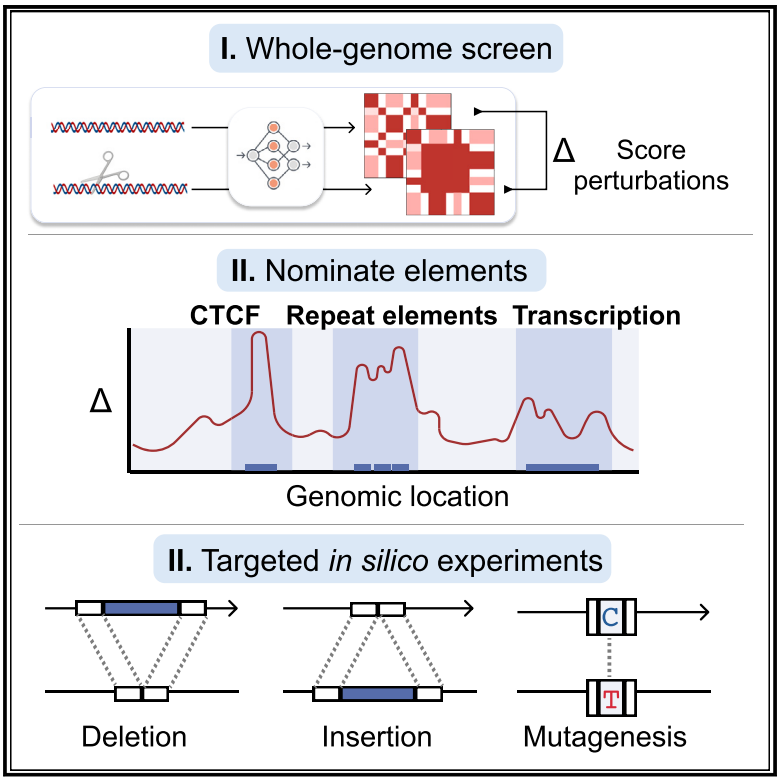
Cell Genomics

Article

In silico discovery of repetitive elements as key sequence determinants of 3D genome folding

Graphical abstract



Highlights

- Mass, unbiased computational screen reveals elements correlated with genome folding
- Paired deletion/insertion experiments disentangle necessary and sufficient sequences
- High-scoring sequences include repeats (Alu, MIR) and RNA genes (tRNA, snRNA)
- Genome folding is also sensitive to CTCF motifs, GC content, and transcription

Authors

Laura M. Gunsalus, Michael J. Keiser, Katherine S. Pollard

Correspondence

katherine.pollard@gladstone.ucsf.edu

In brief

Gunsalus et al. use a deep learning model to screen the human genome for regions where sequence changes have particularly large predicted effects on 3D genome folding. They find that sequence perturbations in CTCF motifs, actively transcribed regions, and Alu, hAT-Charlie, and SVA repeats profoundly influence chromatin interactions. Targeted computational experiments reveal that repetitive elements, sometimes lacking CTCF motifs, provide sequence grammar governing chromatin interactions. This unbiased approach implicates specific repeat families as integral to genome folding.

Gunsalus et al., 2023, Cell Genomics *3*, 100410
October 11, 2023
https://doi.org/10.1016/j.xgen.2023.100410

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

In silico discovery of repetitive elements as key sequence determinants of 3D genome folding

Laura M. Gunsalus,[1,2] Michael J. Keiser,[2,3,4,5,6] and Katherine S. Pollard[1,3,7,8,9,*]
[1]Gladstone Institutes, San Francisco, CA, USA
[2]Institute for Neurodegenerative Diseases, University of California, San Francisco, San Francisco, CA, USA
[3]Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, USA
[4]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, USA
[5]Kavli Institute for Fundamental Neuroscience, University of California, San Francisco, San Francisco, CA, USA
[6]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA, USA
[7]Chan Zuckerberg Biohub, San Francisco, CA, USA
[8]Department of Epidemiology & Biostatistics, University of California, San Francisco, San Francisco, CA, USA
[9]Lead contact
*Correspondence: katherine.pollard@gladstone.ucsf.edu
https://doi.org/10.1016/j.xgen.2023.100410

SUMMARY

Natural and experimental genetic variants can modify DNA loops and insulating boundaries to tune transcription, but it is unknown how sequence perturbations affect chromatin organization genome wide. We developed a deep-learning strategy to quantify the effect of any insertion, deletion, or substitution on chromatin contacts and systematically scored millions of synthetic variants. While most genetic manipulations have little impact, regions with CTCF motifs and active transcription are highly sensitive, as expected. Our unbiased screen and subsequent targeted experiments also point to noncoding RNA genes and several families of repetitive elements as CTCF-motif-free DNA sequences with particularly large effects on nearby chromatin interactions, sometimes exceeding the effects of CTCF sites and explaining interactions that lack CTCF. We anticipate that our disruption tracks may be of broad interest and utility as a measure of 3D genome sensitivity, and our computational strategies may serve as a template for biological inquiry with deep learning.

INTRODUCTION

The human genome gives rise to its own organization in the nucleus, where the folding of chromatin into intricate and hierarchical structures can be reflective and instructive of cell state.[1] Sequence itself contains the information to create some chromatin features. Binding of CTCF proteins to DNA motifs blocks the extrusion of DNA by motor proteins to create topologically associating domains (TADs) spanning hundreds of megabases.[2–5] These dynamic structures permit interaction of elements within their boundaries and limit interaction with elements outside to tune gene expression.[6,7] However, recent reports reveal that CTCF may not be the only factor involved because some contacts remain after CTCF depletion, and interactions across megabases are not affected.[8,9] How exactly sequence informs structure ranging from the highest levels of genome organization—chromosome territories and compartments—to the level of individual enhancer-promoter interactions still remains unclear.

Current approaches relating genome sequence to folding either leverage natural genetic variation or experimentally manipulate particular loci to test specific hypotheses. Applying chromatin capture to genetically diverse individuals has revealed single nucleotide variants associated with loss or gain of chromatin contact.[10] Large structural variants are also rare at domain boundaries in healthy humans but not in patients with autism or developmental delay.[11] To understand the mechanisms underlying these associations, experimental studies have engineered chromatin contact in cells and mice with synthetic tethering[12] and CRISPR systems[13–15] and measured their effects on genome folding and expression of genes such as *Hbb* and *Vcan*. Findings in these individual loci may not apply genome wide and could overlook mechanisms without known precedent. Here, we propose combining the genome-wide power of population genetics with the precision seen in experimental studies. We develop a strategy that leverages deep learning to comprehensively screen the human genome for key regulators of 3D genome folding.

Whereas previous machine learning approaches required domain experts to select the most relevant features, deep learning allows patterns to be learned directly from the data without expert input. Deep learning models perform well in predicting enhancer activity,[16,17] transcription factor binding,[18] gene expression,[19] and genome folding[20,21] from sequence, with newer models increasing scale and incorporating chromatin immunoprecipitation sequencing (ChIP-seq) and ATAC-seq to provide cell-type-specific context.[22–24] The premise for our study is that we can probe these models as computational oracles to predict the behavior of DNA sequence at scales intractable experimentally.[25] Models have been applied to predict

the impact of structural variants on human genome folding,[20,24] confirm the importance of CTCF through computational mutagenesis,[20] and resurrect the folding of Neanderthal genomes.[26] These early reports show that many highly disruptive perturbations lack CTCF or annotated regulatory elements, hinting that there may be sequences left to uncover that encode information needed for genome folding.

Here, we leverage Akita,[20] a convolutional neural network trained to predict genome folding from sequence, to perform unbiased and targeted *in silico* mutagenesis experiments at scale. Applying this approach to a human foreskin fibroblast cell line (HFFc6) with high-resolution micro-C data for model training, we discovered wide variability in how robust genome folding is to sequence perturbations. Investigation of sensitive loci revealed known motifs, like CTCF, and understudied modulators of 3D genome folding, including transposon and RNA gene clusters. These findings were replicated in a human embryonic stem cell line (H1hESC) and supported by experimental Hi-C in loci with human-specific repetitive elements. Thus, our genome-wide screen revealed a diverse vocabulary of DNA elements that collaborate with CTCF to orchestrate TAD-scale chromatin organization.

RESULTS

Genome-wide deletion screen reveals high variability in 3D genome folding

To measure sequence importance to chromatin organization, we developed a deep learning scoring strategy to computationally introduce modifications into the human reference genome and predict their impact on genome folding with Akita.[20] Given an ~1-megabase (Mb) DNA sequence, this model accurately produces a chromatin contact map at ~2-kilobase (kb) resolution, where TADs and DNA loops are visible. Akita has been used previously to perform sequence mutagenesis experiments ranging from one nucleotide to thousands of base pairs.[20,26] To build a flexible *in silico* screening strategy based on Akita, we wrote computationally efficient code that quantifies the impact of a centered sequence variant, which we call disruption, as the log-mean-squared difference between the predicted contact frequency map for the 1-Mb sequence with a sequence alteration compared with that of the reference sequence. If a variant dramatically rearranges how the genome is predicted to fold, then we infer that the altered sequence could regulate chromatin contacts.

In this study, we used disruption scores to perform a variety of genome-wide screens across millions of genetic perturbations, including targeted and unbiased deletions, insertions, and substitutions ranging from 1–500,000 bp (Figure 1A). In contrast to *in vivo* genetic perturbations, our approach enables precise and flexible genome editing at scale. We first assessed all 5-kb deletions tiled across the genome for their impact on folding in HFFc6 cells (n = 574,187). Deletions are highly variable, and around half produce changes to chromatin contact maps that are noticeable by eye (Figure 1B). Some sequence deletions completely rearrange the boundary structure of contact maps, some result in small focal changes (e.g., gain or loss of a loop anchor), and some produce no change at all, suggesting that the chromatin structure is robust to sequence manipulation (Figure 1C). As expected, regions of the genome with many CTCF motifs are particularly sensitive, while regions with no motifs are perturbation resilient (Figure 1D), establishing that our approach identifies known genome folding mechanisms.

Perturbing euchromatin disrupts genome folding

Disruption scores are also correlated with the chromatin compartment, as measured by the first eigenvector of the experimental HFFc6 micro-C contact matrix (Pearson's r = 0.522, $p < 1 \times 10^{-300}$, n = 11,413; Figure 1E).[27] The mean disruption score within gene-rich and open A compartments is 14.6% higher than in compact, inactive B compartments. Motivated by existing work illustrating that gene-rich GC-rich regions fall in A compartments, while GC-poor regions, like lamina-associated domains, are known to self-interact with each other and other GC-poor regions across chromosomes, we next directly evaluated the role of GC content in disrupting genome folding.[28] We observe that high gene density and GC content are associated with peaks in disruption scores (Figures 1D and S1A–S1C). Using HFFc6 total RNA sequencing (RNA-seq),[29] we quantified transcription in each 5-kb window and observed a strong correlation with disruption scores (Pearson's r = 0.366, $p < 1 \times 10^{-300}$, n = 11,413). Other genomic features associated with active chromatin are also more frequent in the most sensitive sequences, including distal and proximal enhancers and promoters (Figures 2A, S1D, and S1E). 62.1% of the most sensitive sequences (top decile of scores) fall within 5 kb of CTCF-bound distal enhancers compared with only 7.3% of the most robust sequences (bottom decile of scores) (Figure 2A). In sum, it is difficult to perturb inactive chromatin and easy to perturb active chromatin.

