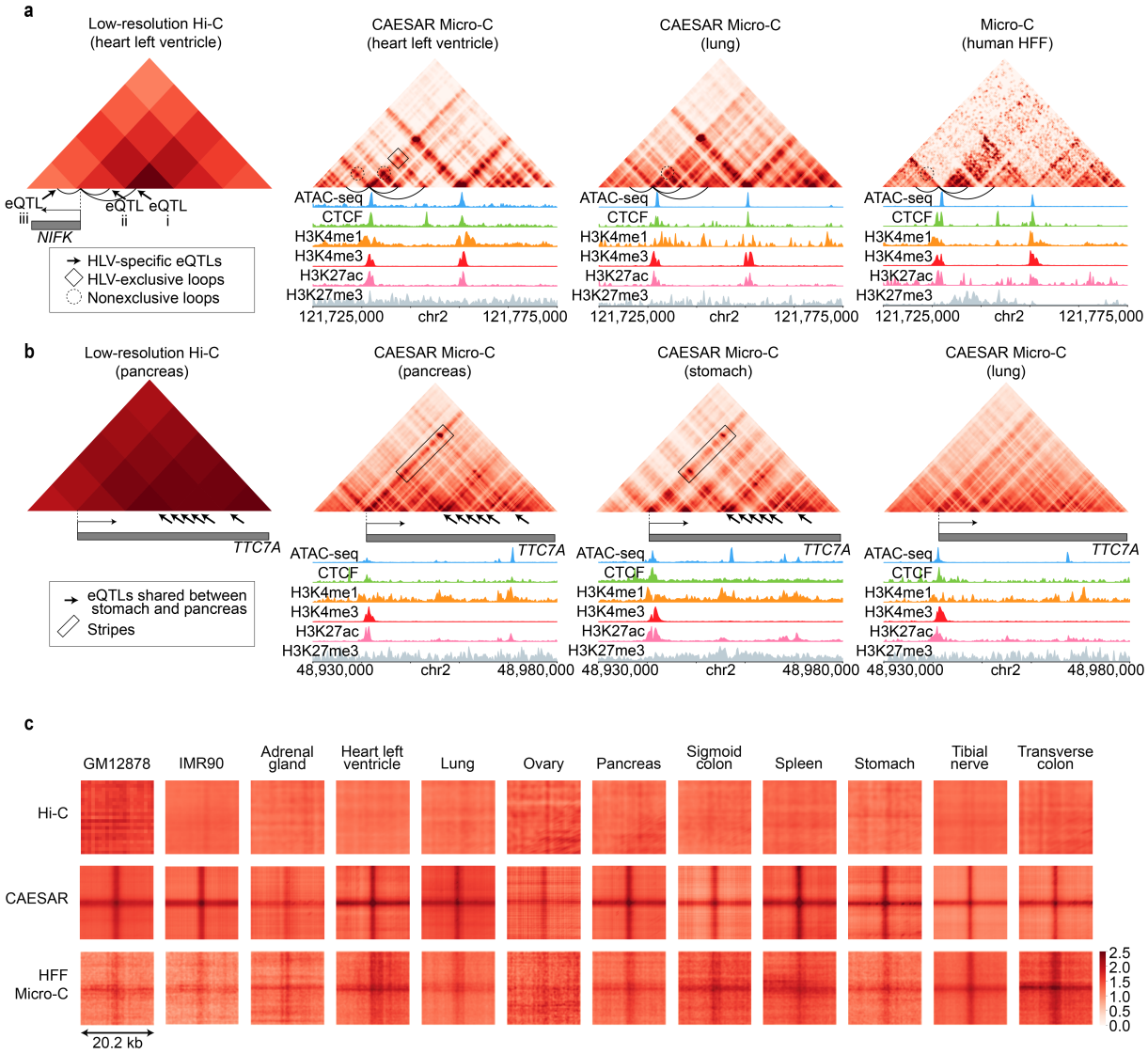

Figure 5: The enrichment of eQTL-gene interactions in CAESAR-imputed contact maps.
a, The loops between gene *NIFK*'s TSS and its three eQTLs specific in heart left ventricle (HLV), which cannot be observed on the low-resolution Hi-C contact map, appear on the CAESAR-imputed contact map of HLV. Although all three eQTLs are HLV-specific, only the loop between *NIFK* TSS and eQTL i is HLV-exclusive; while the other two loops can also be observed on the CAESAR-imputed contact map of lung and the Micro-C contact map of HFF, respectively. **b**, A series of gene *TTC7A*'s eQTLs are shared by stomach and pancreas, and both loops and stripes are observed on the CAESAR-imputed contact maps of the two tissues. As a reference, the contacts are not observed on the low-resolution Hi-C contact map of pancreas and less enriched on the CAESAR-imputed contact maps of lung. **c**, Pile-up analysis of the chromatin contacts between eQTLs and their corresponding gene TSS for 12 different human tissues and cell lines demonstrates highly enriched interactions on the CAESAR-imputed contact maps, but not on original Hi-C contact maps or HFF's Micro-C contact maps.

