Title

Sequence to structure to function: computational strategies for modeling the 3D genome

Author

Gunsalus, Laura Margret

Publication Date

2023

Peer reviewed|Thesis/dissertation

Sequence to structure to function: computational strategies for modeling the 3D genome

by
Laura Gunsalus

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY

in

Biological and Medical Informatics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
Michael Keiser
4DF1BD06D670465...
Michael Keiser
Chair

DocuSigned by:
John Capra
[illegible]
John Capra

DocuSigned by:
Katherine Pollard
[illegible]
Katherine Pollard

DocuSigned by:
Vijay Ramani
E7EC74D06141482...
Vijay Ramani

Committee Members

Acknowledgements

The work of this thesis would not have been possible without the gracious support of friends and mentors. Thank you first to my advisors, Katie Pollard and Mike Keiser, for taking a chance on me and on co-mentorship. Katie's continuous support and expertise has made me a much stronger scientist every year of graduate school. I am grateful to Mike for his generosity and willingness to invest and give considerate feedback on topics new to us both. I am thankful for my encouraging committee: Vijay Ramani and Tony Capra. Collaborating with the Capra lab has been a highlight of my graduate work. Vijay has provided many essential project suggestions and chatting with him always reminds me how exciting regulatory genomics can be. Many thanks to Elphege Nora for his thoughtful and rigorous feedback on my qualifying exam proposal and manuscript drafts, as well as for his enthusiasm for chromatin. I am grateful to have had the chance to learn from many other talented scientists, including Geoff Fudenberg and Evonne McArthur.

I was deeply fortunate to find mentorship before starting graduate school, without which I would surely be on a different path. Thank you to Andreas Pfenning at Carnegie Mellon for such a thoughtful introduction to both computational biology and regulatory genomics. I was fortunate to work alongside SyCom in Boston: thank you to Matthew Eaton, David Orlando, Chris Fiore, and Nisha Rajagopal for answering so many of my questions, cementing my love for regulatory genomics, and helping me feel like a part of a team.

Thank you to Kangway Chuang, whose belief in me and generous spirit has pushed me to grow as a scientist, and whose kindness has eased the many difficulties a PhD brings. He demonstrated what consistent rigorous work looks like, showed that science is most productive and fun together, and always found time to

help others. I can think of no stronger motivators than seeking environments to learn like we found at the whiteboard and making good on all the help he has given me. Thank you to Garrett Gaskins, for reminding me that there are more important things than a PhD. Thanks to Elena Caceres, for her friendship and unwavering, enthusiastic support.

I appreciate all the Pollard lab members who were happy to provide feedback or chat over science. Thank you to Calla Martyn, for her pragmatic levelheadedness when graduate school was most difficult, and for showing me how wonderful San Francisco is. I am not sure how I would have made it without the friendship of Calla, Miriam Goldman, and Matt Johnson as we went through milestones together. I am so glad I had the chance to overlap and work with younger graduate students across labs: Parker Grosjean, Katie Gjoni, Amanda Everitt, and Shiron Drusinsky, whose enthusiasm and curiosity has made me a stronger scientist. I can't wait to see what they accomplish.

Thank you to the friendships outside of graduate school: Meena, Zach, Ann, Preeti, Ariel, Shannon, Patrick, and Aakash. Finally, thank you to my parents, Imke and Rob, and to my brother Timo, for I would be nowhere without them.

Contributions

This dissertation was supervised by Dr. Katherine Pollard and Dr. Michael Keiser.

Chapter 2 contains material from a manuscript in press at Cell Genomics:

> **Laura M. Gunsalus**, Michael J. Keiser, and Katherine S. Pollard. "*In silico* discovery of repetitive elements as key sequence determinants of 3D genome folding." Cell Genomics 3.10 (2023).

Chapter 3 contains material from a manuscript currently under review, available as an open access preprint:

> **Laura M. Gunsalus***, Evonne McArthur*, Ketrin Gjoni, Shuzhen Kuang, Maureen Pittman, John A. Capra, Katherine S. Pollard. "Comparing chromatin contact maps at scale: methods and insights." bioRxiv 2023.04.04.535480.

Chapter 4 contains material from a manuscript currently under review, available as an open access preprint:

> **Laura M. Gunsalus**, Michael J. Keiser, and Katherine S. Pollard. "ChromaFactor: deconvolution of single-cell chromatin organization with NMF." bioRxiv, 2023.11. 22.568268.

* indicates co-first authorship.

Sequence to structure to function: computational strategies for modeling the 3D genome

Laura Gunsalus

Abstract

The spatial organization of chromosomes within the cell nucleus facilitates critical genomic processes including transcription, replication, and repair. Understanding how DNA sequence informs genome folding and how chromatin conformation instructs transcription remains a central challenge. This dissertation presents computational strategies to advance our understanding of the principles governing three-dimensional chromatin structure and their implications for gene regulation. In Chapter 2, I perform large-scale *in silico* mutagenesis using a deep learning model to systematically uncover DNA sequences that encode folding patterns. Chapter 3 introduces new methods to quantify differences between chromatin interaction maps, revealing that integrating simple, map-informed and feature-based strategies provides the most complete perspective on functionally relevant organizational changes. Chapter 4 introduces and applies a non-negative matrix factorization method to decompose single-cell heterogeneity in chromatin structure, linking patterns in cell subpopulations to average folding principles and transcriptional consequences discerned in bulk. Together, these computational methods revealed new biological findings: repetitive elements, sometimes lacking CTCF motifs, provide sequence grammar governing chromatin interactions and the chromatin folding in only a small minority of cells often drives populationwide signals. The work broadly highlights the potential of computational approaches, especially machine learning, to accelerate discovery in genomics. This work provides templates for future studies relating sequence, spatial dynamics, and gene regulation amidst widespread variability in the folded genome.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

bp	base pair
ORCA	Optical reconstruction of chromatin architecture
TAD	Topologically associated domain
TF	Transcription factor
TSS	Transcription start site

1

Introduction

The human genome is intricately folded within the cell nucleus, adopting a complex three-dimensional (3D) structure that is crucial for proper gene regulation and cell function. However, understanding the fundamental principles governing 3D genome organization has remained a central challenge in genomics. While chromosome conformation capture techniques provide a powerful means to probe 3D chromatin interactions genome-wide, realizing the full potential of these data requires new computational strategies tailored to the intricacies of the chromatin folding problem. In this dissertation, I present three complementary projects that aim to advance our understanding of 3D genome organization by probing the sequence determinants of folding, quantifying differences between folding patterns, and unraveling single-molecule heterogeneity.

The sequence of the genome informs its own folding. In **Chapter 2**, I demonstrate how pairing convolutional neural network models of chromatin contacts with large-scale *in silico* mutagenesis can systematically uncover DNA elements governing folding. Our unbiased screen highlighted diverse sequences – including transposons, tRNAs, and GC content shifts – that collaborate with CTCF motifs to architect genome folding. Our findings nominate intriguing sequence candidates outside of known protein binding sites that may orchestrate chromatin interactions.

Comparing folding maps between conditions, individuals, and species has been instrumental for unraveling principles of genome organization. **Chapter 3** reveals that commonly used approaches for quantifying map differences often disagree and overlook functionally relevant changes. We introduce and benchmark several new methods against existing techniques using thousands of *in silico* perturbations, showing that simple, map-informed, and feature-based strategies should be used together. The work provides guidelines, open-source code, and a framework for selecting appropriate comparisons to address diverse research questions in chromatin biology.

The 3D genome exhibits remarkable cell-to-cell variability, obscuring the relationship between folding and function. In **Chapter 4**, I address this challenge by applying non-negative matrix factorization to decompose single-cell chromatin structure datasets into interpretable components. The resulting templates capture salient sources of heterogeneity, enable imputation of noisy single-cell maps, and correlate with key genomic phenotypes like transcription. Subpopulations of cells drive the average folding patterns discerned in bulk. The method and findings provide new avenues to connect 3D genome organization to gene regulation amidst widespread single-cell variation.

Altogether, this body of work presents new computational strategies to tackle several fundamental challenges in connecting form and function of the dynamic 3D genome. The remainder of this introduction reviews key concepts in genome organization and situates the dissertation aims and contributions within the broader field of chromatin organization.

Background on Genome Architecture

The genome is non-randomly arranged within the cell nucleus, exhibiting hierarchical layers of folding implicated in critical nuclear processes like transcription, replication, and repair (Dixon *et al.*, 2015). At the finest scale, 3 billion DNA base pairs wrap around histone protein cores to form nucleosomes, the fundamental repeating unit of chromatin. These "beads on a string" further condense with the assistance of architectural protein complexes like cohesin and condensin (Dixon *et al.*, 2012). Folding at larger scales has been intensively investigated using chromosome conformation capture techniques (Lieberman-Aiden *et al.*, 2009).

Several salient architectural features of genome topology have emerged from population-scale chromatin interaction maps. Chromosomes occupy distinct "territories" dependent on gene density, size, and transcriptional activity (Rao *et al.*, 2014). Within chromosomes, chromatin can segregate into multi-megabase "compartments" of open, gene-rich euchromatin and closed, gene-poor heterochromatin based on patterns of preferential self-interaction (Lieberman-Aiden *et al.*, 2009; Rao *et al.*, 2014). Compartments are further partitioned into topologically-associating domains (TADs), self-interacting regions spanning hundreds of kilobases to a few megabases in size (Dixon *et al.*, 2012). TAD boundaries align with binding of the zinc-finger protein CTCF, which stops the cohesin complex from extruding DNA loops to form insulate domains (Fudenberg *et al.*, 2016). Enhancer-promoter contacts, thought to potentiate transcriptional activation, primarily occur within TADs (Jia *et al.*, 2020). Still longer chromatin loops can link TADs and pull distal loci into spatial proximity (Rao *et al.*, 2014). These layers collectively shape the regulatory landscape of the genome.

Beyond average population folding maps, chromatin architecture exhibits pronounced cell-to-cell variability. Single-cell technologies like single-cell Hi-C and DNA microscopy enable direct observation of cell-to-cell heterogeneity in genome organization (Nagano *et al.*, 2013; Ramani *et al.*, 2017; Mateo *et al.*, 2019; Su *et al.*,

2020). However, these methods suffer from technical artifacts like missing data and limited genomic coverage which complicate interpreting single-cell measurements. Additional work is needed to connect heterogeneous chromatin folding to downstream consequences like variable gene expression.

Notably, the organizational principles described above have all been deduced from population-averaged chromatin interaction maps. Yet, the prevalence of cell-to-cell variability raises questions about how well population maps represent the folding of any individual cell (Krietenstein *et al.*, 2020; Hafner *et al.*, 2022). The complex relationship between population ensembles and single cells highlights a need for strategies tailored to heterogeneity in chromatin structure.

Dissertation Aims

The development of computational strategies to address key challenges in relating genome sequence and folding to function is a common thread across these studies. This dissertation puts forth and applies new machine learning and statistical techniques to tackle three aims:

Aim 1: Identify sequence determinants of 3D genome folding patterns.

While certain architectural proteins like CTCF are known to bind specific sequence motifs to orchestrate folding, a complete accounting of the regulatory genome elements governing chromatin organization has remained elusive. We lack an unbiased systematic approach to exhaustively mine the human genome sequence for additional encodings of chromatin structure outside of known protein binding sites.

Aim 2: Develop and compare methods to accurately quantify differences between 3D genome folding contact maps.

Chromatin interaction maps vary across cell types, individuals, disease states, and species. However, commonly used scoring techniques often disagree on the most dissimilar map pairs. Improved quantification strategies are needed to reliably identify and interpret changes between chromatin folding patterns.

Aim 3: Link cell-to-cell variability in chromatin structure to average genomic patterns and function.

Single-cell genome folding data holds promise for connecting 3D architecture to gene regulation within individual cells. However, pervasive heterogeneity obscures the origins of average folding patterns and their influence on phenotypes like transcription. New approaches are required to deconvolve single-cell variability.

The subsequent sections situate these aims within the context of contemporary chromatin biology. For each aim, I summarize relevant background, current limitations, and how the presented work puts forth computational strategies to drive new insights.

Aim 1: Sequence Determinants of Genome Folding

Background

Chromosome conformation capture techniques like Hi-C uncover principles of chromatin folding by comprehensively profiling physical DNA contacts (Lieberman-Aiden *et al.*, 2009; Rao *et al.*, 2014). However, these approaches sidestep how DNA sequence encoding gives rise to observed architectural patterns. Different sequences fold uniquely, but we have only limited knowledge of the complex relationship between genome sequence and structure. Understanding the genome's sequence-encoded "folding code" could reveal regulators

beyond characterized architectural proteins and provide clues to the molecular driving forces underpinning folding.