The correlation between many of these features reflects an inherent challenge in disentangling which are causal and which are reflective of genome folding (Figure S1C). Indeed, regions that are in A compartments, contain CTCF binding sites, and are actively transcribed are also the most sensitive (Figure 2B). Some elements are overrepresented in A compartments, but the effect of CTCF nonetheless holds in A and B compartments (Figures S2A–S2C), indicating that it is directly associated with sensitive 5-kb bins and not just a proxy for A compartments. However, transcription and compartment are more impactful individually than the presence of CTCF motifs, suggesting that additional rules govern which CTCF sites are in use and which are redundant or decommissioned in a given cell type. Overall, our findings suggest that independent mechanisms at transcriptionally active sites may collaborate to coordinate genome folding.

Transcriptionally active regions modulate folding alongside CTCF

Chromatin contact and transcription are correlated, but which mechanistically precedes the other is currently an area of active investigation. While transcription is classically thought to result from enhancer-promoter interaction constrained by chromatin structure, transcriptional machinery may help to scaffold local chromatin structure as well.[30] CTCF binding, for example, is essential for boundary formation and may also influence the

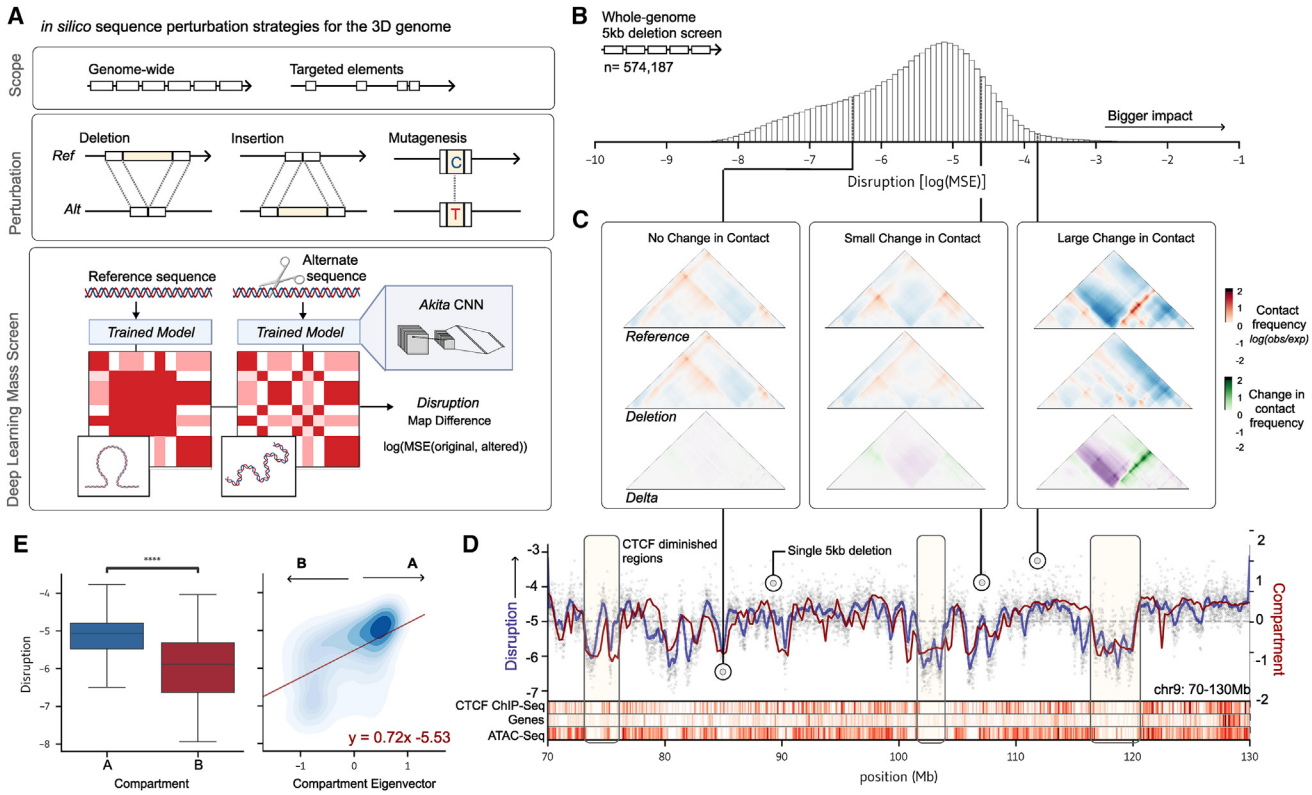


Figure 1. *In silico* deletion screen indicates that the impact of sequence perturbation on 3D genome folding is highly variable

(A) We quantify how important DNA sequence is to genome folding by introducing whole-genome and targeted deletions, insertions, and point mutations and comparing the predicted Hi-C contact maps with maps predicted from the reference sequence. We score disruption as the log-mean-squared difference of the perturbed map relative to the reference map (MSE). Variants with high disruption scores are inferred to contribute to 3D genome folding.

(B) A genome-wide, tiled, 5-kb deletion screen produces a distribution of sequence importance with log(MSE) between −10 and −1 for the HFFc6 cell type.

(C) Genome-wide screens capture a range of disruption scores; some sequences do not change predicted genome folding (left), some produce small focal changes (center), and others dramatically rearrange boundaries (right).

(D) The rolling average of disruption and compartment score across a 60-Mb region of chromosome 4. Peaks correspond to regions sensitive to perturbation, while valleys indicate regions robust to perturbation. Yellow shading highlights genomic regions with relatively few CTCF motifs. These regions have low disruption scores, suggesting that their perturbation has little effect on genome folding.

(E) Sensitivity to disruption correlates strongly with compartment score, as measured by the first eigenvector of HFFc6 micro-C.

See also Figures S1 and S4.

activity of some promoters,[8] and emerging work reveals that RNA polymerase II and transcription may separately influence 3D genome folding.[31,32] To test this hypothesis, we evaluated all single-nucleotide mutations in the 300 bp on either side of the transcription start site (TSS) of the 1,789 highest expressed protein-coding genes in HFFc6[29] and compared disruption scores with expression level in regions where CTCF motifs are present or absent (Figure 2C). In regions flanking a CTCF motif, we observed a strong peak in disruption directly upstream of the TSS (Figures 2D and S3). The periodic pattern is more detailed than underlying CTCF motifs and more precise than a sum of CTCF ChIP-seq peaks around the TSS. Metaplots of the average change in contact reveal that mutations weaken boundaries at the TSS. Our analysis points to a presence of CTCF at the promoters of highly expressed genes, where some CTCF motifs are selectively bound and some are not. We note that, when no CTCF is present, disruption is significantly lower but still slightly elevated upstream of the TSS of highly transcribed genes (Figures 2E and S3). Furthermore, disruption scales with gene expression when CTCF is present and absent (Figures 2F and 2G). These results are consistent with the hypothesis that active transcription may provide an alternate means of stabilizing DNA-DNA interactions in TSSs devoid of CTCF sites through uncharacterized mechanisms, like transcriptional machinery, nascent RNA, or recruited regulatory RNA.

In silico screening approach validates across cell lines

To test the robustness of our approach and findings, we repeated the above analyses in a second cell line. We selected H1hESC because of the availability of micro-C data and the opportunity to compare a pluripotent cell line with a differentiated one. Furthermore, H1hESC is one of the five cell lines for which the Akita model predicts chromatin contacts, enabling us to directly assess the effects of *in silico* disruptions on genome folding patterns in H1hESC alongside HFFc6. This analysis showed that all of the above trends observed in HFFc6, including

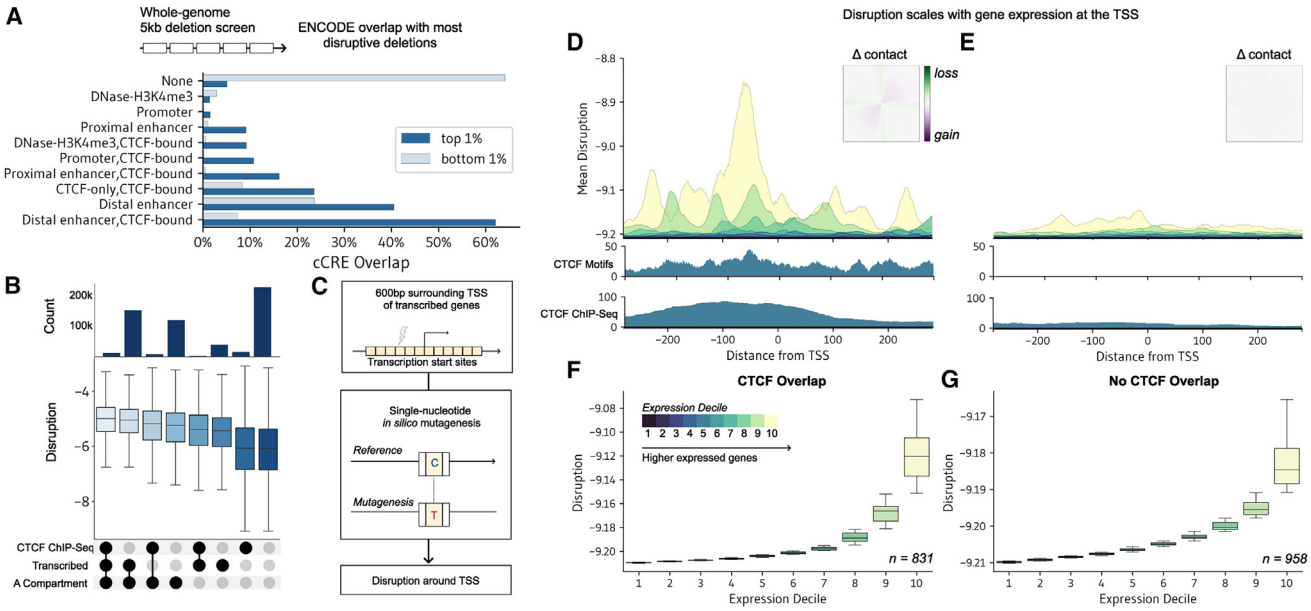


Figure 2. Transcription and CTCF are key modulators of 3D genome folding

(A) Overlap between the top 1% (most disruptive, dark blue) or bottom 1% (least disruptive, light blue) 5-kb sequence deletions and ENCODE candidate *cis*-regulatory elements, quantified as the proportion of deletions with overlap. Each deletion may overlap with more than one regulatory element. See also Figures S1, S2, and S4.

(B) Average disruption score across genomic regions overlapping with CTCF ChIP-seq peaks, A compartments, and/or actively transcribed sequences.

(C) Single-base-pair mutagenesis screen of a 600-bp segment surrounding the transcription start site (TSS) of the most highly transcribed genes in HFFc6 (n = 1,789).

(D and E) Average disruption score of each base at TSS regions with (E) and without (F) a CTCF motif overlap, stratified by expression decile (colors), along with average CTCF motif density and CTCF ChIP-seq. Metaplots (top right) show the average change in contact for the 100 TSSs with the most significant disruption scores. See also Figure S3.

(F and G) Mean disruption score of transcribed genes, stratified by expression level decile (colors), and separated into those whose TSS region overlaps (F) and does not overlap (G) with CTCF sites.

The figures have different scales.

disruptions scores being associated with CTCF motifs, transcription, A compartments, GC content, and the deleted sequence length, are consistent in H1hESC (Figures S4A–S4C).

