PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Interpreting the CTCF-mediated sequence grammar of genome folding with AkitaV2

Paulina N. Smaruj[1], **Fahad Kamulegeya**[2], **David R. Kelley**[3], **Geoffrey Fudenberg**[1]*

1 Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, California, United States of America, **2** Stowers Institute for Medical Research, Kansas City, Missouri, United States of America, **3** Calico Life Sciences LLC, South San Francisco, California, United States of America

* fudenber@usc.edu

OPEN ACCESS

Citation: Smaruj PN, Kamulegeya F, Kelley DR, Fudenberg G (2025) Interpreting the CTCF-mediated sequence grammar of genome folding with AkitaV2. PLoS Comput Biol 21(2): e1012824. https://doi.org/10.1371/journal.pcbi.1012824

Editor: Maxwell Wing Libbrecht, Simon Fraser University, CANADA

Received: August 2, 2024

Accepted: January 24, 2025

Published: February 4, 2025

Data Availability Statement: Scripts used for cross-species AkitaV2 training and model weights are available at https://github.com/calico/basenji/tree/master/manuscripts/akita/v2. General utilities for AkitaV2 are available at https://github.com/Fudenberg-Research-Group/akita_utils. Code to reproduce analyses, including figures, are available at https://github.com/Fudenberg-Research-Group/akitaV2-analyses.

Funding: This work was supported by the National Institute of General Medical Sciences (R35GM143116 to GF, https://www.nigms.nih.gov/).

Abstract

Interphase mammalian genomes are folded in 3D with complex locus-specific patterns that impact gene regulation. CTCF (CCCTC-binding factor) is a key architectural protein that binds specific DNA sites, halts cohesin-mediated loop extrusion, and enables long-range chromatin interactions. There are hundreds of thousands of annotated CTCF-binding sites in mammalian genomes; disruptions of some result in distinct phenotypes, while others have no visible effect. Despite their importance, the determinants of which CTCF sites are necessary for genome folding and gene regulation remain unclear. Here, we update and utilize Akita, a convolutional neural network model, to extract the sequence preferences and grammar of CTCF contributing to genome folding. Our analyses of individual CTCF sites reveal four predictions: (i) only a small fraction of genomic sites are impactful; (ii) impact is highly dependent on sequences flanking the core CTCF binding motif; (iii) core and flanking nucleotides contribute largely additively to the overall impact of a site; (iv) sites created as combinations of different core and flanking sequences have impacts proportional to the product of their average impacts, i.e. they are broadly compatible. Our analysis of collections of CTCF sites make two predictions for multi-motif grammar: (i) insulation strength depends on the number of CTCF sites within a cluster, and (ii) pattern formation is governed by the orientation and spacing of these sites, rather than any inherent specialization of the CTCF motifs themselves. In sum, we present a framework for using neural network models to probe the sequences instructing genome folding and provide a number of predictions to guide future experimental inquiries.

Author summary

Mammalian genomes are spatially organized in 3D with profound consequences for all processes involving DNA. CTCF is a key genome organizer, recognizing numerous sites and creating a variety of contact patterns across the genome. Despite the importance of CTCF, the sequence determinants and grammar of how individual sites collectively instruct genome folding remain unclear. This work leverages the ability of Akita, a deep neural network, to make high-throughput predictions for genome folding after DNA sequence perturbations. Using Akita, we make several experimentally testable predictions.

The funder had no role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Competing interests: D.R.K. is an employee of Calico Life Sciences, LLC. All other authors have declared no competing interests exist.

First, only a minority of annotated sites individually impact folding, and flanking DNA sequences greatly modulate their impact. Second, multiple sites together influence folding based on their number, orientation, and spacing. In sum, we provide a roadmap for interpreting neural networks to better understand genome folding and important considerations for the design of experiments.

Introduction

Mammalian genomes contain thousands of CTCF sites, which collectively instruct locus-specific 3D genome folding [1,2]. These sites influence genome architecture through CTCF's ability to halt loop extrusion by the cohesin complex. Chromosome conformation capture technologies (e.g., Hi-C) have provided numerous insights into these mechanisms via their ability to generate genome-wide maps of pairwise contact frequency. Key features of mammalian Hi-C maps at the megabase scale include topologically associated domains (TADs), which appear as square regions of increased contact frequency, and dots (also termed loops), which appear as focally increased contact frequency between two distal anchor loci. The loss of TADs and dots upon the depletion of either CTCF or cohesin highlights the global importance of these factors [3,4]. Despite these findings, the impact and function of individual CTCF sites remain less understood.

Genome folding is thought to either insulate or facilitate physical interactions between regulatory sequences, thereby modulating enhancer-promoter communication [5]. Localized disruptions at specific CTCF sites have displayed clear changes to genome folding, gene expression, and development. Disrupted folding at the EPHA4 locus is linked to limb phenotypes [6], and changes to individual CTCF sites at the same locus are associated with index finger phenotypes [7]. Targeted deletion of CTCF sites at the Kallikrein locus led to coordinated gene activity associated with prostate cancer [8]. Nevertheless, disruptions to many CTCF sites have little or no impact; for example, [9] observed that the fusion of two adjacent TADs after deletion of multiple CTCF sites at the Kcnj2/Sox9 locus resulted in no detectable phenotype. An understanding of when sequence perturbations at TAD boundaries or CTCF sites would actually disrupt genome folding is a prerequisite for understanding downstream impacts of genome folding on gene regulation.

Which CTCF-binding sites are essential for genome folding, and why are they particularly crucial? Even the highest-resolution mammalian Hi-C and Micro-C assays are challenging to analyze below a resolution of 1kb, much larger than the approximately 20bp-long motif recognized by CTCF. Additionally, individual pixels at dots or boundaries often contain more than one predicted CTCF site. How do collections of CTCF sites instruct the formation of features on Hi-C maps? It has been shown that TAD boundaries are enriched with divergent CTCF sites [10], while dot anchors are enriched with convergent CTCF sites [11]. Altering the orientation of CTCF sites alters local patterns of genome folding [6,12–14]. While experimental techniques can now quantify the impacts of specific DNA sequence perturbations on genome folding and gene expression [15–18], testing throughput is currently limited to dozens of sites in a given genome context.

Computational approaches enable high-throughput screening of hundreds of thousands of flexibly-designed sequences *in silico*. Deep neural network models can now rapidly generate state-of-the-art predictions of genome folding from DNA sequences [19–22]. Using Akita, one of these models, we found that roughly 40% of nucleotides predicted to strongly impact local genome folding are located within the 100 nucleotides flanking CTCF sites [19]. However, these impacts remained poorly characterized. Methods to interpret neural networks trained on genomic data offer promising insight into sequence-based mechanisms [23]. Recent

applications included learning the motif syntax and grammar of transcription factors, such as Nanog [24], OSK, and AP-1 [25]. Together, this argues that further investigation of CTCF flanking sequences by interpreting trained neural networks can provide a deeper understanding of how DNA sequence determines 3D genome folding.

Here we update the Akita framework to jointly leverage mouse and human genome folding data and use this model to quantify the sequence contributions to locus-specific 3D folding (**Fig 1**). We screened millions of sequences *in silico* and quantified their predicted impact on genome folding. We observed a surprisingly low correlation between predicted disruption to chromosome structure upon mutating a CTCF site and CTCF ChIP-seq data, yet a high correlation with the frequency of binding measured by single-molecule footprinting experiments. By inserting thousands of strong CTCF sites into background sequences and assessing their impact, we identified a critical role of flanking sequences for determining the most significant CTCF sites for genome folding. We found that pairs of mutations within CTCF sites are largely additive and that the strength of CTCF site clusters depends on the number of sites, their orientation, and spatial arrangement. Finally, we found that CTCF instructs genome folding in a feature-agnostic manner rather than preferentially forming either dots or TAD boundaries. Collectively these results deepen our understanding of the sequence preferences and grammar through which CTCF contributes to genome folding.

Results

Cross species model

Cross-species training has been shown to improve predictions for sequence-to-profile neural network models for genomic datasets spanning multiple technologies and cell types by drawing upon the additional set of training sequences contributed by a second genome [26]. Following the cross-species Basenji model [26], we thus employed a similar approach to update Akita. We trained an ensemble of eight models, each with a distinct held-out subset of the two genomes. A benefit of the ensemble approach is that predictions that differ between models can be understood to come from limitations of training rather than from true biology. Each joint model was trained on 6 mouse and 5 human high-quality Hi-C and Micro-C datasets as targets (**S1A Fig**) with a slightly increased input sequence length (now 1.3Mb for AkitaV2, up from 1Mb). Trained models predict log2 observed/expected contact frequency at 2048bp resolution for any given input sequence. We observed a modest performance increase for AkitaV2 measured as the correlation of predictions with held-out test data (Pearson R = 0.66 vs. 0.61 previously, **S1B–S1E Fig**). This improvement prompted us to quantify whether increased performance overall also translated to better cell-type specific predictions. For individual regions, those with more accurate predictions also had more accurately predicted differences between cell types (**S1F Fig**). Still, on aggregate AkitaV2 displayed relatively limited cell-type specificity and only slightly more than AkitaV1 (**S2 Fig**). We hence focused analyses on aspects of genome folding that were congruent across cell types.

To quantify the influence of short DNA sequences on genome folding, we defined a disruption score as the square root of the sum of squared differences between predicted maps before and after local sequence perturbations (**Fig 1**), as previously used to interpret Akita's predictions [19,27]. This disruption score is sensitive to gain or loss of boundaries, as well as changes in TAD substructures [28,29]. We leveraged the ensemble of models to validate sequence perturbation approaches at individual predicted CTCF binding sites cataloged in JASPAR [30] by their cross-model stability. We refer to these genomic positions as 'sites' and define their subregions as the 'core motif,' 'upstream flank,' and 'downstream flank.' For AkitaV2, we found that masking (i.e. replacing nucleotides with zeros) was not consistent across models, while

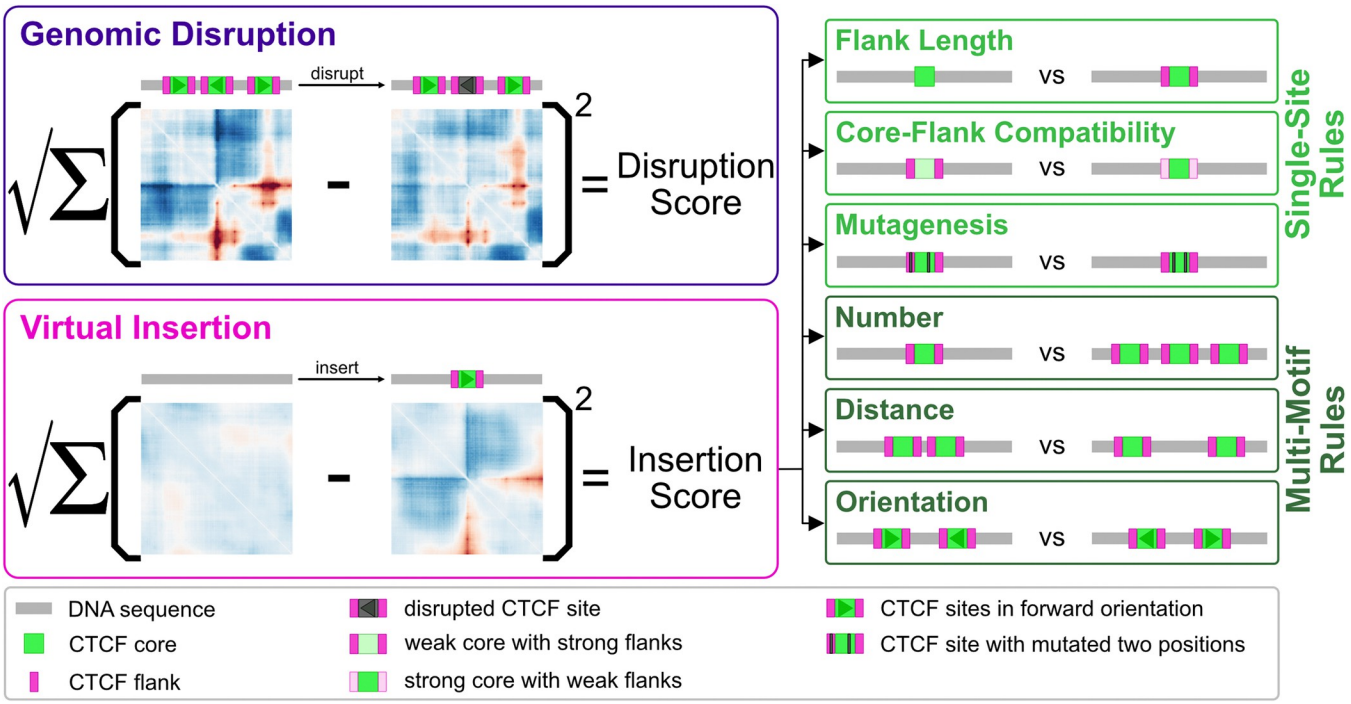

Fig 1. Utilizing AkitaV2 to extract CTCF-directed sequence preferences and grammar. Conceptual summary of the analyses performed in this study. *Left*: two main approaches: genomic disruption and virtual insertion. Genomic disruption involves permuting the nucleotides of a CTCF site within its genomic context, while virtual insertion entails inserting a CTCF site into a feature-less background sequence. Sequences and sites are shown as cartoon sequences with illustrative predicted maps below. *Right*: six types of virtual insertion experiments reveal distinct aspects of CTCF-site grammar. Three experiments tested single-site rules: (1) the impact of flanking sequences, (2) the compatibility between core motifs and their flanking sequences, and (3) nucleotide level mutagenesis. Three experiments tested multi-motif grammar: (4) varying numbers of CTCF sites within a cluster, (5) varying spacing between sites, and (6) varying site orientation. Cartoon sequences represent the parameters tested in these experiments.

https://doi.org/10.1371/journal.pcbi.1012824.g001

disruption by random permutation displayed several favorable properties. Predicted disruption scores by permutation are: (i) highly correlated across random permutations for any given CTCF site; (ii) consistent across models; (iii) robust regardless of whether the perturbed site is in the center or shifted by up to 10kb in the input DNA; (iv) preserved for the reverse-complement (S3A–S3D Fig). These observations argue that disruption by permutation is a robust strategy for extracting predicted impacts of various CTCF sites with Akita. Since disruption scores were highly correlated between human and mouse model outputs (Pearson R = 0.96, S3E Fig), we focused on predictions from the mouse model for subsequent analyses.