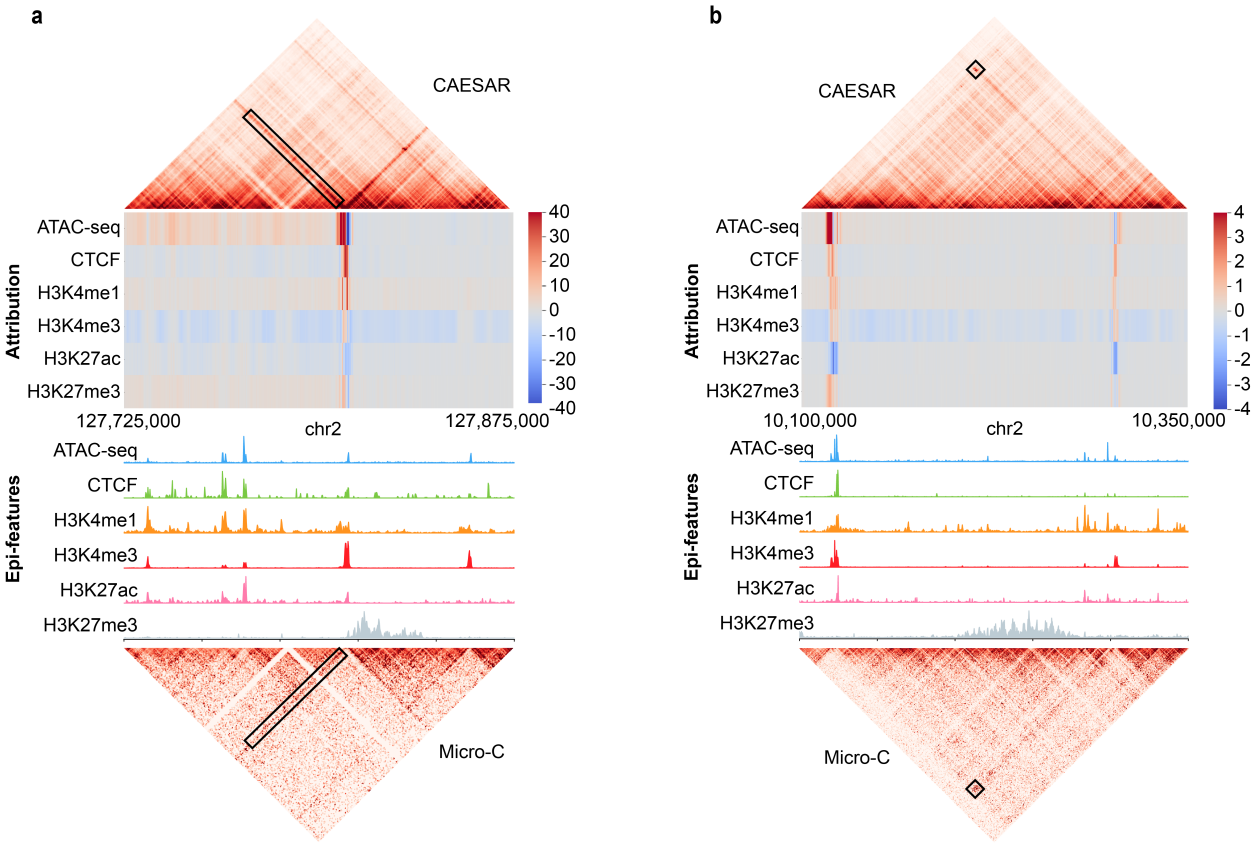

Figure 6: Attributing CAESAR outputs to epigenomic features via *integrated gradient*. Larger attribution magnitudes indicate more contribution to the model's prediction.

a, The significant attribution of the particular stripe are from its anchor. Although all 6 epigenomic features have peaks at the anchor locus, the model predicts the stripe mostly from 1) ATAC-seq and CTCF peaks at the anchor, and 2) H3K4me1 modification surrounding the anchor. **b**, The significant attribution of the particular loop are from its two anchors. Although H3K27ac have peaks at the left anchor locus, its contribution is negative towards predicting the loop. The CTCF binding at the anchors and H3K4me1/H3K4me3 modifications next to the anchors have positive attribution in predicting the loop.