Transposon clusters modulate genome folding independent of CTCF

At the chromosome scale, our unbiased genome-wide screen highlighted clusters of Alu elements and some other repetitive elements alongside peaks in disruption scores, motivating us to explore their role in 3D genome folding (Figure 3A). DNA and RNA transposons replicate and insert themselves into DNA and constitute over 50% of the human genome.[33,34] They are rich in transcription factor binding sites,[35–37] suggesting that some may have been evolutionarily repurposed as regulatory elements. Growing evidence indicates that they provide a source of CTCF motifs across the genome and serve as loop anchors and insulators.[38–40] To measure the impact of different families of repetitive elements on 3D genome sensitivity, we compared disruption of 5-kb windows containing repetitive elements with those with none. Several families exhibit greater sensitivity to perturbation than CTCF-containing regions (e.g., Alu, SVA, small cytoplasmic RNA [scRNA], signal recognition particle RNA [srpRNA]; Figure 3B). Disruption scores of repetitive elements are not correlated with mappability, indicating that poor micro-C read mapping in model training data does not bias this result (Figures S5A–S5F; STAR Methods). As with CTCF,[41] regions with higher numbers of Alu elements are more disruptive upon deletion; the disruption score of 5-kb windows with 5 or more Alu elements is 9.88% higher than that of windows with no elements ($p < 1.54 \times 10^{-291}$; Figure 3C). This clustering effect holds across many repetitive elements, including mammalian-wide interspersed repeat (MIR) and L2 long interspersed nuclear elements (LINEs), as well as across most small non-coding RNA genes (Figure 3C). Many families, like L1 LINEs, show no correlation at all, and trends are consistent across A and B compartments, hinting that clustering is family specific (Figures 3C, S6A, and S6B).

To investigate the contribution of repetitive elements independent of flanking sequence, we next individually deleted over 1 million elements in the RepeatMasker database (Figure 4A). Overall, many elements create large-scale boundary shifts, with some causing increases and others decreases in contact frequency (Figure 4B). Deletions of almost all families are more disruptive than random deletions, and deletions of families such as Alu, small RNAs, SVA, and hAT-Charlie are on par with or exceed deletions of CTCF sites across the genome

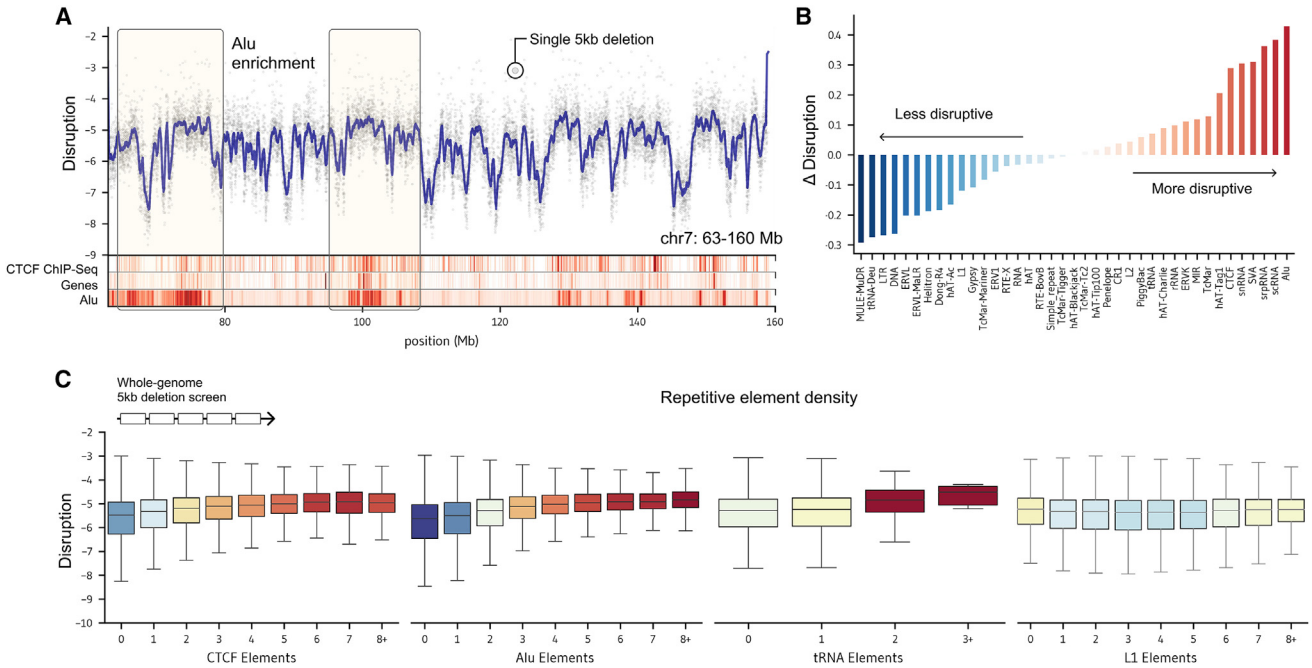


Figure 3. Regions with repetitive elements are sensitive to sequence perturbation

(A) Mean disruption scores of tiled 5-kb deletions across a 100-Mb region of chromosome 7. Tracks below the plot illustrate the density of CTCF motifs, genes, and Alu elements.

(B) Mean difference in disruption scores between windows containing at least one repetitive element and windows containing none, stratified by family.

(C) Disruption scores of 5-kb deletions stratified by the number of Alu elements, tRNAs, L1 LINE elements, and CTCF motifs they contain.

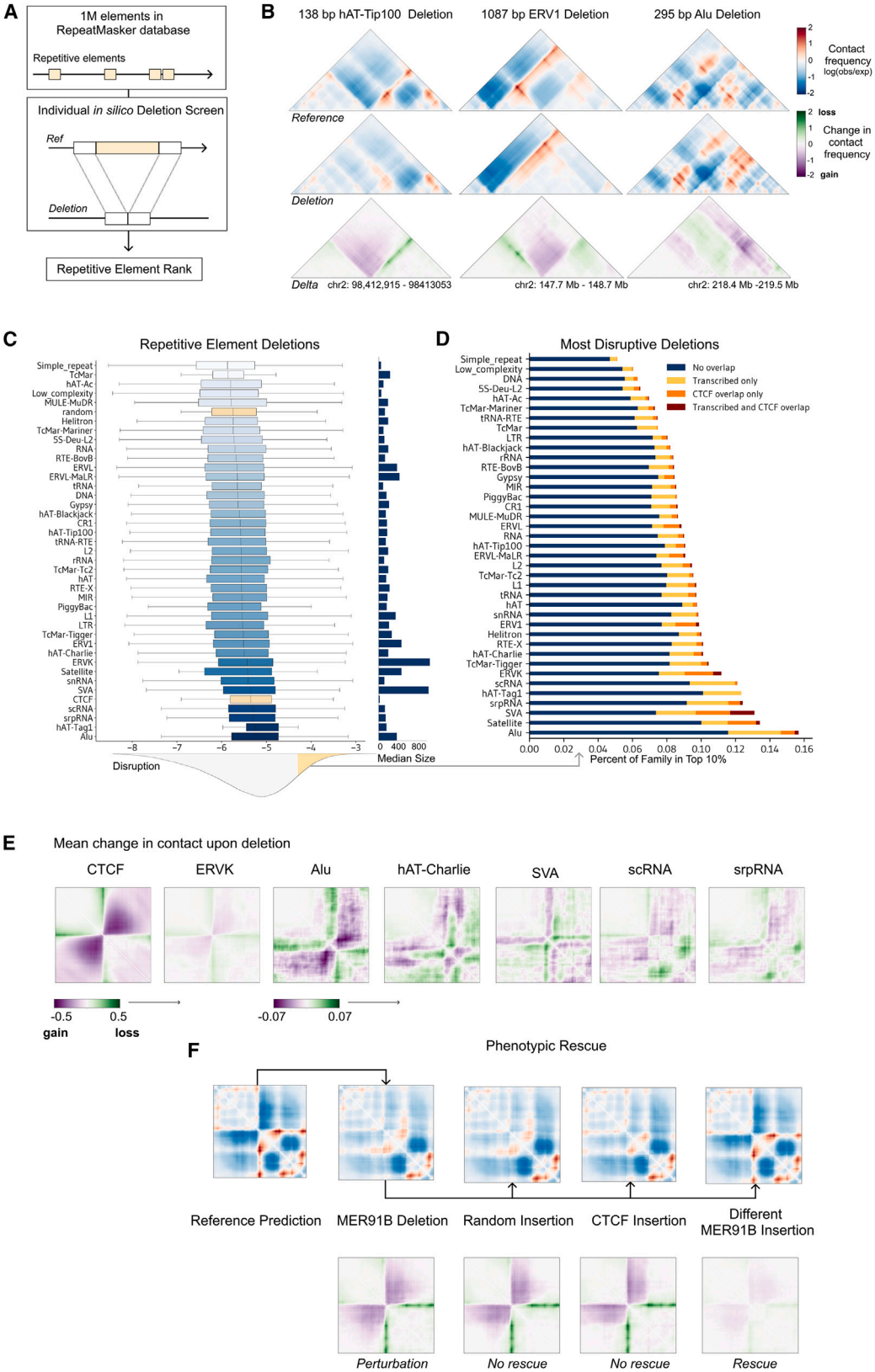
See also Figure S6.

(Figures 4C and S6C). Disruption is moderately correlated with size, but many highly disruptive element families are relatively small and cause unexpectedly large disruptions given their length (Figures S7A, S7B, and 4C). For example, deletion of tRNAs, scRNAs, srpRNAs, and small nuclear RNAs (snRNAs), all under 130 bp on average, have a propensity to drastically alter genome folding.

To experimentally validate these deep-learning-based predictions, we leveraged the natural sequence differences between humans and chimpanzees. Specifically, we examined loci with human-lineage-specific repetitive elements in Hi-C data we generated previously in human and chimpanzee neural progenitor cells.[42] By comparing the experimental data with Akita predictions where the human-specific repetitive element is inserted into the chimpanzee genome and conversely deleted from the human genome, we find that Alu elements unique to humans generate consistent changes to genome folding (Figures S8A–S8C). Thus, experimental data validate our *in silico* screening approach and support the importance of Alu and other repetitive elements in genome folding.

We next explored possible mechanisms through which repetitive elements might influence genome folding. Causality is challenging to untangle because each repetitive element can contain features with known associations to chromatin organization. First, the lengths of repeat clusters are roughly similar to clusters of CTCF motifs at TAD boundaries (Figure 4C). Second, several repeat families are known to harbor CTCF motifs.[43] Third, some repeats have a strong GC bias (e.g., Alu GC% > 50%), potentially allowing them to establish compartments.[44,45] Finally, repetitive elements collectively account for a large amount of total nuclear transcription.[33] To dissect the contributions of CTCF and active transcription versus other features of repetitive elements, we quantified overlap of these two annotations with repetitive elements that have the highest disruption scores. Only 5.86% of the 10% most disruptive elements contain a CTCF motif, while 13.55% are actively transcribed (Figure 4D), so a majority overlap neither. Disruptive repetitive element deletions are enriched at distal enhancers that are not CTCF bound (Figure S9). These findings hint that repetitive elements may aid in genome folding independently and in collaboration with CTCF and transcription.