CTCF ChIP-seq provides poor prioritization for impactful sites as assessed by genomic disruptions. Using our updated model and *in silico* perturbation approach, we computed disruption scores for 9,991 CTCF sites profiled via single-molecule footprinting by Sonmezer et al. [31] and analyzed the relationship between these disruption scores and mouse epigenomic features associated with genome folding (Fig 2A). First, while each of these sequences were reported as predicted CTCF sites, we found only a moderate correlation between disruption scores and CTCF ChIP-seq levels (Fig 2B). This aligns with recent reports that insulator activity displays little or no relationship with CTCF ChIP-seq signal [15,17]. In contrast, we observe a high correlation with cohesin (RAD21) ChIP-seq and with CTCF site occupancy as measured by single-molecule footprinting (SMF) [31].

To quantify the relative importance of each epigenomic feature, we computed their partial correlations with our disruption scores. We observed CTCF SMF occupancy and cohesin ChIP-seq exhibited large positive partial correlations (Fig 2C), indicating they provide

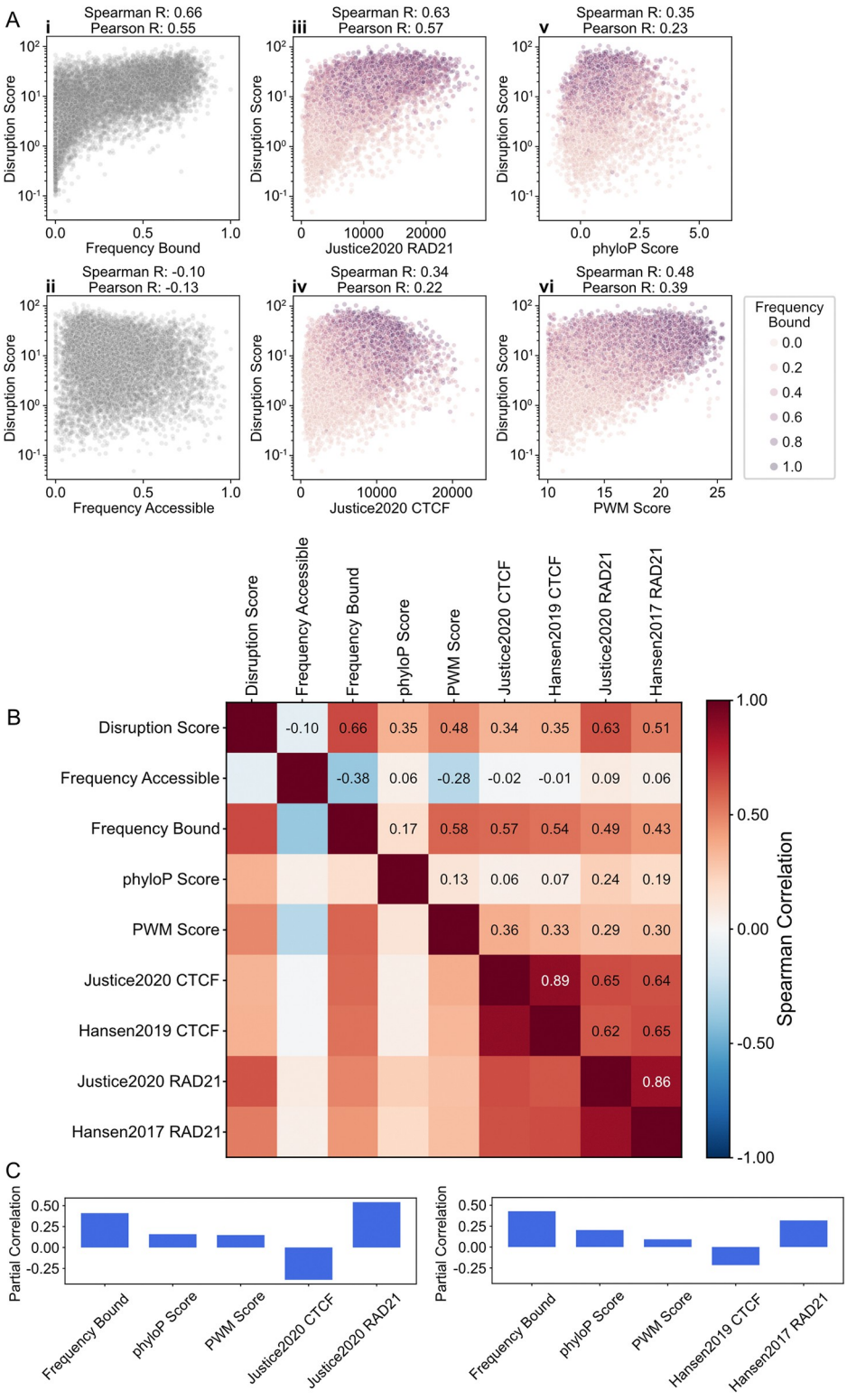

Fig 2. Disruption scores highlight impactful epigenomic features. A) Scatterplots showing disruption scores vs. genomic features at n = 9,991 autosomal CTCF sites, profiled with single-molecule footprinting (SMF) [31] which categorizes sites as: bound, nucleosome occupied, or accessible. We used the complete set of SMF-profiled CTCF sites and disrupted genomic sequence via permutation *in silico*. The first column displays disruption score vs. (i) frequency of being bound or (ii) accessible. The other subplots show disruption scores vs. following genomic features: (iii)

cohesin (Rad21) ChIP-seq signal [60], (iv) CTCF ChIP-seq signal, (v) conservation score (phyloP), and (vi) PWM score, with dots colored by their SMF bound frequency. ChIP-seq signal is quantified as the sum in a ±100bp window around each CTCF site. **B)** Matrix of pairwise correlations between disruption scores and genomic features of n = 9,991 autosomal CTCF sites. **C)** Partial correlation coefficients between disruption scores and subsets of genomic features from panel B, adjusting for mutual influences among these features. Partial correlations computed controlling for CTCF and cohesin ChIP-seq either from [60] (left) or [52,61] (right) are similar qualitatively and quantitatively.

https://doi.org/10.1371/journal.pcbi.1012824.g002

orthogonal sources of information for predicting disruptions. CTCF ChIP-seq, in contrast, displayed a negative partial correlation when accounting for the information from cohesin ChIP-seq and SMF CTCF occupancy. This argues that CTCF ChIP-seq provides redundant information once SMF occupancy is accounted for, and highlights the potential of emerging single-molecule chromatin profiling technologies [32,33] for better understanding 3D genome organization.

Virtual insertions probe CTCF influences independently of genomic context

One possibility for the low observed correlations between epigenomic signals and our CTCF disruption scores were the distinct genomic contexts of each site. For example, redundant TAD boundaries could mask the effects of CTCF perturbations in their native genomic context [6,7]. To quantify CTCF site impacts independent of their genomic context (**Fig 3A**), we developed a virtual insertion screening approach where a large set of genomic CTCF sites are inserted into neutral, largely featureless, background sequences, similar to previous approaches [24,27]. We then computed an "insertion score" for each sequence as the sum of squared differences for predicted maps before versus after the insertion (**Fig 1**).

To generate neutral background sequences, we found that shuffling by 8-mers led to relatively featureless maps (**Methods**, **S4A and S4B Fig**). Shuffling greatly reduced the variance of insertion scores while leaving the mean unchanged (**S4C Fig**), indicating that computationally inexpensive shuffling can reduce the number of neural network predictions needed to make a reliable estimate of a CTCF site's impact. We confirmed the robustness of our virtual insertion strategy by using the ensemble of models provided by AkitaV2. We found that insertion scores are highly consistent across backgrounds (Pearson R $>$ 0.99, **S4D Fig**) and across different models (Pearson R $>$ 0.96, **S4E Fig**).

To focus our analysis on sites with boundary-forming potential, we created a curated set of 7,560 CTCF sites from JASPAR that overlapped TAD boundaries from mESC Hi-C data [34]. Recognizing that predicted CTCF sites can be present yet not bound within repeat elements, like B2 SINEs in the mouse genome [35], we filtered out sites that overlapped repetitive elements. To prevent the inclusion of additional CTCF sites within extended flanking sequences we filtered out CTCF sites located less than 60bp from another site. By removing these potential confounders, we sharpened our focus on individual CTCF site impacts. We found a high correlation (Pearson R $>$ 0.99, **S4F Fig**) between human and mouse predictions of mouse CTCF virtual insertions, demonstrating that a cross-species model with high predictive power utilized very similar patterns at the DNA sequence level.

We observed a strong correlation between virtual insertion scores and genomic disruption scores (Pearson R $>$ 0.91) for the filtered CTCF sites (**Fig 3B**). Despite coming from regions specifying TAD boundaries, the vast majority of CTCF sites exhibited low disruption and insertion scores, and only a small fraction has a considerable predicted impact on genome folding. The great differences between CTCF sites that we predict helps understand experimental observations where deleting different CTCF sites, even within the same boundary, led to distinct outcomes for genome folding and gene expression [7,9,36]. Given the great number

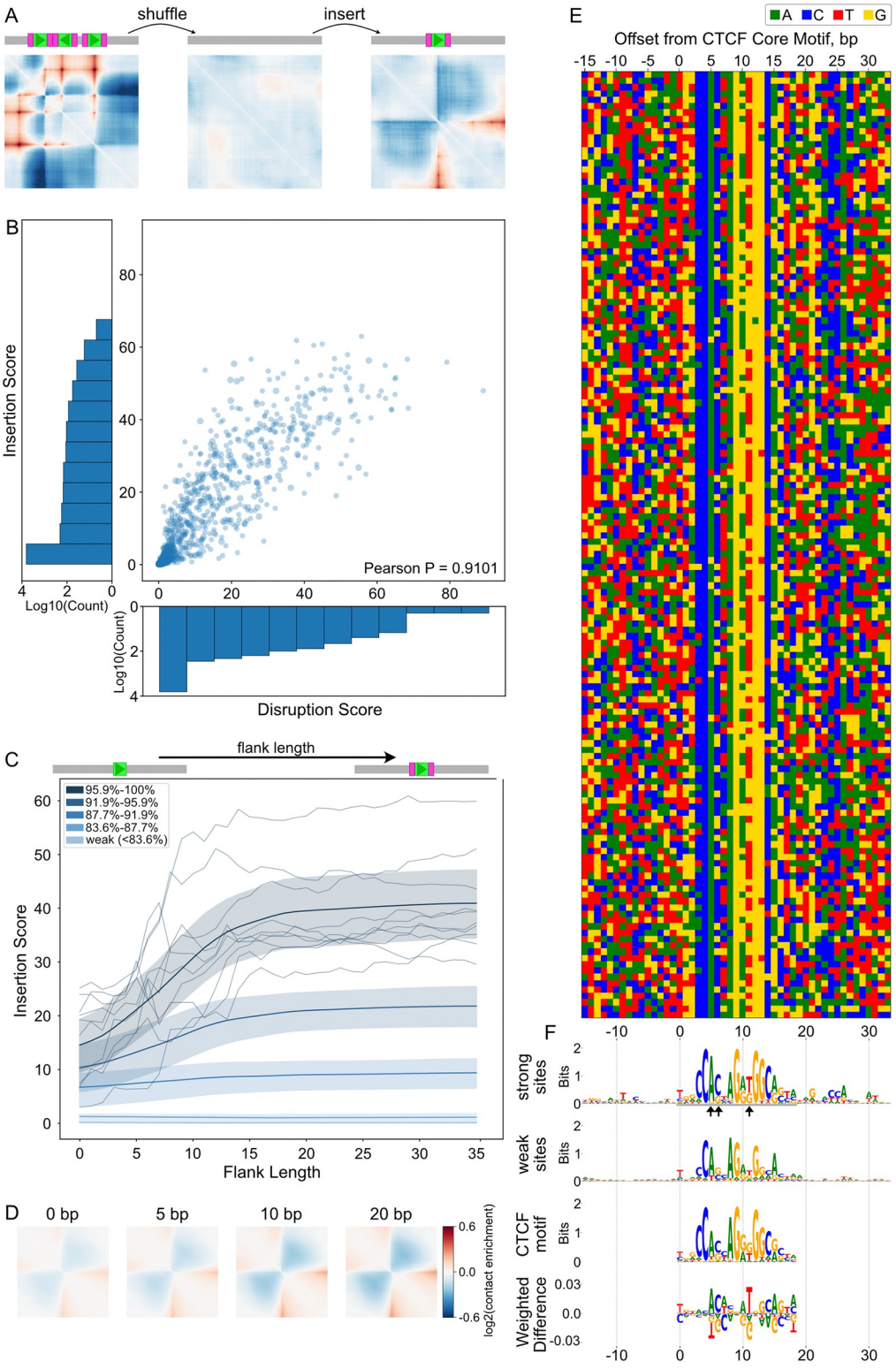

Fig 3. A virtual insertion strategy reveals the impact of flanking sequences. A) Virtual insertion strategy assesses individual CTCF site impacts. We generated background sequences by shuffling genomic sequences such that they produce mostly-featureless predicted maps. A CTCF site (green box) along with its flanking sequences (pink box) is then inserted into these background sequences (in gray). Using the sequence with an insertion as input, we generated predicted maps and quantified the impact as an insertion score. **B)** Scatterplot of insertion versus disruption scores, for