Several lines of evidence indicate instructions for higher-order packing are embedded within the DNA code. First, the genome itself can be reorganized into active and inactive topological domains in the absence of proteins through intrinsic sequence preferences like GC content (Naumova *et al.*, 2013; Dekker and Mirny, 2016). Second, genetic perturbations demonstrate causal links between sequence and chromatin architecture. Structural variants disrupting TAD boundaries have been linked to developmental disorders (Lupiáñez *et al.*, 2015) and deletions of the Xist gene eliminate X chromosome compaction (Giorgetti *et al.*, 2016). Finally, *in silico* mutations to synthetic genomes predictably reconfigure folding, nominating candidate regulatory motifs (Fudenberg, Kelley and Pollard, 2020). These studies revealed the importance of the zinc-finger protein CTCF in facilitating chromatin loop extrusion and TAD boundary formation through motif-driven DNA binding (Fudenberg *et al.*, 2016; Rao *et al.*, 2017) . Several repetitive elements, such as B2 SINEs, have also been shown to provide sequence specificity in local chromatin folding through recruitment of structural proteins and transcriptional regulators (Bourque *et al.*, 2008; Schmidt *et al.*, 2012).

Still, fundamental questions remain about how DNA sequence encodes chromatin topology. We lack knowledge of the spectrum of regulatory elements governing folding genome-wide and the grammar underlying their positional arrangements. A complete framework will likely involve diverse sequences acting in coordination (Fudenberg, Kelley and Pollard, 2020). Current approaches testing targeted loci have inherently limited scope and remain intrinsically biased. Recent advanced neural network architectures including Enformer (Avsec, Agarwal, *et al.*, 2021), Basenji (Kelley *et al.*, 2018), and Basset (Kelley, Snoek and Rinn, 2016) can accurately predict chromatin accessibility, histone marks, and transcription factor binding directly from raw DNA

sequence context. Specialized models like BPNet (Avsec, Weilert, *et al.*, 2021) and Akita (Fudenberg, Kelley and Pollard, 2020) can predict protein binding and 3D genome folding from underlying sequence features. These DNA sequence models open new avenues for mass *in silico* interrogation. Systematically searching genome sequence could uncover underappreciated elements overlooked by studies focused on well-characterized motifs.

Limitations of Current Approaches

Small scale perturbation experiments offer precise mechanistic insight but lack breadth (de Wit *et al.*, 2015; Guo *et al.*, 2015; Morgan *et al.*, 2017). Genome editing techniques like CRISPR permit targeted manipulation of local folding, but are laborious and low throughput (Kubo *et al.*, 2021). Conversely, bioinformatic algorithms predict motifs that correlate with average folding patterns, but rely on known protein binding preferences (Cuartero *et al.*, 2018). Importantly, correlation does not imply causation, and motifs may simply coincide with other folding mechanisms without directly regulating architecture. Lastly, existing mutagenesis strategies interrogate motifs presumed relevant based on limited precedent, constraining the search space (Rao *et al.*, 2017). Open-ended exploration is needed to realize the full potential of the sequence-encoded folding code.

Aim 2: Comparing 3D Chromatin Folding Patterns

Background

Comparisons of Hi-C maps across conditions have been instrumental for discovering principles of chromatin topology (Lieberman-Aiden *et al.*, 2009; Dixon *et al.*, 2012; Rao *et al.*, 2014). Map differences highlight critical genes, boundaries, and compartments altered across cell types, species, and disease states (Dixon *et al.*, 2015; Rao *et al.*, 2017; Eres *et al.*, 2019). Quantitative scoring of map similarities provides a means to distill

genome-wide chromatin interaction changes into interpretable rankings. By pinpointing regions exhibiting significant reorganization, differential mapping focuses attention on loci that may elucidate general organizational mechanisms.

However, commonly used quantitative approaches for scoring map differences often disagree in their results. Basic metrics like Pearson or Spearman correlation summarize differences into a single number but collapse complex 2D map patterns (Rao *et al.*, 2014). Mean squared error alternately weights absolute differences, causing intensity fluctuations to dominate over structural changes (Fudenberg, Kelley and Pollard, 2020). Calling features like loops risks prioritizing noise when signal is ambiguous (Forcato *et al.*, 2017). Multiple strategies likely offer complementary strengths and weaknesses.

The choice of scoring regime impacts downstream interpretation, but no consensus exists on a gold standard quantification approach. Systematically comparing methods using controlled map perturbations could provide guidelines for selecting suitable techniques based on the folding features and differences of interest. However, experimental Hi-C maps intrinsically reflect complex composite changes that prohibit isolating variables. *In silico* perturbations enable precise control over map alterations to rigorously evaluate scoring regimes.

Limitations of Current Approaches

Existing comparative studies focus on reproducibility, not identifying salient differences (Yang *et al.*, 2017; Yardımcı *et al.*, 2019). Methods are assessed using simulated technical noise or not comprehensively benchmarked against each other (Stansfield *et al.*, 2018; Galan *et al.*, 2020). Studies ranking real maps typically apply one approach in isolation without cross-validation (Schwarzer *et al.*, 2017). This strategy risks prioritizing artifacts specific to the chosen metric. No study offers explicit guidelines for method selection tailored to

research aims. Meanwhile, thousands of Hi-C datasets now exist (Ay and Noble, 2015). Comparing maps from distinct samples, conditions, individuals, and perturbations is crucial to extract biological insights from this wealth of data. Systematically characterizing trade-offs between scoring regimes would aid appropriate application to diverse chromatin interaction datasets.

Aim 3: Connecting Single-Cell Genome Folding to Average Patterns

Background

Genome folding maps averaged across cell populations obscure widespread heterogeneity at the single-cell level (Nagano *et al.*, 2017). Still, bulk analyses still provide critical insights into principles of nuclear organization. This raises a question – do ensemble folding patterns seen in bulk studies accurately represent chromatin folding in any individual cell? Or conversely, could a small subpopulation of cells drive average trends not reflected within most cells (Krietenstein *et al.*, 2020)? Single-cell techniques like DNA microscopy and single-cell Hi-C hold unique promise for decoding links between variable chromatin structure and downstream phenotypes like transcription (Ramani *et al.*, 2017; Bintu *et al.*, 2018). However, realizing this potential requires new strategies tailored to pervasive single-cell heterogeneity.

Unraveling single-cell variability holds two key advantages over population studies. First, heterogeneity itself may prove biologically meaningful. Distinct cell subsets could underpin processes like development and disease (Buenrostro *et al.*, 2018). Second, correlating chromatin and phenotypes like gene expression within the same individual cells could help untangle their murky causal relationship. Enhancer-promoter folding may drive transcription – or transcription machinery may conversely stabilize contacts (van Steensel and Furlong, 2019). Teasing apart these possibilities requires connecting behavior in single cells rather than averaging their contact.

Several obstacles currently limit biological insights from single-cell analysis. First, sparse single-cell data exacerbates challenges in accurately identifying features like compartments, TADs, and loops (Forcato *et al.*, 2017). Second, heterogeneous measurements may reflect technical noise rather than meaningful biology. Third, the sheer scale of variation between single cells obscures how subpopulations relate to bulk patterns and behavior. Computational strategies are urgently needed to overcome these hurdles and realize the potential of single-cell chromatin folding data.

Limitations of Current Approaches

Early methods focus on clustering single cells but do not link groups to bulk principles (Nagano *et al.*, 2017). Recent machine learning techniques predict features like compartments directly from contact maps (Zhang, Zhou and Ma, 2022a) (Zhang et al. 2022). However, this approach still analyzes each cell in isolation. Meanwhile, averaging contact maps by phenotype to find differential patterns obscures the cells driving changes.

Another area of active development is imputing missing single-cell data by sharing information across cells (Zhou *et al.*, 2019). Yet, this assumes cells exhibit stereotyped folds, conflicting with known heterogeneity. We currently lack ways to deconvolve single-cell variation into interpretable components explaining global trends. This gap limits connectivity between variable chromatin architecture and downstream consequences.

2

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

2.1 Abstract

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

2.2 Introduction

The human genome gives rise to its own organization in the nucleus, where the folding of chromatin into intricate and hierarchical structures can be reflective and instructive of cell state(Misteli, 2020). Sequence itself contains the information to create some chromatin features. Binding of CTCF proteins to DNA motifs blocks the extrusion of DNA by motor proteins to create topologically associating domains (TADs) spanning hundreds of megabases (Guo *et al.*, 2015; Fudenberg *et al.*, 2016, 2017; Rao *et al.*, 2017). These dynamic structures permit interaction of elements within their boundaries and limit interaction with elements outside to tune gene expression (Nora *et al.*, 2012; Merkenschlager and Nora, 2016). However, recent reports reveal CTCF may not be the only factor involved, as some contacts remain after CTCF depletion, and interactions across megabases are not affected (Nora *et al.*, 2017; Barutcu *et al.*, 2018). How exactly sequence informs structure ranging from the highest levels of genome organization—chromosome territories and compartments—to the level of individual enhancer-promoter interactions, still remains unclear.

Current approaches relating genome sequence to folding either leverage natural genetic variation or experimentally manipulate particular loci to test specific hypotheses. Applying chromatin capture to genetically diverse individuals revealed single nucleotide variants associated with loss or gain of chromatin contact(Gorkin *et al.*, 2019). Large structural variants are also rare at domain boundaries in healthy humans but not in patients with autism or developmental delay(Fudenberg and Pollard, 2019). To understand the mechanisms underlying these associations, experimental studies have engineered chromatin contact in cells and mice with synthetic tethering(Deng *et al.*, 2012) and CRISPR systems(Morgan *et al.*, 2017; Rege *et al.*, 2018; Kubo *et al.*, 2021) and measured their effects on genome folding and expression of genes such as *Hbb* and *Vcan*. Findings in these individual loci may not apply genome-wide and could overlook mechanisms without known precedent. Here, we propose combining the genome-wide power of population genetics with the precision seen in experimental

studies. We develop a strategy which leverages deep learning to comprehensively screen the human genome for key regulators of 3D genome folding.

Whereas previous machine learning approaches required domain experts to select the most relevant features, deep learning allows patterns to be learned directly from the data without expert input. Deep learning models perform well in predicting enhancer activity(de Almeida *et al.*, 2022; Taskiran *et al.*, 2022), transcription factor binding(Avsec, Weilert, *et al.*, 2021), gene expression(Avsec, Agarwal, *et al.*, 2021), and genome folding (Fudenberg, Kelley and Pollard, 2020; Schwessinger *et al.*, 2020) from sequence, with newer models increasing scale and incorporating ChIP-Seq and ATAC-Seq to provide cell type-specific context(Rui Yang *et al.*, 2021; Tan *et al.*, 2022; Zhou, 2022). The premise for our study is that we can probe these models as computational oracles to predict the behavior of DNA sequence at scales intractable experimentally(Yang and Ma, 2022). Models have been applied to predict the impact of structural variants on human genome folding (Fudenberg, Kelley and Pollard, 2020; Zhou, 2021), confirm the importance of CTCF through computational mutagenesis(Fudenberg, Kelley and Pollard, 2020), and resurrect the folding of Neanderthal genomes(McArthur *et al.*, 2022). These early reports show that many highly disruptive perturbations lack CTCF or annotated regulatory elements, hinting that there may be sequences that encode information needed for genome folding left to uncover.

Here, we leverage Akita(Fudenberg, Kelley and Pollard, 2020), a convolutional neural network trained to predict genome folding from sequence, to perform unbiased and targeted *in silico* mutagenesis experiments at scale. Applying this approach to a human foreskin fibroblast cell line (HFFc6) with high-resolution micro-C data for model training, we discovered wide variability in how robust genome folding is to sequence perturbations. Investigation of sensitive loci revealed both known motifs, like CTCF, and understudied modulators of 3D genome folding, including transposon and RNA gene clusters. These findings replicated in a human embryonic stem cell line (H1hESC) and were supported by experimental Hi-C in loci with

human-specific repetitive elements. Thus, our genome-wide screen revealed a diverse vocabulary of DNA elements that collaborate with CTCF to orchestrate TAD-scale chromatin organization.

2.3 Results

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

To measure sequence importance to chromatin organization, we developed a deep-learning scoring strategy to computationally introduce modifications into the human reference genome and predict their impact on genome folding with Akita(Fudenberg, Kelley and Pollard, 2020). Given a ~1-megabase (Mb) DNA sequence, this model accurately produces a chromatin contact map at ~2-kilobase (kb) resolution, where TADs and DNA loops are visible. Akita has previously been used to perform sequence mutagenesis experiments ranging from one nucleotide to thousands of basepairs(Fudenberg, Kelley and Pollard, 2020; McArthur *et al.*, 2022). To build a flexible *in silico* screening strategy based on Akita, we wrote computationally efficient code that quantifies the impact of a centered sequence variant, which we call *disruption*, as the log mean squared difference between the predicted contact frequency map for the 1-Mb sequence with a sequence alteration compared to that of the reference sequence. If a variant dramatically rearranges how the genome is predicted to fold, we infer that the altered sequence could regulate chromatin contacts.

