# Cell type-specific prediction of 3D chromatin architecture

Jimin Tan[1], Javier Rodriguez-Hernaez[2,3], Theodore Sakellaropoulos[2], Francesco Boccalatte[2,4], Iannis Aifantis[2,4], Jane Skok[2,4], David Fenyö[4,5], Bo Xia[1#], Aristotelis Tsirigos[2,3,4#]

[1]Institute for Systems Genetics, New York University School of Medicine, New York, New York
[2]Department of Pathology, New York University School of Medicine, New York, New York
[3]Applied Bioinformatics Laboratories, New York University School of Medicine, New York, New York
[4]Perlmutter Cancer Center, NYU Langone Health, New York, NY
[5]Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, New York

[#]Corresponding authors:
Aristotelis Tsirigos, Aristotelis.Tsirigos@nyulangone.org; Bo Xia, Bo.Xia@nyu.edu

## Abstract:

The mammalian genome is spatially organized in the nucleus to enable cell type-specific gene expression. Investigating how chromatin architecture determines this specificity remains a big challenge. Methods for measuring the 3D chromatin architecture, such as Hi-C, are costly and bears strong technical limitations, restricting their widespread application particularly when concerning genetic perturbations. In this study, we present C.Origami, a deep neural network model for predicting *de novo* cell type-specific chromatin architecture. By incorporating DNA sequence, CTCF binding, and chromatin accessibility profiles, C.Origami achieves accurate cell type-specific prediction. C.Origami enables *in silico* experiments that examine the impact of genetic perturbations on chromatin interactions, and moreover, leads to the identification of a compendium of cell type-specific regulators of 3D chromatin architecture. We expect Origami – the underlying model architecture of C.Origami – to be generalizable for future genomics studies in discovering novel regulatory mechanisms of the genome.

## Introduction:

In mammalian cells, interphase chromosomes are hierarchically organized into large compartments which consist of multiple topologically associating domains (TADs) at the megabase and sub-megabase scale (Dixon et al., 2012). Chromatin looping within TADs functions to restrict enhancer-promoter interactions at the kilobase scale for regulating gene expression (Dixon et al., 2012; Schoenfelder and Fraser, 2019; Tang et al., 2015). The perturbation of TADs, such as disrupting TAD boundary, can lead to aberrant chromatin interactions and changes in gene expression (Kloetgen et al., 2020; Narendra et al., 2015). As a result, mutations that disrupt 3D genome organization can substantially affect developmental programs and play important roles in genetic diseases and cancer (Franke et al., 2016; Lettice et al., 2003; Lupiáñez et al., 2015; Spielmann et al., 2018).

The higher-order organization of the genome is largely determined by intrinsic DNA sequence features known as *cis*-regulatory elements that are bound by *trans*-acting factors in a sequence specific manner (Rowley and Corces, 2018). For example, the location and orientation of CCCTC-binding factor (CTCF) binding sites act as a landmark for defining boundaries of TADs. Other factors, such as the cohesin proteins, act together to regulate chromatin interaction via loop extrusion (Rowley and Corces, 2018). While most TADs are conserved across cell types, a substantial amount (>10%) of TADs are dynamic and vary in different cells (Schmitt et al., 2016). In addition, widespread cell type-specific chromatin-looping contributes to the precise regulation of gene expression (Phillips-Cremins et al., 2013; Tang et al., 2015). These fine-scale chromatin interactions are controlled by chromatin remodeling proteins and cell type-specific transcription factors such as GATA1 and FOX1A (Kagey et al., 2010; Schoenfelder and Fraser, 2019; Weintraub et al., 2017). While the general organization of chromatin architecture is largely well described, the current challenge is to reveal the principles underlying cell type-specific chromatin folding. Chromatin architecture capture technologies, such as Hi-C, are used for examining chromatin structure underlying gene regulation at fine-scales and across cell types (Lieberman-Aiden et al., 2009; Rao et al., 2014). However, these approaches are costly, require large cell numbers, and are unable to distinguish abnormal genome rearrangements, prohibiting their widespread applications in investigating how chromatin architecture determines cell type-specific gene expression, especially in cancer genomes.

Owing to its ability to model complex interactions, deep learning has emerged as a powerful strategy for studying genomic features. Application of deep learning models could minimize the requirement for experimental analyses of chromatin architecture (Eraslan et al., 2019; Zou et al., 2019). Since intrinsic features in DNA sequence of the genome partially determine its general folding principles, an approximate prediction of chromatin architecture can be made using sequence alone (Cao et al., 2021; Fudenberg et al., 2020; Schwessinger et al., 2020). However, different cell types rely on differential compendia of *trans*-acting factors to establish cell type-specific chromatin interactions (Rowley and Corces, 2018). Approaches that rely solely on DNA sequence are unable to predict cell type-specific chromatin interactions (Cao et al., 2021; Fudenberg et al., 2020; Schwessinger et al., 2020). Conversely, methods that rely only on chromatin profiles lack the consideration of DNA sequence features, thus generally requiring multiple epigenomic data to improve predictive power (Belokopytova et al., 2020; Bianco et al., 2018; Di Pierro et al., 2017; Qi and Zhang, 2019; Yang et al., 2021; Zhang et al., 2019). The limitations of current methods make it almost impossible to practically carry out *in silico* experiments for studying how trans-acting factors and DNA seqeunce features work together to shape chromatin architecture for gene expression regulation.

We propose that an accurate prediction of cell type-specific chromatin folding requires a model which effectively recognizes and integrates both DNA sequence features and cell type-specific genomic information. A practical model should also minimize the requirement for input information without performance loss. Based on these principles, we developed C.Origami, a deep neural network that synergistically integrates DNA sequence features and two essential cell type-specific genomic features, CTCF binding profile (CTCF ChIP-seq signal) and chromatin accessibility information (ATAC-seq signal). C.Origami achieved accurate prediction of cell type-specific chromatin architecture in both normal and rearranged genomes. Additionally, the high-performance of C.Origami enables *in silico* genetic perturbation experiments that interrogate the impact on chromatin interactions and moreover, allows the identification of cell type-specific regulators of genomic folding through *in silico* genetic screening. We expect the underlying deep learning architecture, Origami, to be generalizable for predicting genomic features and discovering novel genomic regulations.

## RESULTS:

### Origami: a model architecture for predicting cell type-specific genomic features

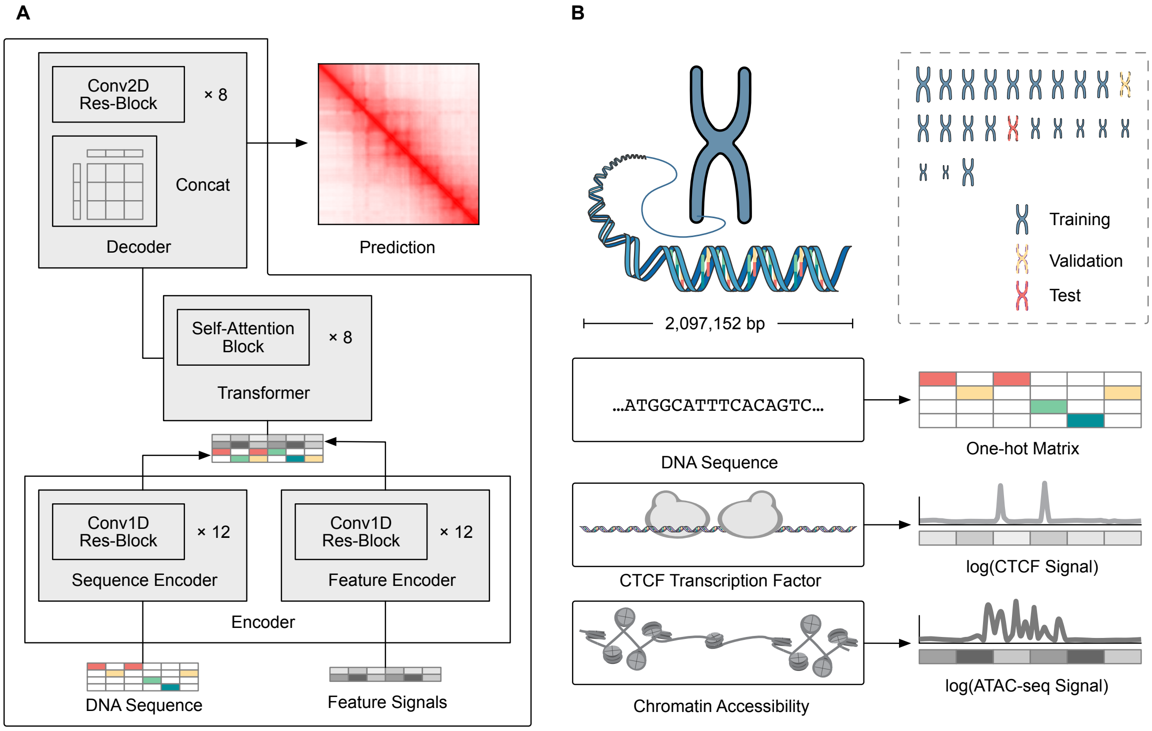


**Figure 1: *de novo* prediction of cell type-specific genomic features with Origami. a**, A schematic of Origami architecture. Origami adopts an encoder-decoder design, separately encoding DNA sequence features and cell type-specific genomic features. The two streams of encoded information are concatenated and processed by a transformer module. The decoder converts the processed 1D information to the final prediction, such as a Hi-C interaction matrix. **b**, Applying Origami model to predicting the Hi-C interaction matrix. The best-practice model integrates DNA sequence, CTCF ChIP-seq signal and ATAC-seq signal as input features to predict Hi-C interaction matrix in 2 Mb windows.

To achieve accurate and cell type-specific prediction of genomic features, we first developed Origami, a general modeling architecture, to synergistically integrate both nucleotide-level DNA sequence and cell type-specific genomic signal (Fig. 1a). In these two streams of information, the former enables recognition of informative sequence motifs, while the later provides cell type-specific features. The Origami architecture consists of two encoders, a transformer module and a decoder (Fig. 1a, see Methods). The two encoders process DNA sequence and genomic features independently. The encoded features are concatenated and further processed by a transformer model (Vaswani et al., 2017), which allows the encoded information to exchange between different genomic regions. The decoder in Origami synthesizes the processed information to make predictions, and depending on the task, can be customized to specific downstream prediction targets. In this study, we deployed a decoder for predicting chromatin architecture represented by Hi-C contact matrices, and therefore named this variant C.Origami.

To cover typical TADs in the genome while maximizing computation efficiency, C.Origami predicts chromatin architecture within a 2 mega-base (2Mb) sized genomic window (Dixon et al., 2012). DNA sequence and genomic features within the 2Mb window were separately encoded as nucleotide-level features (Fig. 1b, see Methods). The model reduces 2Mb wide genomic features down to 256 bins, and output a Hi-C contact matrix with a bin size of 8,192 bp resolution (see Methods). The target Hi-C matrix from the corresponding 2Mb genomic window was processed to have the same bin size. To train the model, we used data from IMR-90 (Rao et al., 2014), a fibroblast cell line isolated from normal lung tissue, and randomly split the chromosomes into training, validation (chromosome 10), and test set (chromosome 15) (Fig. 1b, top right).

To select genomic features as input for cell type-specific chromatin architecture prediction, we considered three criteria: 1) representative for cell type-specific identity; 2) widely available and experimentally robust; 3) minimized number of features to enable broad applicability of the model. CTCF binding is one of the most critical determinants of 3D genome architecture, thus we initially trained the model using DNA sequences and CTCF ChIP-seq signals as the only cell type-specific genomic feature (Supplementary Fig. 2). Our model performed well in most predictions, capturing the TAD structures and chromatin interaction events (Supplementary Fig. 2). However, we found the prediction did not recognize some fine-scale chromatin interaction features, especially in *de novo* prediction on a cell type (Supplementary Fig. 2). These results indicate that integrating DNA sequence with CTCF binding signal alone is not sufficient for optimal prediction of cell type-specific 3D genome conformation.

Previous studies indicate that chromatin accessibility directly or indirectly affects genome conformation with cell type-specific interactions (Stergachis et al., 2014; Thurman et al., 2012). We thus improved the model by including ATAC-seq signals as an extra feature (Fig. 1b). We found that C.Origami trained with nucleotide-level DNA sequence, CTCF ChIP-seq, and ATAC-seq signals provided high-quality predictions for chromatin architecture (Fig. 2). On validation chromosome 10 and test chromosome 15, C.Origami predicted highly accurate contact matrices that emphasized both large topological domains and detailed chromatin looping events (Fig. 2a-c and Supplementary Fig.3). To quantify prediction performance, we calculated the insulation scores from the predicted Hi-C matrix and found a high correlation with the insulation scores calculated from the experimental data (Fig. 2d). C.Origami achieved on average 0.95 and 0.94 Pearson correlation coefficients on validation and test chromosomes, respectively (Fig. 2e). We found that DNA sequence, CTCF binding signal, and chromatin accessibility signal were all required to accurately predict Hi-C contact matrix with high-quality. Compromising any of the signals led to inaccurate prediction (Supplementary Fig. 4).

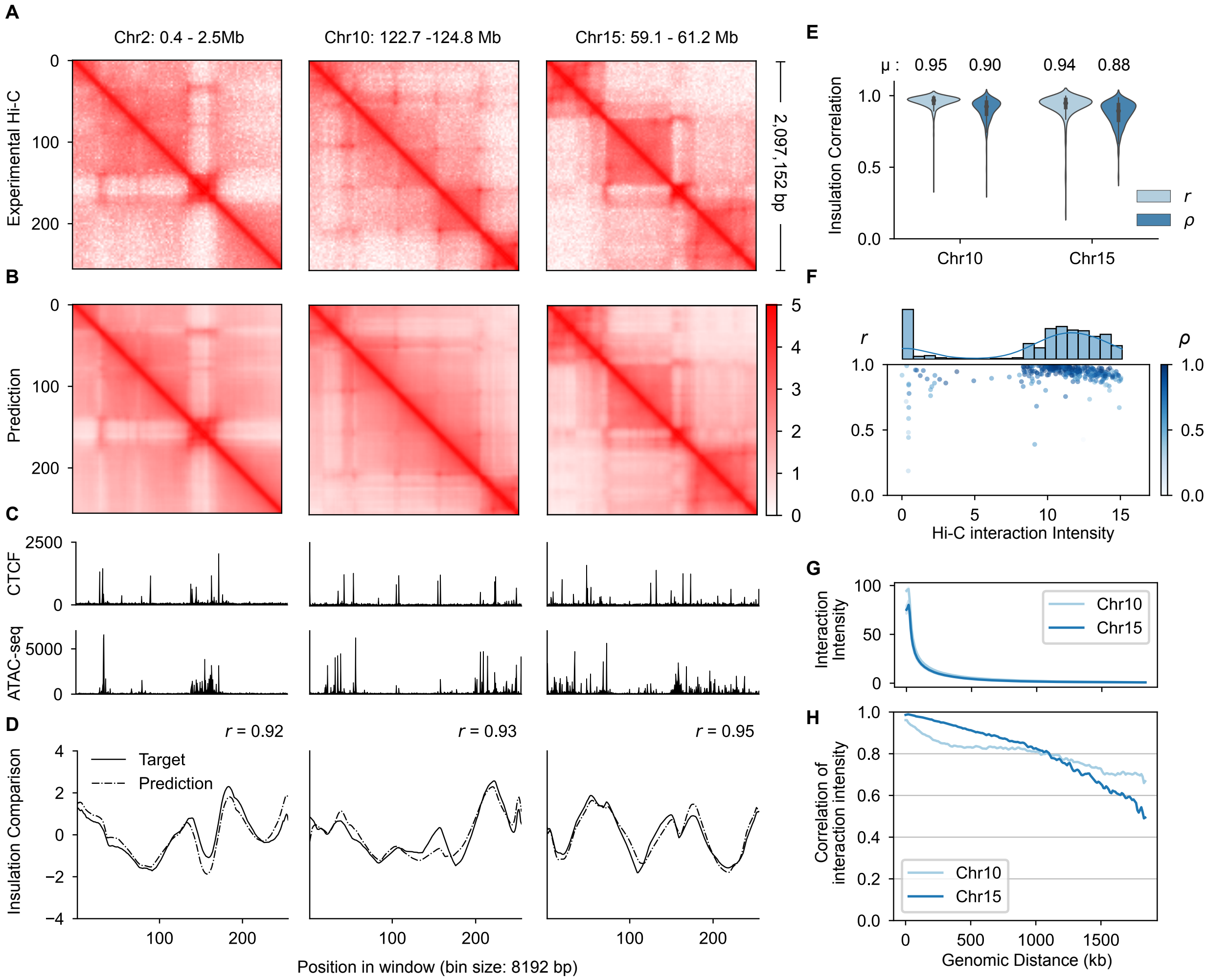


**Figure 2: C.Origami accurately predicts 3D chromatin architecture.** **a-b**, Experimental Hi-C matrices (**a**) and C.Origami predicted Hi-C matrices (**b**) of IMR-90 cell line at chromosome 2 (left), chromosome 10 (middle), and chromosome 15 (right), representing training, validation and test chromosomes, respectively. **c**, Input CTCF binding profiles and chromatin accessibility profiles. **d**, Insulation scores calculated from experimental Hi-C matrices (solid line) and C.Origami predicted Hi-C matrices (dotted line). Pearson correlation coefficients comparing the insulation was indicated in the plots. **e**, Insulation correlation between predicted and experimental Hi-C matrices across all windows in both validation and test chromosomes.