n = 7,560 CTCF sites (PearsonR > 0.91). Sites were obtained by intersecting sites from JASPAR with mESC boundaries and filtering for lack of overlap with repetitive elements within +/- 20bp or other CTCF sites within +/- 60bp. Scores were averaged across all six mouse outputs (i.e. cell types) and all eight models. Insertion scores were additionally averaged over ten background sequences. Histograms show log density along each scatterplot axis, as the majority of sites exhibit both low insertion and disruption scores. Given this, for further analysis we selected the 1250 sites with the highest disruption scores and chose an additional 250 sites randomly from the remaining pool. **C)** Flanking sequence length versus insertion score for the analysis set of n = 1,500 CTCF sites. Flanking sequence was varied from 0bp (19bp core motif only) up to 35bp, depicted as cartoons above the plot. Genomic flanking sequences were symmetrically extracted around each CTCF site. For visualization, sites were divided into five groups based on their insertion score with 30bp flanks. Smoothed lines show the mean for each group, and shaded bands show the 25th to 75th percentiles. To illustrate the variability among sites, we show 10 sites chosen randomly from the strongest group as navy lines. **D)** Predicted contact maps illustrate the impact of increasing flanking sequence lengths for a strong CTCF site. Sequence of inserted CTCF site and flanks obtained from chr15:101,984,508–101,984,527 in the mouse genome. **E)** Heatmap of nucleotide composition around 150 strong CTCF sites (±15bp). Rows ordered by insertion score. **F)** Sequence logos for the sequences with top 150 and bottom 150 insertion scores highlight core motifs and flanking preferences. A gray underline indicates the CTCF core motif in the top 150 logo, with black arrows pointing to positions 6,7, and 12. The CTCF consensus logo from JASPAR (MA0139.1) is aligned below the logos for visual comparison, and the weighted Jensen-Shannon difference between the top 150 CTCF core sequences and the CTCF consensus is displayed for visual comparison of sequence preferences.

https://doi.org/10.1371/journal.pcbi.1012824.g003

of weak sites, for deeper analysis we focused on a set of 1500 sites, including 1250 sites with the highest scores and 250 sites picked randomly from the remaining pool.

Flanking sequences modulate the influence of CTCF sites on genome folding. We next tested how flanking sequences around CTCF sites influenced predicted insertion scores. We defined core motifs as the 19bp-long sequences from JASPAR (MA0139.1), and flanking sequences as the genomic sequences up- and downstream of the core. We inserted individual CTCF sites into neutral background sequences, incrementally extending the flanking genomic sequence around the core motif. We observed that average insertion scores rose sharply with increased flank length up to about 15bp before stabilizing (**Fig 3C and 3D**). Our finding is consistent with experimental observations that highlight the importance of flanking sequences for transcriptional insulation [15,17], binding [37,38], and accessibility [39]. We repeated the flanking sequence insertions with two copies of each CTCF site and observed a similar trend for all four possible orientations (**S5B–S5D Fig**). This argues that the impact of flanking sequence on an individual CTCF site strength is independent of other nearby sites.

Recognizing the strong contribution of flanking sequences, we searched for sequence preferences within these regions. We generated sequence logos for CTCF sites with the highest and lowest insertion scores. In the flanking regions of sites with low insertion scores, we found little sequence preference (**Fig 3F**). In contrast, flanking regions around strong CTCF sites displayed sequence preferences 2 to 13 bp downstream and -15 to -7 bp upstream of the core motif, though these were more subtle than the core motif itself. Influential flanking sequences for genome folding, as predicted by AkitaV2, aligned well with previously documented CTCF binding preferences upstream and/or downstream of the core motif derived from ChIP-seq, MNase-seq, and ChIP-exo [37–42]. The sequence preferences we observed differed from those reported by [15], possibly due to a limited number of experimentally tested sites. Using the weighted Jensen-Shannon difference, we compared the CTCF consensus motif from JASPAR with the sequence logo for sites with the highest insertion scores (**Fig 3F**). This revealed that high scoring sites had relative preferences for A at position 6, C at position 7, and T at position 12, the last of which was also noted in [27]. Finally, our virtual insertion and genomic disruption scores produced largely congruent motif logos (**S5A Fig**).

We hypothesized that subtle sequence preferences might reflect an average over distinct binding modes. We explored multiple methods to order and cluster flanking sequences around strong sites, including by: overall insertion scores (**Fig 3E**), the Hamming distance between upstream, downstream or combined upstream and downstream sequences. None of these revealed clear

clusters of sequence preferences (S6A–S6C Fig). We also performed motif enrichment analysis and *de novo* motif discovery in the flanking sequences with Homer [43]. Neither approach yielded a widely-prevalent flanking sequence motif; the most prevalent *de novo* motif occurred in 20% and the most prevalent known motif in 10% of flanks. Still, strong flanking sequences displayed clear differences in position-wise GC content both upstream and downstream of the core (S6F Fig). We thus tested if strong versus weak flanking sequences were differentially enriched for short k-mers (for k = 2, and k = 3, S6D–S6H Fig). Generally, GC and CG rich k-mers were favored in strong flanks, and AA and TT rich k-mers were more abundant in weaker sequences. However, few k-mers varied more than 2-fold in frequency, and hence enrichment or depletion of any individual k-mer was insufficient to distinguish strong versus weak flanks. Together, our results point towards flanking sequences around CTCF having an important impact on genome folding, albeit one that is not readily characterized as a single position weight matrix.

Core motifs and flanking sequences are broadly compatible. Given the absence of prevalent motifs in the flanking sequences themselves, we investigated whether flanking sequence impacts are contingent on their associated core motif sequences. We assessed predicted core-vs-flank compatibility using 300 CTCF sites classified as strong, medium, or weak based on their overall insertion scores. We computed insertion scores for all pairs of core and flanking sequence combinations, and visualized the resulting matrix (Fig 4A). If compatibility was an important factor, insertion scores for cognate core-flank pairs from the genome would be stronger than synthetic combinations. We found no evident core-flank compatibility: there was no evidence of higher scores for genomic cognate pairs (i.e. no strong diagonal in the pairwise matrix, Fig 4D), and no clear deviation of the distribution of scores for cognate versus synthetic pairs (Fig 4B). Weak cores paired with strong flanks yielded relatively weak sites, while strong cores with weak flanks performed almost on par with medium cores paired with medium flanks. These predictions agree with experiments that quantified the impact of core and flanking sequence from strong versus weak sites on transcriptional insulation [15]. We performed singular value decomposition (SVD) to factorize the matrix of pairwise core-flank combinations into a set of vectors that described the influence of each core or flanking sequence. We found that the product of the first SVD factors for cores and flanks nicely approximated the insertion score for the corresponding core-flank combination (Fig 4C and 4E). This indicates that the rules learned by AkitaV2 for combining core and flanking sequences are largely multiplicative without strict constraints on their compatibility.

Individual nucleotides within a CTCF site contribute additively to genome folding patterns. After observing no strict compatibility requirements at the level of entire core and flanking sequences, we next explored compatibility at the nucleotide level. Given that each zinc finger domain of CTCF is thought to recognize a triplet of DNA base pairs [44], we hypothesized that pairs of mutations within the same triplet might be more detrimental than those spanning different triplets. This would create blocks of low mutation scores along the diagonal of a pairwise mutation matrix. We conducted pairwise mutagenesis on 100 strong CTCF sites that displayed the highest insertion scores. Contrary to our initial hypothesis of an epistatic effect within zinc finger triplets, we observed no clusters of heightened pairwise impacts along the diagonal (Fig 5A). To better understand pairwise impacts, we tested if their impacts deviated from simple additivity. To generate the additive expectation, we performed saturation mutagenesis for single mutations (Fig 5B and 5C) and summed together the scores for pairs of mutations. For weak mutations, the pairwise mutational impact was congruent with the additive expectation. After a strong mutation, however, the impact of additional mutations saturated and diverged from the additive expectation (Fig 5D), suggesting negative epistasis [45,46]. We hypothesize that this arises because the strongest mutations abolish CTCF binding and hence ability to impact genome folding.

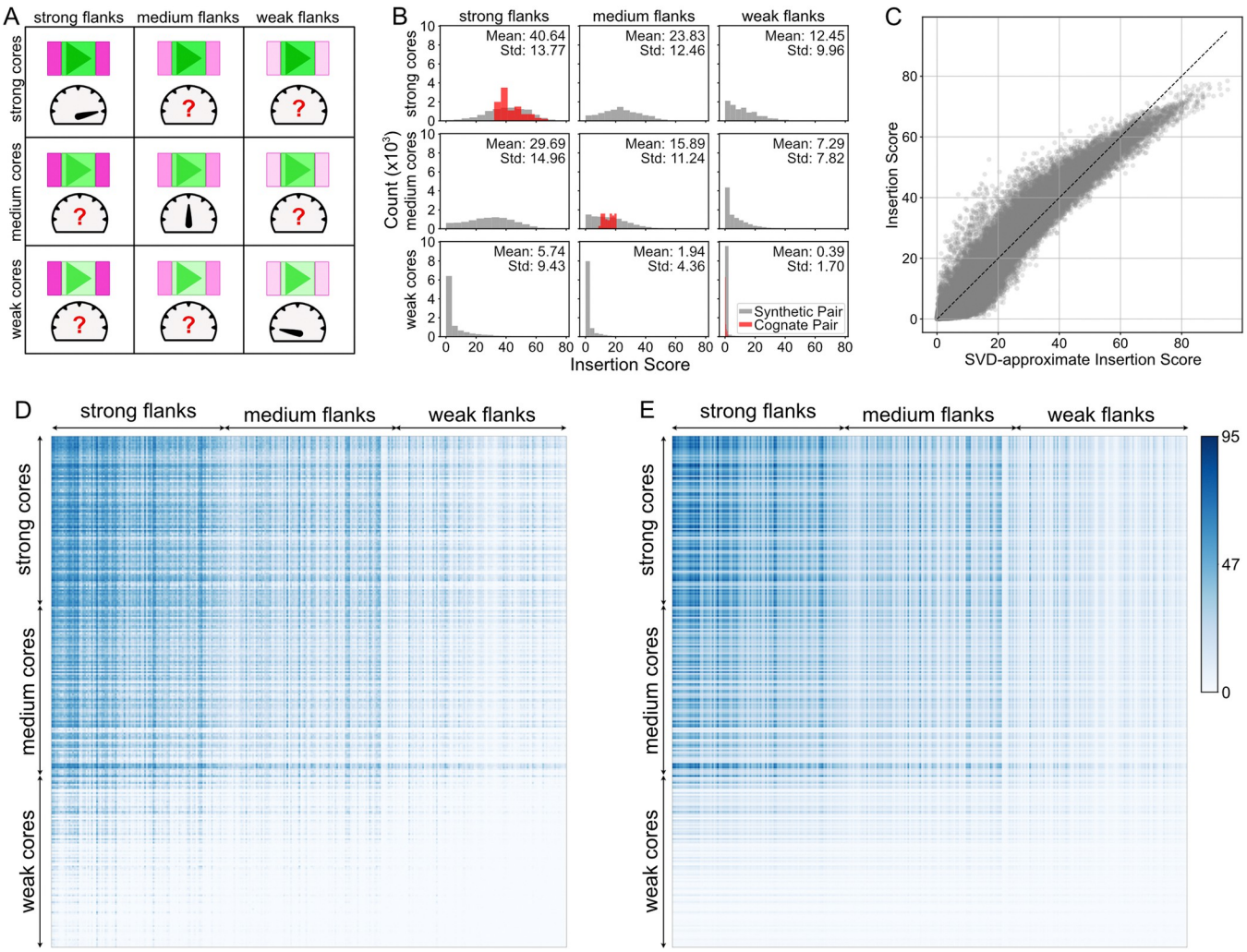

Fig 4. CTCF core and flanking sequences are broadly compatible. A) Illustration of the test for compatibility between core and flanking sequences by assessing all possible combinations of flanks and cores classified into three strength groups. Each row represents a distinct 19bp core motif sequence and each column represents a distinct pair of 30bp flanking sequences adjacent to the core motif. **B)** Distributions of insertion scores for pairs of core and flanking sequences around 100 strong, 100 medium, and 100 weak CTCF sites. Each histogram shows 10,000 (100^2) combinations. Sites were classified as strong, medium, and weak based on their combination of core and flanking sequence seen in the mouse genome. Distributions for original genomic core-flank pairs shown in red (with count scaled by 100), synthetic combinations shown in gray. **C)** Scatterplot of insertion scores (panel D) versus approximate values obtained through SVD (panel E). Their high correspondence indicates that predicted strengths are largely multiplicative and core and flanking sequences are largely compatible. **D)** Heatmap of insertion score for 300 CTCF core and 300 flanking sequence pairs. Each row corresponds to a different core sequence, while each column represents a different flanking sequence. Rows and columns are ordered by the insertion score of the core-flank combination that occurs in the genome (i.e. by values along the diagonal). **E)** Heatmap of approximate insertion strength $M_{i,j}$ obtained via SVD for 300 CTCF core and 300 flanking sequence pairs. $M_{i,j}$ represents the predicted insertion strength for the combination of the core *i* and flank *j*. Using SVD, $M = U\ D\ V^T$, we found $M_{i,j} \approx D^0\ U^0{}_i\ V^0{}_j$, where U^0 and V^0 capture core and flank strengths, respectively. Rows and columns are ordered as panel D.

https://doi.org/10.1371/journal.pcbi.1012824.g004

The positioning and orientation of multiple CTCF sites specifies a broad range of folding patterns

After quantifying the sequence preferences of individual sites, we next turned to deciphering multi-motif CTCF grammar by systematically varying their: (i) number, (ii) spacing, and (iii) orientation.