In this study, we used disruption scores to perform a variety of genome-wide screens across millions of genetic perturbations, including targeted and unbiased deletions, insertions, and substitutions ranging from 1 base pair (bp) to 500,000 bp (**Fig. 2.1a**). In contrast to *in vivo* genetic perturbations, our approach enables precise and flexible genome editing at scale. We first assessed all 5-kb deletions tiled
across the genome for their impact on folding in HFFc6 cells (n=574,187). Deletions are highly variable, and around half produce changes to chromatin contact maps that are noticeable by eye (**Fig. 2.1b**). Some sequence

deletions completely rearrange the boundary structure of contact maps, some result in small focal changes (e.g., gain or loss of a loop anchor), and some produce no change at all, suggesting the chromatin structure is robust to sequence manipulation (**Fig. 2.1c)**. As expected, regions of the genome with many CTCF motifs are particularly sensitive while regions with no motifs are perturbation-resilient (**Fig. 2.1d**). In sum, 62.1% of the most sensitive sequences (top decile of scores) fall within 5 kb of CTCF-bound distal enhancers, compared to only 7.3% of the most robust sequences (bottom decile of scores), establishing that our approach identifies known genome folding mechanisms (**Fig. 2.2a**).

2.3.2 Perturbing euchromatin disrupts genome folding

Disruption scores are also correlated with chromatin compartment, as measured by the first eigenvector of the experimental HFFc6 micro-C contact matrix (Pearson's r = 0.522, $P < 1 \times 10^{-300}$, n = 11,413; **Fig. 2.1e**) (Krietenstein *et al.*, 2020). The mean disruption score within gene-rich and open A compartments is 14.6% higher than in compact, inactive B compartments. Motivated by existing work illustrating gene-rich GC-rich regions fall in A compartments, while GC-poor regions, like lamina-associated domains, are known to self-interact with each other and other GC-poor regions across chromosomes, we next directly evaluated the role of GC content in disrupting genome folding(Naughton *et al.*, 2013). We observe that high gene density and GC content are both associated with peaks in disruption scores (**Fig. 2.1d, Fig. S2.1a-c**). Using HFFc6 total RNA-Seq (ENCODE Project Consortium *et al.*, 2020), we quantified transcription in each 5-kb window and observed a strong correlation with disruption scores (Pearson's r = 0.366, $P < 1 \times 10^{-300}$, n = 11,413). Other genomic features associated with active chromatin are also more frequent in the most sensitive sequences, including distal and proximal enhancers and promoters (**Fig. 2.2a, S2.2**). In sum, it is difficult to perturb inactive chromatin and easy to perturb active chromatin.

The correlation between many of these features reflects an inherent challenge in disentangling which are causal and which are reflective of genome folding (**Fig. S2.1c**). Indeed, regions that are in A compartments, contain CTCF binding sites, and are actively transcribed are also the most sensitive (**Fig. 2.2b**). The effect of CTCF holds within both A and B compartments (**Fig. S2.3a-b**), indicating that it is directly associated with sensitive 5-kb bins and is not just a proxy for A compartments. However, both transcription and compartment are more impactful individually than the presence of CTCF motifs, suggesting additional rules govern which CTCF sites are in use and which are redundant or decommissioned in a given cell type. Overall, our findings suggest that independent mechanisms at transcriptionally active sites may collaborate to coordinate genome folding.

2.3.3 Transcriptionally active regions modulate folding alongside CTCF

Chromatin contact and transcription are correlated, but which mechanistically precedes the other is currently an area of active investigation. While transcription is classically thought to result from enhancer-promoter interaction constrained by chromatin structure, transcriptional machinery may help to scaffold local chromatin structure as well(van Steensel and Furlong, 2019). CTCF binding, for example, is essential for boundary formation and may also influence activity of some promoters(Nora *et al.*, 2017), and emerging work reveals RNA polymerase II and transcription may separately influence 3D genome folding(Busslinger *et al.*, 2017; Zhang *et al.*, 2021). To test this hypothesis, we evaluated all single-nucleotide mutations in the 300 base pairs (bp) on either side of the transcription start site (TSS) of the 1,789 highest expressed protein coding genes in HFFc6(ENCODE Project Consortium *et al.*, 2020) and compared disruption scores to expression level in regions where CTCF motifs are present or absent (**Fig. 2.2c**). In regions flanking a CTCF motif, we observed a strong peak in disruption directly upstream of the TSS (**Fig. 2.2f, Fig. S2.4**). The periodic pattern is more

detailed than underlying CTCF motifs and more precise than a sum of CTCF ChIP-Seq peaks around the TSS. Metaplots of the average change in contact reveal that mutations weaken boundaries at the TSS. Our analysis points to a presence of CTCF at the promoters of highly expressed genes, where some CTCF motifs are selectively bound and some are not. We note that when no CTCF is present, disruption significantly lower but still still slightly elevated upstream of the TSS of highly transcribed genes (**Fig. 2.2g, Fig. S2.4**). Furthermore, disruption scales with gene expression both when CTCF is present and absent (**Fig. 2.2d,e**). These results are consistent with the hypothesis that active transcription may provide an alternate means of stabilizing DNA-DNA interactions in TSS devoid of CTCF sites through uncharacterized mechanisms, like transcriptional machinery, nascent RNA, or recruited regulatory RNA.

2.3.4 In silico screening approach validates across cell lines

To test the robustness of our approach and findings, we repeated the above analyses in a second cell line. We selected H1hESCs due to the availability of micro-C data and the opportunity to compare a pluripotent cell line to a differentiated one. Furthermore, H1hESC is one of the five cell lines for which the Akita model predicts chromatin contacts, enabling us to directly assess the effects of *in silico* disruptions on genome folding patterns in H1hESC alongside HFFc6. This analysis showed that all of the above trends observed in HFFc6, including disruptions scores being associated with CTCF motifs, transcription, A compartments, GC content, and the deleted sequence length, are consistent in H1hESC (**Fig. S2.5**).

2.3.5 Transposon clusters modulate genome folding independently of CTCF

At the chromosome scale, our unbiased genome-wide screen highlighted clusters of Alu elements and some other repetitive elements alongside peaks in disruption scores, motivating us to explore their role in 3D genome folding (**Fig. 2.3a**). DNA and RNA transposons replicate and insert themselves into DNA, and constitute over 50% of the human genome(Trigiante, Blanes Ruiz and Cerase, 2021; Nurk *et al.*, 2022). They are rich in transcription factor binding sites(Wang *et al.*, 2007; Bourque *et al.*, 2008; Kunarso *et al.*, 2010), suggesting that some may have been evolutionarily repurposed as regulatory elements. Growing evidence indicates they provide a source of CTCF motifs across the genome and serve as both loop anchors and insulators(Raviram *et al.*, 2018; Choudhary *et al.*, 2020; Diehl, Ouyang and Boyle, 2020). To measure the impact of different families of repetitive elements on 3D genome sensitivity, we compared disruption of 5-kb windows containing repetitive elements to those with none. Several families exhibit greater sensitivity to perturbation than CTCF containing regions (e.g., Alu, SVA, scRNA, srpRNA; **Fig. 2.3b**). Disruption scores of repetitive elements are not correlated with mappability, indicating that poor micro-C read mapping in model training data does not bias this result (**Fig. S2.6, Supplemental Note**). As with CTCF(Kentepozidou *et al.*, 2020), regions with higher numbers of Alu elements are more disruptive upon deletion: the disruption score of 5-kb windows with 5 or more Alu elements is 9.88% higher than that of windows with no elements ($P < 1.54 \times 10^{-291}$; **Fig. 2.3c**). This clustering effect holds across many repetitive elements, including MIR and L2 LINE elements, as well as across most small, non-coding RNA genes (**Fig. 3c**). Many families, like L1 LINE elements, show no correlation at all, and trends are consistent across both A and B compartments, hinting that clustering is family specific (**Fig. 2.3c**, **Fig. S2.3c-d**).

To investigate the contribution of repetitive elements independently of flanking sequence, we next individually deleted over 1 million elements in the RepeatMasker database (**Fig. 2.4a**). Overall, many elements create large-scale boundary shifts, with some causing increases and others decreases in contact frequency (**Fig. 2.4b)**. Deletions of almost all families are more disruptive than random deletions, and deletions of families such as Alu, small RNAs, SVA, and hAT-Charlie are on par with or exceed deletions of CTCF sites across the genome (**Fig. 2.4c**). Disruption is moderately correlated with size, but many highly disruptive element families are relatively small and cause unexpectedly large disruptions given their length (**Fig. S2.1d-e, Fig. 2.4c**). For example, deletion of tRNAs, scRNAs, srpRNAs, and snRNAs–all under 130 bp on average–have a propensity to drastically alter genome folding.

In order to experimentally validate these deep-learning based predictions, we leveraged the natural sequence differences between humans and chimpanzees. Specifically, we examined loci with human lineage specific repetitive elements in Hi-C data that we previously generated in human and chimpanzee neural progenitor cells(Keough *et al.*, 2023). By comparing the experimental data to Akita predictions where the human-specific repetitive element is inserted into the chimpanzee genome and conversely deleted from the human genome, we find that Alu elements unique to humans generate consistent changes to genome folding (**Fig. S2.7**). Thus, experimental data validates our *in silico* screening approach and supports the importance of Alu and other repetitive elements in genome folding.

We next explored possible mechanisms through which repetitive elements might influence genome folding. Causality is challenging to untangle since each repetitive element can contain features with known associations to chromatin organization. First, the lengths of repeat clusters are roughly similar to clusters of CTCF motifs at TAD boundaries (**Fig. 2.4c**). Second, several repeat families are known to harbor CTCF motifs(Schmidt *et al.*, 2012). Third, some repeats have a strong GC bias (e.g., Alu GC% > 50%), potentially allowing them to establish compartments(Su *et al.*, 2014; Lu *et al.*, 2021). Finally, repetitive elements collectively

account for a large amount of total nuclear transcription(Trigiante, Blanes Ruiz and Cerase, 2021). To dissect the contributions of CTCF and active transcription versus other features of repetitive elements, we quantified overlap of these two annotations with repetitive elements that have the highest disruption scores. Only 5.86% of the 10% most disruptive elements contain a CTCF motif while 13.55% are actively transcribed (**Fig. 2.4d**), so a majority overlap neither. Disruptive repetitive element deletions are enriched at distal enhancers that are not CTCF bound (**Fig. S2.3f**). These findings hint that repetitive elements may aid in genome folding independently and in collaboration with CTCF and transcription.

To understand the folding phenotypes of element deletions, we next averaged the changes in contact frequency for the top-scoring elements of each family (**Fig. 2.4e**). ERVK elements behaved like CTCF sites: their deletion led to a strong and centered loss of a chromatin boundary. Other repeat families created an off-diagonal gain in contact, as seen with Alu and hAT-Charlie, dispersed focal disruptions, as with non-coding RNAs, and stripes, as with SVA elements. To demonstrate that the model is internally consistent, we performed a *phenotypic rescue*, where we deleted an individual hAT-Tip100 element to produce a large change in contact and attempted to restore the original folding pattern with a different sequence (**Fig. 2.4f**). While introducing random DNA or a CTCF motif did not recreate the original contact, inserting a related MER91B hAT-Tip100 element did. We conclude that repetitive element families are associated with distinct chromatin contact map features, and elements within a family are generally functionally interchangeable.

2.3.6 Insertion of repetitive elements leads to distinct folding phenotypes

Our deletion experiments do not distinguish between repetitive elements that collaborate with CTCF to weaken or strengthen nearby TAD boundaries and those that separately create chromatin contact. To isolate the effects of repetitive elements, we next designed *in silico* insertion experiments. We first engineered a “blank canvas” with

no predicted structure by depleting a randomly generated 1 Mb DNA sequence of all CTCF-like motifs (**Fig. 2.5a, Fig. S2.8**). We then inserted one or more copies of any query sequence into this 1 Mb and quantified newly arising chromatin contacts. We easily recreated a division closely resembling a TAD boundary by inserting multiple copies of the canonical CTCF motif (**Fig. 2.5b**), validating this approach in creating chromatin contact phenotypes.