Each group included both Pearson correlation (*r*) and Spearman correlation (*ρ*) coefficients. **f**, The distribution of experimental Hi-C intensity scores by insulation correlation (Pearson's *r*) between prediction and experiment. Each point represents a 2Mb genomic window in chromosome 15 (test). Colormap indicates the Spearman's *ρ* of insulation correlation between prediction and experiment. **g**, Average intensity of the interaction matrix across genomic distances. **h**, Distance-stratified interaction correlation (Pearson) between prediction and experiment.

We carried out multiple different measurements to further evaluate the performance of C.Origami. First, by plotting the insulation correlation between prediction and experiment against Hi-C data intensity, we found that the predictions in the test set maintain uniform high performance across different clusters, demonstrating the robustness of the model (Fig. 2f). The few data points with low intensity are regions corresponding to unmappable or repeat sequences such as centromeres and telomeres (Fig. 2f and Supplementary Fig. 5). Second, our predicted Hi-C contact map followed the exponential decay pattern that are generally present in experimental Hi-C data (Fig. 2g). Third, we plotted the distance-stratified interaction correlation (Pearson) between prediction and experiment. C.Origami achieved correlation above 0.8 within 1Mb region and 0.6 within 1.5Mb (Fig. 2h). Last, we found that predictions from C.Origami were highly consistent across neighboring regions (Supplementary Fig. 6). Thus, C.Origami can be used to construct chromosome-wide prediction of Hi-C contact matrix by joining predictions across sliding windows. Together, the results demonstrate that C.Origami can accurately predict 3D chromatin architecture with minimum input data.

### ***De novo*** **prediction of cell type-specific chromatin architecture**

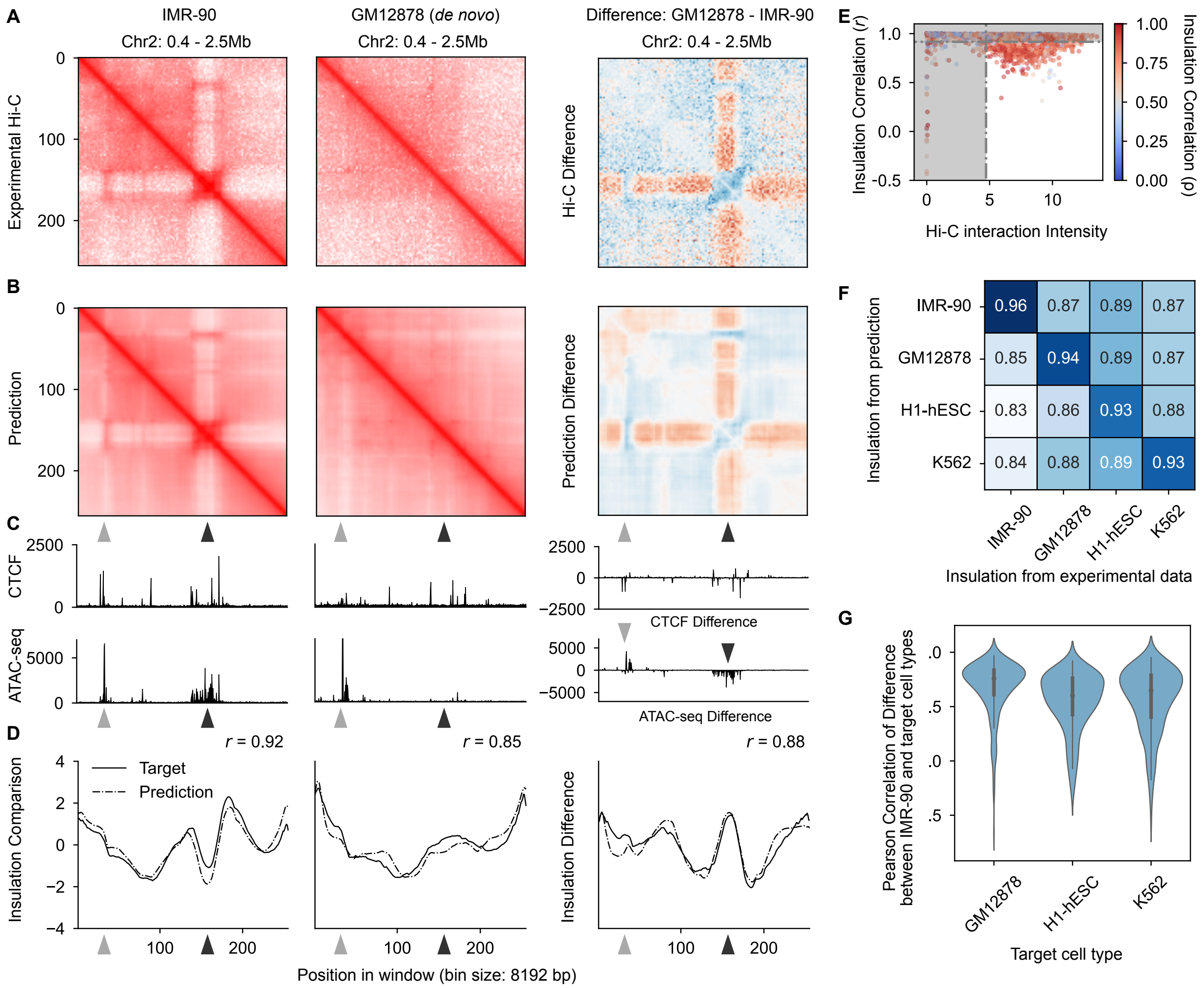


**Figure 3: Cell type-specific *de novo* prediction of chromatin structure. a**, Experimental Hi-C matrices from IMR-90 (left) and GM12878 (middle) cell lines at chromosome 2, highlighting cell type-specific chromatin differences (right). **b**, C.Origami-predicted Hi-C matrices of IMR-90 (left) and GM12878 (middle), precisely recapitulated the experimental Hi-C matrices (**a**). The arrow heads highlighted differential chromatin interactions between the two cell types. **c**, CTCF binding profiles and chromatin accessibility profiles of IMR-90 (left), GM12878 (middle) and their difference (right). **d**, Insulation scores calculated from experimental Hi-C matrices (solid line) and C.Origami predicted Hi-C matrices (dotted line) of IMR-90 (left), GM12878 (middle) and their difference (right). **e**, The distribution of interaction intensity by insulation correlation (Pearson) between the experimental Hi-C matrices of IMR-90 and GM12878. Colormap indicates the corresponding Spearman correlation coefficient (*ρ*). Dotted lines denote the filtering criteria in selecting representative loci with cell-type specificity. **f**, Pearson correlation between insulation scores calculated from predicted and experimental Hi-C matrices across cell types. Prediction from each cell type was similar to the corresponding experimental data. **g**, Pearson's *r* of predicted insulation difference and experimental insulation difference between IMR-90 and other cell types. The correlation was calculated as:

Pearson(*Insu*(IMR-90_pred) - *Insu*(Target_pred), *Insu*(IMR-90_data) - *Insu*(Target_data)). High correlation indicates that our model detected cell types-specific features applicable across different cell types.

We next tested whether our model generalizes to *de novo* predict of chromatin architecture in new cell types. GM12878, a lymphoblastoid cell line, differs substantially from IMR-90 in its chromatin architecture (Rao et al., 2014), as exemplified at locus Chr2:400,000-2,497,152 (Fig. 3a). Specifically, we highlighted a cell type-specific interaction related to chromatin accessibility changes (black arrowhead) and a distal interaction that associates with both CTCF and ATAC-seq signal changes (gray arrowhead, Fig. 3c). These cell type-specific features were clearly demonstrated by differences in their signal intensity in Hi-C and genomic tracks (Fig. 3a and 3c, right). To evaluate how C.Origami performs in *de novo* predicting cell type-specific chromatin architecture, we applied the prediction to both cell types at this locus. We found that the cell type-specific chromatin interactions were accurately captured in our prediction, and matched with the experimental Hi-C contact matrix in both cell types(Fig. 3b). The calculated insulation scores from the predicted Hi-C matrix were also highly correlated with the scores of the experimental data from both cell types (Fig. 3d, left and middle). In addition, the difference between insulation scores of the two cell types were highly correlated (Fig. 3d, right). We further expanded the *de novo* chromatin architecture prediction to two more cell lines, embryonic H1-hESC and erythroleukemia K562. Again, our model achieved accurate predictions of cell type-specific chromatin architecture with high specificity, demonstrating the robustness of C.Origami in *de novo* prediction and its practical potential for general application (Supplementary Fig. 7).

To systematically evaluate our model, we next assessed its performance across the genome. Although we presented accurate prediction results of multiple loci that have cell type-specific chromatin structures, most TAD boundaries are conserved across cell types (Schmitt et al., 2016). Therefore, we aimed to test the model on a subset of 2Mb loci with differential chromatin structures between IMR-90 and GM12878. Regions with normal intensity (> 10% intensity quantile) and low similarity (< 20% insulation difference) between the experimental Hi-C matrices of the two cell types were selected. In total, ~15% of the entire genome (~450Mb) were included for evaluating the performance of cell type-specific Hi-C prediction (Fig. 3e).

We calculated the correlation coefficient between the insulation scores of the predicted and experimental Hi-C matrices across all four cell types (Supplementary Fig. 7). In line with observations from the single locus experiment (Fig. 3a-d, Supplementary Fig. 7), we found that predictions using input features from one cell type has the highest correlation coefficients with the experimental Hi-C data of the same cell type (Fig. 3f, scores at the diagonal line). The correlation coefficients between prediction and experimental data from different cell types were lower, consistent with the expectation that the model predicts cell type-specific chromatin interactions (Fig. 3f). Similarly, these results were recapitulated by correlation analysis using pixel-level Observed/Expected contact matrices (Supplementary Fig. 8a-b). As a control, we performed a similar analysis using structurally conserved genomic regions, characterized by normal intensity (> 10% intensity quantile) and high similarity (> 20% insulation difference), between IMR-90 and GM12878 (Supplementary Fig. 8c). As expected, we found the prediction in these regions was highly correlated with the experimental data across all cell types (Supplementary Fig. 8d-e).

To quantify the performance of C.Origami in predicting cell type-specific chromatin architecture across the genome, we calculated the insulation difference between Hi-C matrices of IMR-90 to that of the three other cell lines using predicted or experimental data (Fig. 3g). We then computed the correlation between the cell-type insulation differences calculated from prediction and that from the experimental data. We found that all comparisons yielded high correlations between prediction and experimental data (Fig. 3g), indicating that C.Origami accurately detected the chromatin architecture difference across cell types comparable to that detected from experimental Hi-C technique.

We further compared the performance of C.Origami to Akita, a deep learning model trained on DNA sequence alone for predicting Hi-C contact matrix (Fudenberg et al., 2020). We found C.Origami outperformed Akita and made accurate cell type-specific predictions regardless of loci (Supplementary Fig. 9). Together, our results indicate that C.Origami trained with DNA sequence, CTCF binding and chromatin accessibility signals performs optimal in *de novo* predicting high-quality Hi-C contact matrix, and sensitively captures cell type-specific chromatin folding features.

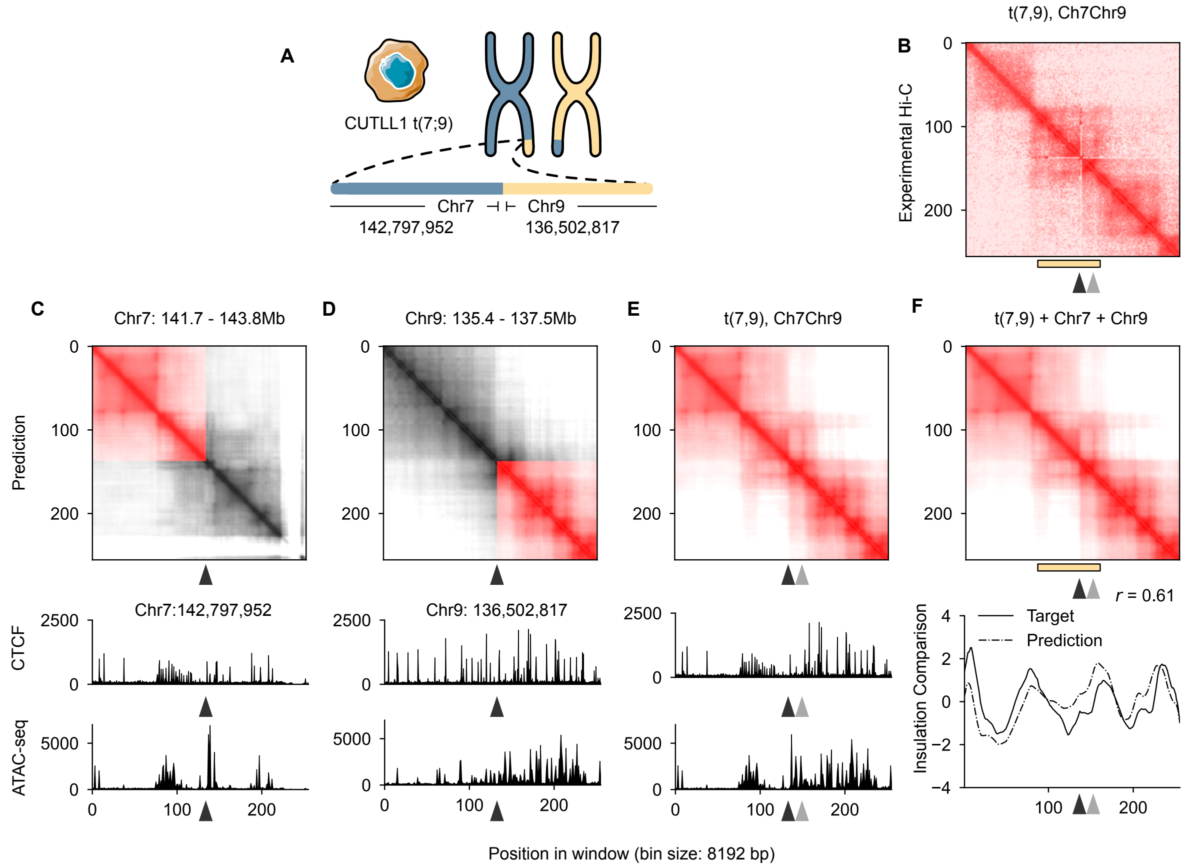


**Figure 4: C.Origami enables allele-specific prediction of 3D chromatin architecture in rearranged cancer genome. a**, Chromosomal translocation between chromosome 7 and chromosome 9 in CUTLL1 T cell leukemia cells (Palomero et al., 2006). **b**, Experimental Hi-C data mapped to a custom reference chromosome with t(7,9) translocation (Kloetgen et al., 2020). **c-d**, C.Origami prediction of chromatin architecture of chromosome 7 (**c**) and chromosome 9 (**d**) in CUTLL1 cells. The windows represented intact chromosomal loci around the translocation sites in CUTLL1 cells. **e**, C.Origami prediction of chromatin architecture at the t(7,9) translocation locus. **f**, A simulated Hi-C contact matrix using prediction for mimicking of experimental mapping results. The simulated result was averaged from the prediction of both normal and translocated alleles. The simulated Hi-C matrix was aligned to the experimental Hi-C matrix (**b**), with highlights for the neo-TAD at the translocation locus (yellow bar). Black arrowhead indicates the translocation site. The grey arrowhead indicates a stripe in the neo-TAD.