We found that larger numbers of inserted sites produced correspondingly larger predicted insertion scores (Fig 6A). This aligns with experimental observations that insulation strength

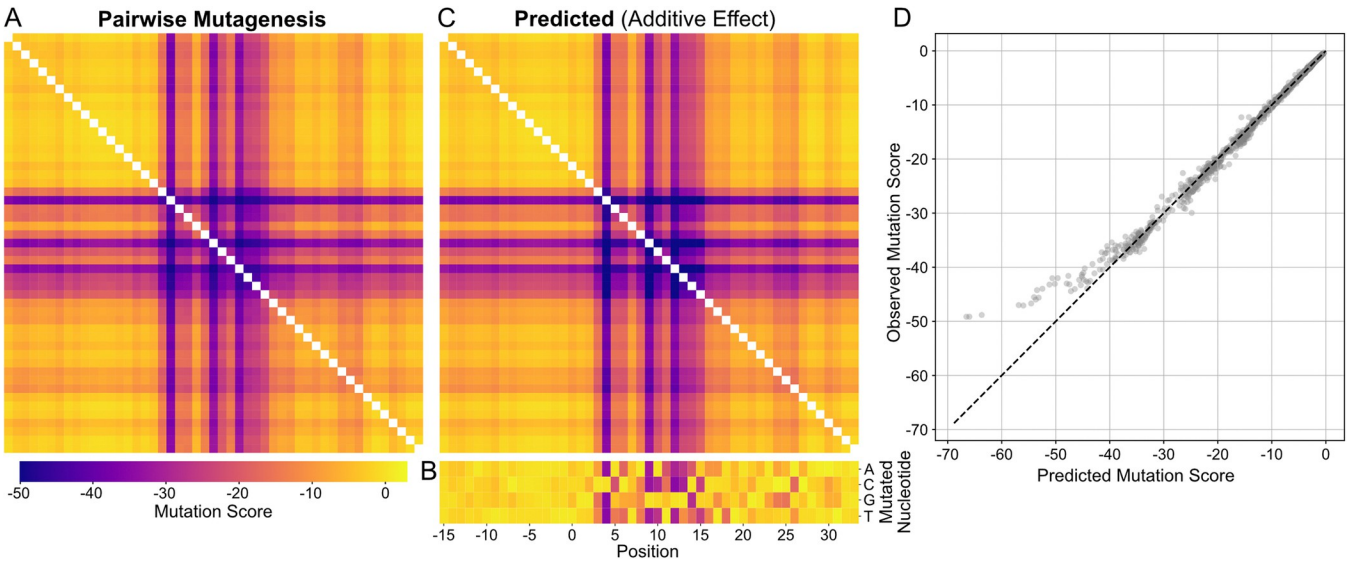

Fig 5. Pairwise nucleotide dependencies are largely additive in core CTCF motifs and their flanking sequences. A) Pairwise mutagenesis. Mutation score for pairs of mutations in the 19bp core motifs +/- 15bp flanks for 100 strong CTCF sites. Mutation score is calculated as the difference between insertion scores for the mutant versus the unperturbed sequence. The heatmap shows the average mutation score for each pair of positions. **B)** Single-nucleotide saturation mutagenesis of the 19bp core motifs +/- 15bp flanks for the same sequences in A. The heatmap presents the average over all CTCF sites for each possible substitution. **C)** Predicted additive impact of pairwise mutagenesis. The predicted additive pairwise impact is the sum of the average single-nucleotide impacts (panel B). Note the shared color scale across panels A-C. **D)** Scatterplot of predicted additive and observed pairwise mutagenesis effects from panels A,B. For pairs of weak mutations, impacts are largely additive (up to mutation scores of -40). Higher impact mutations (i.e. more negative mutation scores) appear to saturate and diverge from this linear trend.

https://doi.org/10.1371/journal.pcbi.1012824.g005

increases with more inserted CTCF sites [15] and that stronger TAD boundaries contain greater numbers of CTCF sites [47,48]. We found that dosage-dependent insulation requires strong CTCF sites, as the insertion scores for weak sites remained low. Similarly, experiments show that tandem arrays of CTCF sites from non-boundary regions do not function as insulators [15]. Our observation indicates that clusters of strong CTCF sites could be used to make stronger TAD boundaries.

We explored how genomic context modulates CTCF grammar by inserting hundreds of strong pairs of sites at varying genomic distances and in four distinct orientations. The most notable differences in predicted maps occurred between convergent and divergent orientations. Divergent sites exhibited their highest impacts at a relatively short distance (~70kb) before declining, whereas convergent sites displayed high impacts at two distances, including at a much greater secondary distance (~600kb) (**Fig 6B**). Visual inspection indicated that this second maxima corresponded to dot patterns in the predicted maps (**Fig 6C**). We also observed an initial dip in insertion score at an inter-motif distance of ~170bp, consistent across all four orientations (**Fig 6D**). Interestingly, this is close to the nucleosome repeat length estimated from MNase-seq [40], and strong CTCF sites are often flanked by arrays of phased nucleosomes [49,50]. The profiles for tandem left and tandem right sites are closely aligned, confirming the model's strand independence. These observations demonstrate the complexity of genome folding, as even two CTCF sites can generate a diversity of features within predicted maps (**Fig 6C**).

Motivated by the distinct maps for convergent versus divergent pairs of CTCF sites, we tested whether there are categories of CTCF sites that exhibit feature specialization. To test this hypothesis, we designed an *in silico* screen using pairs of sites positioned in two distinct scenarios: (i) convergent sites (><) spaced 400kb apart, to test their dot-formation ability; (ii)

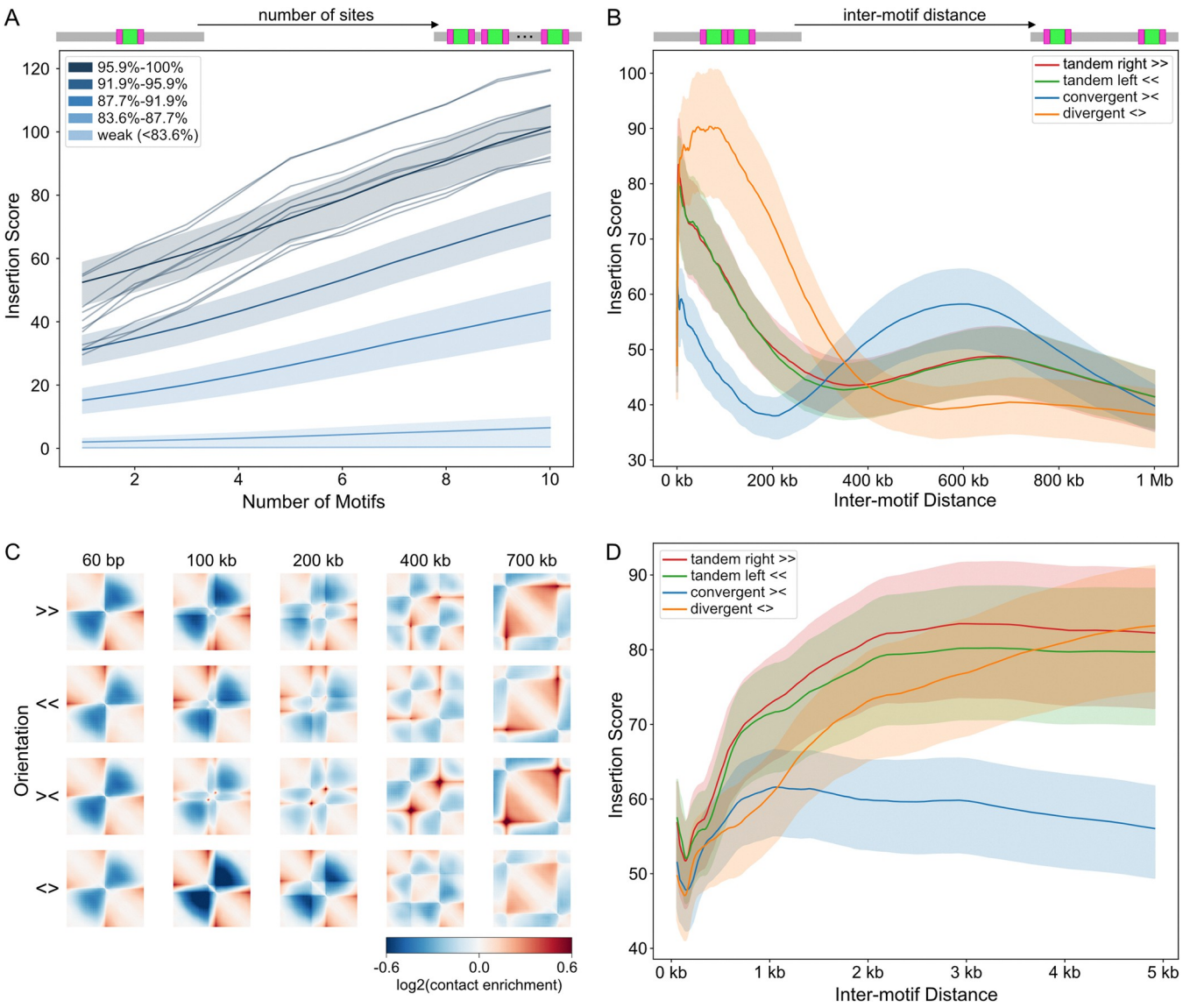

Fig 6. CTCF grammar depends on number and spacing. A) Insertion score versus number of inserted CTCF sites. Averages over five groups of n = 1,500 CTCF sites plotted as in Fig 3C. Shaded areas indicate 25-75th percentile for each group. Variability among sites is highlighted using 10 randomly chosen sites from the strongest group (dark navy lines). All sites inserted in a rightward orientation with 30bp flanks and 180bp spacing between cores. Note that with this spacing, 10 inserted sites constitute about 2kb or one bin. **B)** Insertion score as a function of spacing for four possible orientations for 300 CTCF sites (the strongest 20% from A), also with 30bp flanks. Average across sites shown for each orientation, with variability indicated by 25–75 percentile bands. **C)** Predicted log-transformed observed/expected contact frequency maps by CTCF pair orientation and spacing for insertion of a representative CTCF site (sequence from chr2:93,199,043–93,199,062). **D)** As in B), zooming into the 0-5kb region.

https://doi.org/10.1371/journal.pcbi.1012824.g006

divergent sites (<>) with a smaller spacing of 180bp, to test their boundary-formation ability (Fig 7A). From predicted maps, boundary strength was estimated using the insertion score, while dot strength was assessed as the enrichment between dot anchors versus surrounding regions (Methods). Collectively, CTCF sites followed a consistent trend for dot versus boundary strength (Fig 7B and 7C), arguing against the feature specialization hypothesis. We also found that insertion scores from boundary scenarios were the same magnitude as those for dot scenarios (S6A Fig), suggesting that while the number of sites influences the overall strength of the predicted map, the spacing and orientation determine which features are visible. To rule