After introducing the 1,000 most disruptive repetitive elements in our deletion screen into a blank canvas, we found a majority also changed contact with insertion, including 80.3% of Alu elements and 86.0% of ERVK elements (**Fig**. **2.5c**). Additional copies strengthened impact, and fewer copies were needed to induce a chromatin boundary compared to the CTCF motif (**Fig. 2.5e**, **Fig. S2.9a)**. Clustering the insertion maps revealed hAT/MIR insertions produced distinct folding patterns from ERV/SVA element insertions (**Fig. 2.5d**). Alu elements consistently produced focal changes at the site of insertion that appeared unlike CTCF-like boundaries. Curiously, repetitive elements seem to produce two distinct modifications to 3D structure upon insertion. Some elements create CTCF-like domain boundaries which increase in strength as more elements are inserted (**Fig. 2.5f**). Other elements, like the Alu and SVA families, form a pattern resembling a cross, with increased contact both upstream and downstream from the insertion point. This cross-like pattern increases in size with more element insertions. Insertions of tRNA genes did not create new boundaries, suggesting that their effect on 3D genome folding may be context dependent.

Some repetitive elements harbor CTCF motifs and overlap with CTCF ChIP-Seq peaks, strongly suggesting that the Akita model predicted their importance because they contain CTCF binding sites. To test this hypothesis, we performed saturation mutagenesis across a number of high scoring repetitive elements (**Fig. 2.5e**). Screening an ERVK element, for example, revealed that the single nucleotides predicted to have the highest importance for contacts lie directly at the center of a CTCF binding site (**Fig. 2.5e**). Overall, the closer a sequence is to matching the canonical CTCF motif, the larger the predicted impact of its insertion (**Fig.**

S2.10a-b). Still, most of the elements that produced contact changes had no CTCF overlap, and the 5 to 50-bp motifs within these elements with the greatest impact did not resemble CTCF motifs (**Fig. 2.5e, Fig. S2.10c-d**). Therefore, insertions support the hypothesis that repetitive elements contain sequence determinants of 3D genome folding beyond CTCF motifs.

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

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

First, we examined CTCF motifs. Fudenberg et al. mutated all motifs in the JASPAR transcription factor database and determined that CTCF and CTCFL are most sensitive to sequence perturbation(Fudenberg, Kelley and Pollard, 2020). To complement this work, we inserted all motifs into a blank map. We find that regardless of motif spacing CTCF and CTCFL are the transcription factor motifs best able to induce genome folding independently of any surrounding genomic context, followed by HAND2, Ptf1A, and YY2 (**Fig. 2.6a**, **Fig. S2.11a)**. YY1 scores relatively lower, perhaps due to its less stable binding or its binding with co-factors(Hsieh *et al.*, 2022). Sampling and inserting motifs from the CTCF position weight matrix, we found that the consensus sequence creates a stronger boundary than 99.50% of CTCF variants (**Fig. 2.6b**). However, a small minority of CTCF "super-motifs" with a T at positions 8 and 12 outperformed the canonical motif, hinting that the most common CTCF motifs may not be the most strongly insulating ones. The super-motif sequences also produced stronger boundaries in experimental Hi-C than do the CTCF consensus sequence (**Fig. S2.11b-c)**, and they are equally likely to overlap CTCF ChIP-seq peaks. These results

illustrate that Akita can be used to interpret the function of CTCF and other transcription factor binding sites at single nucleotide resolution.

Next, we dissected Charlie 7, a 367-bp AT-rich (29% GC) hAT-Charlie element on chromosome 11. Deleting Charlie 7 eliminates chromatin interactions (**Fig. 2.5c; Fig. 2.6c**). Inserting twenty tandem copies of Charlie 7 creates a CTCF-like boundary, despite no subsequence resembling a CTCF motif. This boundary could not be reproduced by inserting a shuffled Charlie7 sequence or a random sequence of the same length. We therefore shuffled individual 10-bp segments of Charlie 7 to destroy local sequence grammar before reinserting the element into the blank canvas. Shuffling the final 60 bp had the same effect as shuffling the entire element, revealing that this end of the element is necessary for boundary creation (**Fig. 2.6d**). We then created sliding windows of 10 bp, 50 bp, and 100 bp along the element and inserted each subsequence into the blank canvas. No individual subsequence was sufficient to reproduce the effect of the entire element (**Fig. 2.6e**). However, shuffling the first 307 bp while maintaining the last 60 bp intact did create a strong boundary. Since the GC content of Charlie7 is unusually low, we next replaced parts of the element with random GC-matched sequence. A length-matched sequence with a GC content below 30% and the final 60 bp of the Charlie7 element was sufficient to create a boundary (**Fig. 2.6f**). Completely random insertions with a GC content below 30% and above 60% are also highly impactful (**Fig. S2.9c-d**). Based on these *in silico* experiments, we conclude that GC content along with sequence syntax could be critical for the insulating behavior of Charlie7. Looking across all disruptive retrotransposons, we identify several families with very extreme average GC content (**Fig. S2.8b**), suggesting the intriguing hypothesis that abrupt shifts in GC content resulting from repetitive element insertions into genomic DNA contribute to genome folding.

2.4 Discussion

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

This study contributes to a growing body of evidence showing that transposable elements modulate genome folding(Zhang *et al.*, 2019) and replication timing(Yang *et al.*, 2022). It has long been hypothesized that transposons may have been evolutionarily co-opted as regulatory elements(Bourque *et al.*, 2008; Kunarso *et al.*, 2010). Most transposable elements are decommissioned by chromatin modifications(Slotkin and Martienssen, 2007), but functional escape can change genome conformation(Huda, Mariño-Ramírez and Jordan, 2010). We observe both loss and gain of contact upon transposable element deletion, supporting the idea that these elements can both establish new boundaries by installing CTCF-like motifs and inhibit ancient CTCF binding sites to block contact(Choudhary *et al.*, 2020). Our results are also consistent with previous findings that specific MIR elements and tRNAs can serve as insulators (Van Bortle *et al.*, 2014; Wang *et al.*, 2015), while Alu and hAT provide loop anchors(Ferrari *et al.*, 2020; Choudhary *et al.*, 2022), and hint that repetitive elements may

work in tandem(Lu *et al.*, 2021). Cao et al., for example, identified that many transposable element families, but MIR SINE elements and L2 LINE elements in particular, are enriched for binding sites and active chromatin marks, appear in the vicinity of tissue-specific gene expression, and interact with each other extensively to collaborate as enhancers or repressors(Cao *et al.*, 2019). In future work, it would be exciting to test coordination of transposable elements as *shadow loop anchors*, theorized by Choudhary et al. to act as redundant regulatory material supporting CTCF(Choudhary *et al.*, 2020). We anticipate that comparing disruption to element age and species divergence will help us to understand the evolutionary mechanisms of transposable element deprogramming and selection in gene regulation.

Although we did not focus on CTCF specifically, a similar targeted *in silico* approach could directly address why the majority of CTCF motifs are not active(Kim *et al.*, 2007; Chen *et al.*, 2012), and if methylation sensitivity of CTCF motifs containing CpGs tunes folding specificity(Hark *et al.*, 2000). We also anticipate future *in silico* experiments and investigation of the model with activation maximization(Shrikumar, Greenside and Kundaje, 2017) will refine the spacing and orientation rules of neighboring and redundant CTCF elements and reveal how CTCF coordinates with flanking proteins and transposable elements.

It is important to emphasize that our *in silico* strategy, while demonstrated here and previously to be highly accurate(Fudenberg, Kelley and Pollard, 2020), is a screening and hypothesis-generating tool. Model predictions, especially those that implicate novel sequence elements or mechanisms, will require further experimental validation. We view this as a strength of our approach, not a weakness. Our ability to test millions of mutations efficiently and in an unbiased manner enables us to develop hypotheses and prioritize genomic loci that would not otherwise have been considered for functional characterization. It is now a high priority to apply massively parallel reporter assays, epitope devices, and genome engineering to explore how hAT, MIR, ERV and SVA elements function in the context of 3D genome folding. We advocate for deep learning as a powerful

strategy for driving experimental innovation which can be used iteratively with wet lab technologies to accelerate discovery.

Our conclusions rest heavily upon the Akita model, which only considers a limited genomic window. Future work could apply the approach presented here with other deep-learning models to test the robustness of our findings and potentially discover additional sequence features missed in our work. Our method scores the entire 1Mb contact map and weights all regions equally, which may be too insensitive to capture small changes to specific loci. Filtering or weighting regions of the predicted contact maps by overlap with functional genomic annotations during score computations could also help to selectively test specific hypotheses. Our study is further limited by the quality of the hg38 reference genome, and we anticipate that extending to the new telomere-to-telomere human genome assembly will enable a better understanding of near-identical repetitive elements(Nurk *et al.*, 2022). Finally, in order to leverage the best quality data currently available, we only made predictions across HFFc6 and H1hESC, but features of the 3D genome can be cell-type specific(Schmitt *et al.*, 2016). As very high-resolution and single-cell measurements of chromatin contacts, gene expression, and accessibility are generated for more cell types, it will be exciting to search for sequences that are necessary and sufficient for chromatin contacts in each cell type and to explore how variable these sequence determinants are across cellular contexts.

In our investigation, we develop a toolkit of *in silico* experimental strategies, including: unbiased and targeted deletion screens, phenotypic rescue, insertions into synthetic sequence, sampling around known sequence motifs, and sequence contribution tracks across tens of basepairs to megabases. We hope that the variety of experiments profiled here may serve as a template for foundational biological research with deep learning. We also anticipate that our released disruption tracks will provide useful annotations for genome sensitivity and yield further insights into chromatin biology (**Supplementary Data Table 2.1**). In sum, our

work highlights the potential of deep learning models as powerful tools for biological hypothesis generation and discovery in regulatory genomics.

2.5 Methods

2.5.1 Akita model and datasets

Throughout this analysis, we use the published convolutional neural network Akita to predict log(observed/expected) chromatin contact maps from ~1 Mb (1,048,576 bp) of real, altered, or synthetic DNA sequence(Fudenberg, Kelley and Pollard, 2020) (https://github.com/calico/basenji/tree/master/manuscripts/akita). All types of mutations, including deletions, insertions, inversions and substitutions, may be scored as long as they are smaller than 1 Mb. Akita's predictions have been shown to mirror experimental results with deletions across scales of thousands of base pairs (bp) to single nucleotides. Fudenberg et al. originally trained Akita across six cell-types simultaneously, and we made all predictions in this work in the cell-type with the highest resolution of training data, human foreskin fibroblasts (HFFc6). We find that disruption in H1hESC is highly correlated (**Fig. S2.5**). The experimental Micro-C maps from HFFc6 (Krietenstein *et al.*, 2020) are used in visualizations. All chromatin and transcriptomic data were generated in HFFc6 and downloaded from public repositories. The source of all public data, including Micro-C, ATAC-Seq, RNA-Seq, ChIP-Seq, and compartment calls, can be found in the key resources table. All analyses use the hg38 genome build. We downloaded centromere locations from UCSC Table Browser (Kent *et al.*, 2002).

2.5.2 Computing 3D genome folding disruption scores and deletion screens

The location of deletions and insertions are centered such that the start position of the variant is always introduced halfway through the 1-Mb sequence at 2^{19} bp. For deletions, we pull additional sequence from the right to pad the input to 2^{20} bp. We remove sequences from our analysis which overlap centromeres (Miga *et al.*, 2014), ENCODE blacklisted regions (Amemiya, Kundaje and Boyle, 2019), and regions with an N content greater than 5%. Evaluating predictions on GPU (NVIDIA GeForce GTX 1080 Ti, NVIDIA TITAN Xp, NVIDIA GeForce RTX 2080 Ti) decreased the time per variant from 1.58 seconds to 262 ms, on average.

We score disruption as the log of the mean squared error between reference and perturbed maps. Mean squared error captures large-scale contact map changes, and has been used previously to rank predictions (Fudenberg, Kelley and Pollard, 2020). Pearson/Spearman correlation is also an appropriate choice (McArthur *et al.*, 2022).

Mass deletion screens

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

1. *5 kb, whole genome* (n = 562,743).
2. *10,000 (10k) random CTCF deletions*. CTCF locations are pulled from JASPAR 2022 (Castro-Mondragon *et al.*, 2022).
3. *10k 100-bp random deletions.* Start locations are randomly sampled from the genome.

4. *Randomly sized deletions*, ranging from 1 bp to 100 kb (n = 41,207). Start locations are randomly sampled from the genome.
5. *RepeatMasker database deletions* (n = 1,164,107) *(Smit, AFA, Hubley, R & Green, P., no date)*. RepeatMasker downloaded from UCSC Table Browser. We exclude ambiguous elements (containing '?' in the label). We initially sample 10,000 elements per family or up to the total number of elements in the family, whichever is less. Thereafter, we randomly sample from the database.
6. *TSS deletions*. (n = 1,073,329 mutations across 1,789 genes).

A full summary as well as the location of these results can be found in **Supplementary Data Table 2.1**.