## Allele-specific prediction in rearranged cancer genomes

Chromosomal translocations and other structural variants generate novel recombined DNA sequences, subsequently inducing new chromatin interactions which may be critical in tumorigenesis and progression (Rabbitts, 1994; Spielmann et al., 2018). However, the allelic effect of translocation and structural variations frequently seen in cancer genomes makes it challenging to distinguish the chromatin architecture of the variant chromosome from a normal one. For example, CUTLL1, a T cell leukemia cell line, incorporated a heterozygous t(7,9) translocation where the end of chromosome 7 is recombined with chromosome 9 (Palomero et al., 2006) (Fig. 4a). The translocation introduces new CTCF binding signals from chromosome 9 to chromosome 7 (Kloetgen et al., 2020). Experimental Hi-C in CUTLL1 cells detected the formation of a neo-TAD at the translocation locus when mapped to a custom CUTLL1 reference genome (Fig. 4b). However, due to the limitation in mapping sequencing data to the reference genome, experimental Hi-C measures chromatin architecture allele-agnostically, and is thus unable to quantify allele-specific translocation.

To examine the performance of C.Origami in predicting chromatin architecture from recombined cancer genomes, we applied the model to 2Mb windows centered at the translocation breakpoint in CUTLL1 cells (Fig. 4c-e). We first predicted the Hi-C contact matrices referring to normal alleles at chromosome 7 and chromosome 9 (Fig. 4c-d). Since the input CTCF ChIP-seq and ATAC-seq profiles can only be mapped allele-agnostically, our prediction used these inputs as an approximation. Then we simulated the translocation by fusing DNA sequences at the breakpoint in Chromosome 7 (q34) to the Chromosome 9 (q34) breakpoint together with all genomic features (see Methods). The predicted Hi-C map from translocation detected a neo-TAD forming between the two recombined chromosomes (Fig. 4e). Specifically, we found a stripe extending from translocated chromosome 9 to chromosome 7, indicating a novel regulation in the recombined chromosome (Fig. 4e, gray arrowhead). We next averaged the Hi-C contact matrix from normal and translocated alleles, mimicking the allele-agnostic Hi-C mapping in the experimental data, and found a high correlation between the two (Fig. 4b and 4f, see Methods). The high-accuracy in prediction underscores the potential of applying C.Origami in future cancer genomics studies.

### Transferring knowledge learned from human genome to predict mouse chromatin architecture

The mouse genome differs from human in its genomic components but the two share similar mechanisms in 3D chromatin organization (Cheng et al., 2014; Dixon et al., 2012; Stergachis et al., 2014). We sought to test whether C.Origami could apply knowledge learned from human genome to a different species. In an initial trial, we found that our model trained with DNA sequences and dense genomic features (e.g. bigwig tracks) did not achieve good performance. We hypothesized that the background intensity in dense features can be highly specific to species and thus such knowledge learned from dense profiles in human made it challenging to transfer to the mouse.

We expect sparse features such as peaks to be less specific, and more consistent across species. To achieve cross-species prediction using a model trained with human data, we modified our input data by performing a peak-calling step on the CTCF CHIP-seq and ATAC-seq profiles and used such sparse genomic features as input for training and prediction (see Methods). We confirmed that using sparse input genomic features did not significantly undermine the model's prediction performance in human (Supplementary Fig. 10). Testing the model trained on sparse features of human IMR-90 cell line for mouse prediction, we found it capable of predicting mouse chromatin architecture with good quality, indicating the power of C.Origami for transferring the conserved genomic features learned from different species (Supplementary Fig. 11). Notwithstanding the good performance, the accuracy of C.Origami can be further improved by training on mouse data to adapt to mouse sequence and genomic features.

## High-accuracy prediction of C.Origami enables cell type-specific *in silico* genetic experiments

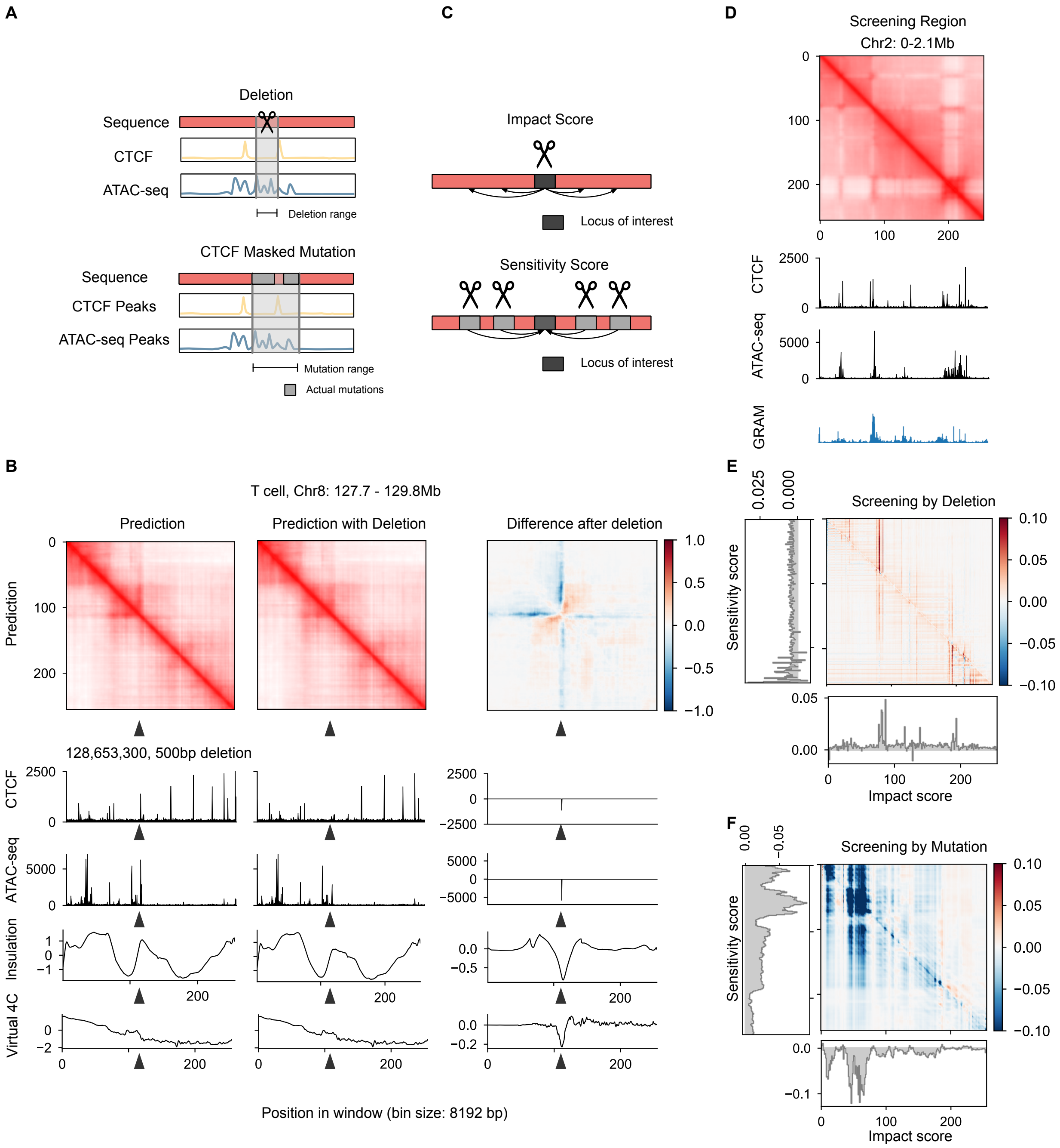


**Figure 5, *In silico* genetic experiments for identifying *cis*-regulatory elements determining chromatin architecture. a**, Schematic of *in silico* deletion and masked mutation experiments. A deletion experiment completely removed both DNA sequences and genomic signals, while a masked mutation experiment shuffled DNA sequence but not the genomic peaks and their underlying DNA sequences. **b**, A 500bp deletion in chromosome 8 led to chromatin looping changes in T cells. The presented 2Mb window starts at the promoter region of *MYC*, and the experimental deletion perturbed a CTCF binding site at the arrowhead location (Kloetgen et al., 2020). The presented results include C.Origami prediction of the Hi-C contact matrices with (middle) or without (left) the deletion, and their difference (right). The virtual 4C signal, calculated from the predicted Hi-C matrices, is shown at the bottom. **c**, Schematic of impact score that indicates how perturbation of one locus affected the local chromatin folding, and sensitivity score that indicates how sensitive a locus is to genetic perturbations in neighboring areas. **d**, GRAM score, indicating the contribution of genomic location to the predicted Hi-C matrix. **e-f**, Sliding-window deletion screening (**e**) and CTCF-masked mutation screening (**f**) across a 2Mb window corresponding to **d**. Impact and sensitivity scores were shown on the horizontal and vertical axis, respectively. CTCF peak and its DNA sequences were masked to prevent disruption of CTCF signal. Arrowhead in **f** indicates a potential regulatory elements free of CTCF binding and ATAC-seq signals.

The high accuracy of C.Origami allowed us to perform cell type-specific *in silico* experiments, and therefore enabled studying how chromatin interaction may be altered upon genetic perturbation. Deletions and mutations are two common types of perturbations in genetic studies. Deletion removes all three types of input features at the perturbed locus, and can lead to a TAD merge event in experiments (Narendra et al., 2015) (Fig. 5a, top). Instead of experimentally performing such genetic studies, we modelled deletions of TAD boundary sequences in IMR-90 cells *in silico*, and subsequently predicted local chromatin interaction maps with C.Origami. We found that *in silico* deletion at TAD boundaries led to TAD merging events of the originally insulated adjacent TADs and a sharp drop in insulation score (Supplementary Fig. 12), indicating the impact of this genetic alteration.

To further investigate the validity of *in silico* genetic experiments, we applied C.Origami to predict chromatin interactions surrounding the *MYC* locus which was experimentally perturbed in T cells (Kloetgen et al., 2020). Our previous study showed that disrupting a CTCF-binding site near *MYC* reduced the chromatin looping efficiency in T cells, resulting in a reduced insulation score (Kloetgen et al., 2020). Applying C.Origami at the locus, we found a stripe in the predicted Hi-C matrix (Fig. 5b, left, arrowhead), while a 500bp *in silico* deletion covering the perturbed CTCF-binding signal attenuated such interaction (Fig. 5b, middle and right). Based on our predicted Hi-C matrices, we calculated virtual 4C profiles after perturbing the CTCF binding site and found them to be consistent with the experimental data (Supplementary Fig. 7E in Kloetgen, *et al*)(Kloetgen et al., 2020).

### Cell type-specific *in silico* genetic screen of *cis*-regulatory elements

To determine whether C.Origami could be used to identify *cis*-regulatory elements affecting chromatin folding using *in silico* genetic screening, we developed two different approaches:

gradient-based scoring and perturbation-based approaches (Fig. 5c-f). In the gradient-based approach, we defined a GRAM (Gradient-weighted Regional Activation Mapping) score to estimate how significant each genomic site contributed to the prediction of the final Hi-C matrix (Fig. 5c, see Methods). We found GRAM score precisely captured important genomic regions that determine 3D genome structure such as TAD boundaries (Fig. 5d).

To orthogonally demonstrate the capability of C.Origami in discovering novel regulation of chromatin architecture, we carried out *in silico* genetic screening experiments with systematic perturbation. We divided the window into 256 perturbation regions of ~8kb, followed by deletion and prediction across the whole 2Mb window (see Methods). This process produced a mapping of intensity shift at each perturbed region. We defined the impact score to measure the contribution of a locus on chromatin architecture within the 2Mb window (Fig. 5c, top). This was calculated as the average intensity change of the entire 2Mb window after perturbation of a given locus. We also defined a sensitivity score to measure how sensitive a locus is to the perturbations of its surrounding region (Fig. 5c, bottom). We calculated it as the average intensity change of one locus when every region in a 2Mb window is perturbed. We found that deletion at TAD boundaries with enriched CTCF ChIP-seq peaks had the highest impact on chromatin folding in the *in silico* screening experiment (Fig. 5d-e). This result is consistent with the fact that CTCF binding is a key signal in determining TAD boundaries, and its deletion can lead to alteration of TAD structure, thereby changing the overall intensity of neighboring regions (Kloetgen et al., 2020; Narendra et al., 2015).

To discover CTCF-independent factors regulating chromatin interaction, we performed an *in silico* screening through CTCF-masked mutagenesis (referred to as mutation) experiment. We first selected a perturbation region and masked the CTCF peaks and their underlying DNA sequences. We then performed the mutation experiment of the given region by shuffling unmasked DNA sequences, followed by a prediction from C.Origami on the 2Mb genomic window (see Methods). We then calculated the impact and sensitivity scores similar to the *in silico* deletion screening. By masking CTCF peaks and its underlying sequence, mutation screening allowed us to identify multiple CTCF-independent genomic elements that might be critical for chromatin architecture, including regions free of ATAC-seq signal (Fig. 5f, arrowhead). In contrast, we found sensitivity scores were more similar for loci within the same TADs than those across different TADs, consistent with the expectation that the deletion perturbation is likely to cause intensity shifts within the TAD (Fig. 5f). Together, our data show that C.Origami can be used to systematically identify how *cis*-regulatory elements affect chromatin folding in high-throughput *in silico* genetic screening.