To understand the folding phenotypes of element deletions, we next averaged the changes in contact frequency for the top-scoring elements of each family (Figure 4E). Endogenous retrovirus-K (ERVK) elements behaved like CTCF sites: their deletion led to a strong and centered loss of a chromatin boundary. Other repeat families created an off-diagonal gain in contact, as seen with Alu and hAT-Charlie; dispersed focal disruptions, as with non-coding RNAs; and stripes, as with SVA elements. To demonstrate that the model is internally consistent, we performed a phenotypic rescue, where we deleted an individual hAT-Tip100 element to produce a large change in contact and attempted to restore the original folding pattern with a different sequence (Figure 4F). While introducing random DNA or a



(legend on next page)

CTCF motif did not recreate the original contact, inserting a related MER91B hAT-Tip100 element did. We conclude that repetitive element families are associated with distinct chromatin contact map features and that elements within a family are generally functionally interchangeable.

Insertion of repetitive elements leads to distinct folding phenotypes

Our deletion experiments do not distinguish between repetitive elements that collaborate with CTCF to weaken or strengthen nearby TAD boundaries and those that separately create chromatin contact. To isolate the effects of repetitive elements, we next designed *in silico* insertion experiments. We first engineered a "blank canvas" with no predicted structure by depleting a randomly generated 1-Mb DNA sequence of all CTCF-like motifs (Figures 5A and S10). We then inserted one or more copies of any query sequence into this 1 Mb and quantified newly arising chromatin contacts. We easily recreated a division closely resembling a TAD boundary by inserting multiple copies of the canonical CTCF motif (Figure 5B), validating this approach to creating chromatin contact phenotypes.

After introducing the 1,000 most disruptive repetitive elements in our deletion screen into a blank canvas, we found that the majority also changed contact with insertion, including 80.3% of Alu elements and 86.0% of ERVK elements (Figure 5C). Additional copies strengthened the impact, and fewer copies were needed to induce a chromatin boundary compared with the CTCF motif (Figures 5D and S11). Clustering the insertion maps revealed that hAT/MIR insertions produced distinct folding patterns from ERV/SVA element insertions (Figure 5E). Alu elements consistently produced focal changes at the site of insertion that appeared unlike CTCF-like boundaries. Curiously, repetitive elements seem to produce two distinct modifications to 3D structure upon insertion. Some elements create CTCF-like domain boundaries that increase in strength as more elements are inserted (Figure 5F). Other elements, like the Alu and SVA families, form a pattern resembling a cross, with increased contact upstream and downstream of the insertion point. This cross-like pattern increases in size with more element insertions. Insertions of tRNA genes did not create new boundaries, suggesting that their effect on 3D genome folding may be context dependent.

Some repetitive elements harbor CTCF motifs and overlap with CTCF ChIP-seq peaks, strongly suggesting that the Akita model predicted their importance because they contain CTCF binding sites. To test this hypothesis, we performed saturation mutagenesis across a number of high-scoring repetitive elements (Figure 5D). Screening an ERVK element, for example, revealed that the single nucleotides predicted to have the highest importance for contacts lie directly at the center of a CTCF binding site (Figure 5D). Overall, the closer a sequence is to matching the canonical CTCF motif, the larger the predicted impact of its insertion (Figures S12A and S12B). Still, most of the elements that produced contact changes had no CTCF overlap, and the 5- to 50-bp motifs within these elements with the greatest impact did not resemble CTCF motifs (Figures 5D, S12C, and S12D). Therefore, insertions support the hypothesis that repetitive elements contain sequence determinants of 3D genome folding beyond CTCF motifs.

Necessary vs. sufficient: A 60-bp segment of Charlie7 is sufficient to induce a CTCF-like boundary

Mutating individual nucleotides can be enough to disturb protein binding and profoundly impair 3D folding. By contrast, creating a boundary, loop, or domain from scratch is more challenging, and it is fundamentally unclear what minimum sequence is sufficient. We next extended our screening approach to explore which subsequences can produce the *de novo* contact of a full element.

First, we examined CTCF motifs. Fudenberg et al.[20] mutated all motifs in the JASPAR transcription factor database and determined that CTCF and CTCFL are most sensitive to sequence perturbation. To complement this work, we inserted all motifs into a blank map. We find that, regardless of motif spacing, CTCF and CTCFL are the transcription factor motifs best able to induce genome folding independent of any surrounding genomic context, followed by HAND2, Ptf1A, and YY2 (Figures 6A and S13A). YY1 scores relatively lower, perhaps because of its less stable binding or its binding with co-factors.[46] Sampling and inserting motifs from the CTCF position weight matrix, we found that the consensus sequence creates a stronger boundary than 99.50% of CTCF variants (Figure 6B). However, a small minority of CTCF "super-motifs" with a T at positions 8 and 12 outperformed the canonical motif, hinting that the most common CTCF motifs may not be the most strongly insulating ones. The super-motif sequences also produced stronger boundaries in experimental Hi-C than the CTCF consensus sequence (Figures S13B and S13C), and they are equally likely to overlap CTCF ChIP-seq peaks (Figure S13D). These results illustrate that Akita can be used to interpret the function of CTCF and other transcription factor binding sites at single-nucleotide resolution.

Figure 4. Repetitive element deletions impact genome folding

(A) Strategy to individually delete over 1 million elements from the RepeatMasker database.

(B) Representative examples from chromosome 2, showing how deletion of a hAT-Tip100 element, an ERV1 element, and an Alu element *in silico* significantly alter contact maps. Single elements are predicted to disrupt genome folding.

(C) Distribution of disruption scores across each repetitive element family (n = 1,164,108). The distributions of disruptions from 100,000 CTCF deletions (positive control) and 100,000 100-bp random deletions (negative control) are shown in yellow. The median size in base pairs of deleted elements for each family is shown on the right. See also Figure S6.

(D) The top 10% most disruptive elements across the screen by repetitive element family. Most elements do not overlap a CTCF motif or a region actively transcribed in the HFFc6 cell line.

(E) Average changes in contact maps for the top 100 elements per family.

(F) Phenotypic rescue. We showcase a 138-bp MER91B hAT-Tip100 element whose deletion produces a loss of a boundary. Inserting a random size-matched sequence and a CTCF motif does not change the disturbed contact map, but introducing an MER91B element from the same family restores the original genome folding.

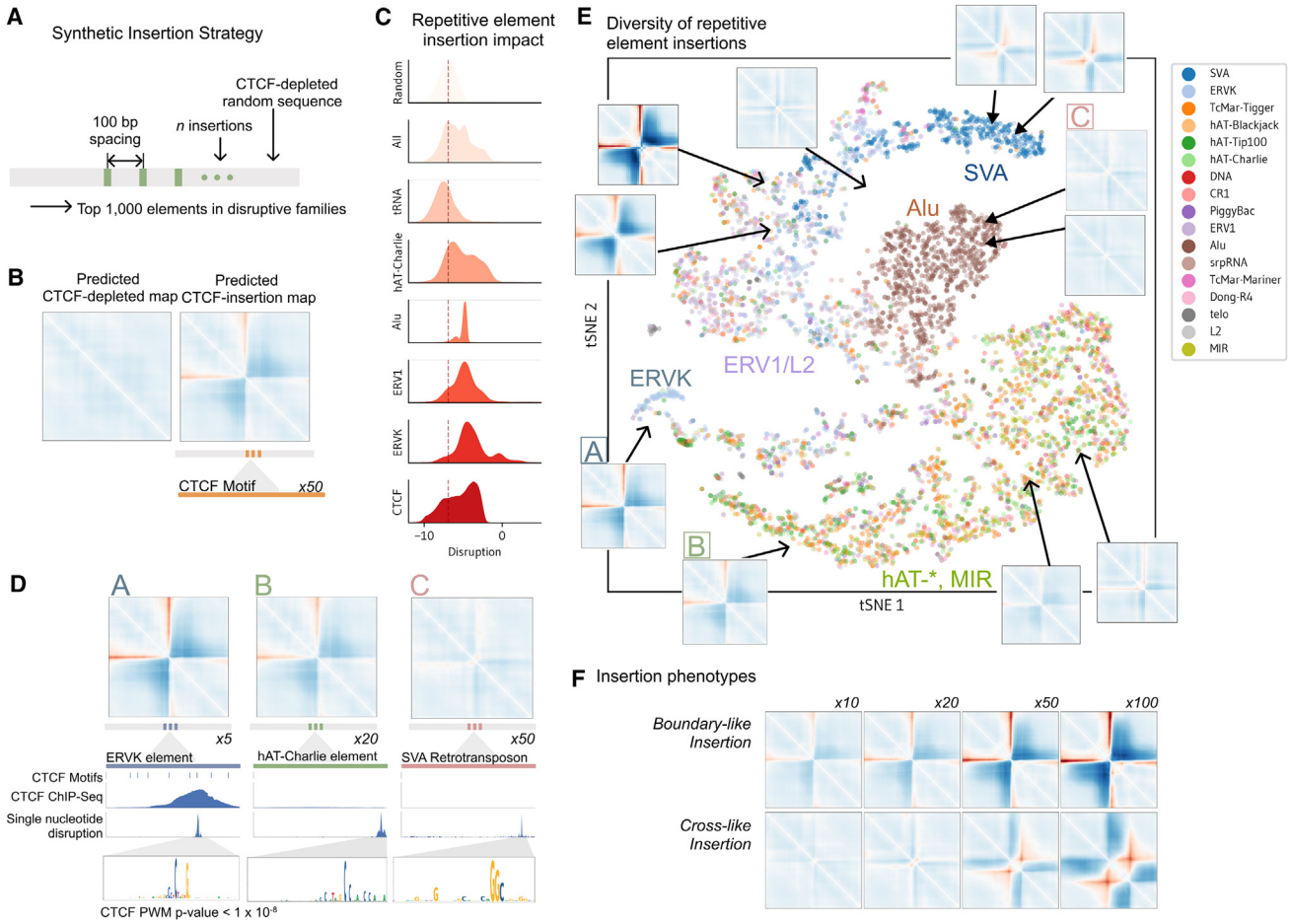


Figure 5. *In silico* insertion screen reveals that repetitive elements can induce different boundary types

(A) Insertion screen strategy. For each of the 1,000 most disruptive elements, up to 100 individual copies (green) are inserted 100 bp apart, centered in a 1-Mb random DNA sequence depleted of CTCF sites.

(B) The map predicted from the CTCF-depleted random sequence (left) provides a blank canvas against which we can measure the impact of insertions. A CTCF site insertion into the middle of the sequence produces boundaries in the predicted maps (right). Disruption is measured as the mean squared difference between the blank map and the predicted post-insertion map. See also Figure S10.

(C) Distribution of disruption scores across repetitive element insertions (n = 14,514). The score distributions of 10,000 100-bp random insertions (negative control) and of 10,000 CTCF motif insertions (positive control) are shown. See also Figure S12.

(D) We highlight three repetitive elements that are highly disruptive when deleted and inserted. We overlay overlapping annotated CTCF motifs and CTCF sites confirmed by ChIP-seq in HFFc6 cells. We also show the disruption score of each nucleotide across the element following single-base-pair *in silico* mutagenesis, highlighting the motif within the repetitive element responsible for the element's high disruption score.

(E) t-SNE visualization of all predicted maps from repetitive element insertions with a disruption score above −5.5. Predicted maps are colored by element family.

(F) We observe two primary classes of insertions. CTCF-like boundary insertions are common across ERVK and ERV1 elements, and cross-like insertions are common across SVA and Alu elements.