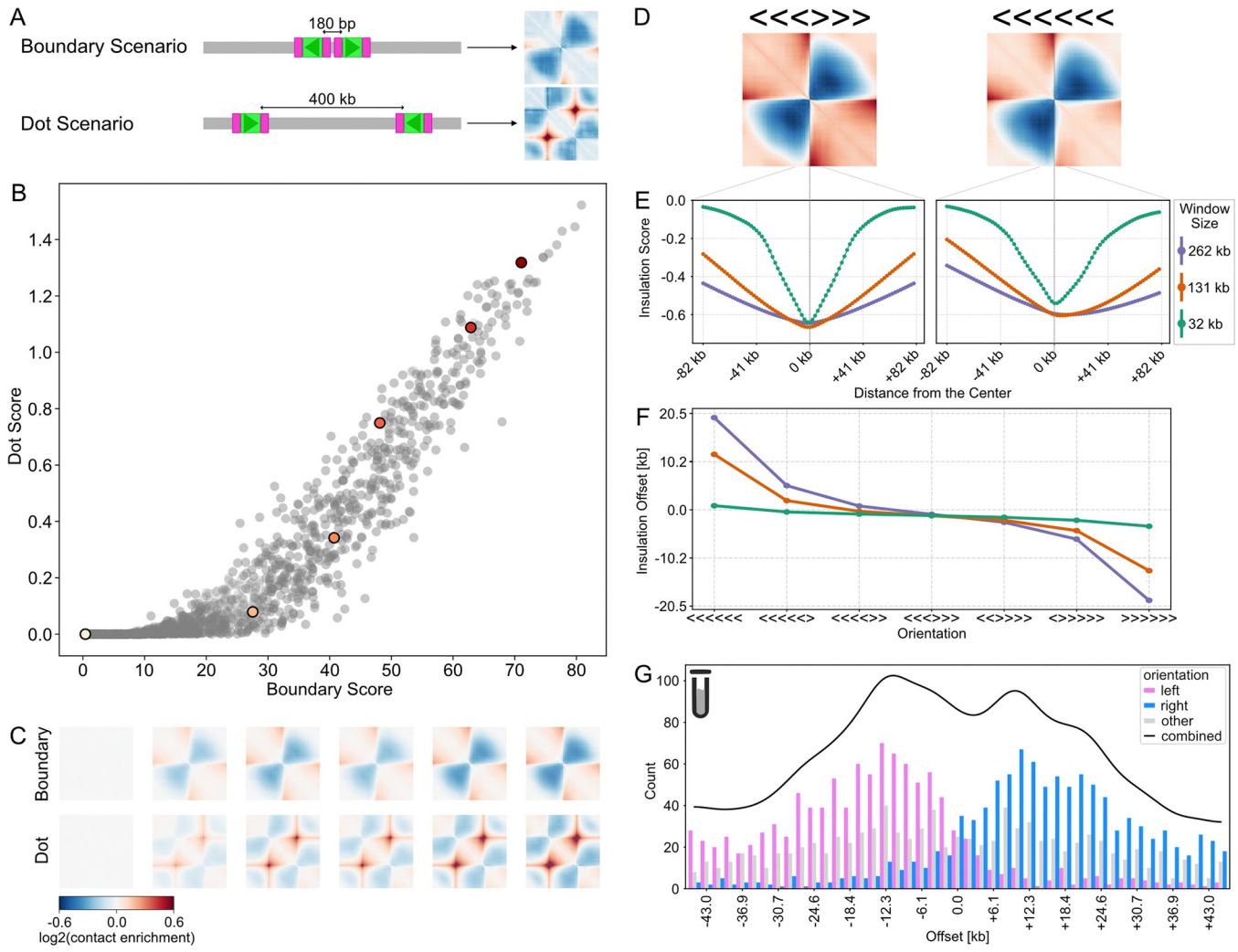

Fig 7. CTCF sites do not mediate feature-specific genome folding. A) Illustration of the test for CTCF feature specialization using two distinct layouts: (i) a 'boundary' with two divergent sites, 180bp apart, versus (ii) a 'dot' with two convergent sites, 400kb apart. CTCF insertions are shown as green rectangles (core motifs) with pink flanks (30bp), arrows indicate the orientation of the CTCF site. **B)** Scatterplot of boundary vs. dot strength (n = 1,500 CTCF sites). Boundary strength is the overall intensity of the map; dot strength is the local average signal within versus around the dot. Six CTCF sites spanning a range of strengths are highlighted with colored dots. **C)** Predicted maps for boundary and dot scenarios for the six highlighted CTCF sites in panel B. **D)** Predicted maps for symmetric (<<<>>>) and asymmetric (<<<<<<) insertions for a cassette of six CTCF sites into the middle of a background sequence. **E)** Insulation scores calculated using sliding diamond windows of three sizes (32.7kb, 131kb, 262.1kb), shown for the central 163,84kb of the map. Note that insulation minima display an offset for the asymmetric case, and the same coloring used for window sizes in E and F. **F)** Insulation minima offset for indicated CTCF cassette insertions. Insulation offset is the position of the insulation score minima relative to the center of the sequence (window sizes of 32.7kb, 131kb, 262.1kb). Each point represents the average across 100 strong CTCF site insertions. Note the insulation offset increases with the asymmetry of the inserted CTCF site configuration and is more pronounced for larger window sizes. **G)** Histogram of bins around disruption-sensitive TAD boundaries stratified by their orientation. To obtain bin orientation, we: disrupted sequences in non-overlapping 2048 bp bins around each TAD boundary, took the bin with the highest disruption score, and assigned a left (pink) or right (blue) orientation if all CTCF sites within a bin are aligned in the same orientation. Bins without assigned orientations shown in grey. The black line shows the smoothed total number of disruption-sensitive bins (across all orientations).

https://doi.org/10.1371/journal.pcbi.1012824.g007

out bias from our selection of inserted sites, we repeated the analysis with CTCF sites specifically overlapping dot anchors, as identified by *MUSTACHE* [51]. We found a largely similar trend in terms of dot versus boundary strength (**S6B Fig**). Given we did not find sets of CTCF sites with strong preferences for dot versus boundary formation, we conclude that individual sites have a versatile role in specifying chromatin architecture without feature-specific specialization.

For two CTCF sites inserted in tandem, we observed a slight asymmetry in the predicted maps of co-oriented sites (**Figs 6C and S5D**). When we inserted an asymmetric cassette of six strong sites, we observed even more asymmetry (**Fig 7D**). To investigate this quantitatively, we examined the insulation profile of the resulting contact maps. To our surprise, for this cluster of six co-oriented sites, the insulation minima did not align with the center of the inserted cluster, and the misalignment increased with larger insulation window sizes (**Fig 7E**). By inserting cassettes with various orientations of the six sites, we found that both the direction and magnitude of the insulation offset depended on the asymmetry of the inserted cluster (**Fig 7F**). In summary, our analysis revealed that the orientation of CTCF site insertions significantly affects the asymmetry and insulation properties of predicted maps.

Influential boundary sequences are often outside insulation minima. Motivated by the potential mismatch between insulation minima and influential CTCF sites in synthetic sequences (**Fig 7D–4F**), we returned to experimental mESC insulation profiles to determine if this offset might be relevant there as well. We computed the predicted sensitivity to disruption using non-overlapping 2048bp bins tiled across the ±45kb region around 4,474 mouse boundaries determined with standard thresholds on the insulation profile (**S7C Fig**). Few boundaries were sensitive right at the insulation minima (~6%, ±5kb). In contrast, nearly half of boundaries were sensitive in a nearby bin (43.4%, ±28kb), with visually confirmed loss of TAD boundaries in predicted maps (**S7E Fig**). The remaining 50.6% of boundaries with low disruption scores were enriched for missing data, depleted for sequencing read counts, and had poorer predictions for the reference targets (**S7D and S7F–S7H Fig**). For each sensitive boundary we selected the bin with the highest disruption score and determined its orientation. We overlapped these bins with CTCF sites from the JASPAR database, and assigned a left or right orientation to bins where all sites were co-oriented. Notably, bins around sensitive boundaries had a distinct orientation asymmetry that mirrored the offset seen in the virtual insertions (**Fig 7G**). Left-oriented bins were enriched upstream of boundaries, whereas right-oriented bins were enriched downstream of boundaries. Bins where an orientation was not assigned were distributed more uniformly relative to the boundary. These results suggest that sequences specifying TAD boundaries may not always align with insulation minima and the design and interpretation of experiments perturbing TAD boundaries can benefit from predictions generated by deep learning models.

Discussion

In summary, we utilized an updated version of the Akita deep neural network to screen millions of *in silico* DNA sequence perturbations for their impact on genome folding. We quantified the predicted sequence preferences and grammar of CTCF sites with insertion and disruption scores.

For individual CTCF sites, we found a surprisingly low correlation between CTCF ChIP-seq data and *in silico* disruption scores. This aligns with recent observations that CTCF ChIP-seq does not correlate with differential insulation activity [15,17]. Similar to [17], we observed only moderate correlations between the effectiveness of CTCF sites and their resemblance to the core motif or their degree of conservation. In contrast, we observed a higher correlation between impact and CTCF occupancy by SMF. We speculate that this either stems from technical limitations of ChIP-seq for CTCF, or that CTCF ChIP-seq quantifies a population of CTCFs that are chromatin associated, yet not productively bound in a way to engage cohesin, e.g. via their RNA binding domain [52]. Indeed, in simulations of loop extrusion, sufficient CTCF occupancy is required to observe accumulation of cohesin at barriers and impact on genome folding [53]. Thus, the strong correlation between disruption scores and cohesin

ChIP-seq that we observe is consistent with a large role of cohesin-mediate loop extrusion for organizing genomes at the megabase scale.

Our analysis highlights the role of 15bp flanking sequences for specifying strong CTCF sites for genome folding. This generalizes and refines experimental observations about the importance of flanking sequence around core CTCF motifs for insulation activity [15,17]. Strong sites displayed sequence preferences both upstream and downstream of the core motif, albeit with lower information content. While not pronounced enough to be readily extracted by motif-discovery algorithms, the flanking sequences identified by AkitaV2 show similarities to those reported as important for CTCF binding and DNA accessibility at CTCF sites [37–42,54,55]. Similarly to us, [41] and [40] used neural network models to extract sequence preferences around CTCF sites, albeit starting from predicted DNA accessibility instead of predicted genome folding. Differences with sequence preferences reported in [15] likely come from limitations to the number of sites that could be experimentally assayed. [56] reported that accessible sites bound by CTCF can be differentiated from unbound sites by the enrichment of transcription factor binding sites in close proximity to the CTCF motif. However, if co-binding does occur, our results argue for a variable and versatile set of transcription factors.

Our observations argue for two principles of CTCF multi-motif grammar: (i) boundary strength is influenced by the number of CTCF sites within a cluster, and (ii) pattern formation is determined by site orientation and spacing, without intrinsic specialization of individual CTCF sites. The first principle aligns with positive correlations between the number of sites and TAD boundary strengths observed in genomic data [47,48], as well as the number of sites and impact on transcription in synthetic sequences [15,16]. The second principle aligns with the correspondence between convergently oriented sites and dots [4] versus divergently oriented sites and boundaries [10] in the genome, as well as orientation-specific impacts on transcription in synthetic sequences [16]. Indeed, we predict that CTCF sites do not preferentially specify dots or boundaries. Experimentally, this is supported by the emergence of new dots between pairs of loci that previously displayed boundaries after the deletion of intervening CTCF sites [7,9]. More broadly, we predict that specific contact patterns emerge from the positions and arrangements of CTCF sites rather than specialized pattern-specific motifs.

A central limitation of our approach is that the sequence preferences and grammar we can identify must have been extracted by the deep neural network we use, AkitaV2. Model performance, architecture, and training scheme could each contribute to what can be learned via our approach [23,57]. While including training data from more cell types and species, we found that AkitaV2's overall performance and cell-type specificity only slightly improved relative to AkitaV1. New training strategies, including two-stage approaches [58], and transfer learning more broadly [58,59], present opportunities to make predictions more cell type-specific and enhance overall performance. Indeed, a two-stage training procedure, where models are first trained on average accessibility and then fine-tuned for specific cell types, has been beneficial for ATAC-seq models but has not yet been explored for contact map models. A further limitation of the current architecture is that the input sequence length restricts the maximum distance over which any grammar can be extracted. Only patterns that occur with sufficient regularity have a chance of being learned by the model. For example, repetitive elements that are rare or make species-specific contributions to genome folding are unlikely to be reliably extracted by our approach. Similarly, because large clusters of CTCF sites are relatively rare in the genome (e.g., only 0.36% of boundaries have >10 CTCF sites), our approach may overestimate their impact. Finally, while we primarily used feature-agnostic scores over the full predicted megabase region, feature-specific scores [29] could enable extraction of additional insight.

Collectively, our observations provide a roadmap for the design of experiments hoping to test the sequence determinants of genome folding and downstream consequences for communication between enhancers and promoters. Successful experimental designs will consider both the content of core and flanking CTCF sequences as well as their positioning relative to other regulatory sequences.

Methods

Data preprocessing

We followed the preprocessing described in prior research using the Akita framework [19] for the 6 mouse and 5 human datasets in **S1 Table (data from [4, 11, 34, 62, 63]).** Briefly, we reprocessed these datasets using the distiller pipeline (https://github.com/open2c/distiller-nf, [64]), extracting contacts with pairtools [65], binning each dataset to 2,048bp cooler files (https://github.com/open2c/cooler, [66]) and performing genome-wide iterative correction [67]. Individual target matrices were extracted from genome-wide cooler files for regions corresponding to 1,310,720bp of input sequence, 25% larger than the original 1,048,576bp. As previously, the following steps were applied to matrices for individual regions in the training and test sets: adaptive coarse-graining, normalization for distance-dependence, natural log, clipping to (-2,2), linear interpolation of missing bins, and convolving with a small 2D gaussian filter. The first and third steps used cooltools (https://github.com/open2c/cooltools, [68]).