## Genome-wide *in silico* screening revealed canonical and novel regulators of chromatin folding

We next asked whether C.Origami could identify a compendium of *trans*-acting regulators determining the  chromatin interactions in a cell-type specific scenario. We first systematically scanned through the whole genome to discover genomic loci that were critical for predicting chromatin architecture in IMR-90 cells. We separately applied *in silico* deletion and mutation experiments across the entire genome and calculated the impact score at each 20Kb locus. The DNA sequence of the perturbed loci with high impacts – positive or negative – were designated as potential functional elements for subsequent analysis with LOLA (Locus OverLap Analysis for enrichment of genomic ranges) (Sheffield and Bock, 2016) (Fig. 6a).

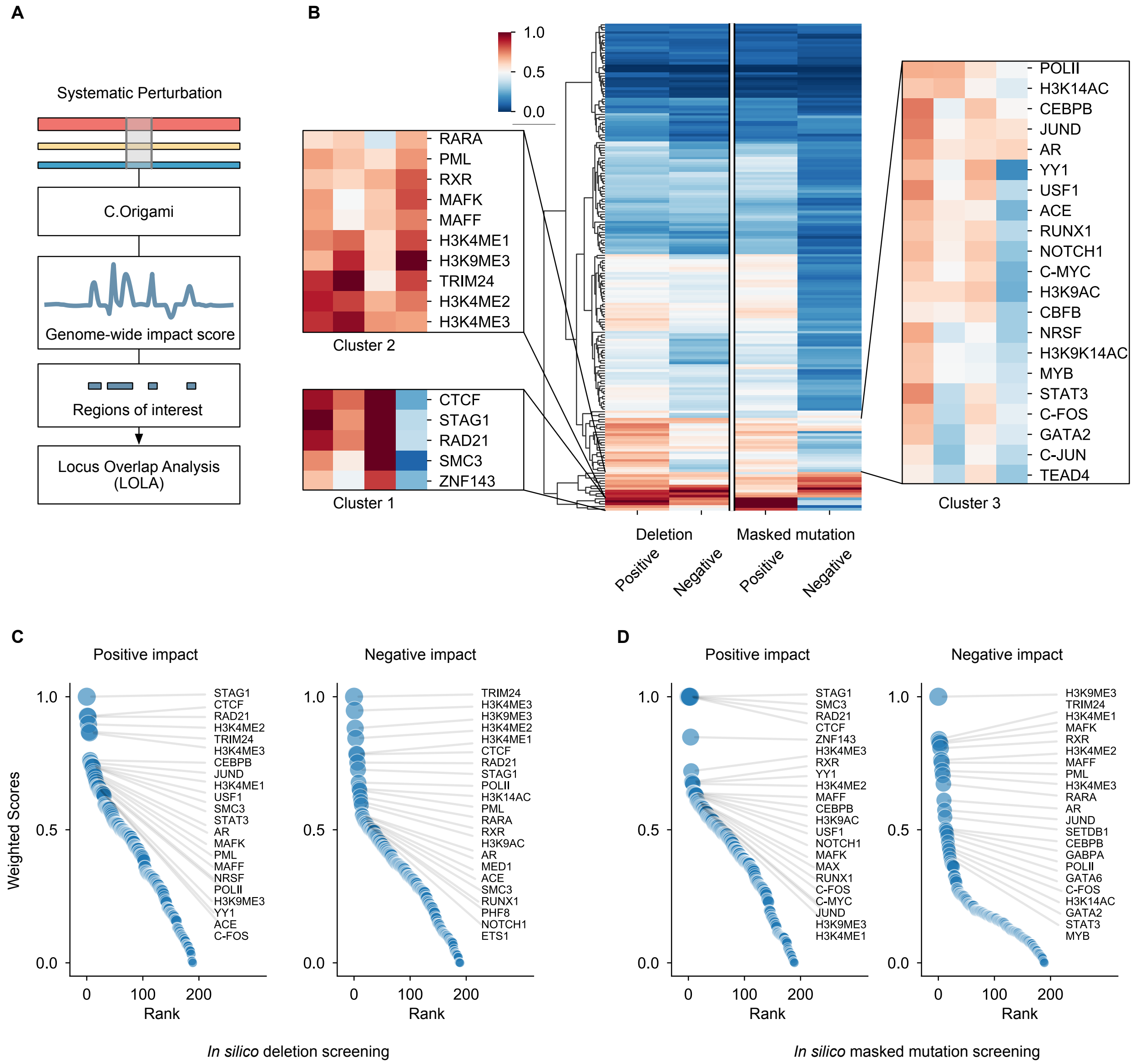


**Figure 6: Genome-wide *in silico* screening uncovers *trans*-regulators of chromatin folding. a**, Schematic of whole-genome *in silico* screening process. **b**, A heatmap of weighted scores across the four categories of *in silico* screen-determined contributing factors. The plot highlights three major clusters of contributing factors. **c-d**, *In silico* screening-identified contributing factors ranked by their weighted scores in each of the four categories as defined in **b**.

Scanning throughout the genome separately in the two types of *in silico* screening allowed us to identify *trans*-acting factors important for chromatin structure (Fig. 6b). As expected, CTCF, together with other canonical factors such as RAD21, STAG1 and SMC3, were significantly enriched in the positive impact score categories due to their role in determining TAD boundaries (Fig. 6b, cluster 1). These factors did not stand out in the negative score category of mutation screening due to CTCF masking, acting as a negative control for the results.

In contrast to the category enriched in the positive impact score group, we identified a cluster of factors which strongly associated with both positive and negative impacts on chromatin folding in the screening experiments (Fig. 6b, cluster 2). Of note, this cluster was enriched in several histone modifications represented by H3K4me1/2/3, identifying active chromatin marks that are known to contribute to enhancer-promoter looping (Zhao et al., 2019). This cluster is also enriched for H3K9me3, a mark of constitutive heterochromatin, which is involved in shaping chromatin compartment boundaries (Feng et al., 2020).

In addition, the *in silico* screening identified multiple transcription factors which may function to modulate fine-scale chromatin interactions. The positive impact score categories enriched for many transcription factors (Fig. 6b, cluster 3), such as YY1, NOTCH, and GATA2, indicating that the *in silico* screening precisely identified these as critical factors for chromatin interactions, in line with previous studies (Petrovic et al., 2019; Weintraub et al., 2017; Wu et al., 2014). Beyond this, cluster 3 identified factors that were not previously known to have a role in modulating chromatin interactions, such as the stress response transcription factors JUND and C-JUN. Interestingly, other AP-1 family proteins such as FOS, have been reported to alter chromatin interactions of their targeting genes (Beagan et al., 2020). Together, our *in silico* genetic screen confidently recognized critical chromatin architecture regulators, highlighting its potential for identifying a compendium of *trans*-acting factors and discovering novel regulation in determining chromatin interactions.

## Discussion:

Cell type-specific gene expression profiles require unique chromatin folding patterns. In this study, we developed a novel deep neural network model, C.Origami, that synergistically incorporates both DNA sequence and cell type-specific genomic features for *de novo* prediction of 3D genome architecture. We found that CTCF binding together with DNA sequence was not sufficient for accurately predicting cell type-specific chromatin architecture. Additional features such as cell type-specific chromatin states play an essential role in chromatin interactions (Stergachis et al., 2014; Thurman et al., 2012). Consistent with this, we found that incorporating chromatin accessibility data into C.Origami provided enough information for accurately predicting chromatin architecture, mirroring the results of a high-quality Hi-C experiment. The C.Origami model achieves high accuracy in *de novo* predicting cell type-specific chromatin architecture. This high performance and minimal requirement on input data make it practical for *de novo* prediction of Hi-C contact maps. The predicted Hi-C contact matrices can be further analyzed and interpreted through other available computational tools for inferring TADs, enhancer-promoter interactions, and higher-order chromosomal structures (Forcato et al., 2017; Lu et al., 2020; Szabo et al., 2018).

C.Origami model learned critical features from DNA sequences and cell type-specific information from the CTCF binding and ATAC-seq profiles, thus achieving high performance in *de novo* prediction of cell type-specific chromatin architecture. Other methods for predicting chromatin architecture either lack cell type-specificity or require substantial amount of input data, making them not practical for studying chromatin architecture underlying gene expression regulation. It is worth mentioning that, while preparing the manuscript, another method, Epiphany, was developed for cell type-specific prediction of Hi-C contact matrices using five input genomic profiles (Yang et al., 2021). Compared with Epiphany, C.Origami achieved high-quality prediction with minimal input data.

With highly accurate prediction of chromatin architecture, our model enables *in silico* genetic perturbation as a tool to study how *cis*-regulatory elements determine 3D chromatin architecture in a cell type-specific manner. C.Origami is able to accurately simulate the changes in chromatin architecture upon genetic perturbation within seconds and without the need to perform experimental studies. The low cost and high speed of C.Origami simulation make it useful in studies requiring frequent measurement of chromatin architecture, such as cancer genomics involving widespread genome rearrangement and synthetic regulatory genomics with *de novo* regulatory circuit  construction (Pinglay et al., 2021; Rabbitts, 1994; Spielmann et al., 2018).

Expanding the throughput of *in silico* genetic perturbations, we performed genome-wide *in silico* screening of features using deletion and masked mutation experiments in IMR-90 cells. This screening allowed us to determine the compendium of *trans*-acting regulators determining the chromatin architecture in a cell type-specific manner. This compendium not only includes canonical factors for determining chromatin architecture, such as CTCF, RAD21, STAG1 and SMC3, but also transcription factors that potentially function through modulating fine-scale chromatin structure for the regulation of gene expression. Meanwhile, the *in silico* screening identified *cis*-regulatory elements free of CTCF binding and ATAC-seq signals, indicating potential uncharacterized regulatory sequences in the genome. We postulate that systematic *in silico* screening could be generally applicable in discovering novel 3D genome regulatory mechanisms and identifying the specific compendium of regulators across different cell types.

We demonstrated that by integrating cell type-specific genomic features and DNA sequence features, C.Origami model is capable of predicting complex genomic features such as 3D chromatin architecture with high accuracy. The underlying architecture of our model, Origami, is generalizable beyond 3D genome structure prediction. Origami can be trained with appropriate genomic datasets for predicting cell type-specific genomic features, such as epigenetic modifications. Ultimately, we expect future genomics study to shift towards using tools that leverage high-capacity machine learning models to perform *in silico* experiments for discovering novel genomic regulation.

## Acknowledgement

A.T. is supported by the NCI/NIH P01CA229086, NCI/NIH R01CA252239, NCI/NIH R01CA260028 and NIH/NCI R01CA140729. I.A. is supported by the NIH R01CA266212, R01CA242020, R01CA228135 and P01CA229086. JS is supported by the R35GM122515, P01 CA229086 and P30CA016087. We would like to thank the Genome Technology Center (GTC) for expert library preparation and sequencing, and the Applied Bioinformatics Laboratories (ABL) for providing bioinformatics support and helping with the analysis and interpretation of the data. GTC and ABL are shared resources partially supported by the Cancer Center Support Grant P30CA016087 at the Laura and Isaac Perlmutter Cancer Center. This work has used computing resources at the NYU School of Medicine High Performance Computing (HPC) Facility. We would like to thank Sudarshan Pinglay, Jef Boeke, Huiyuan Zhang, and the members of the Tsirigos lab for suggestions and discussion.

## Author contribution

J.T. and B.X. conceived the project. J.T., B.X. and A.T. designed the experiments and interpreted the results. J.T. designed, implemented and optimized the neural network, and performed all the downstream computational analysis. J.R. helped with processing the sequencing data. F.B. generated ATAC-seq for CUTLL1. J.T. prepared figures with inputs from B.X., A.T. and D.F. T.S., J.S., I.A. and D.F. contributed to discussion. B.X., J.T. and A.T. wrote the manuscript with input from all authors.

## Competing interests

A.T. is a scientific advisor to Intelligencia AI. I.A. is a consultant for Foresite Labs. J.T, B.X and A.T are inventors on a filed patent covering the models and tools reported herein. All other authors declare no competing interests.

## Methods:

### Hi-C data:

We used seven human and mouse Hi-C profiles in this study: IMR-90, GM12878, H1-hESC, K562, CUTLL1, T cell, Mouse ESC (Supplemental Table 1). All the data are available on GEO (www.ncbi.nlm.nih.gov/geo) and 4D Nucleome Data Portal (https://data.4dnucleome.org).

| Cell Type | Enzyme | Accession Number | Reference |
|---|---|---|---|
| IMR-90 | MboI | GSE63525 | Rao et al. |
| GM12878 | MboI | GSE63525 | Rao et al. |
| H1-hESC | Arima | 4DNESFSCP5L8 | Calandrelli et al. |
| K562 | MboI | GSE63525 | Rao et al. |
| CUTLL1 | Arima | GSE115896 | Kloetgen et al. |
| T cell | Arima | GSE115896 | Kloetgen et al. |
| Mouse ESC | Arima | GSE140363 | Nishana et al. |

Supplementary Table 1

### Hi-C data preprocessing:

To minimize bias in preprocessing, we obtained counts data in raw fastq format. The reads from human cell lines were aligned to GRCh38 human reference genome and mouse cell lines are aligned to mm10 mouse genome. The alignments were filtered at 10kb resolution and iteratively corrected with HiC-bench (Lazaris et al., 2017). To ensure the compatibility of prediction result with downstream softwares, we only used the a reversible natural log transform to process the Hi-C prediction targets. Prediction from C.Origami with exponential transformation can be directly used as Hi-C data for any downstream analysis.

### CTCF ChIP-seq and ATAC-seq data:

All the CTCF ChIP-seq and ATAC-seq data for all cell-types are publicly available online from GEO (www.ncbi.nlm.nih.gov/geo) and ENCODE data portal (www.encodeproject.org/). CUTLL1 ATAC-seq is sequenced according to standard method (Buenrostro et al., 2015). Details on accession number are listed in Supplemental Table 2. To maintain signal consistency across different cell lines, we aggregated fastq data from different replicates and subsampled them down to 40 million reads. The reads were processed by Seq-N-Slide to generate bigWig files (https://doi.org/10.5281/zenodo.6308846). The bigWig was used as regular, dense inputs to our model. To prepare an alternative sparse input format, we used MACS2 to perform peak calling on the intermediate bam files to obtain sparse peaks for CTCF and ATAC-seq (Zhang et al., 2008). The sparse narrowPeak file was converted back to bigWig with ucscutils. We took the natural log of both dense and sparse bigWig files and used them as inputs to the model.

| Cell Type | CTCF ChIP-seq | ATAC-seq |
|---|---|---|
| IMR-90 | ENCSR000EFI | ENCSR200OML |
| GM12878 | ENCSR000AKB | ENCSR095QNB |
| H1-hESC | ENCSR000AMF | GSE85330 |
| K562 | ENCSR000AKO | ENCSR483RKN |
| CUTLL1 | GSE115893 | see Methods CTULL1 |
| T cell | GSE115893 | GSE168880 |
| Mouse ESC | GSE140363 | GSE140363 |

Supplementary Table 2

**DNA sequence**

We used the reference DNA-sequence from UCSC. The original fasta file includes four types of nucleotides and “n” for unknown type with upper- and lower-case letters which represent (repeat sequences). We retained the ‘n’ category and encoded each nucleotide as a 5 channel one-hot vector representing ATCGN. The same sequence is used for all cell types.