Next, we dissected Charlie7, a 367-bp AT-rich (29% GC) hAT-Charlie element on chromosome 11. Deleting Charlie7 eliminates chromatin interactions (Figures 5D and 6C). Inserting 20 tandem copies of Charlie7 creates a CTCF-like boundary despite no subsequence resembling a CTCF motif. This boundary could not be reproduced by inserting a shuffled Charlie7 sequence or a random sequence of the same length. We therefore shuffled individual 10-bp segments of Charlie7 to destroy local sequence grammar before reinserting the element into the blank canvas. Shuffling the final 60 bp had the same effect as shuffling the entire element, revealing that this end of the element is necessary for boundary creation (Figure 6D). We then created sliding windows of 10 bp, 50 bp, and 100 bp along the element and inserted each subsequence into the blank canvas. No individual subsequence was sufficient to reproduce the effect of the entire element (Figure 6E). However, shuffling the first 307 bp while maintaining the last 60 bp intact did create a strong boundary. Because the GC content of Charlie7 is unusually low, we next replaced parts of the element with random GC-matched sequence. A length-matched sequence with a GC content below 30% and the final 60 bp of the Charlie7 element was sufficient to create a boundary (Figure 6F). Completely random insertions

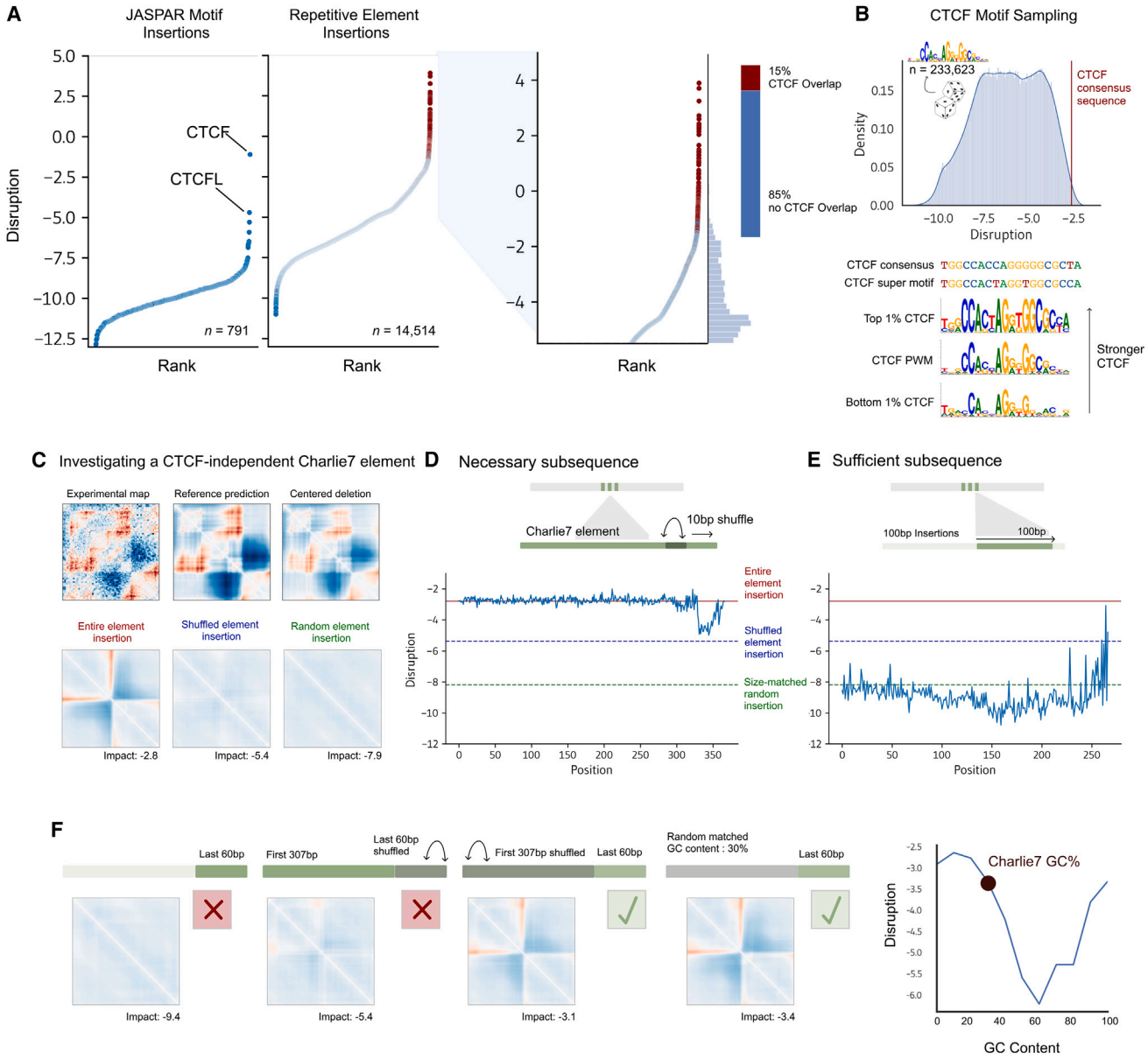


Figure 6. *In silico* investigation of sequence features necessary and sufficient for repetitive element Charlie7 to create a boundary

(A). We insert every JASPAR motif into a CTCF-depleted random sequence, as well as 14,514 repetitive elements, and rank them according to their disruption score. 85% of the most impactful insertions (score > −5.5) do not overlap a CTCF motif. See also Figures S12 and S13.

(B) We generate CTCF motif variants with frequencies sampled from the CTCF motif position weight matrix (PWM) and insert them into the random reference sequence (n = 326,177), finding that 0.50% of motifs produce stronger predicted boundaries when inserted than the CTCF consensus sequence. These "super-motifs" share Ts at positions 8 and 12.

(C) We investigate a 367-bp disruptive Charlie7 hAT-Charlie element that does not overlap a CTCF motif or ChIP-seq peak. Shown in the top row are the experimental micro-C contact map around the locus of the Charlie7 insertion, the map of the locus predicted by Akita, and the predicted map following deletion of the entire element. Shown in the bottom row are the predicted maps after insertion into the reference, CTCF-depleted sequence of the Charlie7 element (left), a version of the element with a shuffled sequence (center), and a random sequence of equal length (right).

(D) We shuffle each 10-bp subsequence along the element to determine which one is necessary to produce the boundary seen from introducing the whole element.

(E) We introduce 100-bp segments scanning the entire element into the reference sequence and find that none is sufficient to produce a strong boundary.

(F) A DNA sequence matching the GC content of Charlie7's first 307 bp combined with the last 60 bp is sufficient to recreate a boundary. Right: the first 307 bp of Charlie7 were replaced with randomly generated sequence across a range of GC content.

with a GC content below 30% and above 60% are also highly impactful (Figures S14A and S14B). Based on these *in silico* experiments, we conclude that GC content along with sequence syntax could be critical for the insulating behavior of Charlie7. Looking across all disruptive retrotransposons, we identify several families with very extreme average GC content (Figure S14C), suggesting the intriguing hypothesis that abrupt shifts in GC content resulting from repetitive element insertions into genomic DNA contribute to genome folding.

DISCUSSION

In summary, we present a whole-genome, unbiased survey of the sequence determinants of 3D genome folding using a flexible deep learning strategy for scoring the effect of genetic variants on local chromatin interactions. Our study utilized synthetic mutations ranging from large deletions tiled across hundreds of megabases down to single-nucleotide perturbations within sequence motifs. Leveraging the high throughput of this *in silico* screening strategy, we showed that sensitivity to 3D genome disruption is associated with A compartments, extreme GC content, CTCF motif density, and active transcription. We identified clusters of retrotransposons and RNA genes important for 3D genome folding, as modulating their sequences disrupted chromatin contacts on par with or more than modulating CTCF sites. Many of the repetitive elements with the largest effects on 3D genome folding when deleted and inserted do not contain CTCF and have not been implicated previously in chromatin architecture, but they often have different GC content from the sequences into which they are inserted.

This study contributes to a growing body of evidence showing that transposable elements modulate genome folding[47] and replication timing.[48] It has long been hypothesized that transposons may have been evolutionarily co-opted as regulatory elements.[36,37] Most transposable elements are decommissioned by chromatin modifications,[49] but functional escape can change genome conformation.[50] We observe loss and gain of contact upon transposable element deletion, supporting the idea that these elements can establish new boundaries by installing CTCF-like motifs and inhibit ancient CTCF binding sites to block contact.[38] Our results are also consistent with previous findings showing that specific MIR elements and tRNAs can serve as insulators,[51,52] while Alu and hAT provide loop anchors,[53–55] and hint that repetitive elements may work in tandem.[44] Cao et al.,[56] for example, identified that many transposable element families, but MIR short interspersed nuclear elements (SINEs) and L2 LINEs in particular, are enriched for binding sites and active chromatin marks, appear in the vicinity of tissue-specific gene expression, and interact with each other extensively to collaborate as enhancers or repressors. Liang et al.[55] showed that complementary RNAs from Alu sequences at enhancers and promoters promote chromatin interactions. In future work, it would be exciting to test coordination of transposable elements as shadow loop anchors, theorized by Choudhary et al.[38] to act as redundant regulatory material supporting CTCF.[38] We anticipate that comparing disruption with element age and species divergence will help us to understand the evolutionary mechanisms of transposable element deprogramming and selection in gene regulation.

Although we did not focus on CTCF specifically, a similar targeted *in silico* approach could directly address why the majority of CTCF motifs are not active[57,58] and whether methylation sensitivity of CTCF motifs containing CpGs tunes folding specificity.[59] We also anticipate that future *in silico* experiments and investigation of the model with activation maximization[60] will refine the spacing and orientation rules of neighboring and redundant CTCF elements and reveal how CTCF coordinates with flanking proteins and transposable elements.

Limitations of the study

It is important to emphasize that our *in silico* strategy, while demonstrated here and previously to be highly accurate,[20] is a screening and hypothesis-generating tool. Model predictions, especially those that implicate novel sequence elements or mechanisms, will require further experimental validation. We view this as a strength of our approach, not a weakness. Our ability to test millions of mutations efficiently and in an unbiased manner enables us to develop hypotheses and prioritize genomic loci that would not otherwise have been considered for functional characterization. It is now a high priority to apply massively parallel reporter assays, epitope devices, and genome engineering to explore how hAT, MIR, ERV, and SVA elements function in the context of 3D genome folding. We advocate for deep learning as a powerful strategy for driving experimental innovation that can be used iteratively with wet lab technologies to accelerate discovery.

Our conclusions rest heavily on the Akita model, which only considers a limited genomic window. Future work could apply the approach presented here with other deep learning models to test the robustness of our findings and potentially discover additional sequence features missed in our work. Our method scores the entire 1-Mb contact map and weights all regions equally, which may be too insensitive to capture small changes to specific loci. Filtering or weighting regions of the predicted contact maps by overlap with functional genomic annotations during score computations could also help to selectively test specific hypotheses. Our study is further limited by the quality of the hg38 reference genome, and we anticipate that extending to the new telomere-to-telomere human genome assembly will enable a better understanding of near-identical repetitive elements.[34] Finally, to leverage the best-quality data currently available, we only made predictions across HFFc6 and H1hESC, but features of the 3D genome can be cell type specific.[61] As very high-resolution and single-cell measurements of chromatin contacts, gene expression, and accessibility are generated for more cell types, it will be exciting to search for sequences that are necessary and sufficient for chromatin contacts in each cell type and to explore how variable these sequence determinants are across cellular contexts.