Cross-species model

We used the same neural network structure and weights as described for Akita, with the following modification to the last layer: instead of a single dense layer, either a 5-unit dense layer was appended for predicting the 5 human targets or a separate 6 unit dense layer was appended for predicting the 6 mouse targets. We implemented this model using TensorFlow [69]. See https://github.com/calico/basenji/tree/master/manuscripts/akita/v2 permission for full specification of model weights, learning rate, and other hyperparameters.

Cross-species training

As for cross-species Basenji training [26], we aimed to avoid leakage between training and test sets by jointly assigning orthologous human and mouse sequences to the same training, validation, or test fold. Briefly this involved: dividing the genome in 5 Mb regions, constructing a bipartite graph if they have >500kb of aligning sequence, and partitioning connected components into 8 distinct folds. We trained an ensemble of models, in which model i used fold i as its test set, i+1 as its validation set, and the remaining folds as its training set. Weights for each model were updated depending on the combined loss over all targets; the only difference between models is which portion of the genome falls into the training, validation and test set. A benefit of this approach is that any differences for predicted perturbations between models would be related to limitations of training rather than true biology. During training, we alternated between batches of 2 human and 2 mouse sequences and Hi-C targets, updating weights for the corresponding final mouse or human dense layer. We trained using stochastic gradient descent with 0.98 momentum and the 1cycle learning rate schedule, in which the learning rate linearly increases from an initial value 0.002 to a maximum value 0.04 over 56 epochs, followed by a linearly decrease back to 0.002 over the next 56 epochs, concluded by dropping the learning rate to 0.0003 for 2 final epochs. We chose the final model weights from the epoch where the validation Pearson's R reached its max.

SMF CTCF Sites

We computed disruption scores for the full set of 9,991 CTCF sites profiled by Sönmezer et al. [31]. Since some of these CTCF sites overlap, we note that correlations between disruption score and SMF "bound" frequency do not change after filtering for a minimum distance of 200 bp or 500 bp between CTCF sites.

TAD boundary CTCF sites

To profile a set of CTCF sites capable of strongly impacting genome folding we took the following steps: extract Jaspar [30] mm10 CTCF site positions (MA0139.1) that overlap with TAD boundaries from mESC Hi-C data [34] obtained at 10kb resolution. These sites were then filtered to exclude overlaps with other CTCF sites (+/- 60bp) or repeat elements (+/- 20bp, RMSK table: https://genome.ucsc.edu/cgi-bin/hgTables?db=mm10&hgta_table%20=%20rmsk), resulting in 7,560 CTCF sites. Since most of these sites displayed little impact on genome folding, analyses from **Fig 3C–7F** were performed with the strongest subset of these sites (n = 1,500).

Visualizing predicted maps

For visualization of predicted insertions, maps were averaged over ten background sequences and all mouse outputs for the first model, unless specified.

Disruption scores

To generate disruption scores, a sub-sequence (e.g., the 19 bp CTCF core) was replaced with a random permutation of the same sub-sequence while keeping the rest of the input genomic sequence fixed. Permutation maintains the nucleotide composition, allowing us to assess the impact of disrupting the specific sequence on the model predictions. Disruption scores are then quantified as mean squared pixelwise differences between predicted maps before and after the permutation (as previously [19]). Unless noted, this score is averaged across all six cell types and four models.

Insertion score

Insertion scores are calculated similarly to disruption scores, where the difference is taken with the background map prediction before insertion. Averaged across six cell types, four models, and ten background sequences per model unless noted.

Background generation

We generated backgrounds for each model by iteratively shuffling genomic sequences until the resulting maps achieved a uniformly flat profile, assessed by the predicted map signal strength (i.e. sum of squared values for the predicted map). Each sequence was shuffled repeatedly until its signal strength fell below a predetermined threshold of 35. Sequences that exceeded this threshold after a maximum of 20 iterations were discarded.

Weighted Jensen-Shannon Difference

We computed the Weighted Jensen-Shannon difference to quantify position-specific dissimilarities between two motifs. This metric captures changes in probability distributions and information content, with symbol contributions weighted by normalized probability

differences. Increased probabilities in the first motif are shown as upward bars, while decreases are shown as downward bars, implemented following [70].

SVD for core-flank compatibility

We generated the core-flank matrix $M_{i,j}$ where the (*i*,*j*) entry corresponded to the predicted insertion scores for a sequence made from core sequence *i* and flank sequence *j*. We then used SVD to decompose $M = U\ D\ V^T$, where U is the matrix of left singular vectors corresponding to cores, V is a matrix of right singular vectors corresponding to flanks, and D is a diagonal matrix of singular values. Elements of M can be re-written as a sum over singular values indexed by k, $M_{i,j} = \Sigma_k\ D^k\ U^k_i\ V^k_j$. As $M_{i,j} \approx D^0\ U^0_i\ V^0_j$ the core-flank matrix can be well reconstructed by the outer product of the first singular vectors, where U^0 corresponds to core strength and V^0 corresponds to flank strength.

Dot and boundary scores

The boundary score is determined by the global insertion score from a 'boundary' scenario insertion. Conversely, the dot score is a localized measure calculated by applying a 13x13 bin kernel to a map patch where a dot is expected. This kernel features a 3x3 bin center over the anticipated dot location, surrounded by four 10x3 bin arms. Bonafide dots have lower contact frequency values in these surrounding regions. As predicted maps correspond to log observed/expected contact frequencies, the dot score is computed as the difference in the average square root of the sum of squared values between the center and the arms of the kernel.

Dot anchors

We identified CTCF sites overlapping dot anchors using mm10 ESC data at 10kb resolution [34] using *MUSTACHE* [51], similarly to how we identified those overlapping boundaries. We initially found 39,226 CTCF sites overlapping dot anchors and refined this by excluding 2,278 sites that also overlapped TAD boundaries, resulting in a distinct set of 36,948 CTCF sites.

Insulation offset

The insulation score is derived from the average value within a sliding diamond window along the main diagonal of the map, similar to the method used in [66]. The insulation offset is the distance between the insulation score minima and the map's center point. This offset was averaged across the tested set of CTCF sites, all cell types, ten different background sequences, and four distinct models.

TAD disruption sensitivity

Predicted disruption sensitivity was calculated for 2048bp non-overlapping bins spanning ±45kb around 4,474 mouse TAD boundaries, identified using standard insulation profile thresholds at 10kb resolution. Disruption was simulated by permuting the DNA subsequences corresponding to each bin, rearranging their order randomly. Disruption scores were calculated as previously described. The proportion of TAD boundaries classified as disruption-sensitive or disruption-resilient was determined using bins within ±28kb of the called boundaries.

Asymmetry analysis

For each disruption-sensitive boundary, the bin with the highest disruption score was identified, and orientation was determined based on overlap with CTCF sites annotated in the JASPAR database. Bins were categorized as left-oriented or right-oriented if all overlapping CTCF

sites were co-oriented; bins without consistent orientation were classified as "others." The smoothed total number of disruption-sensitive bins was calculated using a combination of spline fitting and Gaussian kernel smoothing.

Statistics and software

The neural network has been implemented using python (v3.7) and tensorflow (v2.4). The main text and Fig legends indicate the statistical tests used in the comparisons. Pearson R and Spearman R were calculated with SciPy 1.11.4 [71]. Analyses were performed using: NumPy 1.23.5 [72], pandas 2.1.4 [73], matplotlib 3.8.4 [74], and seaborn 0.13.0 [75], and additionally made use of h5py 3.10.0 [76], and pysam 0.22.0 [77].

Supporting information

S1 Table. Data used for AkitaV2 cross-species training.
(XLSX)

S1 Fig. AkitaV2 enables mouse and human predictions via a cross-species training approach. A) AkitaV2 architecture. This model inputs ~1.3 million base pairs of DNA to predict log(observed/expected) pairwise contact frequencies. The model employs a shared trunk and two distinct prediction heads: one for six mouse cell types and another for five human cell types. **B)** Scatterplot of MSE for AkitaV1 vs. the AkitaV2 human predictions for each genomic region in the overlap between AkitaV1 and AkitaV2 test sets, making use of how each AktiaV2 region was in the test set for one of the eight AkitaV2 models. We overlapped the test set from AkitaV1 (413 sequences) and that for all models from AkitaV2 (5841 sequences) using an inner join, yielding 400 regions with substantial overlap. We then selected the prediction for the AkitaV2 model that had this region in its test set. For a conservative comparison we cropped AkitaV2's prediction to match the size of AkitaV1's predictions, as AkitaV2 generates larger contact maps by design, though this slightly underestimates the increased performance for AkitaV2. AkitaV2 displays enhanced performance (0.131 vs. 0.139). Pink and green dots highlight two representative genomic regions: one where AkitaV2 prediction is comparable with AkitaV1, and one where the AkitaV2 prediction outperformed AkitaV1. Predicted maps for the two highlighted regions are shown below. **C)** Scatterplot of Spearman correlation coefficients for AkitaV1 vs AkitaV2 for regions in the same test set as in (B). Akita displayed improved Spearman R (0.59 vs. 0.56) and Pearson R (0.66 vs. 0.62) across the test set. Colored dots as in (B). **D)** Visual comparison of log(observed/expected) contact frequencies for a genomic window with minimal improvement. From left to right: the experimental target map, the prediction by AkitaV1, and the prediction by AkitaV2, all for the human HFF model output. **E)** Comparative analysis of AkitaV2 predictions for mESCs and cortical neurons (CN), both from Bonev et al., 2017 [34]. In the central scatterplot, each point represents correlations for one genomic region (i). The x-axis shows the correlation between experimental target maps and AkitaV2 predictions for the mESC and CN outputs, Pearson R(preds(i,j,[mECS,CN]), targets(i,j,[mESC,CN])). The y-axis shows the correlation between cell-type-specific differences observed in targets (Δtargets) and predictions (Δpreds), Pearson R(Δpreds, Δtargets) where Δtargets≡(targets(i,j,mESC)—targets(i,j,CN)), and Δpreds is defined similarly. Note this quantifies how well the model captures the cross cell-type differences. Correlations are computed per region *i* in the model 0 test set across all pixels *j* per region. For each selected region (red dots), a 2x2 grid of contact maps is displayed: the top row shows experimental target maps, and the bottom row shows AkitaV2 predictions. The first column corresponds to mESCs and the second column to CNs. We observed that better predictions tend to also have higher cell

type specificity and often corresponded to genomic regions with higher dynamic range.
(TIFF)

S2 Fig. Akita exhibits some limited cell-type specificity in its predictions. A) Predicted vs. experimental log(observed/expected) values for each bin pair across all regions in the test set for mouse model 0, shown separately for each target. The plot demonstrates a correlation between predictions and experimental data across cell types. Colors represent the log10 number of bin pairs for each set of predicted vs. experimental values, and Pearson R is provided as a measure of correlation. Cell types are abbreviated as: mouse embryonic stem cells (mESC), cortical neurons (CN), neocortex cortical neurons (NCN), neural progenitor cells (NPC), neocortex neural progenitor cells (NNPC). **B)** Across all regions in the mouse model 0 test set for different cell types, we observe the following: *Left*: model predictions are highly correlated between cell types (Pearson R(pred(i,j,c1), pred(i,j,c2)), where c1 and c2 denote cell types, and the correlation is computed across all genomic regions i and pixels j). *Middle*: experimentally assayed genome folding shows correlations between cell types, though these are weaker (Pearson R(targets(i,j,c1), targets(i,j,c2)). Right: predicted cell-type differences from our models show weak correlations with observed differences (Pearson R(pred(i,j,c1)—pred(i,j,c2), targets (i,j,c1)—targets(i,j,c2)). Note that the scales for Pearson R differ across the panels. **C)** As in A), for human model 0. **D)** As in B), for human model 0.
(TIFF)

S3 Fig. DNA sequence disruption by permutation offers a robust strategy for computing predicted impacts of CTCF. A) Disruption scores are highly correlated for random CTCF permutations. Scatterplot of disruption scores for n = 7,560 individual CTCF sites subjected to random permutations, where each point represents the predicted disruption score of an individual CTCF site. Disruption scores were computed twice (with model 0) for each CTCF-binding site overlapping a TAD boundary. **B)** Inter-model consistency in CTCF site disruption. Scatterplot of disruption scores for n = 7,560 individual CTCF sites compared between model 0 and model 1. Disruption scores are highly correlated across all pairs of models 0–7 (PearsonR $>$ 0.955). **C)** Position of disrupted CTCF site relative to the prediction window. This plot explores the effect of shifting the predictive window by +10kb. We tested shifts of ±10kb, ±1kb, ±100bp, ±10bp, and ±1bp, comparing the scores for each shift with the scores from the centered permutation. All correlation coefficients between the disruption scores for shifted permutations and the centered permutation were consistently high, exceeding 0.997. **D)** Consistency in CTCF sites disruption across DNA strands. Disruption scores do not depend on the input sequence orientation. **E)** Inter-species consistency of CTCF site disruption. 18Mb of mouse chromosome 1 (ch1:3,653,632–21,776,376) was disrupted by permuting a sliding ~200bp window and disruption scores were calculated for either the human (hESC) or mouse (mESC) output. Scatterplot shows median disruption score across four models for each 200bp window for the human versus the mouse output (Pearson R $>$ 0.96).
(TIFF)