**Training data:**

The training data consists of DNA sequence, CTCF signal, ATAC-seq signal and Hi-C matrix on the IMR-90 cell line. The input data to the model is sequence, CTCF ChIP-seq signal, ATAC-seq signal at a 2,097,152 bp region and the output target is the Hi-C matrix at the corresponding regions. The original Hi-C matrix was originally called at 10Kb resolution and downscaled 8,192 bp to match the model output resolution.  To generate batches of training data, we defined 2Mb sliding windows across the genome with 40kb steps. Windows that have overlap with telomere or centromere were removed. We split training, validation and test set by chromosome. Chromosome 10 is used as the validation set and Chromosome 15 as the test set. The rest of the chromosomes are used as the training set.

**Model Architecture:**

The model is implemented with the PyTorch framework. Our model consists of two 1D convolutional encoders, a transformer module and a 2D convolutional decoder. To adapt to input channels of sequence and genomic features. The sequence encoder has 5 input channels, and the genomic feature encoder has 2 input channels. The two encoders have similar structures otherwise. Each encoder starts with a 1D convolution header with stride 2 to half the size of the 2m bp input before it goes to convolution blocks to reduce memory cost. To reduce the input length down to 256, we deployed 12 convolution modules each of which consists of a residual block and a scaling block. The residual block has 2 sets of convolution layers with kernel width 5 and same padding. Batch normalization and ReLU nonlinearity follows each conv layer, and the start and end position of the residual block is connected by a residual connection. Residual blocks keep the same dimension of inputs and promote information propagation. The scaling block consists of a 1D convolutional layer with kernel size 5 and stride 2 followed by batch normalization and ReLU activation. The scaling block reduces input length by a factor of 2 and increases the number of hidden layers. We increase the hidden size according to this schedule: 32, 32, 32, 32, 64, 64, 128, 128, 128, 128, 256, 256. The output from the last scaling module has length 256 with 256 channels.

The transformer module is built with 8 customized attention layers adopted from Huggingface Bert implementation(Devlin et al., 2018). Specifically, we set the number of hidden layers to 256, ReLU as the activation function and used 8 attention heads. We used relative key query as positional embedding and set the maximum length to be 256.

After the transformer module, we concatenate each position in the 256 bins to every other position to form a 256 by 256 interaction map. The concatenation function takes the 256-bin sequence from the feature extraction module and outputs a 256 by 256 grid where location (i, j) is a concatenation of the features at i and j position. Then a 1-dimensional distance matrix is calculated and appended to the grid. The distance matrix value at location (i, j) is the Manhattan Distance between point (i, i) and (j, j) on the grid divided by 2. Since each bin has 256 channels, after concatenation and addition of the distance matrix, we arrived at an output of 256 by 256 with 513 channels. The decoder consists of 5 dilated residual networks. We set the dilation factor to be 2, 4, 8, 16, 32 so that the receptive field at the last layer covers the input space. At the end of the decoder, we use a Conv2D layer with 1x1 kernel to combine 256 channels down to 1 channel and the output is a 256 by 256 matrix with one channel.

The 256x256 output from the model is compared with ground truth Hi-C map via a mean squared error (MSE) loss. The loss is back propagated through the whole network for gradient updates.

### Data augmentation

To avoid overfitting, we implemented 3 types of data augmentations. 1) During training, we dynamically selected the 2Mb window with random shifts between plus and minus 0.36 mb range. 2) We reverse complemented the sequence and flipped the target Hi-C matrix with 0.5 chance. 3) We added gaussian noise to sequence, CTCF and ATAC-seq signal with zero mean and 0.1 standard deviation.

### Model Training:

To train the model we used a training batch size of 8 and Adam optimizer with learning rate 0.002. The cosine learning rate scheduler with 200 epoch period is used for stabilizing training. The minimal validation loss is achieved when the model is trained for 54 epochs. We trained the model for 18 hours on a GPU cluster with 4 NVIDIA Tesla V100 GPUs with 320GB RAM to store training data. To prevent bottlenecking from the data loading process, we used 8 CPU workers to load data and assigned 10 CPU cores in total for the training procedure. Model inference with a mobile NVIDIA RTX 2060 GPU can be achieved in under 1 second and inference on an Intel i7-8750H CPU is around 3 seconds.

**Insulation Score:**

Insulation score is implemented as the ratio of maximum left and right region average intensity and the middle region intensity. We also added a pseudo-count calculated from chromosome wide average intensity to prevent division by zero in unmappable regions. The insulation score can be formulated as follows:

Insulation = (max(avg(Left Region), avg(Right Region)) + pseudocount) / (avg(Center Region) + pseudocount)

**Fused chromosome prediction:**

Most downstream analysis on Hi-C is conducted on Hi-C contact matrices at the level of a chromosome. To bridge the gap between our 2Mb window prediction and over 100mb chromosome, we applied window fusion to construct chromosome wide prediction from individual 2Mb predictions windows. We run the prediction in a sliding window of step side 262,144 bp which is 1/8 of the 2Mb prediction window. All predictions are in-painted to their corresponding location on the contact map. Most regions are covered by prediction for 8 times, and regions like the beginning of the chromosome are only covered for 1 time. To correct for different levels of overlap, we calculated times of overlap for every pixel and applied corresponding scaling factors. The resulting chromosome wide prediction can be directly used for downstream analysis tasks like insulation score (Supplementary Fig. 6).

**Stratified intensity and correlation**

Stratified intensity and correlation are based on fused chromosome prediction. Stratified intensity at distance i is calculated by aggregating the line that is parallel to the diagonal with offset of i. Stratified correlation is calculated as Pearson's *r* between the shifted diagonal line of prediction and ground truth.

**CUTLL1 translocation**

CUTLL1 translocation is heterozygous, and this property adds more complexity to its corresponding Hi-C matrix. Hi-C matrix is called from interactions between two genomics loci but we do not have information on which chromatid this loci is located, so there is no way to call Hi-C matrix for only the translocation. Since only one chromatid has translocation, the measured Hi-C matrix is a combination of both translocation and normal state. To align with this hybrid Hi-C map, we predicted the Hi-C map for Chr7Chr9 translocation chromatid and Chr7 and Chr9 without translocation. The interaction between Chromosome 7 and Chromosome 9 is an average of the interaction in the Chr7Chr9 in the translocated chromatid and the inter-chromosomal interaction between Chromosome 7 and Chromosome 9. We do not count the inter-chromosomal interaction because it is relatively weak compared to interaction at the translocation. The predicted interaction on Chromosome 7 until breakpoint chr7:142,797,952 is averaged with the translocated prediction. Similarly, predicted interaction on chr9 starting 136,502,817 is also averaged with translocation prediction.

### Mouse prediction

For mouse prediction, we trained the model with sparse genomic features as inputs. To obtain sparse features, we called peaks for CTCF ChIP-seq and ATAC-seq with MACS2 from the bam files generated by the Seq-N-Slide pipeline.

### In silico genetic deletion experiment

We conducted genetic screening on the 2Mb window by systematically removing segments from model inputs. We selected deletion windows of 8192 bp or 1 bin on the predicted matrix. To scan the entire region, we performed 256 deletion experiments at each bin and calculated the prediction difference map before and after deletion. Deletion reduces the input length from 2,097,152 bp to 2,088,960 bp. To maintain input shape, we appended 8192 bp of the following region.

### Reducing impact and sensitivity score from 3D voxels

Screening by deletion produces a 3D voxel with coordinates (i, j, k) where the first two dimensions (i, j) correspond to the Hi-C matrix difference and the third dimension k denotes deletion locus. Under this framework, the impact score can be defined as reducing the first two dimensions (i, j) with mean or sum, denoting the overall intensity shift with respect to deletion. The sensitivity score can be defined as the result of reducing either of the first two dimensions (i or j) and the third deletion dimension k. From another perspective, sensitivity score of a locus denotes average intensity shift over all deletions with respect to its location.

### GRAM (Gradient-weighted Regional Activate Mapping)

This scoring system is a generalized version of Grad-CAM on 2D outputs (Selvaraju et al., 2017). Instead of taking a single output, GRAM operates on a region *r* in the output space and runs backpropagation on all pixels within *r*. GRAM on region *r* in network layer *m* is defined as follows:

$$GRAM_m^r = \sum_k |\alpha_k^r||A_k^r|$$

Where $\alpha_k^r$ is the activation weight for channel *k* and region *r*, is calculated by the average gradient at the layer *m*. $A_k^r$ is the activation in channel k at layer *m*. In this study, we choose *r* to be the full output space.

### CTCF-masked mutation

For the given mutation range, we randomly change the nucleotides at all locations. The region that is under a CTCF ChIP-seq peak is kept unchanged. To accommodate the peak signal used in this task, we used the sparse model for this screening experiment.

### *In silico* genome-wide genetic screen

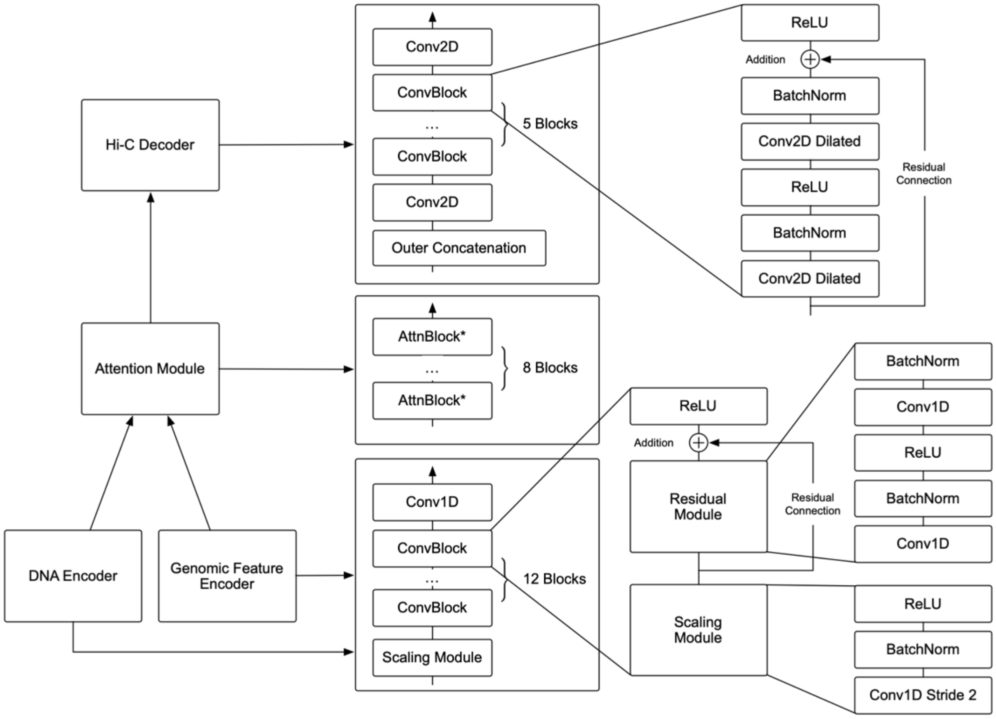
For both deletion and masked mutation, we performed saturated editing with 20Kb width and step size. Specifically, we defined a 20Kb edit region at the center of the 2Mb window. The inputs within the 20Kb region are modified and we predict the Hi-C matrix from the modified inputs. Then we measure the intensity shift of the entire 2Mb window and move to the next window which is downstream with a 20Kb offset. After whole genome screening, we obtain a genome-wide impact score for every 20Kb perturbation.

LOLA (Locus Overlap Analysis) takes a genomic region set and compares it to a set of core databases and calculates enrichment score for every feature in the database (Sheffield and Bock, 2016). The enrichment score is calculated with fisher's exact test on a contingency table. The two sets of conditions of the contingency table are defined as present/absent and query/database. The query region is the genomic region we are testing and database regions are from a target database feature that we are comparing against. LOLA also requires a universe set which we choose to be the whole genome with 20Kb widths.

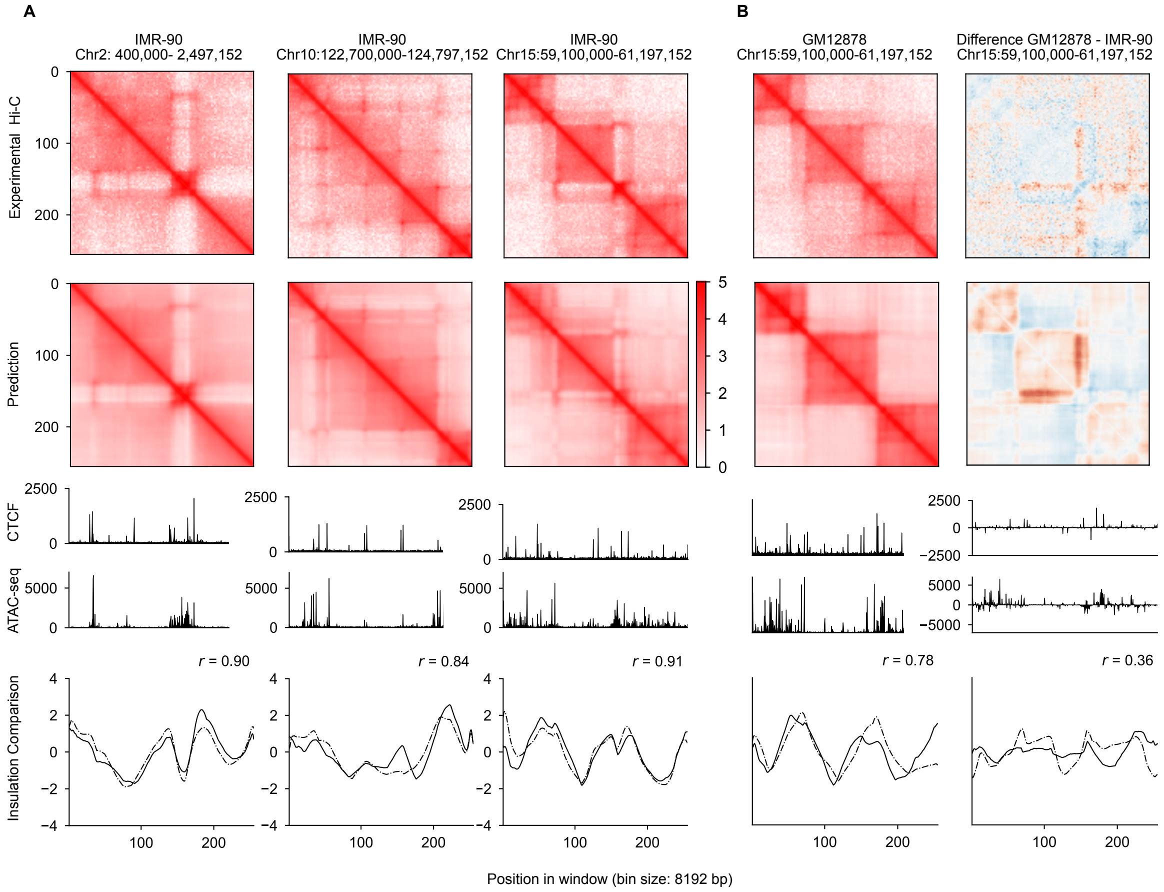
To generate a set of genomic regions from our impact score, we choose a sliding window of size 2Mb and step 20Kb across the genome and aggregate the region with the highest impact scores. These regions are then merged to continuous regions and formatted to a bed file as input (query set in LOLA) to LOLA. The background input (universe set in LOLA) to LOLA is selected as the entire genome with offsets of 20kb. Since high impact can be either positive and negative, we also generated regions with lowest impact scores and tested its enrichment.