To experimentally validate the importance of repetitive elements in 3D genome folding, one could use CRISPR interference (CRISPRi) to acutely perturb select repetitive elements predicted to be highly disruptive by our computational screen. One could design guide RNAs targeting Alu elements predicted to be highly disruptive upon deletion, SVA elements predicted to

induce stripe-like patterns upon insertion, hAT-Charlie elements computationally shown to be sufficient to induce boundaries, and size- and chromosome-matched control repetitive elements predicted to have minimal impact. Instead of repressing transcription of targeted repetitive elements, one could alternatively perform a CRISPR deletion to disentangle the effect of transcription and other mechanisms. Significant disruption of contacts or boundaries measured by Hi-C specifically at the Alu, SVA, and hAT-Charlie elements targeted by CRISPRi or CRISPR compared with control elements would provide strong experimental support for the importance of these families of repetitive elements in establishing local chromatin architecture.

Conclusions

In our investigation, we develop a toolkit of *in silico* experimental strategies, including unbiased and targeted deletion screens, phenotypic rescue, insertions into synthetic sequence, sampling around known sequence motifs, and sequence contribution tracks across tens of base pairs to megabases. We hope that the variety of experiments profiled here may serve as a template for foundational biological research with deep learning. We also anticipate that our released disruption tracks will provide useful annotations for genome sensitivity and yield further insights into chromatin biology (Table S1). In sum, our work highlights the potential of deep learning models as powerful tools for biological hypothesis generation and discovery in regulatory genomics.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Akita model and datasets
  - Computing 3D genome folding disruption scores
  - Mass deletion screens
  - Genomic tracks
  - Overlap with genomic annotations
  - Mappability
  - *In silico* mutagenesis at the TSS
  - Repetitive elements
  - Phenotypic rescue
  - Insertion screens
  - Considering sequence mappability
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Disruption score significance
  - Motif significance

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2023.100410.

ACKNOWLEDGMENTS

We gratefully acknowledge Vijay Ramani, Elphege Nora, Tony Capra, and Geoff Fudenberg for valuable scientific insights and project guidance. We thank Evonne McArthur for generous discussion and the idea to insert motifs into CTCF-depleted random sequence. We additionally thank members of the Pollard and Keiser labs for useful comments and manuscript feedback. This work was supported by the NIH 4D Nucleome Project (grant U01HL157989 to K.S.P.), grant DAF2018-191905 (https://doi.org/10.37921/550142lkcjzw) from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation (funder https://doi.org/10.13039/100014989) (to M.J.K.), and a UCSF Achievement Reward for College Scientists (ARCS) Scholarship (to L.M.G.).

AUTHOR CONTRIBUTIONS

L.M.G., K.S.P., and M.J.K. conceptualized the work, designed experiments, and guided the project direction. L.M.G. conducted all analyses and prepared figures with feedback from K.S.P. and M.J.K. L.M.G. and K.S.P. drafted the original manuscript. L.M.G., K.S.P., and M.J.K. reviewed and edited the manuscript. K.S.P. and M.J.K. acquired funding.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: October 12, 2022
Revised: November 8, 2022
Accepted: August 31, 2023
Published: September 25, 2023

REFERENCES

1. Misteli, T. (2020). The Self-Organizing Genome: Principles of Genome Architecture and Function. Cell *183*, 28–45. https://doi.org/10.1016/j.cell.2020.09.014.
2. Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of Chromosomal Domains by Loop Extrusion. Cell Rep. *15*, 2038–2049. https://doi.org/10.1016/j.celrep.2016.04.085.
3. Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., et al. (2015). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900–910. https://doi.org/10.1016/j.cell.2015.07.038.
4. Rao, S.S.P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.-R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., et al. (2017). Cohesin Loss Eliminates All Loop Domains. Cell *171*, 305–320.e24. https://doi.org/10.1016/j.cell.2017.09.026.
5. Fudenberg, G., Abdennur, N., Imakaev, M., Goloborodko, A., and Mirny, L.A. (2017). Emerging Evidence of Chromosome Folding by Loop Extrusion. Cold Spring Harbor Symp. Quant. Biol. *82*, 45–55. https://doi.org/10.1101/sqb.2017.82.034710.
6. Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385. https://doi.org/10.1038/nature11049.
7. Merkenschlager, M., and Nora, E.P. (2016). CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. Annu. Rev. Genom. Hum. Genet. *17*, 17–43. https://doi.org/10.1146/annurev-genom-083115-022339.

8. Nora, E.P., Goloborodko, A., Valton, A.-L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A., and Bruneau, B.G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell *169*, 930–944.e22. https://doi.org/10.1016/j.cell.2017.05.004.

9. Barutcu, A.R., Maass, P.G., Lewandowski, J.P., Weiner, C.L., and Rinn, J.L. (2018). A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. Nat. Commun. *9*, 1444. https://doi.org/10.1038/s41467-018-03614-0.

10. Gorkin, D.U., Qiu, Y., Hu, M., Fletez-Brant, K., Liu, T., Schmitt, A.D., Noor, A., Chiou, J., Gaulton, K.J., Sebat, J., et al. (2019). Common DNA sequence variation influences 3-dimensional conformation of the human genome. Genome Biol. *20*, 255. https://doi.org/10.1186/s13059-019-1855-4.

11. Fudenberg, G., and Pollard, K.S. (2019). Chromatin features constrain structural variation across evolutionary timescales. Proc. Natl. Acad. Sci. USA *116*, 2175–2180. https://doi.org/10.1073/pnas.1808631116.

12. Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P.D., Dean, A., and Blobel, G.A. (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell *149*, 1233–1244. https://doi.org/10.1016/j.cell.2012.03.051.

13. Morgan, S.L., Mariano, N.C., Bermudez, A., Arruda, N.L., Wu, F., Luo, Y., Shankar, G., Jia, L., Chen, H., Hu, J.-F., et al. (2017). Manipulation of nuclear architecture through CRISPR-mediated chromosomal looping. Nat. Commun. *8*, 15993. https://doi.org/10.1038/ncomms15993.

14. Kubo, N., Ishii, H., Xiong, X., Bianco, S., Meitinger, F., Hu, R., Hocker, J.D., Conte, M., Gorkin, D., Yu, M., et al. (2021). Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation. Nat. Struct. Mol. Biol. *28*, 152–161. https://doi.org/10.1038/s41594-020-00539-5.

15. Kim, J.H., Rege, M., Valeri, J., Dunagin, M.C., Metzger, A., Titus, K.R., Gilgenast, T.G., Gong, W., Beagan, J.A., Raj, A., and Phillips-Cremins, J.E. (2019). LADL: light-activated dynamic looping for endogenous gene expression control. Nat. Methods *16*, 633–639. https://doi.org/10.1038/s41592-019-0436-5.

16. de Almeida, B.P., Reiter, F., Pagani, M., and Stark, A. (2022). DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. Nat. Genet. *54*, 613–624. https://doi.org/10.1038/s41588-022-01048-5.

17. Taskiran, I.I., Spanier, K.I., Christiaens, V., Mauduit, D., and Aerts, S. (2022). Cell type directed design of synthetic enhancers. Preprint at bioRxiv. https://doi.org/10.1101/2022.07.26.501466.

18. Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., and Zeitlinger, J. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat. Genet. *53*, 354–366. https://doi.org/10.1038/s41588-021-00782-6.

19. Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., and Kelley, D.R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nat. Methods *18*, 1196–1203. https://doi.org/10.1038/s41592-021-01252-x.

20. Fudenberg, G., Kelley, D.R., and Pollard, K.S. (2020). Predicting 3D genome folding from DNA sequence with Akita. Nat. Methods *17*, 1111–1117. https://doi.org/10.1038/s41592-020-0958-x.

21. 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. https://doi.org/10.1038/s41592-020-0960-3.

22. Yang, R., Das, A., Gao, V.R., Karbalayghareh, A., Noble, W.S., Bilmes, J.A., and Leslie, C.S. (2023). Epiphany: predicting Hi-C contact maps from 1D epigenomic signals. Genome Biol. *24*, 134. https://doi.org/10.1186/s13059-023-02934-9.

23. Tan, J., Shenker-Tauris, N., Rodriguez-Hernaez, J., Wang, E., Sakellaropoulos, T., Boccalatte, F., Thandapani, P., Skok, J., Aifantis, I., Fenyö, D., et al. (2023). Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening. Nat. Biotechnol. *41*, 1140–1150. https://doi.org/10.1038/s41587-022-01612-8.

24. Zhou, J. (2022). Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. Nat. Genet. *54*, 725–734. https://doi.org/10.1038/s41588-022-01065-4.

25. Yang, M., and Ma, J. (2022). Machine Learning Methods for Exploring Sequence Determinants of 3D Genome Organization. J. Mol. Biol. *434*, 167666. https://doi.org/10.1016/j.jmb.2022.167666.

26. McArthur, E., Rinker, D.C., Gilbertson, E.N., Fudenberg, G., Pittman, M., Keough, K., Pollard, K.S., and Capra, J.A. (2022). Reconstructing the 3D genome organization of Neanderthals reveals that chromatin folding shaped phenotypic and sequence divergence. Preprint at bioRxiv. https://doi.org/10.1101/2022.02.07.479462.

27. Krietenstein, N., Abraham, S., Venev, S.V., Abdennur, N., Gibcus, J., Hsieh, T.-H.S., Parsi, K.M., Yang, L., Maehr, R., Mirny, L.A., et al. (2020). Ultrastructural Details of Mammalian Chromosome Architecture. Mol. Cell *78*, 554–565.e7. https://doi.org/10.1016/j.molcel.2020.03.003.

28. Naughton, C., Avlonitis, N., Corless, S., Prendergast, J.G., Mati, I.K., Eijk, P.P., Cockroft, S.L., Bradley, M., Ylstra, B., and Gilbert, N. (2013). Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. Nat. Struct. Mol. Biol. *20*, 387–395. https://doi.org/10.1038/nsmb.2509.

29. ENCODE Project Consortium; Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710. https://doi.org/10.1038/s41586-020-2493-4.

30. van Steensel, B., and Furlong, E.E.M. (2019). The role of transcription in shaping the spatial organization of the genome. Nat. Rev. Mol. Cell Biol. *20*, 327–337. https://doi.org/10.1038/s41580-019-0114-6.

31. Busslinger, G.A., Stocsits, R.R., van der Lelij, P., Axelsson, E., Tedeschi, A., Galjart, N., and Peters, J.-M. (2017). Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. Nature *544*, 503–507. https://doi.org/10.1038/nature22063.

32. Zhang, S., Übelmesser, N., Josipovic, N., Forte, G., Slotman, J.A., Chiang, M., Gothe, H.J., Gusmao, E.G., Becker, C., Altmüller, J., et al. (2021). RNA polymerase II is required for spatial chromatin reorganization following exit from mitosis. Sci. Adv. *7*, eabg8205. https://doi.org/10.1126/sciadv.abg8205.

33. Trigiante, G., Blanes Ruiz, N., and Cerase, A. (2021). Emerging roles of repetitive and repeat-containing RNA in nuclear and chromatin organization and gene expression. Front. Cell Dev. Biol. *9*, 735527. https://doi.org/10.3389/fcell.2021.735527.

34. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44–53. https://doi.org/10.1126/science.abj6987.

35. Wang, T., Zeng, J., Lowe, C.B., Sellers, R.G., Salama, S.R., Yang, M., Burgess, S.M., Brachmann, R.K., and Haussler, D. (2007). Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. Proc. Natl. Acad. Sci. USA *104*, 18613–18618. https://doi.org/10.1073/pnas.0703637104.