S4 Fig. DNA Sequence Shuffling and Insertion Score Robustness. A) Boxplots constructed from predicted map signal strength scores of n = 590 shuffled genomic windows show DNA sequence shuffling impact on contact matrix strength, with predicted map signal strength for model 0 across shuffled sequences (1, 2, 4, 8, 16, 32 nucleotides). Lower scores denote weaker maps, with k = 8 shuffling resulting in the most neutral maps. **B)** Scatterplot of predicted map signal strength versus GC content for shuffled genomic sequences. Points represent scores from model 0 for n = 590 genomic windows shuffled using k = 8, and show no trend between GC content and SCD. **C)** Scatterplot comparing insertion scores for the insertion of a strong

CTCF site into n = 590 genomic sequences, both original and shuffled once with k = 8. While shuffling does not alter the mean, it remarkably reduces the variance of the insertion scores. **D)** Scatterplot of virtual insertion score for background sequence 0 vs. background sequence 1 for model 0 across n = 7,560 CTCF sites. Virtual insertion scores are highly correlated across backgrounds (PearsonR > 0.94 for any pair of background sequences). **E)** Scatterplot of insertion scores between model 0 and model 1 for n = 7560 CTCF sites. Virtual insertion scores are highly correlated across pairs of models (PearsonR > 0.96). **F)** Scatterplot of insertion scores between mouse and human predictions for n = 7560 mouse CTCF sites inserted into shuffled mouse background sequences. Insertion scores are highly correlated (PearsonR > 0.99).
(TIFF)

S5 Fig. Sequence Preferences within Flanking Sequences and the Double-Site Virtual Insertion. A) Sequence logos for the strongest 150 CTCF sites ranked by either insertion (top) or disruption (bottom) scores. Red boxes highlight weak sequence preferences upstream and downstream of the CTCF core motif. **B)** Illustration of a double CTCF site insertion. Two CTCF sites (green boxes) are virtually inserted symmetrically around a background sequence's midpoint (gray rectangle) with constant 180bp spacing (as in [15]) with flanking regions (pink boxes). The impact is quantified by the squared contact difference between maps with and without inserted CTCF sites (insertion score). **C)** Insertion score versus flanking sequence length for insertions of tandem CTCF sites in four orientations (left, right, convergent, divergent). Grouping and shading as in **Fig 6B**. Insertions of tandem sites show a similar trend to single sites, and similar trends across all orientations. **D)** Predicted maps for double CTCF insertions with increasing flank length for different orientations. Inserted CTCF sequence extracted from chr7:37,357,852–37,357,871. The maps are arranged in a grid by flank length (columns) and site orientation (left, right, convergent, divergent) in rows, demonstrating similar strength impacts across orientations but slight asymmetry for sites inserted in tandem.
(TIFF)

S6 Fig. No evidence of clear clusters in flanking sequences. A) Heatmap of nucleotide composition around 150 strong CTCF sites (±30bp), with rows sorted according to the Hamming distance between their sequences. **B)** Same as panel A except rows are ordered by the Hamming distance between upstream flanking sequences. **C)** Same as panel A and B, but with rows arranged by the Hamming distance between downstream flanking sequences. **D)** Scatterplot showing the fraction of strong and weak upstream flanking sequences containing k-mers (k = 2). Strong and weak sequences are defined as the 150 CTCF sites with the highest and lowest insertion scores, respectively. Red dots indicate significantly enriched or depleted k-mers in strong versus weak flanking regions. Bootstrap sampling (n = 100,000) was used to generate distributions of differences in k-mer presence fractions. These distributions were used to assess significance (at 0.05 corrected by the number of k-mers tested) of k-mers enrichment (above zero) or depletion (below zero). Grey lines represent 2-fold up and down ratios. **E)** Same as panel D, but for k = 3. **F)** Average GC content (%) for 150 strong (blue) and 150 weak (orange) CTCF sites. **G)** Same as panel D, but for k-mer enrichment in downstream flanking regions. **H)** Same as panel E, but for k-mer enrichment in downstream flanking regions.
(TIFF)

S7 Fig. Feature-Specificity of CTCF Sites. A) Scatterplot of insertion scores from boundary versus dot scenarios shows a high correlation (PearsonR > 0.99), indicating that the global metric of predicted map strength does not significantly vary with the insertion scenario. **B)** CTCF sites with differing genomic origins have similar dot and boundary strengths. Scatterplot shows dot versus boundary strengths for two sets of CTCF sites, either: overlapping TAD

boundaries (orange dots, n = 1,500), or dot anchors (blue dots, n>36,900) called in experimental Hi-C maps. These sets of CTCF sites are disjoint. Uniform distribution across the plot shows all sites behave similarly in the experiment, regardless of their genomic origin. This suggests that CTCF's role in chromatin architecture does not inherently differ between those overlapping with TAD boundaries and those at dot anchors. **C)** TAD disruption analysis in mouse embryonic stem cells (mESCs) at 10-kb resolution. The Akita model was used to evaluate the impact of permuting 2048-bp sequences within or near 4,474 TAD boundaries. The histogram depicts the distribution of maximum disruption scores per boundary, with the red dashed line representing the Li threshold. This threshold separates disruption-sensitive boundaries (high scores) from disruption-resilient boundaries (low scores). Of the disruption-sensitive boundaries, 751 overlapped with transcription start sites (TSSs), compared to 828 overlaps in the disruption-resilient group. **D)** Disruption-resilient boundaries (2,263 total) lacked evidence of TAD boundary disappearance and often displayed higher levels of missing data and less accurate predictions. **E)** Disruption-sensitive boundaries (2,211 total) were characterized by the disappearance of TAD boundaries in predicted Hi-C maps. **F)** Disruption-resilient boundaries were enriched for regions with missing bins. **G)** Disruption-resilient boundaries showed enrichment in areas with low sequencing coverage. **H)** Disruption-resilient boundaries corresponded to regions lacking distinct TAD boundaries in predicted maps, as reflected by elevated insulation scores.
(TIFF)

Acknowledgments

The authors thank Elphège Nora, Erika Anderson, and Katherine Pollard for feedback.

Author Contributions

Conceptualization: David R. Kelley, Geoffrey Fudenberg.

Data curation: Geoffrey Fudenberg.

Formal analysis: Paulina N. Smaruj, David R. Kelley, Geoffrey Fudenberg.

Funding acquisition: Geoffrey Fudenberg.

Investigation: Paulina N. Smaruj, Geoffrey Fudenberg.

Methodology: Paulina N. Smaruj, Fahad Kamulegeya.

Project administration: Geoffrey Fudenberg.

Resources: Fahad Kamulegeya, David R. Kelley.

Software: Paulina N. Smaruj, David R. Kelley, Geoffrey Fudenberg.

Supervision: Geoffrey Fudenberg.

Writing – original draft: Paulina N. Smaruj, Geoffrey Fudenberg.

Writing – review & editing: Paulina N. Smaruj, David R. Kelley, Geoffrey Fudenberg.

References

1. McCord RP, Kaplan N, Giorgetti L. Chromosome Conformation Capture and Beyond: Toward an Integrative View of Chromosome Structure and Function. Mol Cell. 2020; 77: 688–708. https://doi.org/10.1016/j.molcel.2019.12.021 PMID: 32001106
2. van Ruiten MS, Rowland BD. On the choreography of genome folding: A grand pas de deux of cohesin and CTCF. Cell Nucl. 2021; 70: 84–90. https://doi.org/10.1016/j.ceb.2020.12.001 PMID: 33545664

3. Nora EP, Goloborodko A, Valton A-L, Gibcus JH, Uebersohn A, Abdennur N, et al. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. Cell. 2017; 169: 930–944.e22. https://doi.org/10.1016/j.cell.2017.05.004 PMID: 28525758
4. Rao SSP, Huang S-C, Glenn St Hilaire B, Engreitz JM, Perez EM, Kieffer-Kwon K-R, et al. Cohesin Loss Eliminates All Loop Domains. Cell. 2017; 171: 305–320.e24. https://doi.org/10.1016/j.cell.2017.09.026 PMID: 28985562
5. Dekker J, Mirny L. The 3D Genome as Moderator of Chromosomal Communication. Cell. 2016; 164: 1110–1121. https://doi.org/10.1016/j.cell.2016.02.007 PMID: 26967279
6. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. Cell. 2015; 161: 1012–1025. https://doi.org/10.1016/j.cell.2015.04.004 PMID: 25959774
7. Anania C, Acemel RD, Jedamzick J, Bolondi A, Cova G, Brieske N, et al. In vivo dissection of a clustered-CTCF domain boundary reveals developmental principles of regulatory insulation. Nat Genet. 2022; 54: 1026–1036. https://doi.org/10.1038/s41588-022-01117-9 PMID: 35817979
8. Khoury A, Achinger-Kawecka J, Bert SA, Smith GC, French HJ, Luu P-L, et al. Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. Nat Commun. 2020; 11: 54. https://doi.org/10.1038/s41467-019-13753-7 PMID: 31911579
9. Despang A, Schöpflin R, Franke M, Ali S, Jerković I, Paliou C, et al. Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. Nat Genet. 2019; 51: 1263–1271. https://doi.org/10.1038/s41588-019-0466-z PMID: 31358994
10. Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom DT, Tanay A, et al. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. Cell Rep. 2015; 10: 1297–1309. https://doi.org/10.1016/j.celrep.2015.02.004 PMID: 25732821
11. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell. 2014; 159: 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021 PMID: 25497547
12. de Wit E, Vos ESM, Holwerda SJB, Valdes-Quezada C, Verstegen MJAM, Teunissen H, et al. CTCF Binding Polarity Determines Chromatin Looping. Mol Cell. 2015; 60: 676–684. https://doi.org/10.1016/j.molcel.2015.09.023 PMID: 26527277
13. Guo Y, Xu Q, Canzio D, Shou J, Li J, Gorkin DU, et al. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell. 2015; 162: 900–910. https://doi.org/10.1016/j.cell.2015.07.038 PMID: 26276636
14. Kraft K, Magg A, Heinrich V, Riemenschneider C, Schöpflin R, Markowski J, et al. Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations. Nat Cell Biol. 2019; 21: 305–310. https://doi.org/10.1038/s41556-019-0273-x PMID: 30742094
15. Huang H, Zhu Q, Jussila A, Han Y, Bintu B, Kern C, et al. CTCF mediates dosage- and sequence-context-dependent transcriptional insulation by forming local chromatin domains. Nat Genet. 2021; 53: 1064–1074. https://doi.org/10.1038/s41588-021-00863-6 PMID: 34002095
16. Stolper RJ, Tsang FH, Georgiades E, Hansen LLP, Downes DJ, Harrold CL, et al. Loop extrusion by cohesin plays a key role in enhancer-activated gene expression during differentiation. bioRxiv. 2023; 2023.09.07.556660. https://doi.org/10.1101/2023.09.07.556660
17. Tsang FH, Stolper RJ, Hanifi M, Cornell LJ, Francis HS, Davies B, et al. The characteristics of CTCF binding sequences contribute to enhancer blocking activity. bioRxiv. 2023; 2023.09.06.556325. https://doi.org/10.1101/2023.09.06.556325
18. Zuin J, Roth G, Zhan Y, Cramard J, Redolfi J, Piskadlo E, et al. Nonlinear control of transcription through enhancer–promoter interactions. Nature. 2022; 604: 571–577. https://doi.org/10.1038/s41586-022-04570-y PMID: 35418676
19. Fudenberg G, Kelley DR, Pollard KS. Predicting 3D genome folding from DNA sequence with Akita. Nat Methods. 2020; 17: 1111–1117. https://doi.org/10.1038/s41592-020-0958-x PMID: 33046897
20. Schwessinger R, Gosden M, Downes D, Brown RC, Oudelaar AM, Telenius J, et al. DeepC: predicting 3D genome folding using megabase-scale transfer learning. Nat Methods. 2020; 17: 1118–1124. https://doi.org/10.1038/s41592-020-0960-3 PMID: 33046896
21. Tan J, Shenker-Tauris N, Rodriguez-Hernaez J, Wang E, Sakellaropoulos T, Boccalatte F, et al. Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening. Nat Biotechnol. 2023; 41: 1140–1150. https://doi.org/10.1038/s41587-022-01612-8 PMID: 36624151
22. Zhou J. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. Nat Genet. 2022; 54: 725–734. https://doi.org/10.1038/s41588-022-01065-4 PMID: 35551308

23. Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat Rev Genet. 2023; 24: 125–137. https://doi.org/10.1038/s41576-022-00532-2 PMID: 36192604

24. Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nat Genet. 2021; 53: 354–366. https://doi.org/10.1038/s41588-021-00782-6 PMID: 33603233

25. Nair S, Ameen M, Sundaram L, Pampari A, Schreiber J, Balsubramani A, et al. Transcription factor stoichiometry, motif affinity and syntax regulate single-cell chromatin dynamics during fibroblast reprogramming to pluripotency. bioRxiv. 2023; 2023.10.04.560808. https://doi.org/10.1101/2023.10.04.560808 PMID: 37873116

26. Kelley DR. Cross-species regulatory sequence activity prediction. PLOS Comput Biol. 2020; 16: e1008050. https://doi.org/10.1371/journal.pcbi.1008050 PMID: 32687525