The output from LOLA is processed by merging and filtering different features. For features with the same antibody name, only the highest ranked one was kept for analysis. Features without antibody name are removed. Then we filtered out the features with odds ratio less than 2 in all four categories: deletion postive/negative and mutation positive/negative. We collected 191 relevant factors and ranked them according to by a weighted score defined as min-max normalized -log10(q-value). We then visualized the relationship between different transcription factors with heatmaps and hierarchical clustering.

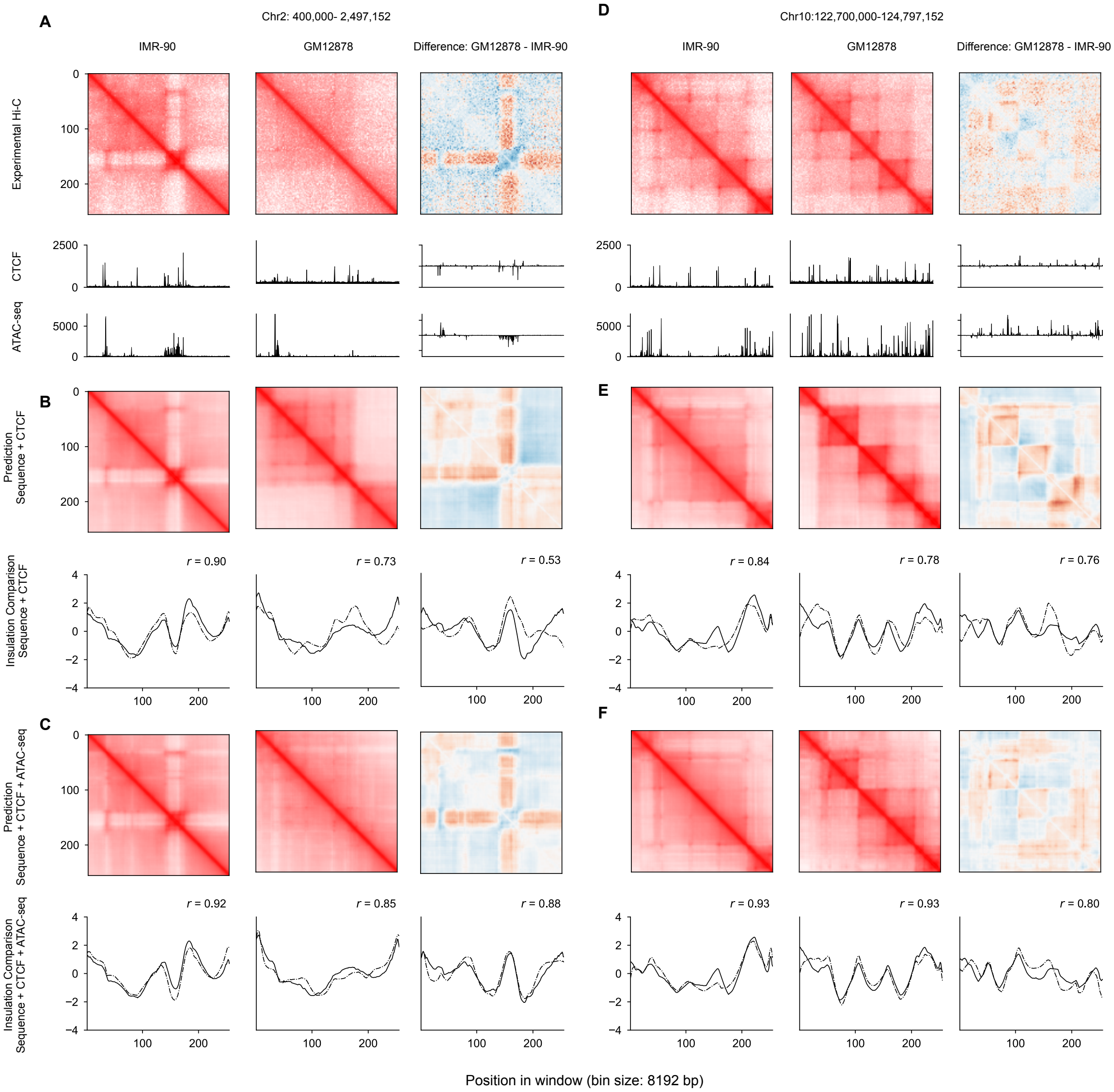
## Supplementary Figures:



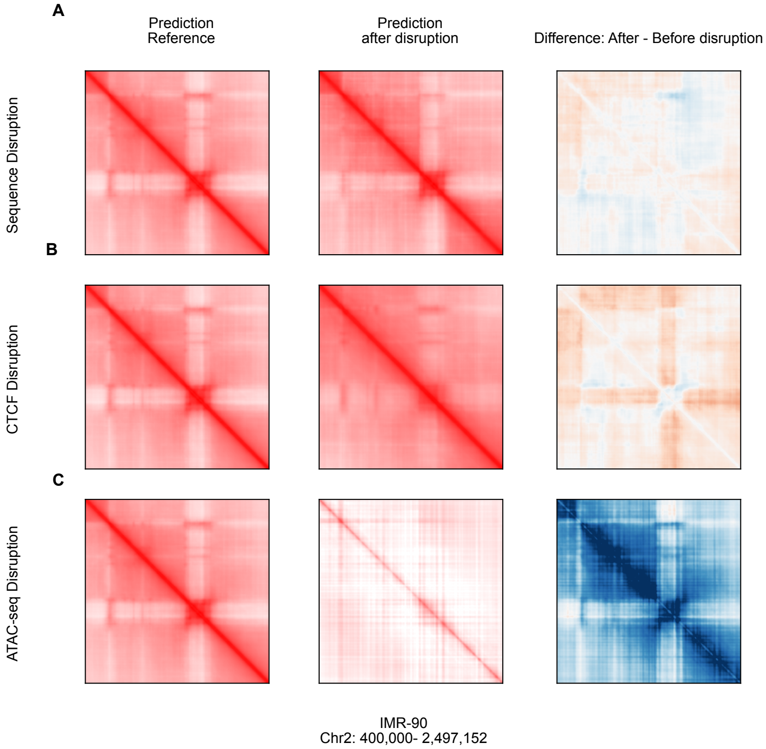
**Supplementary Figure 1: C.Origami model structure and module components.** A detailed schematic of C.Origami model architecture. The DNA encoder and Genomic Feature encoder have similar architectures and they only different in input channels where DNA encoder has 5 and Feature encoder has 2. To encoder data, we built the encoder with 12 convolution blocks, each consisting of a scaling module and residual module. The scaling module downscales input features by a factor of 2 with a stride-2 1D convolution layer. The residual module promotes information propagation in very deep networks (REF Deep Residual Learning for Image Recognition). The number of modules was carefully chosen such that we scale the 2,097,152 input down to 256 bins at the end of the encoder. To enhance interactions within the 2Mb window, we used an attention module that consists of 8 attention blocks modified from the transformer architecture. Each position of the output is concatenated with every other position to form a 2D matrix, resembling a vector outer-product process. To refine the final prediction, we used a 5-layer dilated 2D convolutional network as decoder. We deliberately chose the dilation parameters to ensure that every position at the last layer has a receptive field covering the input range.



**Supplementary Figure 2: Performance of C.Origami trained with DNA sequence and CTCF binding profiles. a**, Predicting chromatin architecture using a model trained with DNA sequence and CTCF binding profiles. The plots were organized the same as Fig. 2 **a-d**. **b**, *De novo* predicting chromatin architecture of the chromosome 15 locus in GM12878 using the model trained with DNA sequence and CTCF binding profiles. The difference between IMR-90 and GM12878 were presented on the right. While C.Origami trained with DNA sequence and CTCF profile achieved good performance in validation and test set in IMR-90 (**a**), it missed predicting some fine-scale chromatin structures in GM12878.

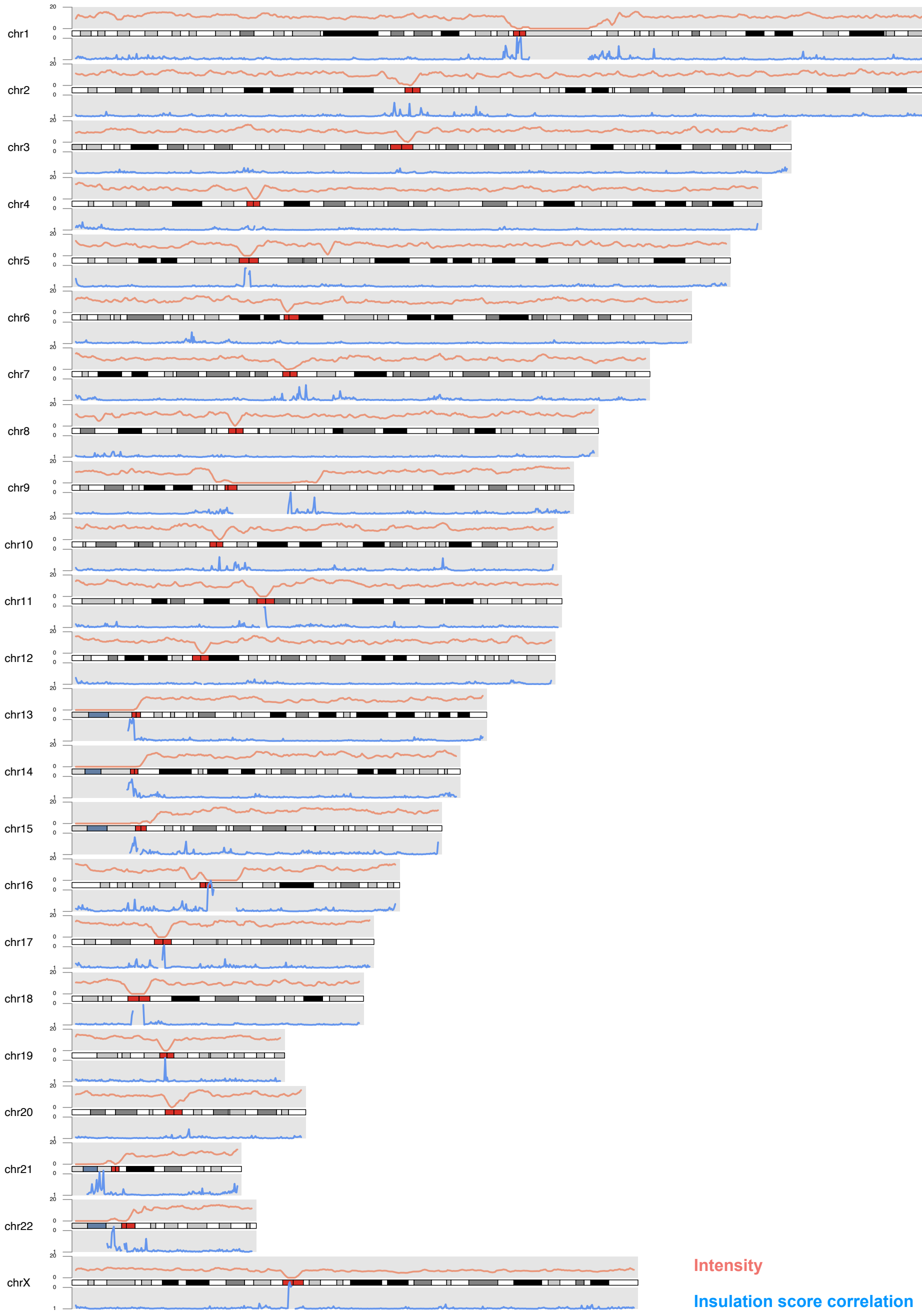


**Supplementary Figure 3: C.Origami trained with DNA sequence, CTCF binding, and chromatin accessibility profiles performed optimally. a**, Experimental Hi-C matrices and genomic profiles of IMR-90 and GM12878 cells at chr2:400,000-2,497,152. The difference between the two cell lines were presented on the right. **b-c**, Cell type-specific prediction of the chromatin architecture at the same locus using C.Origami models trains with DNA sequence and CTCF binding (**b**) or DNA sequence, CTCF binding, and chromatin accessibility profiles (**c**). **d-e**, Same as **a-c** at a difference locus, chr10:122,700,000-122,797,152.