36. Bourque, G., Leong, B., Vega, V.B., Chen, X., Lee, Y.L., Srinivasan, K.G., Chew, J.-L., Ruan, Y., Wei, C.-L., Ng, H.H., and Liu, E.T. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res. *18*, 1752–1762. https://doi.org/10.1101/gr.080663.108.

37. Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat. Genet. *42*, 631–634. https://doi.org/10.1038/ng.600.

38. Choudhary, M.N., Friedman, R.Z., Wang, J.T., Jang, H.S., Zhuo, X., and Wang, T. (2020). Co-opted transposons help perpetuate conserved higher-order chromosomal structures. Genome Biol. *21*, 16. https://doi.org/10.1186/s13059-019-1916-8.

39. Raviram, R., Rocha, P.P., Luo, V.M., Swanzey, E., Miraldi, E.R., Chuong, E.B., Feschotte, C., Bonneau, R., and Skok, J.A. (2018). Analysis of 3D genomic interactions identifies candidate host genes that transposable elements potentially regulate. Genome Biol. *19*, 216. https://doi.org/10.1186/s13059-018-1598-7.

40. Diehl, A.G., Ouyang, N., and Boyle, A.P. (2020). Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes. Nat. Commun. *11*, 1796. https://doi.org/10.1038/s41467-020-15520-5.

41. Kentepozidou, E., Aitken, S.J., Feig, C., Stefflova, K., Ibarra-Soria, X., Odom, D.T., Roller, M., and Flicek, P. (2020). Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. Genome Biol. *21*, 5. https://doi.org/10.1186/s13059-019-1894-x.

42. Keough, K.C., Whalen, S., Inoue, F., Przytycki, P.F., Fair, T., Deng, C., Steyert, M., Ryu, H., Lindblad-Toh, K., Karlsson, E., et al. (2023). Three-dimensional genome rewiring in loci with human accelerated regions. Science *380*, eabm1696. https://doi.org/10.1126/science.abm1696.

43. Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Gonçalves, A., Kutter, C., Brown, G.D., Marshall, A., Flicek, P., and Odom, D.T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell *148*, 335–348. https://doi.org/10.1016/j.cell.2011.11.058.

44. Lu, J.Y., Chang, L., Li, T., Wang, T., Yin, Y., Zhan, G., Han, X., Zhang, K., Tao, Y., Percharde, M., et al. (2021). Homotypic clustering of L1 and B1/Alu repeats compartmentalizes the 3D genome. Cell Res. *31*, 613–630. https://doi.org/10.1038/s41422-020-00466-6.

45. Su, M., Han, D., Boyd-Kirkup, J., Yu, X., and Han, J.-D.J. (2014). Evolution of Alu elements toward enhancers. Cell Rep. *7*, 376–385. https://doi.org/10.1016/j.celrep.2014.03.011.

46. Hsieh, T.-H.S., Cattoglio, C., Slobodyanyuk, E., Hansen, A.S., Darzacq, X., and Tjian, R. (2022). Enhancer-promoter interactions and transcription are largely maintained upon acute loss of CTCF, cohesin, WAPL or YY1. Nat. Genet. *54*, 1919–1932. https://doi.org/10.1038/s41588-022-01223-8.

47. Zhang, Y., Li, T., Preissl, S., Amaral, M.L., Grinstein, J.D., Farah, E.N., Destici, E., Qiu, Y., Hu, R., Lee, A.Y., et al. (2019). Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. Nat. Genet. *51*, 1380–1388. https://doi.org/10.1038/s41588-019-0479-7.

48. Yang, Y., Wang, Y., Zhang, Y., and Ma, J. (2022). Concert: Genome-Wide Prediction of Sequence Elements That Modulate DNA Replication Timing. In Research in Computational Molecular Biology (Springer International Publishing), pp. 358–359. https://doi.org/10.1007/978-3-031-04749-7_27.

49. Slotkin, R.K., and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. *8*, 272–285. https://doi.org/10.1038/nrg2072.

50. Huda, A., Mariño-Ramírez, L., and Jordan, I.K. (2010). Epigenetic histone modifications of human transposable elements: genome defense versus exaptation. Mobile DNA *1*, 2. https://doi.org/10.1186/1759-8753-1-2.

51. Wang, J., Vicente-García, C., Seruggia, D., Moltó, E., Fernandez-Miñán, A., Neto, A., Lee, E., Gómez-Skarmeta, J.L., Montoliu, L., Lunyak, V.V., and Jordan, I.K. (2015). MIR retrotransposon sequences provide insulators to the human genome. Proc. Natl. Acad. Sci. USA *112*, E4428–E4437. https://doi.org/10.1073/pnas.1507253112.

52. Van Bortle, K., Nichols, M.H., Li, L., Ong, C.-T., Takenaka, N., Qin, Z.S., and Corces, V.G. (2014). Insulator function and topological domain border strength scale with architectural protein occupancy. Genome Biol. *15*, R82. https://doi.org/10.1186/gb-2014-15-5-r82.

53. Ferrari, R., de Llobet Cucalon, L.I., Di Vona, C., Le Dilly, F., Vidal, E., Lioutas, A., Oliete, J.Q., Jochem, L., Cutts, E., Dieci, G., et al. (2020). TFIIIC Binding to Alu Elements Controls Gene Expression via Chromatin Looping and Histone Acetylation. Mol. Cell *77*, 475–487.e11. https://doi.org/10.1016/j.molcel.2019.10.020.

54. Choudhary, M.N.K., Quaid, K., Xing, X., Schmidt, H., and Wang, T. (2023). Widespread contribution of transposable elements to the rewiring of mammalian 3D genomes. Nat. Commun. *14*, 634. https://doi.org/10.1038/s41467-023-36364-9.

55. Liang, L., Cao, C., Ji, L., Cai, Z., Wang, D., Ye, R., Chen, J., Yu, X., Zhou, J., Bai, Z., et al. (2023). Complementary Alu sequences mediate enhancer-promoter selectivity. Nature *619*, 868–875. https://doi.org/10.1038/s41586-023-06323-x.

56. Cao, Y., Chen, G., Wu, G., Zhang, X., McDermott, J., Chen, X., Xu, C., Jiang, Q., Chen, Z., Zeng, Y., et al. (2019). Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions. Genome Res. *29*, 40–52. https://doi.org/10.1101/gr.235747.118.

57. Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231–1245. https://doi.org/10.1016/j.cell.2006.12.048.

58. Chen, H., Tian, Y., Shu, W., Bo, X., and Wang, S. (2012). Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. PLoS One *7*, e41374. https://doi.org/10.1371/journal.pone.0041374.

59. Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M., and Tilghman, S.M. (2000). CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. Nature *405*, 486–489. https://doi.org/10.1038/35013106.

60. Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. Preprint at arXiv. https://doi.org/10.48550/arXiv.1704.02685.

61. Schmitt, A.D., Hu, M., Jung, I., Xu, Z., Qiu, Y., Tan, C.L., Li, Y., Lin, S., Lin, Y., Barr, C.L., and Ren, B. (2016). A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. Cell Rep. *17*, 2042–2059. https://doi.org/10.1016/j.celrep.2016.10.061.

62. Karimzadeh, M., Ernst, C., Kundaje, A., and Hoffman, M.M. (2018). Umap and Bismap: quantifying genome and methylome mappability. Nucleic Acids Res. *46*, e120. https://doi.org/10.1093/nar/gky677.

63. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006. https://doi.org/10.1101/gr.229102.

64. Miga, K.H., Newton, Y., Jain, M., Altemose, N., Willard, H.F., and Kent, W.J. (2014). Centromere reference models for human chromosomes X and Y satellite arrays. Genome Res. *24*, 697–707. https://doi.org/10.1101/gr.159624.113.

65. Bembom, O.. seqlogo: Sequence Logos for DNA Sequence Alignments. R Package Version 1.48.0. https://bioconductor.org/packages/release/bioc/html/seqLogo.html.

66. Open2C, Abdennur, N., Fudenberg, G., Flyamer, I., Galitsyna, A.A., Goloborodko, A., Imakaev, M., and Venev, S.V. (2022). Bioframe: Operations on Genomic Intervals in Pandas Dataframes. Preprint at bioRxiv. https://doi.org/10.1101/2022.02.16.480748.

67. Sherman, M.D. seqlogo: Python Port of the R Bioconductor 'seqLogo' Package. https://github.com/betteridiot/seqlogo.

68. Khan, A. (2021). pyJASPAR: A pythonic interface to JASPAR transcription factor motifs. Zenodo. https://doi.org/10.5281/zenodo.4509415.

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

70. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

71. Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME Suite. Nucleic Acids Res. *43*, W39–W49. https://doi.org/10.1093/nar/gkv416.

72. Akgol Oksuz, B., Yang, L., Abraham, S., Venev, S.V., Krietenstein, N., Parsi, K.M., Ozadam, H., Oomen, M.E., Nand, A., Mao, H., et al. (2021). Systematic evaluation of chromosome conformation capture assays. Nat. Methods *18*, 1046–1055. https://doi.org/10.1038/s41592-021-01248-7.

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

74. Smit, A.F.A., Hubley, R., and Green, P. RepeatMasker Open-4.0. http://www.repeatmasker.org.

75. Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Sci. Rep. *9*, 9354. https://doi.org/10.1038/s41598-019-45839-z.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Micro-C, HFF	4DN, Krietenstein et al.[28]	4DNESWST3UBH
RNA-Seq, HFF	ENCODE,[29] GEO:GSE188028	ENCFF262XNE, ENCFF381OAF
ChIP-Seq, CTCF, HFF	ENCODE, Lou et al.[62]	ENCFF465OQP
ATAC-Seq, HFF	4DN, Oksuz et al.[63]	4DNESMBA9T3L
Compartment, HFF	4DN, Oksuz et al.[63]	4DNES9X112GZ
RepeatMasker	RepeatMasker Open-4.0[60]	http://www.repeatmasker.org; RRID:SCR_012954
ENCODE blacklist v2	Amemiya et al.[58]	https://github.com/Boyle-Lab/Blacklist/tree/master/lists
Software and algorithms		
Akita	Fudenberg et al.[20]	https://github.com/calico/basenji/tree/master/manuscripts/akita
bioframe	Open2C et al.[64]	https://bioframe.readthedocs.io/en/latest/
matplotlib	Hunter et al.[65]	http://matplotlib.sourceforge.net/; RRID:SCR_008624
MEME	Bailey et al.[66]	https://meme-suite.org/; RRID:SCR_001783
numpy	Harris et al.[67]	http://www.numpy.org/; RRID:SCR_008633
pandas	Pandas development team[68]	https://pandas.pydata.org/; RRID:SCR_018214
pyJASPAR	Khan[69]	https://zenodo.org/record/4509415
Python 3.7	Python Software Foundation	https://www.python.org/; RRID:SCR_008394
seaborn	Waskom et al.[70]	https://seaborn.pydata.org/; RRID:SCR_018132
scikit-learn	Pedregosa et al.[71]	http://scikit-learn.org/; RRID:SCR_002577
scipy	Virtanen et al.[72]	https://scipy.org/; RRID:SCR_008058
seqlogo	Bembom[73]	https://bioconductor.org/packages/release/bioc/html/seqLogo.html
seqlogo, python port	Sherman[74]	https://github.com/betteridiot/seqlogo
Disruption scoring code and screening results	This paper	https://github.com/keiserlab/3d-genome-disruption-paper; https://doi.org/10.5281/zenodo.8280391