Genomic tracks

We smoothed the disruption scores of 5-kb deletions with a rolling average of 50 bp to create disruption tracks (**Fig. 2.1d**, **Fig. 2.3a**). We additionally visualize the density of the following elements at 5-kb resolution:

1. Reference genes, hg38, GENCODE v39 (Frankish *et al.*, 2021), downloaded from UCSC Table Browser.
2. ENCODE hg38 v3 candidate cCREs, ENCODE Project (ENCODE Project Consortium *et al.*, 2020) , downloaded from UCSC Table Browser.
3. CTCF motifs (MA0139.1), JASPAR 2022 (Castro-Mondragon *et al.*, 2022), downloaded from http://expdata.cmmt.ubc.ca/JASPAR/downloads/UCSC_tracks/2022/hg38/.
4. ATAC-Seq peaks in HFFc6 (Akgol Oksuz *et al.*, 2021).
5. Alu, L1, and L2 elements, RepeatMasker database, v. 4.1.2 (Smit, AFA, Hubley, R & Green, P., no date), downloaded from UCSC Table Browser.

Overlap with genomic annotations

We used pre-computed compartment scores generated from the HFFc6 Micro-C dataset originally employed for training Akita (Krietenstein *et al.*, 2020). To calculate the overlap between disruption scores for 5-kb deletions and compartment scores generated at 50-kb resolution, we merged both measures by genomic location, filled missing disruption values with linear interpolation, and calculated the overlap across A compartments with a compartment score greater than 0 and B compartments with a compartment score less than 0.

We intersected deleted windows and transposable elements with ENCODE cCREs using bioframe (Open2C, Abdennur, Fudenberg, *et al.*, 2022) to calculate the percentage overlap. We use the same strategy to calculate overlap with JASPAR CTCF motifs, ATAC-Seq peaks, and transcribed elements. When quantifying transcription of repetitive elements unannotated as genes, we calculated overlap with RNA-seq BigWigs, summed across both strands.

Mappability

Per nucleotide mappability was measured using 24-kmer multi-read mappability, where mappability is the probability that a randomly selected read of length k in a given region is uniquely mappable (Karimzadeh *et al.*, 2018). Mappability tracks were downloaded from the Hoffman lab (https://bismap.hoffmanlab.org). In this study, mappability averaged across 5 kb deletions, repetitive element families, and Alu element types in a 100 Mb subset of chromosome 1 from 100 Mb to 200 Mb.

In silico mutagenesis at the TSS

We examined behavior at the TSS using *in silico* mutagenesis. We individually randomly mutated each nucleotide 300 bp upstream to 300 bp downstream of the top 1,789 highest expressed protein coding genes via total RNA-Seq and quantified the MSE between mutated and reference predicted maps. We observed that 1,015 genes fell in A compartments, while 63 fell in B compartments. To produce tracks in **Fig. 2.2f-g**, we averaged the disruption of each nucleotide by position and smoothed using a rolling average of 20 bp. We used the same strategy across select repetitive elements to identify which nucleotides most contribute to entire-element disruption scores (**Fig. 2.5e**). To create metaplots, we selected the highest scoring nucleotide change for each gene, and filtered all genes with a maximum disruption score above -7. We then averaged the difference between reference and perturbed maps for these genes.

Repetitive elements

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

Phenotypic Rescue

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

1. A MER91B hAT-Tip100 element at position chr2:98412915-98413053.

   SWA score: 392, Divergence: 27%. Disruption from reference = -2.65.

2. A size-matched 138-bp random DNA sequence.

Disruption from deletion = -2.55.

3. The canonical CTCF motif (TGGCCACCAGGGGGCGCTA).

   Disruption = -2.68.

4. A MER91B element at position chr12:51824097-51824219.

   SWA score: 245, Divergence: 20.9%. Disruption = -5.28.

2.5.3 Insertion screens

CTCF depletion: We created a simulated Hi-C contact map without structure as a blank canvas for insertion experiments. We first generated a random DNA sequence of length 2^{20} bp. By chance, predicted maps from random sequence will contain some above background contact frequencies. To remove all structure, we incremented across this sequence one nucleotide at a time with a 12-bp sliding window. For each position, we computed the edit distance to the consensus CTCF motif. If the edit distance fell below a set threshold, we inserted a random DNA sequence of length 12 until the subsequence was sufficiently different from CTCF. Experimenting with edit distances, we found that a distance of 7 produces predicted maps which lack structure but do not result in artificial model predictions (**Fig. S2.8**). We call this a "blank canvas" 1-Mb sequence.

***CTCF insertion*:** We inserted the CTCF motif into the blank canvas and predicted expected contact frequencies with Akita. We quantify insertion impact as the log mean squared error between the predicted maps of the blank canvas and the insertion. If more than one motif was added, the insertions were centered and separated by an arbitrary 100 bp. To sample the CTCF motif, we drew frequencies from the CTCF position weight matrix (Castro-Mondragon *et al.*, 2022). To create a baseline, we inserted 5,000 CTCF motifs drawn from locations in the genome. Sequence motifs were visualized with a python port of the seqLogo package(Bembom, no date; Sherman, no date).

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

***Additional genomic tracks*:** In **Fig. 2.5e**, we visualized CTCF ChIP-Seq and CTCF motif locations in the element's original genomic context. Along with deleting the entire element, we performed mutagenesis to a random nucleotide across the length of the element to create a 'disruption track' of nucleotides most sensitive to perturbation. We highlight the most sensitive bases.

***JASPAR Insertions*:** We inserted the forward and reverse complement of each JASPAR motif(Castro-Mondragon *et al.*, 2022) into CTCF-depleted sequence with 100-bp spacing (n = 842). JASPAR motifs were pulled and coordinated with pyJASPAR (Khan, 2021).

2.5.4 Quantification and statistical analysis

Disruption score significance

Pearson correlation coefficient of disruption scores compared to several other genomic annotations was calculated using scipy.stats.linregress (**Fig 2.1e, 2.1d**, **Supplemental Fig. 2.1c**) (Virtanen *et al.*, 2020). To assess

if the relationship between disruption and additional annotations was significant, we performed a two-sided Mann-Whitney-Wilcoxon test with compartment annotations (**Results**, **Fig. 2.1e**) and transcription level using HFFc6 total RNA-Seq (**Results**) in 5-kb genome windows genome-wide using scipy.stats. The number of bins considered (*n*) and p-values are provided in the Results section. A two-sided test was performed because no directionality was assumed. A two-sided Mann-Whitney-Wilcoxon test was also used to assess the significance of disruption between genomic windows containing no Alu elements and windows with 5 or more (**Results**, **Fig. 2.3c**).

Motif significance

We evaluated the presence of CTCF in deleted and inserted transposable elements with overlap of CTCF ChIP-Seq, overlap of annotated CTCF motifs, and hamming distance to the canonical CTCF motif. Significance of a CTCF match was evaluated using FIMO from the MEME suite (Bailey *et al.*, 2015).

2.6 Figures

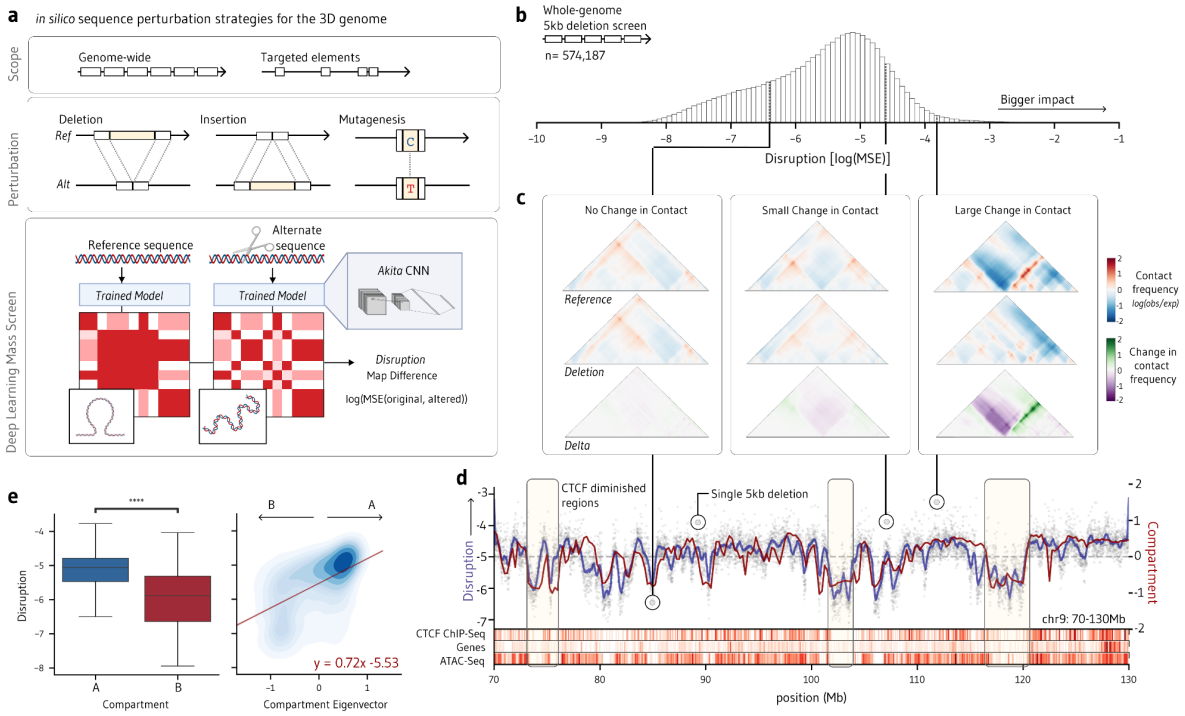


Figure 2.1: In silico deletion screen indicates the impact of sequence perturbation on 3D genome folding is highly variable.

a. We quantify how important DNA sequence is to genome folding by introducing whole-genome and targeted deletions, insertions, and point mutations and comparing the predicted Hi-C contact maps to maps predicted from the reference sequence. We score disruption as the log mean squared difference of the perturbed map relative to the reference map (MSE). Variants with high disruption scores are inferred to contribute to 3D genome folding. **b.** A genome-wide tiled 5-kb deletion screen produces a distribution of sequence importance with log(MSE) between -10 and -1 for the HFFc6 cell type. **c.** Genome-wide screens capture a range of disruption scores; some sequences do not change predicted genome folding (left panel), some produce small focal changes (middle panel), and others dramatically rearrange boundaries (right panel). **d.** The rolling average of disruption and compartment score across a 60-Mb region of chromosome 4. Peaks correspond to regions sensitive to perturbation, while valleys indicate regions robust to perturbation. Yellow shading highlights genomic regions with relatively few CTCF motifs. These regions have low disruption scores, suggesting that their perturbation has little effect on genome folding. **e.** Sensitivity to disruption correlates strongly with compartment score, as measured by the first eigenvector of HFFc6 micro-C.

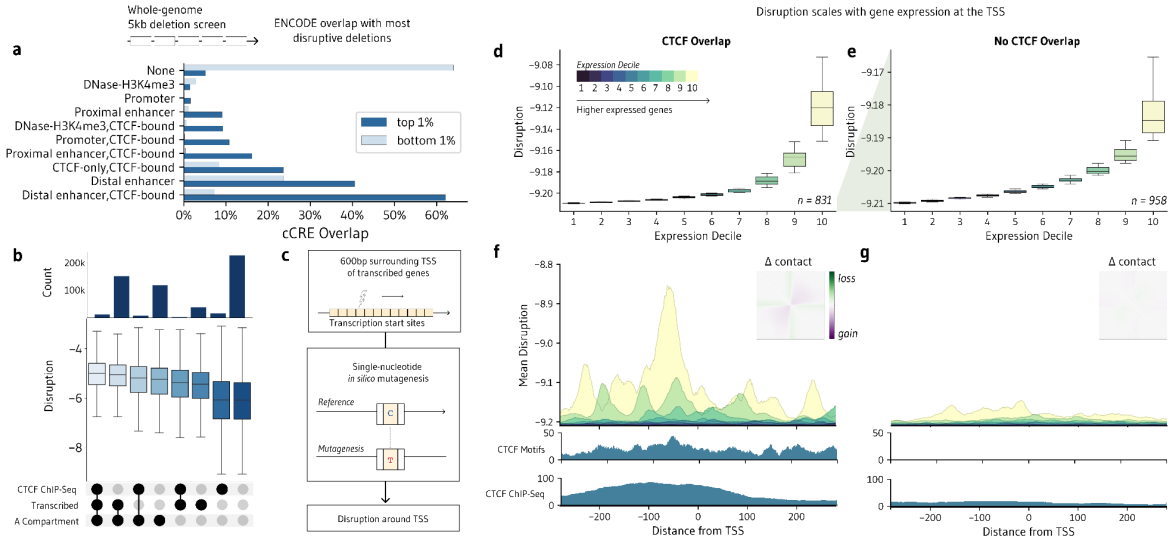


Figure 2.2: Transcription and CTCF are key modulators of 3D genome folding.

a. Overlap between top 1% (most disruptive; dark blue) or bottom 1% (least disruptive; light blue) 5-kb sequence deletions and ENCODE candidate cis-regulatory elements, quantified as the proportion of deletions with overlap. Each deletion may overlap with more than one regulatory element. **b.** Average disruption score across genomic regions overlapping with CTCF ChIP-seq peaks, A compartments, and/or actively transcribed sequences. **c.** Single base-pair mutagenesis screen of a 600-bp segment surrounding the transcription start site (TSS) of the most highly transcribed genes in HFFc6 (n=1,789). **d-e.** Mean disruption score of transcribed genes, stratified by expression level decile (colors), and separated into those whose TSS region overlaps (**d**) versus does not overlap (**e**) with CTCF sites. The figures have different scales. **f-g.** Average disruption score of each base at TSS regions with (**f**) and without (**g**) a CTCF motif overlap, stratified by expression decile (colors), along with average CTCF motif density and CTCF ChIP-seq. Metaplots (upper right) show the average change in contact for the 100 TSSs with the most significant disruption scores.