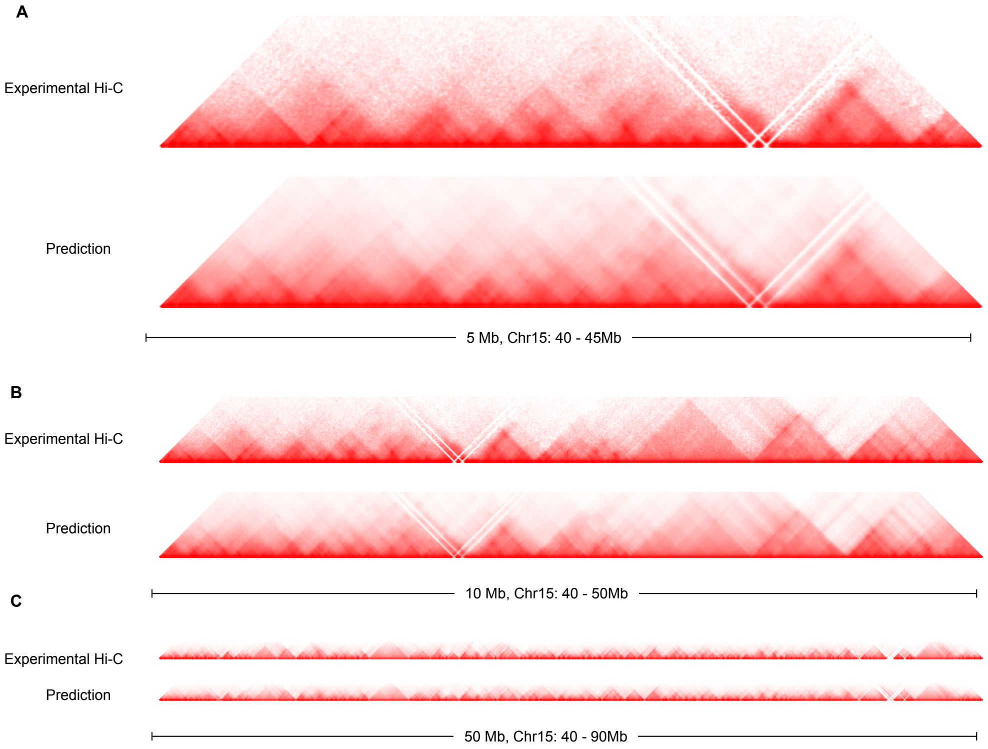


809

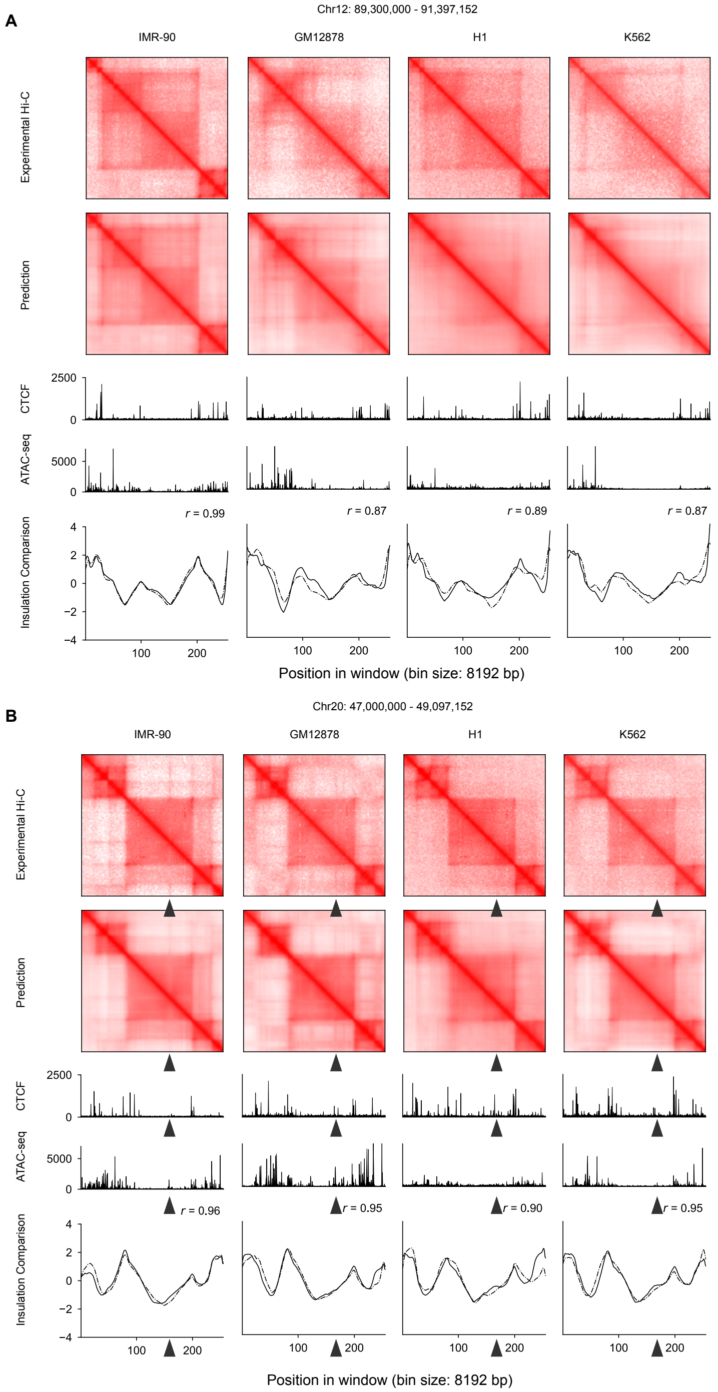
810 **Supplementary Figure 4: Ablation study on different input features.** Using the C.Origami model trained
811 with DNA sequence, CTCF binding, and chromatin accessibility profiles, the experiments was performed
812 by random shuffling DNA sequences at base pair level (**a**), random shuffling CTCF signal (**b**), and random
813 shuffling ATAC-seq signal (**c**). From left to right, reference prediction with all inputs (left), prediction with
814 sequence shuffled (middle), difference between perturbed prediction and reference prediction (right).



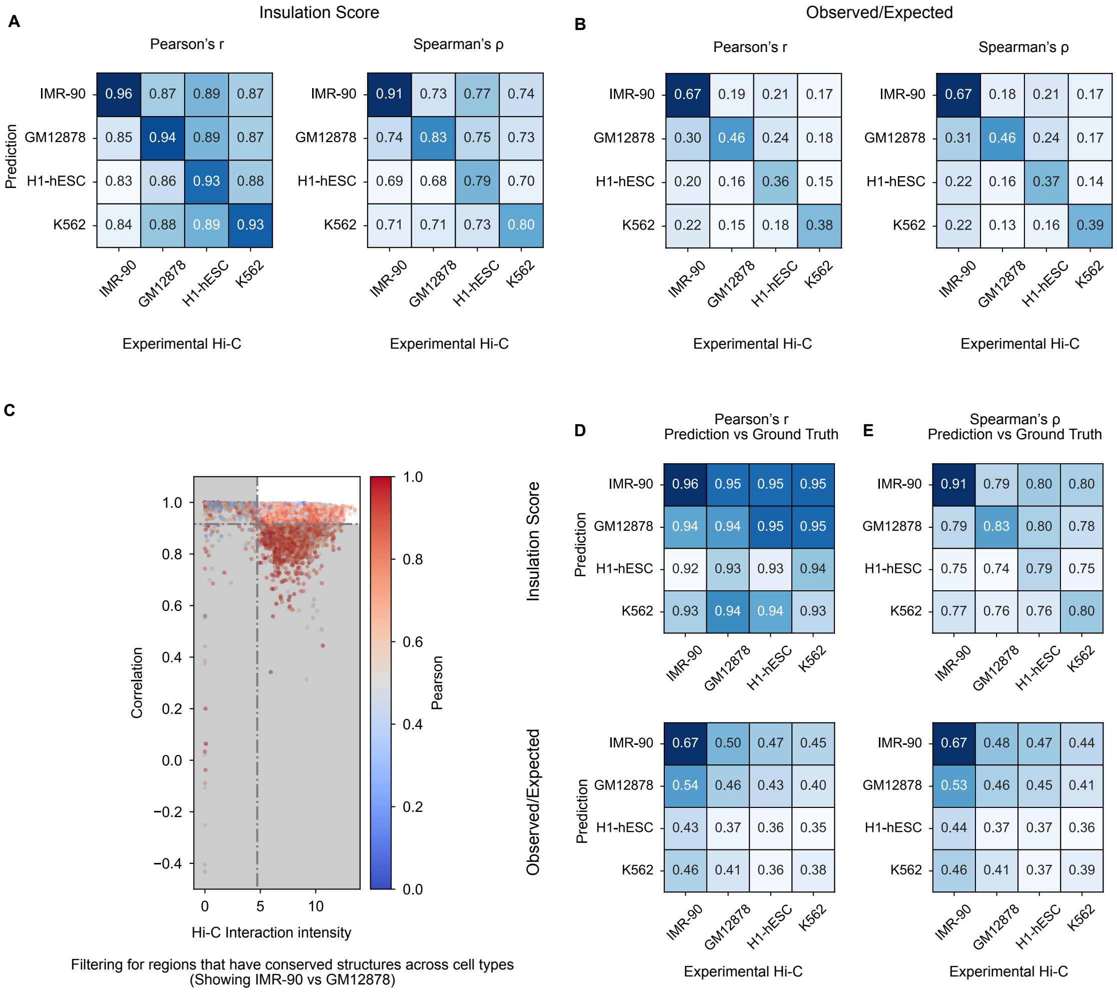
**Supplementary Figure 5: Chromosome karyotype visualization along with chromosome-wide Hi-C intensity and correlation of insulation scores.** The results were visualized using karyoploteR (Gel and Serra, 2017). Chromosome 1 to chromosome X were plotted to visualize the Pearson correlation coefficients of insulation scores calculated from prediction and that from experimental Hi-C. Average intensity of 2Mb windows were plotted in red. Centromere regions were denoted with red segments on the genome.



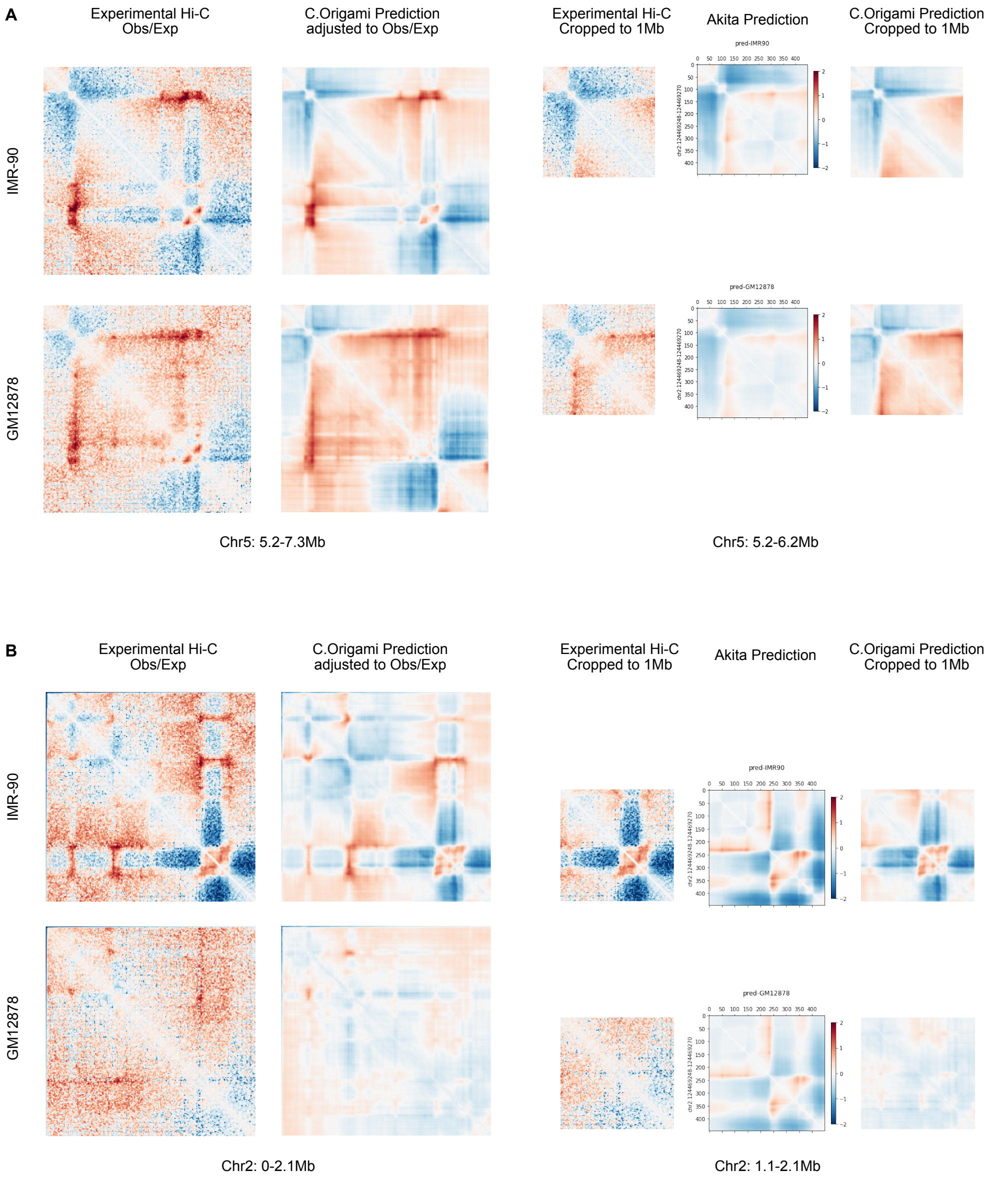
**Supplementary Figure 6: Fusing C.Origami-predicted 2Mb Hi-C maps into larger interaction maps.** The predicted 2Mb Hi-C maps were fused to 5Mb (**a**), 10Mb (**b**), and 50Mb (**c**) on chromosome 15, all with the same starting site at 40 Mb.



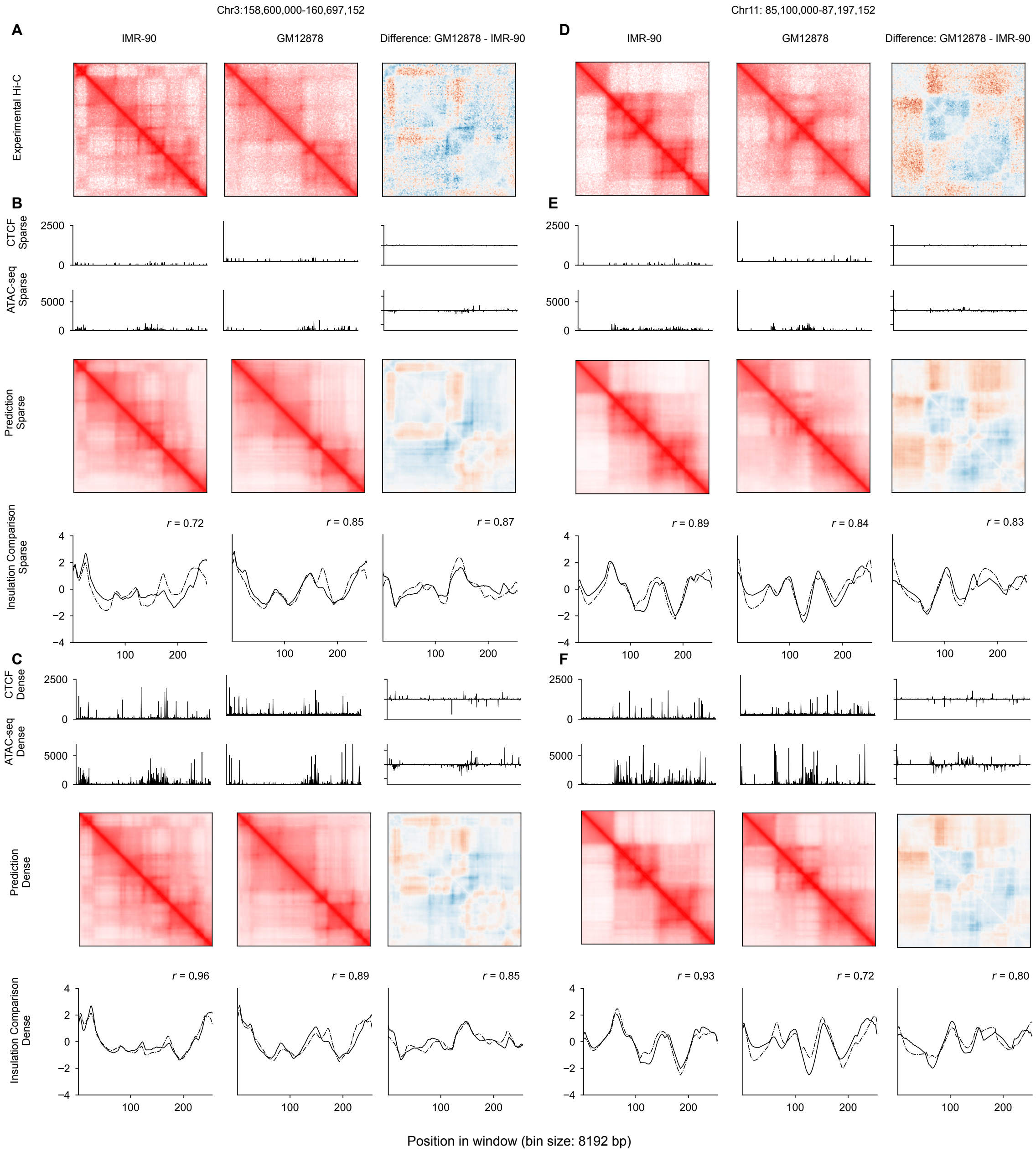
**Supplementary Figure 7: C.Origami predicts chromatin architectures across multiple cell types.** Two representative loci were separately presented across IMR-90, GM12878, H1-hESCs, and K562 in **a** and **b**. From top to bottom, each panel included experimental Hi-C matrix, predicted Hi-C matrix, CTCF and ATAC-seq signals, and insulation scores calculated from experimental and predicted Hi-C data.



**Supplementary Figure 8: Genome-wide statistics on cell type-specific prediction performance. a-b**, Pearson's *r* (left) and Spearman's *ρ* (right) between prediction (row) and experimental data (column) for different cell types with insulation score (**a**) and observed/expected score (**b**) as metrics. The scores were calculated based on the differentially structured loci defined in Fig. 3. The correlation between Observed/Expected contact matrices was lower due to higher background noise. **c**, selecting structurally conserved loci across different cell types. Conserved subset accounts for ~60% of the data. **d-e**, Same as **a-b** but for the structurally conserved loci across different cell types.