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Katherine Pollard (katherine.pollard@gladstone.ucsf.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- All original code has been deposited at github.com/keiserlab/3d-genome-disruption-paper and is publicly available as of the date of publication under the https://doi.org/10.5281/zenodo.8280391.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Akita model and datasets

Throughout this analysis, we use the published convolutional neural network Akita to predict log(observed/expected) chromatin contact maps from ~1 Mb (1,048,576 bp) of real, altered, or synthetic DNA sequence[20] (https://github.com/calico/basenji/tree/master/manuscripts/akita). All types of mutations, including deletions, insertions, inversions and substitutions, may be scored as long as they are smaller than 1 Mb. Akita's predictions have been shown to mirror experimental results with deletions across scales of thousands of base pairs (bp) to single nucleotides. Fudenberg et al. originally trained Akita across six cell-types simultaneously, and we made all predictions in this work in the cell-type with the highest resolution of training data, human foreskin fibroblasts (HFFc6).[20] We find that disruption in H1hESC is highly correlated (Figures S4 and S6C). The experimental Micro-C maps from HFFc6[27] are used in visualizations. All chromatin and transcriptomic data were generated in HFFc6 and downloaded from public repositories. The source of all public data, including Micro-C, ATAC-Seq, RNA-Seq, ChIP-Seq, and compartment calls, can be found in the key resources table. All analyses use the hg38 genome build. We downloaded centromere locations from UCSC Table Browser.[63]

Computing 3D genome folding disruption scores

The location of deletions and insertions are centered such that the start position of the variant is always introduced halfway through the 1-Mb sequence at 2^{19} bp. For deletions, we pull additional sequence from the right to pad the input to 2^{20} bp. We remove sequences from our analysis which overlap centromeres,[64] ENCODE blacklisted regions,[75] and regions with an N content greater than 5%. Evaluating predictions on GPU (NVIDIA GeForce GTX 1080 Ti, NVIDIA TITAN Xp, NVIDIA GeForce RTX 2080 Ti) decreased the time per variant from 1.58 s to 262 ms, on average.

We score disruption as the log of the mean squared error between reference and perturbed maps. Mean squared error captures large-scale contact map changes, and has been used previously to rank predictions.[20] Pearson/Spearman correlation is also an appropriate choice.[26]

Mass deletion screens

Along with controls, we perform the following large-scale deletion screens.

1. *5 kb*, *whole genome* (n = 562,743).
2. *10,000 (10k) random CTCF deletions*. CTCF locations are pulled from JASPAR 2022.[73]
3. *10k 100-bp random deletions*. Start locations are randomly sampled from the genome.
4. *Randomly sized deletions*, ranging from 1 bp to 100 kb (n = 41,207). Start locations are randomly sampled from the genome.
5. *RepeatMasker database deletions* (n = 1,164,107).[74] RepeatMasker downloaded from UCSC Table Browser. We exclude ambiguous elements (containing '?' in the label). We initially sample 10,000 elements per family or up to the total number of elements in the family, whichever is less. Thereafter, we randomly sample from the database.
6. *TSS deletions*. (n = 1,073,329 mutations across 1,789 genes).

A full summary as well as the location of these results can be found in Data Table S1.

Genomic tracks

We smoothed the disruption scores of 5-kb deletions with a rolling average of 50 bp to create disruption tracks (Figures 1D and 3A). We additionally visualize the density of the following elements at 5-kb resolution.

1. Reference genes, hg38, GENCODE v39,[69] downloaded from UCSC Table Browser.
2. ENCODE hg38 v3 candidate cCREs, ENCODE Project,[29] downloaded from UCSC Table Browser.
3. CTCF motifs (MA0139.1), JASPAR 2022,[73] downloaded from http://expdata.cmmt.ubc.ca/JASPAR/downloads/UCSC_tracks/2022/hg38/.
4. ATAC-Seq peaks in HFFc6.[72]
5. Alu, L1, and L2 elements, RepeatMasker database, v. 4.1.2,[74] downloaded from UCSC Table Browser.

Overlap with genomic annotations

We used pre-computed compartment scores generated from the HFFc6 Micro-C dataset originally employed for training Akita.[27] To calculate the overlap between disruption scores for 5-kb deletions and compartment scores generated at 50-kb resolution, we merged both measures by genomic location, filled missing disruption values with linear interpolation, and calculated the overlap across A compartments with a compartment score greater than 0 and B compartments with a compartment score less than 0.

We intersected deleted windows and transposable elements with ENCODE cCREs using bioframe[66] to calculate the percentage overlap. We use the same strategy to calculate overlap with JASPAR CTCF motifs, ATAC-Seq peaks, and transcribed elements. When quantifying transcription of repetitive elements unannotated as genes, we calculated overlap with RNA-seq BigWigs, summed across both strands.

Mappability

Per nucleotide mappability was measured using 24-kmer multi-read mappability, where mappability is the probability that a randomly selected read of length k in a given region is uniquely mappable.[62] Mappability tracks were downloaded from the Hoffman lab (https://bismap.hoffmanlab.org). In this study, mappability averaged across 5 kb deletions, repetitive element families, and Alu element types in a 100 Mb subset of chromosome 1 from 100 Mb to 200 Mb.

In silico mutagenesis at the TSS

We examined behavior at the TSS using *in silico* mutagenesis. We individually randomly mutated each nucleotide 300 bp upstream to 300 bp downstream of the top 1,789 highest expressed protein coding genes via total RNA-Seq and quantified the MSE between mutated and reference predicted maps. We observed that 1,015 genes fell in A compartments, while 63 fell in B compartments. To produce tracks in Figures 2D and 2E, we averaged the disruption of each nucleotide by position and smoothed using a rolling average of 20 bp. We used the same strategy across select repetitive elements to identify which nucleotides most contribute to entire-element disruption scores (Figure 5D). To create metaplots, we selected the highest scoring nucleotide change for each gene, and filtered all genes with a maximum disruption score above −7. We then averaged the difference between reference and perturbed maps for these genes.

Repetitive elements

Repetitive element density was calculated as the number of elements across the entire RepeatMasker database overlapping each 5-kb genomic bin. We quantified enrichment as the log fold change of the mean disruption across 10% of genomic windows per family compared to all windows. To create metaplots, we average the difference between maps for the top 100 repetitive element deletions per family, along with CTCF deletions.

Phenotypic rescue

We profiled the following elements in our proof-of-concept phenotype rescue screen.

1. A MER91B hAT-Tip100 element at position chr2:98412915-98413053.
SWA score: 392, Divergence: 27%. Disruption from reference = −2.65.
2. A size-matched 138-bp random DNA sequence.
Disruption from deletion = −2.55.
3. The canonical CTCF motif (TGGCCACCAGGGGGCGCTA).
Disruption = −2.68.
4. A MER91B element at position chr12:51824097-51824219.
SWA score: 245, Divergence: 20.9%. Disruption = −5.28.

Insertion screens

CTCF depletion

We created a simulated Hi-C contact map without structure as a blank canvas for insertion experiments. We first generated a random DNA sequence of length 2^{20} bp. By chance, predicted maps from random sequence will contain some above background contact frequencies. To remove all structure, we incremented across this sequence one nucleotide at a time with a 12-bp sliding window. For each position, we computed the edit distance to the consensus CTCF motif. If the edit distance fell below a set threshold, we inserted a random DNA sequence of length 12 until the subsequence was sufficiently different from CTCF. Experimenting with edit distances, we found that a distance of 7 produces predicted maps which lack structure but do not result in artificial model predictions (Figure S10). We call this a "blank canvas" 1-Mb sequence.

CTCF insertion

We inserted the CTCF motif into the blank canvas and predicted expected contact frequencies with Akita. We quantify insertion impact as the log mean squared error between the predicted maps of the blank canvas and the insertion. If more than one motif was added, the insertions were centered and separated by an arbitrary 100 bp. To sample the CTCF motif, we drew frequencies from the CTCF position weight matrix.[73] To create a baseline, we inserted 5,000 CTCF motifs drawn from locations in the genome. Sequence motifs were visualized with a python port of the seqLogo package.[65,67]

Repetitive element insertions

We selected the top 1,000 most disruptive repetitive elements per family by the deletion screen to insert back into the blank canvas sequence. We inserted both the forward and reverse complement of each sequence, and selected the direction with the highest score. For an initial screen, we inserted all elements 100x with 100-bp spacing. As an additional baseline, we inserted 1,000 201-bp randomly generated sequences, as the median repetitive element size in our insertion screen was 201 bp. To perform clustering with t-SNE, we decreased the resolution of the 448x448 pixel maps to 100x100 pixels and flatten them to 1D vectors before clustering.

Additional genomic tracks

In Figure 5D, we visualized CTCF ChIP-Seq and CTCF motif locations in the element's original genomic context. Along with deleting the entire element, we performed mutagenesis to a random nucleotide across the length of the element to create a 'disruption track' of nucleotides most sensitive to perturbation. We highlight the most sensitive bases.

JASPAR insertions

We inserted the forward and reverse complement of each JASPAR motif[73] into CTCF-depleted sequence with 100-bp spacing (n = 842). JASPAR motifs were pulled and coordinated with pyJASPAR.[68]

Considering sequence mappability

One concern is sequence mappability potentially confounding model training. Repetitive elements are, by nature, highly conserved and present inherent difficulties assigning multi-mapped reads. Before training the model, large gaps were excluded from the training dataset and missing Hi-C bins were linearly interpolated.[20] If repetitive elements were systematically removed or imputed, the model may behave unreliably when predicting unseen repetitive element sequences.

To investigate this confounder, we examined how sequence mappability compares to disruption score (Figure S6). In general, we observe no correlation between deletions of 5-kb windows and mappability, indicating that poorly mappable sequences do not have unusually high or low disruption scores. Mappability of individual elements is also uncorrelated with disruption.

We do find that Alu elements have particularly low sequence mappability and particularly high predicted importance. Many Alu elements are still active and recently inserted into DNA, and therefore have high sequence similarly, presenting a challenge in mapping. It is also possible that the highly conserved nature of recent Alu elements contributes to their utility in shaping the 3D genome. The correlation with mappability is expected and may or may not indicate a bias; it is difficult to disentangle these two possibilities easily. Relatively low negative correlation between disruption score and mappability for individual elements within the Alu class suggests that many of the highly disruptive Alus are not in regions of low mappability.

QUANTIFICATION AND STATISTICAL ANALYSIS

Disruption score significance

Pearson correlation coefficient of disruption scores compared to several other genomic annotations was calculated using scipy.stats.linregress (Figures 1E, 1D, and S1C).[70] To assess if the relationship between disruption and additional annotations was significant, we performed a two-sided Mann-Whitney-Wilcoxon test with compartment annotations (results, Figure 1E) and transcription level using HFFc6 total RNA-Seq (results) in 5-kb genome windows genome-wide using scipy.stats. The number of bins considered (*n*) and p values are provided in the results section. A two-sided test was performed because no directionality was assumed. A two-sided Mann-Whitney-Wilcoxon test was also used to assess the significance of disruption between genomic windows containing no Alu elements and windows with 5 or more (results, Figure 3C).

Motif significance

We evaluated the presence of CTCF in deleted and inserted transposable elements with overlap of CTCF ChIP-Seq, overlap of annotated CTCF motifs, and hamming distance to the canonical CTCF motif. Significance of a CTCF match was evaluated using FIMO from the MEME suite.[71]