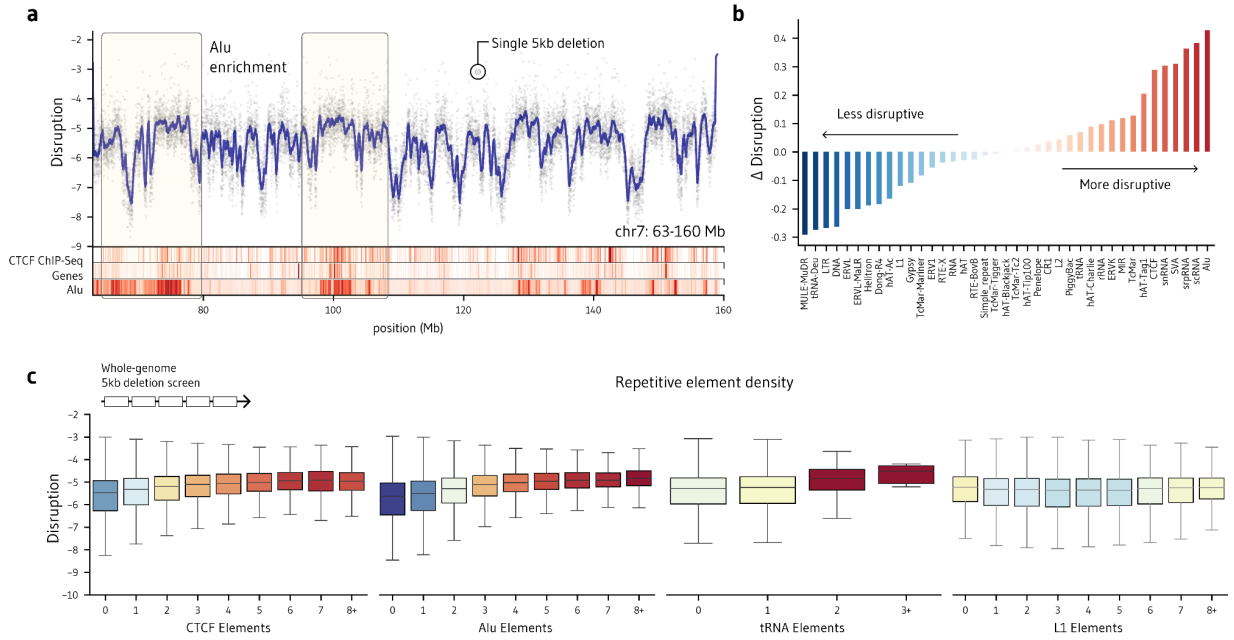


Figure 2.3: Regions with repetitive elements are sensitive to sequence perturbation.

a. Mean disruption scores of tiled 5-kb deletions across a 100-Mb region of chromosome 7. Tracks below the plot illustrate the density of CTCF motifs, genes, and Alu elements. **b.** Mean difference in disruption scores between windows containing at least one repetitive element and windows containing none, stratified by family. **c.** Disruption scores of 5-kb deletions stratified by the number of Alu elements, tRNAs, L1 LINE elements, and CTCF motifs they contain.

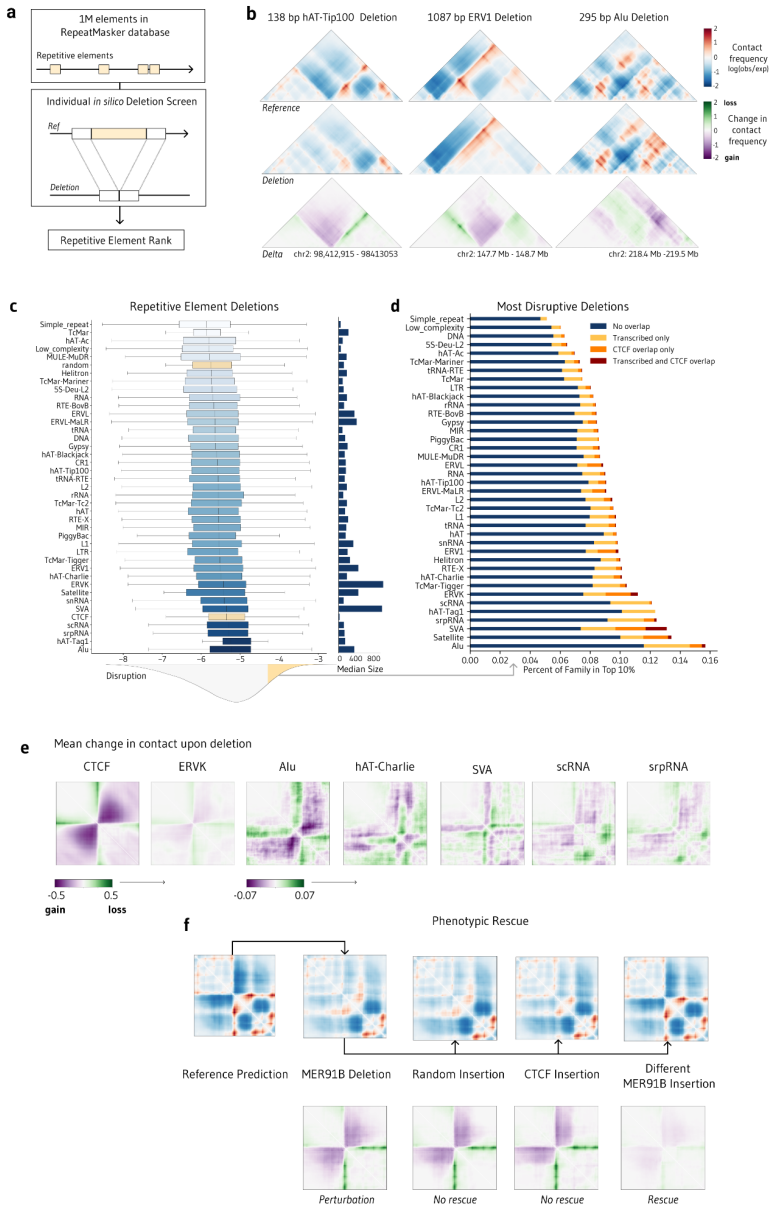


Figure 2.4: Repetitive element deletions impact genome folding.

a. Strategy to individually delete over 1 million elements from the RepeatMasker database. **b.** Representative examples from chromosome 2 showing how the deletion of a hAT-Tip100 element, an ERV1 element and an Alu element *in silico* significantly alter contact maps. Single elements are predicted to disrupt genome folding. **c.** Distribution of disruption scores across each repetitive element family (n = 1,164,108). The distribution of disruptions from 100,000 CTCF deletions (positive control) and 100,000 100-bp random deletions (negative control) are shown in yellow. The median size in base pairs of deleted elements for each family is shown on the right. **d.** The top 10% most disruptive elements across the screen by repetitive element family. Most elements do not overlap a CTCF motif or a region actively transcribed in the HFFc6 cell line. **e.** Average changes in contact maps for the top 100 elements per family. **f.** Phenotypic rescue. We showcase a 138-bp MER91B hAT-Tip100 element whose deletion produces a loss of a boundary. Inserting a random size-matched sequence and a CTCF motif does not change the disturbed contact map, but introducing an MER91B element from the same family restores the original genome folding.

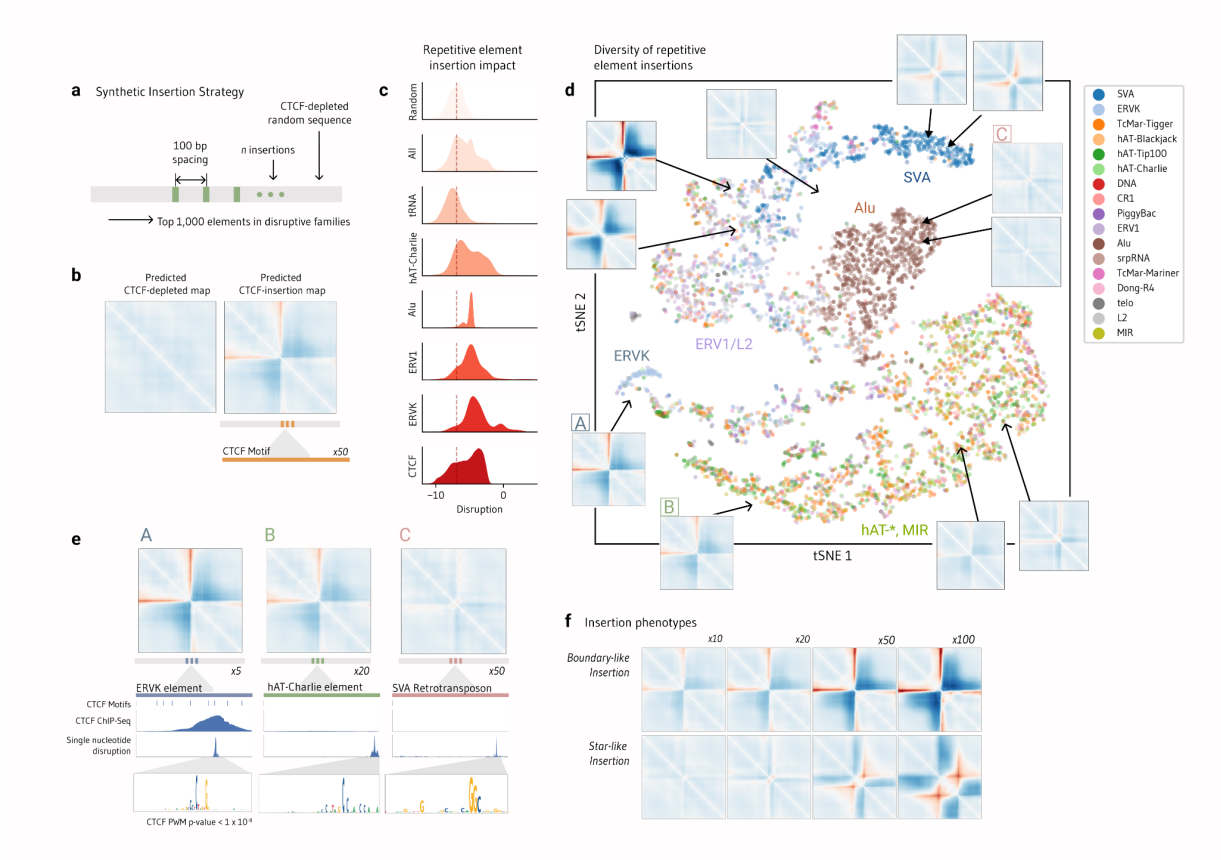


Figure 2.5: In silico insertion screen reveals repetitive elements can induce different boundary types.

a. Insertion screen strategy. For each of the 1,000 most disruptive elements, up to 100 individual copies (green) are inserted 100 bp apart centered in a 1-Mb random DNA sequence depleted of CTCF sites. **b.** The map predicted from the CTCF-depleted random sequence (left panel) provides a blank canvas against which we can measure the impact of insertions. A CTCF site insertion into the middle of the sequence produces boundaries in the predicted maps (right panel). Disruption is measured as the mean squared difference between the blank map and the predicted post-insertion map. **c.** Distribution of disruption scores across repetitive element insertions (n = 14,514). The score distributions of 10,000 100-bp random insertions (negative control) and of 10,000 CTCF motif insertions (positive control) are shown. **d.** t-SNE visualization of all predicted maps from repetitive element insertions with an disruption score above -5.5. Predicted maps are colored by element family. **e.** We highlight three repetitive elements which are highly disruptive both when deleted and inserted. We overlay overlapping annotated CTCF motifs and CTCF sites confirmed by ChIP-Seq in HFFc6 cells. We also show the disruption score of each nucleotide across the element following single base pair *in silico* mutagenesis, highlighting the motif within the repetitive element responsible for the element's high disruption score. **f.** We observe two primary classes of insertions: CTCF-like boundary insertions are common across ERVK and ERV1 elements and star-like insertions are common across SVA and Alu elements.