27. Gunsalus LM, Keiser MJ, Pollard KS. In silico discovery of repetitive elements as key sequence determinants of 3D genome folding. Cell Genomics. 2023; 3: 100410. https://doi.org/10.1016/j.xgen.2023.100410 PMID: 37868032

28. Gjoni K, Pollard KS. SuPreMo: a computational tool for streamlining in silico perturbation using sequence-based predictive models. Bioinformatics. 2024; 40: btae340. https://doi.org/10.1093/bioinformatics/btae340 PMID: 38796686

29. Gunsalus LM, McArthur E, Gjoni K, Kuang S, Pittman M, Capra JA, et al. Comparing chromatin contact maps at scale: methods and insights. bioRxiv. 2023; 2023.04.04.535480. https://doi.org/10.1101/2023.04.04.535480 PMID: 37066196

30. Rauluseviciute I, Riudavets-Puig R, Blanc-Mathieu R, Castro-Mondragon JA, Ferenc K, Kumar V, et al. JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2024; 52: D174–D182. https://doi.org/10.1093/nar/gkad1059 PMID: 37962376

31. Sönmezer C, Kleinendorst R, Imanci D, Barzaghi G, Villacorta L, Schübeler D, et al. Molecular Co-occupancy Identifies Transcription Factor Binding Cooperativity In Vivo. Mol Cell. 2021; 81: 255–267.e6. https://doi.org/10.1016/j.molcel.2020.11.015 PMID: 33290745

32. Abdulhay NJ, McNally CP, Hsieh LJ, Kasinathan S, Keith A, Estes LS, et al. Massively multiplex single-molecule oligonucleosome footprinting. Dekker J, Barkai N, Dekker J, editors. eLife. 2020; 9: e59404. https://doi.org/10.7554/eLife.59404 PMID: 33263279

33. Stergachis AB, Debo BM, Haugen E, Churchman LS, Stamatoyannopoulos JA. Single-molecule regulatory architectures captured by chromatin fiber sequencing. Science. 2020; 368: 1449–1454. https://doi.org/10.1126/science.aaz1646 PMID: 32587015

34. Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, et al. Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell. 2017; 171: 557–572.e24. https://doi.org/10.1016/j.cell.2017.09.043 PMID: 29053968

35. Kaaij LJT, Mohn F, van der Weide RH, de Wit E, Bühler M. The ChAHP Complex Counteracts Chromatin Looping at CTCF Sites that Emerged from SINE Expansions in Mouse. Cell. 2019; 178: 1437–1451.e14. https://doi.org/10.1016/j.cell.2019.08.007 PMID: 31491387

36. Hanssen LLP, Kassouf MT, Oudelaar AM, Biggs D, Preece C, Downes DJ, et al. Tissue-specific CTCF–cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. Nat Cell Biol. 2017; 19: 952–961. https://doi.org/10.1038/ncb3573 PMID: 28737770

37. Nakahashi H, Kwon K-RK, Resch W, Vian L, Dose M, Stavreva D, et al. A Genome-wide Map of CTCF Multivalency Redefines the CTCF Code. Cell Rep. 2013; 3: 1678–1689. https://doi.org/10.1016/j.celrep.2013.04.024 PMID: 23707059

38. Rhee HS, Pugh BF. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. Cell. 2011; 147: 1408–1419. https://doi.org/10.1016/j.cell.2011.11.013 PMID: 22153082

39. Ramani V, Qiu R, Shendure J. High Sensitivity Profiling of Chromatin Structure by MNase-SSP. Cell Rep. 2019; 26: 2465–2476.e4. https://doi.org/10.1016/j.celrep.2019.02.007 PMID: 30811994

40. Iurlaro M, Masoni F, Flyamer IM, Wirbelauer C, Iskar M, Burger L, et al. Systematic assessment of ISWI subunits shows that NURF creates local accessibility for CTCF. Nat Genet. 2024; 56: 1203–1212. https://doi.org/10.1038/s41588-024-01767-x PMID: 38816647

41. Kaplow IM, Banerjee A, Foo CS. Neural network modeling of differential binding between wild-type and mutant CTCF reveals putative binding preferences for zinc fingers 1–2. BMC Genomics. 2022; 23: 295. https://doi.org/10.1186/s12864-022-08486-9 PMID: 35410161

42. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves Â, Kutter C, et al. Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. Cell. 2012; 148: 335–348. https://doi.org/10.1016/j.cell.2011.11.058 PMID: 22244452

43. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell. 2010; 38: 576–589. https://doi.org/10.1016/j.molcel.2010.05.004 PMID: 20513432

44. Yang J, Horton JR, Liu B, Corces VG, Blumenthal RM, Zhang X, et al. Structures of CTCF–DNA complexes including all 11 zinc fingers. Nucleic Acids Res. 2023; 51: 8447–8462. https://doi.org/10.1093/nar/gkad594 PMID: 37439339

45. Anderson DW, McKeown AN, Thornton JW. Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. Weigel D, editor. eLife. 2015; 4: e07864. https://doi.org/10.7554/eLife.07864 PMID: 26076233

46. Starr TN, Thornton JW. Epistasis in protein evolution. Protein Sci. 2016; 25: 1204–1218. https://doi.org/10.1002/pro.2897 PMID: 26833806

47. Chang L-H, Ghosh S, Papale A, Luppino JM, Miranda M, Piras V, et al. Multi-feature clustering of CTCF binding creates robustness for loop extrusion blocking and Topologically Associating Domain boundaries. Nat Commun. 2023; 14: 5615. https://doi.org/10.1038/s41467-023-41265-y PMID: 37699887

48. Harmston N, Ing-Simmons E, Tan G, Perry M, Merkenschlager M, Lenhard B. Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. Nat Commun. 2017; 8: 441. https://doi.org/10.1038/s41467-017-00524-5 PMID: 28874668

49. Barisic D, Stadler MB, Iurlaro M, Schübeler D. Mammalian ISWI and SWI/SNF selectively mediate binding of distinct transcription factors. Nature. 2019; 569: 136–140. https://doi.org/10.1038/s41586-019-1115-5 PMID: 30996347

50. Fu Y, Sinha M, Peterson CL, Weng Z. The Insulator Binding Protein CTCF Positions 20 Nucleosomes around Its Binding Sites across the Human Genome. PLOS Genet. 2008; 4: e1000138. https://doi.org/10.1371/journal.pgen.1000138 PMID: 18654629

51. Roayaei Ardakany A, Gezer HT, Lonardi S, Ay F. Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. Genome Biol. 2020; 21: 256. https://doi.org/10.1186/s13059-020-02167-0 PMID: 32998764

52. Hansen AS, Hsieh T-HS, Cattoglio C, Pustova I, Saldaña-Meyer R, Reinberg D, et al. Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region in CTCF. Mol Cell. 2019; 76: 395–411.e13. https://doi.org/10.1016/j.molcel.2019.07.039 PMID: 31522987

53. Rahmaninejad H, Xiao Y, Tortora MMC, Fudenberg G. Dynamic barriers modulate cohesin positioning and genome folding at fixed occupancy. bioRxiv. 2024; 2024.10.08.617113. https://doi.org/10.1101/2024.10.08.617113 PMID: 39416077

54. Tsang Felice H., Stolper Rosa J., Hanifi Muhammad, Cornell Lucy J., Francis Helena S., Davies Benjamin, et al. The characteristics of CTCF binding sequences contribute to enhancer blocking activity. bioRxiv. 2023; 2023.09.06.556325. https://doi.org/10.1101/2023.09.06.556325

55. Zuo Z, Billings T, Walker M, Petkov PM, Fordyce PM, Stormo GD. On the dependent recognition of some long zinc finger proteins. Nucleic Acids Res. 2023; 51: 5364–5376. https://doi.org/10.1093/nar/gkad207 PMID: 36951113

56. Do C, Jiang G, Cova G, Katsifis CC, Narducci DN, Yang J, et al. Brain and cancer associated binding domain mutations provide insight into CTCF's relationship with chromatin and its ability to act as a chromatin organizer. bioRxiv. 2024; 2024.01.11.575070. https://doi.org/10.1101/2024.01.11.575070 PMID: 38370764

57. Koo PK, Majdandzic A, Ploenzke M, Anand P, Paul SB. Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. PLOS Comput Biol. 2021; 17: e1008925. https://doi.org/10.1371/journal.pcbi.1008925 PMID: 33983921

58. Lal A, Karollus A, Gunsalus L, Garfield D, Nair S, Tseng AM, et al. Decoding sequence determinants of gene expression in diverse cellular and disease states. bioRxiv. 2024; 2024.10.09.617507. https://doi.org/10.1101/2024.10.09.617507

59. de Almeida BP, Schaub C, Pagani M, Secchia S, Furlong EEM, Stark A. Targeted design of synthetic enhancers for selected tissues in the Drosophila embryo. Nature. 2024; 626: 207–211. https://doi.org/10.1038/s41586-023-06905-9 PMID: 38086418

60. Justice M, Carico ZM, Stefan HC, Dowen JM. A WIZ/Cohesin/CTCF Complex Anchors DNA Loops to Define Gene Expression and Cell Identity. Cell Rep. 2020; 31. https://doi.org/10.1016/j.celrep.2020.03.067 PMID: 32294452

61. Hansen AS, Pustova I, Cattoglio C, Tjian R, Darzacq X. CTCF and cohesin regulate chromatin loop stability with distinct dynamics. Sherratt D, editor. eLife. 2017; 6: e25776. https://doi.org/10.7554/eLife.25776 PMID: 28467304

62. Hsieh T-HS, Cattoglio C, Slobodyanyuk E, Hansen AS, Rando OJ, Tjian R, et al. Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. Mol Cell. 2020; 78: 539–553.e8. https://doi.org/10.1016/j.molcel.2020.03.002 PMID: 32213323

63. Krietenstein N, Abraham S, Venev SV, Abdennur N, Gibcus J, Hsieh T-HS, et al. Ultrastructural Details of Mammalian Chromosome Architecture. Mol Cell. 2020; 78: 554–565.e7. https://doi.org/10.1016/j.molcel.2020.03.003 PMID: 32213324

64. Goloborodko Anton, Venev Sergey, Spracklin George, Abdennur Nezar, Galitsyna Aleksandra, Shaytan Alexey, et al. open2c/distiller-nf: v0.3.4. https://doi.org/10.5281/zenodo.1490628

65. Open2C, Abdennur N, Fudenberg G, Flyamer IM, Galitsyna AA, Goloborodko A, et al. Pairtools: From sequencing data to chromosome contacts. PLOS Comput Biol. 2024; 20: e1012164. https://doi.org/10.1371/journal.pcbi.1012164 PMID: 38809952

66. Abdennur N, Mirny LA. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics. 2020; 36: 311–316. https://doi.org/10.1093/bioinformatics/btz540 PMID: 31290943

67. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 2012; 9: 999–1003. https://doi.org/10.1038/nmeth.2148 PMID: 22941365

68. Open2C, Abdennur N, Abraham S, Fudenberg G, Flyamer IM, Galitsyna AA, et al. Cooltools: Enabling high-resolution Hi-C analysis in Python. PLoS Comput Biol. 2024; 20(5): e1012067. https://doi.org/10.1371/journal.pcbi.1012067

69. Abadi Martín, Barham Paul, Chen Jianmin, Chen Zhifeng, Davis Andy, Dean Jeffrey, Devin Matthieu, Ghemawat Sanjay, Irving Geoffrey, Isard Michael, Kudlur Manjunath, Levenberg Josh, Monga Rajat, Moore Sherry, Murray Derek G., Steiner Benoit, Tucker Paul, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, Google Brain. Tensorflow: A system for large-scale machine learning. 12th USENIX Symp Oper Syst Des Implement OSDI. 2016; 16: 265–283.

70. Nettling M, Treutler H, Grau J, Keilwagen J, Posch S, Grosse I. DiffLogo: a comparative visualization of sequence motifs. BMC Bioinformatics. 2015; 16: 387. https://doi.org/10.1186/s12859-015-0767-x PMID: 26577052

71. Virtanen Pauli and Gommers Ralf and Oliphant, Travis E. and, Haberland Mattand Reddy Tylerand Cournapeau Davidand, Burovski Evgeniand Peterson Pearuand Weckesser Warrenand, Bright Jonathanand {van der Walt} St{\'e}fan J. and, Brett Matthewand Wilson Joshuaand Jarrod Millman, K. and, Mayorov Nikolayand Nelson, Andrew R. J.and Jones Eric and, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nat Methods. 2020; 17: 261–272. https://doi.org/10.1038/s41592-019-0686-2 PMID: 32015543

72. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020; 585: 357–362. https://doi.org/10.1038/s41586-020-2649-2 PMID: 32939066

73. The pandas development team. pandas-dev/pandas: Pandas. Zenodo. 2023. https://doi.org/10.5281/zenodo.3509134

74. Hunter J. D. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007; 9: 90–95. https://doi.org/10.1109/MCSE.2007.55

75. Waskom, Michael L. seaborn: statistical data visualization. J Open Source Softw. 2021; 6: 3021.

76. Collette Andrew. Python and HDF5. O'Reilly; 2013.

77. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021; 10: giab008. https://doi.org/10.1093/gigascience/giab008 PMID: 33590861