**Supplementary Figure 9: Comparing the performance of C.Origami with Akita in cell-type specific prediction.** Two represented loci were presented (**a-b**). Each locus includes the experimental Hi-C matrix together with the C.Origami prediction in IMR-90 cells and GM12878 cells (lef). Akita predicted chromatin architectures in windows of 1Mb, thus fractioned Hi-C matrices were presented on the right for comparison.



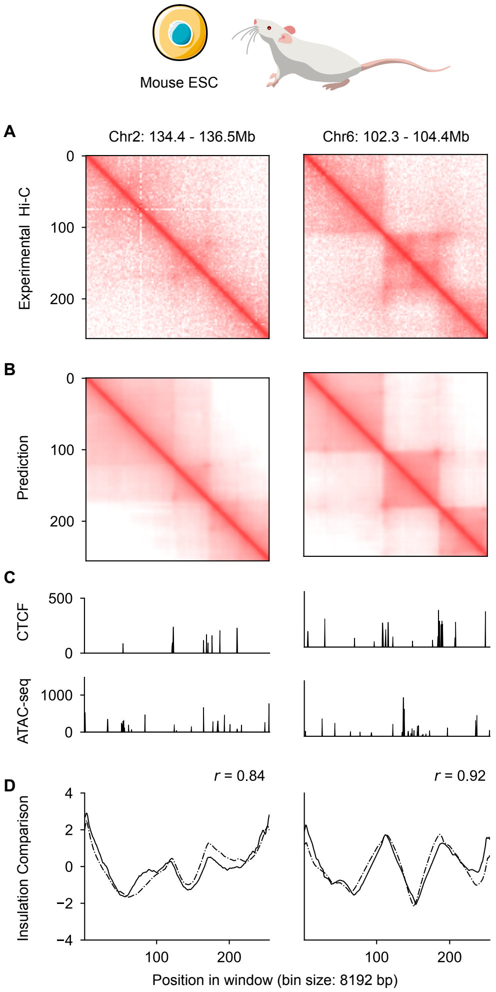
855

856 **Supplementary Figure 10: Performance comparison of C.Origami models trained with sparse information and dense information. a**, Experimental Hi-C matrices and genomic profiles of IMR-90 and GM12878 cells at chr3: 158,600,000-160,697,152. The difference between the two cell lines were presented on the right. **b-c**, Cell type-specific prediction of the chromatin architecture at the same locus using C.Origami models trains  with sparse genomic information (**b**) or dense genomic information (**c**). **d-e**, Same as **a-c** at a difference locus, chr10: 85,100,000-87,197,152.

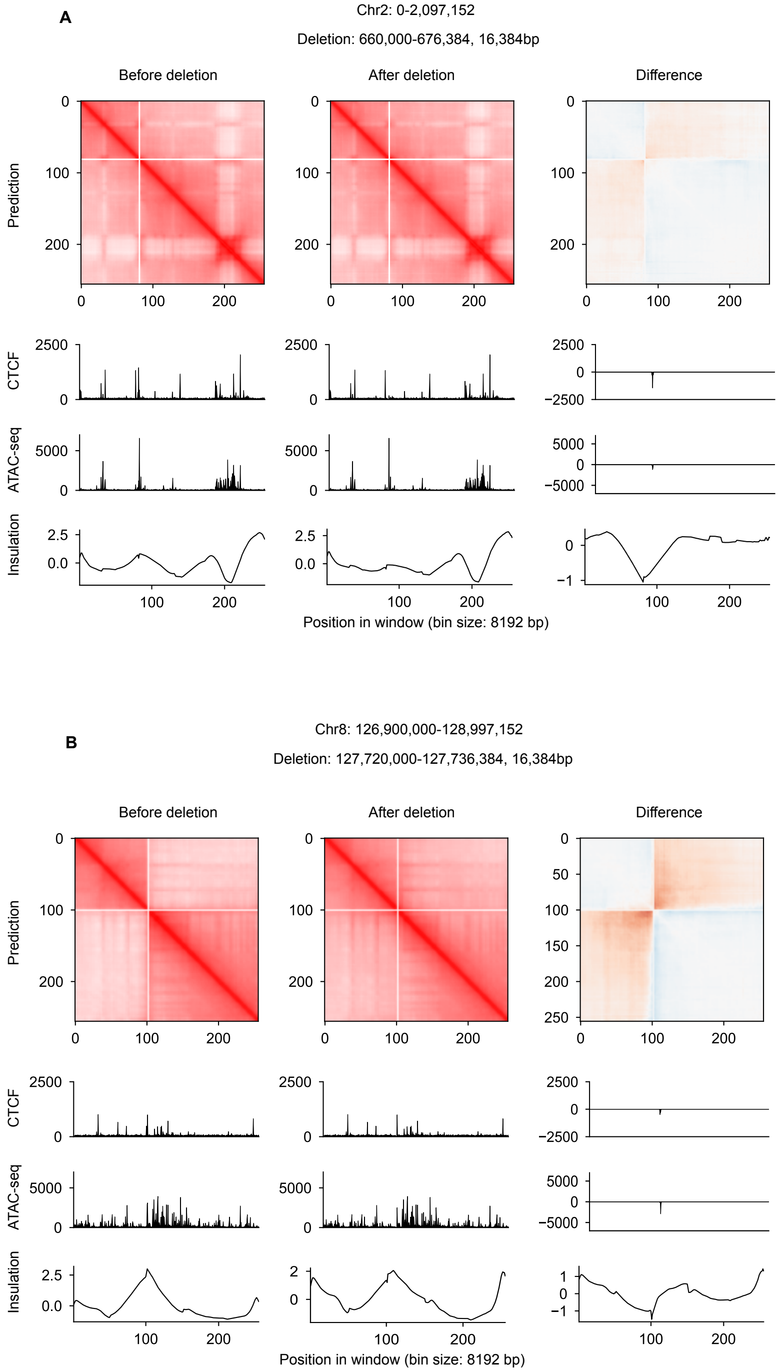
862

863

864



**Supplementary Figure 11: Mouse chromatin architecture prediction using C.Origami trained with human data.** Experimental Hi-C matrices (**a**), predicted Hi-C matrices (**b**), CTCF and ATAC-seq signals (**c**), and insulation scores calculated from experimental and predicted Hi-C data (**d**) were presented from top to bottom, each with two different loci.



**Supplementary Figure 12: *In silico* genetic experiments performed on IMR-90 cells.** Two *in silico* deletion experiments were separately represented in **a** and **b**. Each experiment included the prediction before (left) and after deletion (middle). The difference in chromatin folding after deletion were presented on the right.

## References

Beagan, J.A., Pastuzyn, E.D., Fernandez, L.R., Guo, M.H., Feng, K., Titus, K.R., Chandrashekar, H., Shepherd, J.D., and Phillips-Cremins, J.E. (2020). Three-dimensional genome restructuring across timescales of activity-induced neuronal gene expression. Nat. Neurosci. *23*, 707–717.

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.

Bianco, S., Lupiáñez, D.G., Chiariello, A.M., Annunziatella, C., Kraft, K., Schöpflin, R., Wittler, L., Andrey, G., Vingron, M., Pombo, A., et al. (2018). Polymer physics predicts the effects of structural variants on chromatin architecture. Nat. Genet. *50*, 662–667.

Buenrostro, J.D., Wu, B., Chang, H.Y., and Greenleaf, W.J. (2015). ATAC-seq: a method for assaying chromatin accessibility genome-wide. Curr. Protoc. Mol. Biol. *109*, 21.29.1–21.29.9.

Cao, F., Zhang, Y., Cai, Y., Animesh, S., Zhang, Y., Akincilar, S.C., Loh, Y.P., Li, X., Chng, W.J., Tergaonkar, V., et al. (2021). Chromatin interaction neural network (ChINN): a machine learning-based method for predicting chromatin interactions from DNA sequences. Genome Biol. *22*, 226.

Cheng, Y., Ma, Z., Kim, B.-H., Wu, W., Cayting, P., Boyle, A.P., Sundaram, V., Xing, X., Dogan, N., Li, J., et al. (2014). Principles of regulatory information conservation between mouse and human. Nature *515*, 371–375.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F.J. (2019). Deep learning: new computational modelling techniques for genomics. Nat. Rev. Genet. *20*, 389–403.

Feng, Y., Wang, Y., Wang, X., He, X., Yang, C., Naseri, A., Pederson, T., Zheng, J., Zhang, S., Xiao, X., et al. (2020). Simultaneous epigenetic perturbation and genome imaging reveal distinct roles of H3K9me3 in chromatin architecture and transcription. Genome Biol. *21*, 296.

Forcato, M., Nicoletti, C., Pal, K., Livi, C.M., Ferrari, F., and Bicciato, S. (2017). Comparison of computational methods for Hi-C data analysis. Nat. Methods *14*, 679–685.

Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W.-L., et al. (2016). Formation of new chromatin domains determines pathogenicity of genomic duplications. Nature *538*, 265–269.

Fudenberg, G., Kelley, D.R., and Pollard, K.S. (2020). Predicting 3D genome folding from DNA sequence with Akita. Nat. Methods *17*, 1111–1117.

Gel, B., and Serra, E. (2017). karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics *33*, 3088–3090.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kloetgen, A., Thandapani, P., Ntziachristos, P., Ghebrechristos, Y., Nomikou, S., Lazaris, C., Chen, X., Hu, H., Bakogianni, S., Wang, J., et al. (2020). Three-dimensional chromatin landscapes in T cell acute lymphoblastic leukemia. Nat. Genet. *52*, 388–400.

Lazaris, C., Kelly, S., Ntziachristos, P., Aifantis, I., and Tsirigos, A. (2017). HiC-bench: comprehensive and reproducible Hi-C data analysis designed for parameter exploration and benchmarking. BMC Genomics *18*, 22.

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. *12*, 1725–1735.

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.

Lu, L., Liu, X., Huang, W.-K., Giusti-Rodríguez, P., Cui, J., Zhang, S., Xu, W., Wen, Z., Ma, S., Rosen, J.D., et al. (2020). Robust Hi-C Maps of Enhancer-Promoter Interactions Reveal the Function of Non-coding Genome in Neural Development and Diseases. Mol. Cell *79*, 521–534.e15.

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.

Narendra, V., Rocha, P.P., An, D., Raviram, R., Skok, J.A., Mazzoni, E.O., and Reinberg, D. (2015). CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. Science *347*, 1017–1021.

Palomero, T., Barnes, K.C., Real, P.J., Glade Bender, J.L., Sulis, M.L., Murty, V.V., Colovai, A.I., Balbin, M., and Ferrando, A.A. (2006). CUTLL1, a novel human T-cell lymphoma cell line with t(7;9) rearrangement, aberrant NOTCH1 activation and high sensitivity to gamma-secretase inhibitors. Leukemia *20*, 1279–1287.

Petrovic, J., Zhou, Y., Fasolino, M., Goldman, N., Schwartz, G.W., Mumbach, M.R., Nguyen, S.C., Rome, K.S., Sela, Y., Zapataro, Z., et al. (2019). Oncogenic Notch Promotes Long-Range Regulatory Interactions within Hyperconnected 3D Cliques. Mol. Cell *73*, 1174–1190.e12.

Phillips-Cremins, J.E., Sauria, M.E.G., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S.K., Ong, C.-T., Hookway, T.A., Guo, C., Sun, Y., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell *153*, 1281–1295.

Di Pierro, M., Cheng, R.R., Lieberman Aiden, E., Wolynes, P.G., and Onuchic, J.N. (2017). De novo prediction of human chromosome structures: Epigenetic marking patterns encode genome architecture. Proc. Natl. Acad. Sci. USA *114*, 12126–12131.

Pinglay, S., Bulajic, M., Rahe, D.P., Huang, E., Brosh, R., German, S., Cadley, J.A., Rieber, L., Easo, N., Mahony, S., et al. (2021). Synthetic genomic reconstitution reveals principles of mammalian Hox cluster regulation. BioRxiv.

Qi, Y., and Zhang, B. (2019). Predicting three-dimensional genome organization with chromatin states. PLoS Comput. Biol. *15*, e1007024.

Rabbitts, T.H. (1994). Chromosomal translocations in human cancer. Nature *372*, 143–149.

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

Rowley, M.J., and Corces, V.G. (2018). Organizational principles of 3D genome architecture. Nat. Rev. Genet. *19*, 789–800.

Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., Barr, C.L., et al. (2016). A compendium of chromatin contact maps reveals spatially active regions in the human genome. Cell Rep. *17*, 2042–2059.

Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer-promoter contacts in gene expression control. Nat. Rev. Genet. *20*, 437–455.

Schwessinger, R., Gosden, M., Downes, D., Brown, R.C., Oudelaar, A.M., Telenius, J., Teh, Y.W., Lunter, G., and Hughes, J.R. (2020). DeepC: predicting 3D genome folding using megabase-scale transfer learning. Nat. Methods *17*, 1118–1124.

Selvaraju, R.R., Cogswell, M., and Das, A. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the ….

Sheffield, N.C., and Bock, C. (2016). LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics *32*, 587–589.

Spielmann, M., Lupiáñez, D.G., and Mundlos, S. (2018). Structural variation in the 3D genome. Nat. Rev. Genet. *19*, 453–467.

Stergachis, A.B., Neph, S., Sandstrom, R., Haugen, E., Reynolds, A.P., Zhang, M., Byron, R., Canfield, T., Stelhing-Sun, S., Lee, K., et al. (2014). Conservation of trans-acting circuitry during mammalian regulatory evolution. Nature *515*, 365–370.

Szabo, Q., Jost, D., Chang, J.-M., Cattoni, D.I., Papadopoulos, G.L., Bonev, B., Sexton, T., Gurgo, J., Jacquier, C., Nollmann, M., et al. (2018). TADs are 3D structural units of higher-order chromosome organization in Drosophila. Sci. Adv. *4*, eaar8082.

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.

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.

Vaswani, A., Shazeer, N., and Parmar, N. (2017). Attention is all you need. … neural information ….

Weintraub, A.S., Li, C.H., Zamudio, A.V., Sigova, A.A., Hannett, N.M., Day, D.S., Abraham, B.J., Cohen, M.A., Nabet, B., Buckley, D.L., et al. (2017). YY1 Is a Structural Regulator of Enhancer-Promoter Loops. Cell *171*, 1573–1588.e28.

Wu, D., Sunkel, B., Chen, Z., Liu, X., Ye, Z., Li, Q., Grenade, C., Ke, J., Zhang, C., Chen, H., et al. (2014). Three-tiered role of the pioneer factor GATA2 in promoting androgen-dependent gene expression in prostate cancer. Nucleic Acids Res. *42*, 3607–3622.

Yang, R., Das, A., Gao, V.R., Karbalayghareh, A., Noble, W.S., Bilmes, J.A., and Leslie, C.S. (2021). Epiphany: predicting Hi-C contact maps from 1D epigenomic signals. BioRxiv.

Zhang, S., Chasman, D., Knaack, S., and Roy, S. (2019). In silico prediction of high-resolution Hi-C interaction matrices. Nat. Commun. *10*, 5449.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhao, L., Wang, S., Cao, Z., Ouyang, W., Zhang, Q., Xie, L., Zheng, R., Guo, M., Ma, M., Hu, Z., et al. (2019). Chromatin loops associated with active genes and heterochromatin shape rice genome architecture for transcriptional regulation. Nat. Commun. *10*, 3640.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A primer on deep learning in genomics. Nat. Genet. *51*, 12–18.