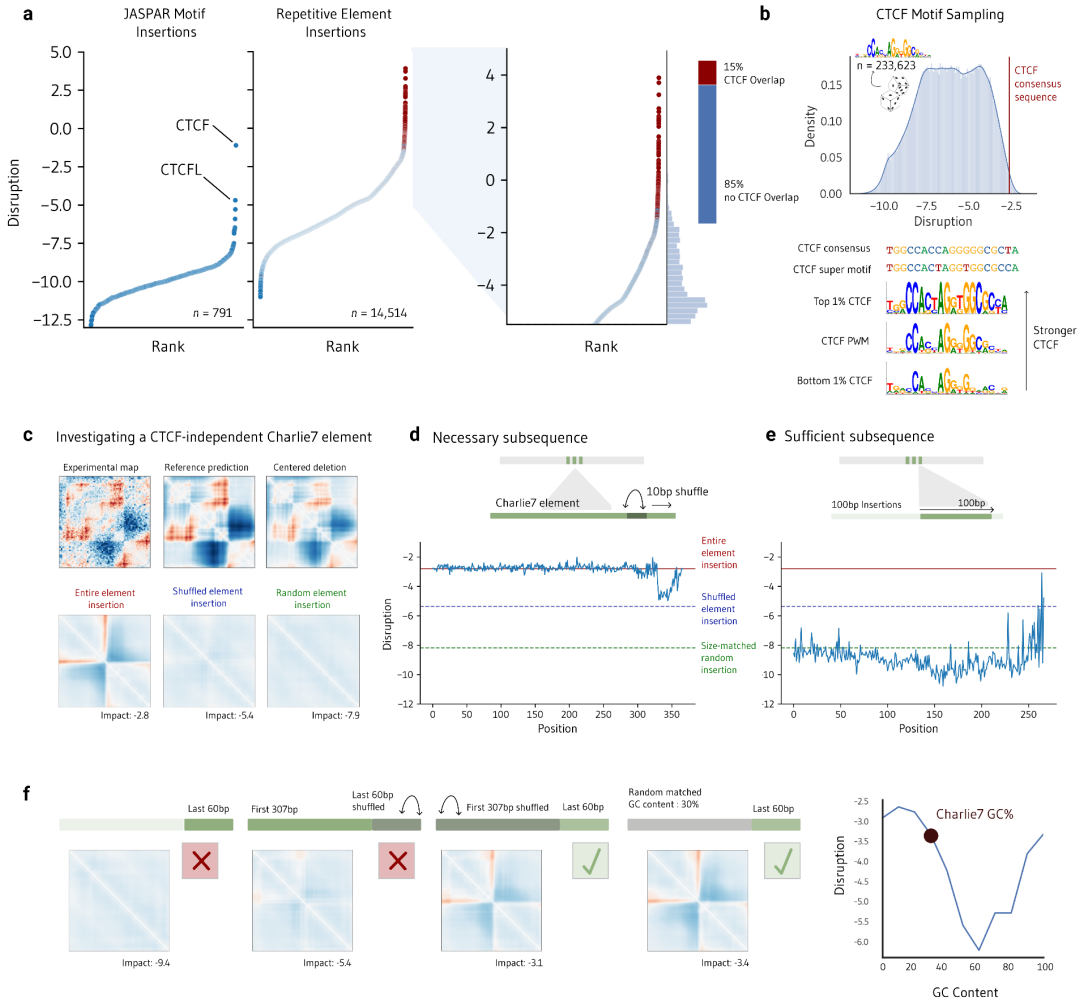


Figure 2.6: *In silico* investigation of sequence features necessary and sufficient for repetitive element Charlie7 to create a boundary.

a. We insert every JASPAR motif into a CTCF-depleted random sequence, as well as 14,514 repetitive elements, and rank them according to their disruption score. 85% of the most impactful insertions (score > -5.5) do not overlap a CTCF motif. **b.** We generate CTCF motif variants with frequencies sampled from the CTCF motif position weight matrix (PWM) and insert them into the random reference sequence (n = 326,177), finding that 0.50% of motifs produce stronger predicted boundaries when inserted than the CTCF consensus sequence. These 'super motifs' share Ts at positions 8 and 12. **c.** We investigate a 367-bp disruptive Charlie7 hAT-Charlie element which does not overlap a CTCF motif or ChIP-Seq peak. Shown in the top row are the experimental micro-C contact map around the locus of the Charlie7 insertion, the map of the locus predicted by Akita, and the predicted map following the deletion of the entire element. Shown in the bottom row are the predicted maps after insertion into the reference, CTCF-depleted sequence of the Charlie7 element (left), a version of the element with a shuffled sequence (middle) and a random sequence of equal length (right). **d.** We shuffle each 10-bp subsequence along the element to determine which one is necessary to produce the boundary seen from introducing the whole element. **e.** We introduce 100-bp segments scanning the entire element into the reference sequence and find that none is sufficient to produce a strong boundary. **f.** A DNA sequence matching the GC content of Charlie7's first 307 bp combined with the last 60 bp is sufficient to recreate a boundary. Right panel: The first 307 bp of Charlie7 was replaced with randomly generated sequence across a range of GC content.

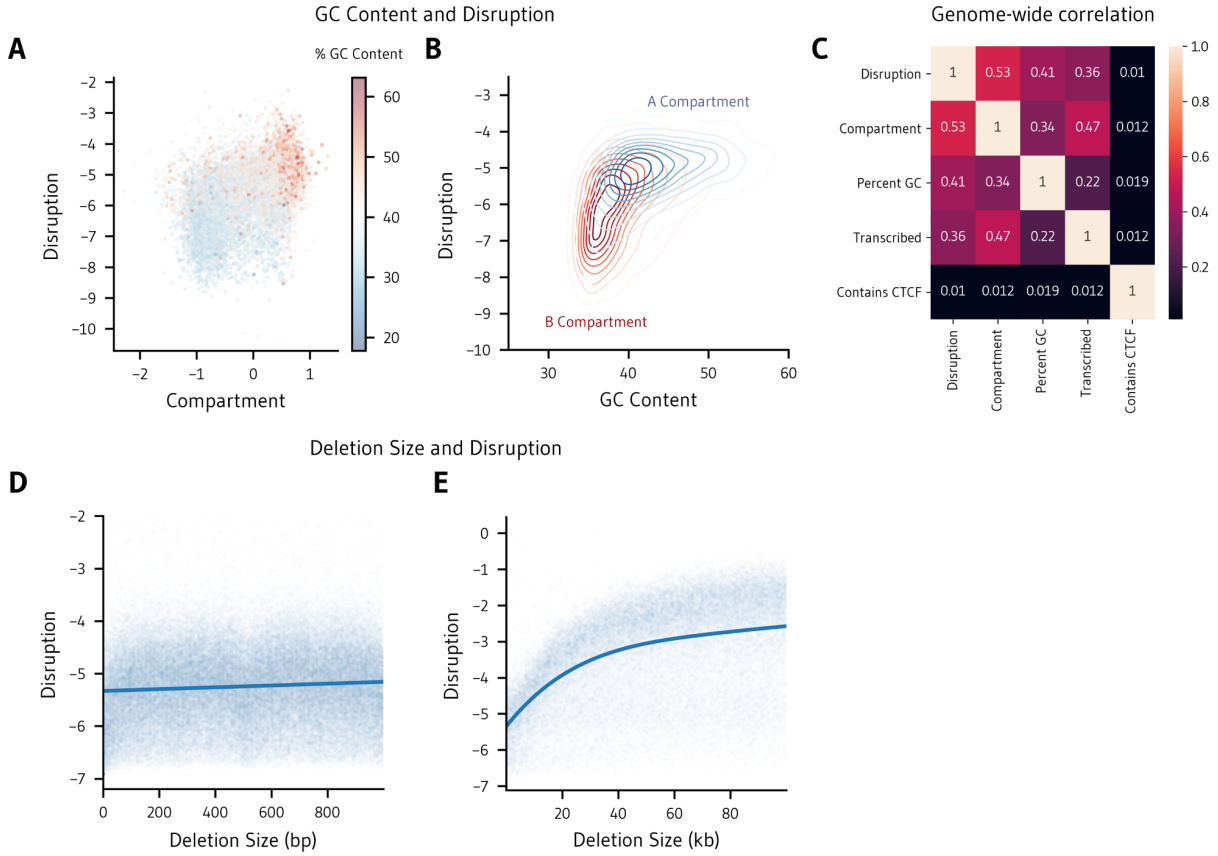
2.7 Supplemental Note

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

To investigate this confounder, we examined how sequence mappability compares to disruption score **(Fig. S 2.6).** In general, we observe no correlation between deletions of 5-kb windows and mappability, indicating that poorly mappable sequences do not have unusually high or low disruption scores. Mappability of individual elements is also uncorrelated with disruption.

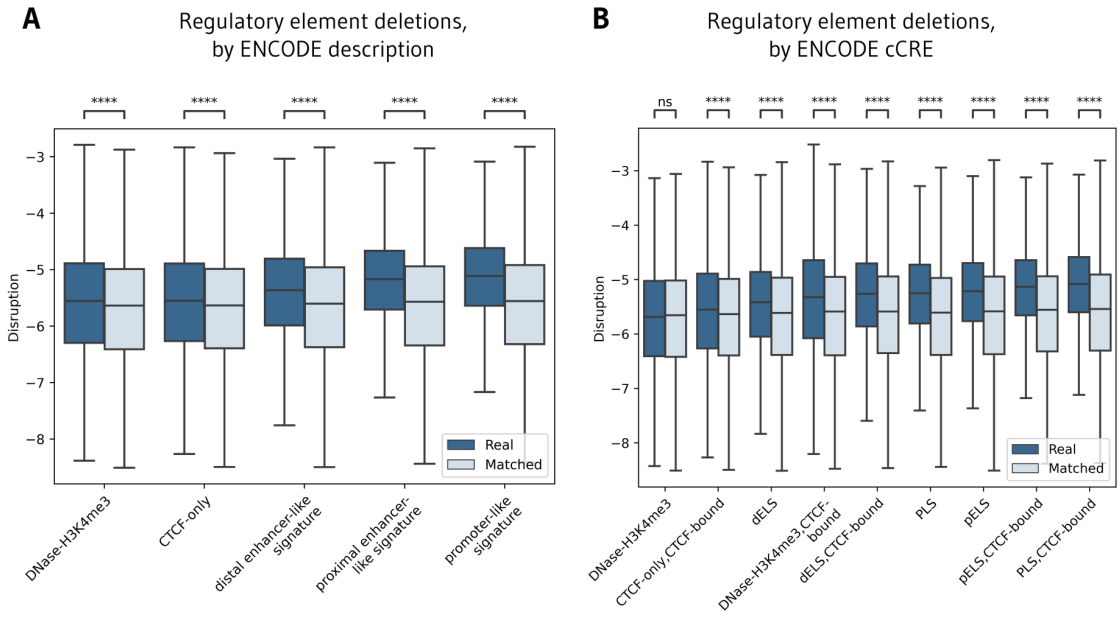
We do find that Alu elements have particularly low sequence mappability and particularly high predicted importance. Many Alu elements are still active and recently inserted into DNA, and therefore have high sequence similarly, presenting a challenge in mapping. It is also possible that the highly conserved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

2.8 Supplemental Figures



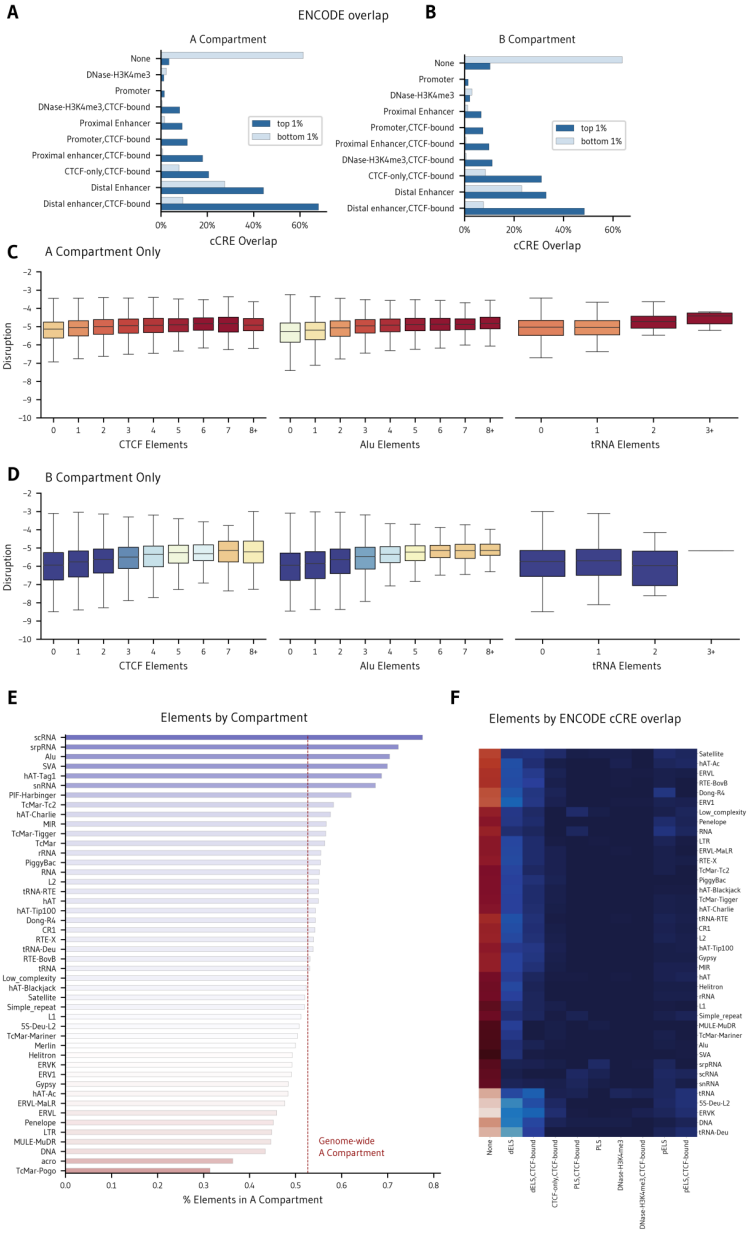
Supplemental Figure 2.1. Disruption is correlated with GC content and deletion size.

a. Disruption scores across the 5 kb whole-genome deletion screen compared to compartment score, as defined as the first eigenvector of the experimental micro-C contact matrix in HFFc6. **b.** GC Content across the 5-kb screen compared to disruption score. **c.** Disruption scores across a deletion screen of random sized genomic segments ranging from 1 bp to 1,000 bp across chromosome 17 (n = 2,000). **d.** Disruption scores across a deletion screen of random sizes ranging from 1 bp to 100,000 bp across chromosome 1 (n = 39,207).



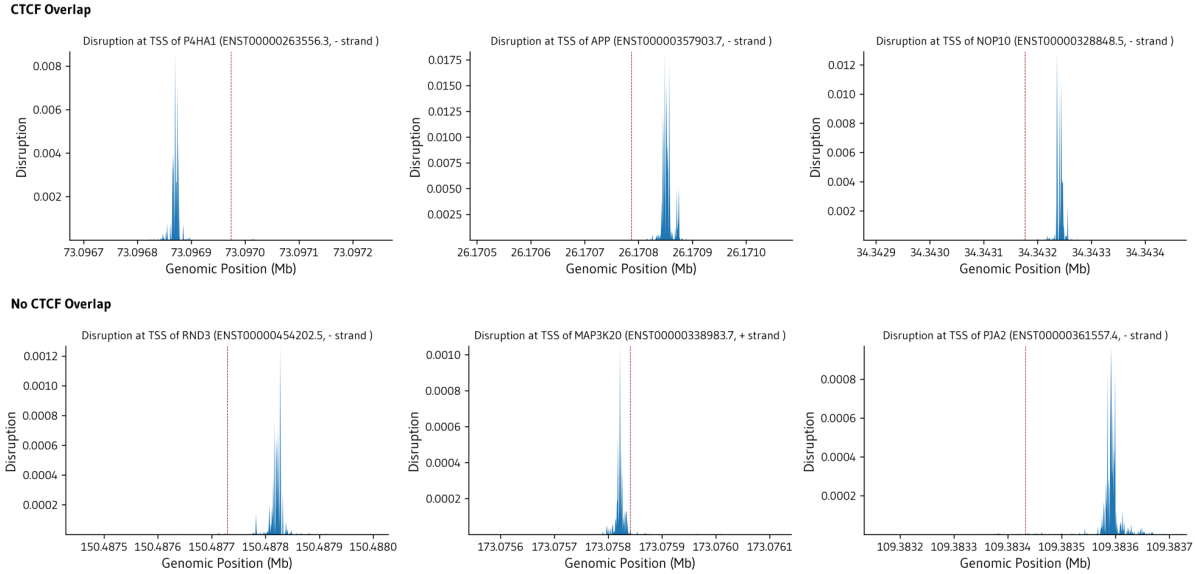
Supplemental Figure 2.2. Enhancer deletions.

a-b. Predicted impact of deletion of 177,049 ENCODE regulatory elements, ranging from 150 to 350 bp, along with size-matched genomic fragments randomly selected from the same chromosome. Regulatory element deletions are significantly more disruptive than deletions of their random counterparts by a two-sided Mann-Whitney-Wilcoxon test.



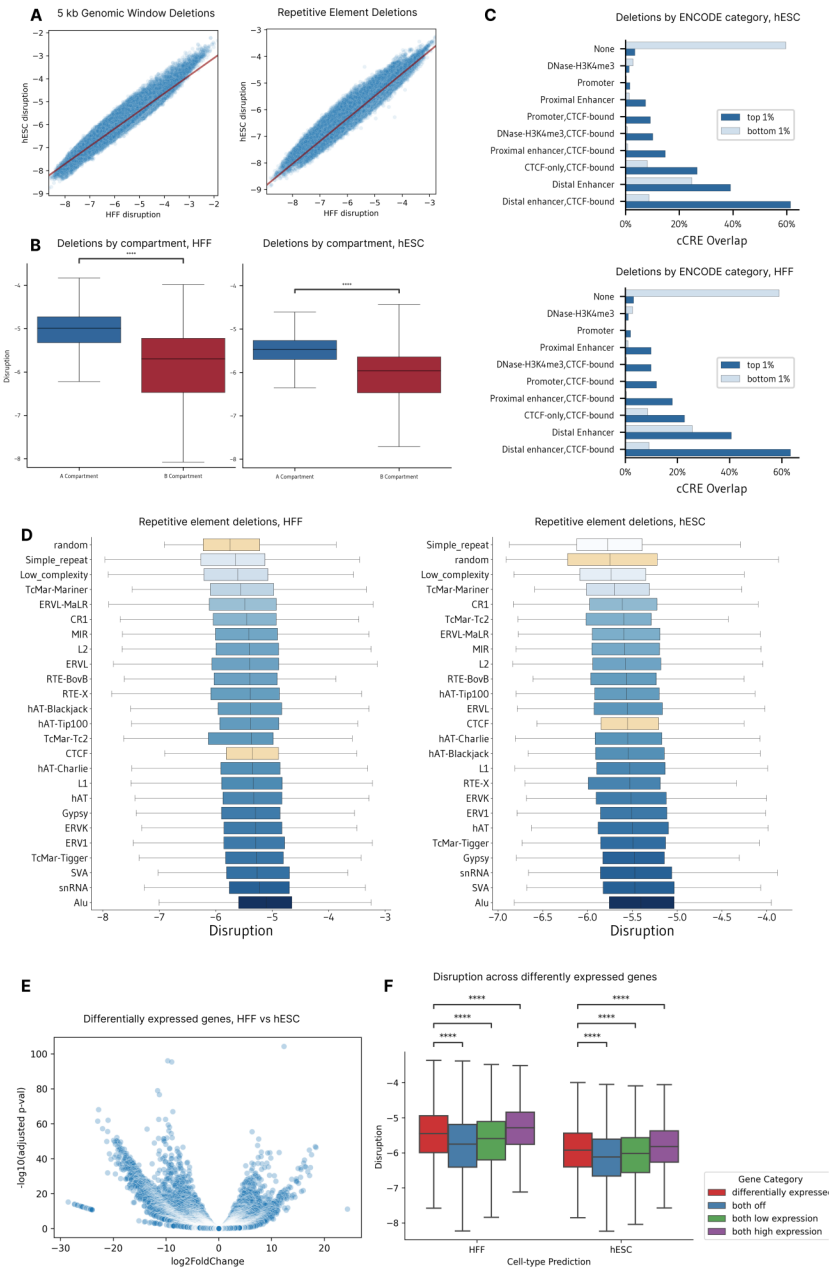
Supplemental Figure 2.3. Disruption across compartment and regulatory region.

CTCF Enrichment in A and B Compartments. CTCF-bound regions are enriched within the top 1% most disruptive 5 kb regions compared to the bottom 1% in both A compartments (**a**) and B compartments (**b**). **c-d.** Disruption by repetitive element count within 5 kb genomic windows, by compartment. **e.** Percent of each sampled repetitive element family found within the A compartment, as defined by the first eigenvector of experimental micro-C in HFFc6. **f.** Overlap of top 10% most disruptive repetitive elements by the deletion screen stratified by overlap with ENCODE candidate regulatory elements (cCREs).



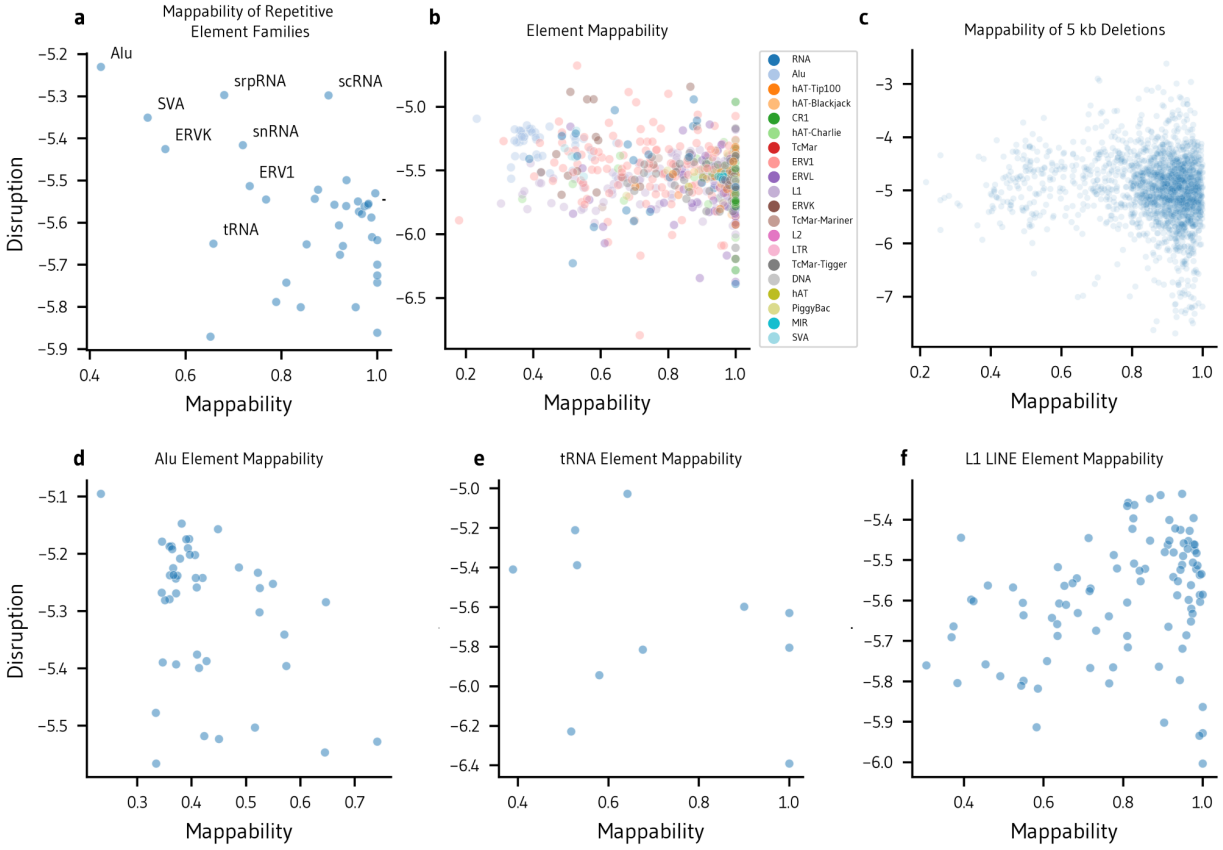
Supplemental Figure 2.4. Transcription tracks.

Individual single-nucleotide disruption tracks around the TSS of highly expressed genes which overlap CTCF (top) and do not overlap CTCF (bottom). The location of the TSS is marked in red.



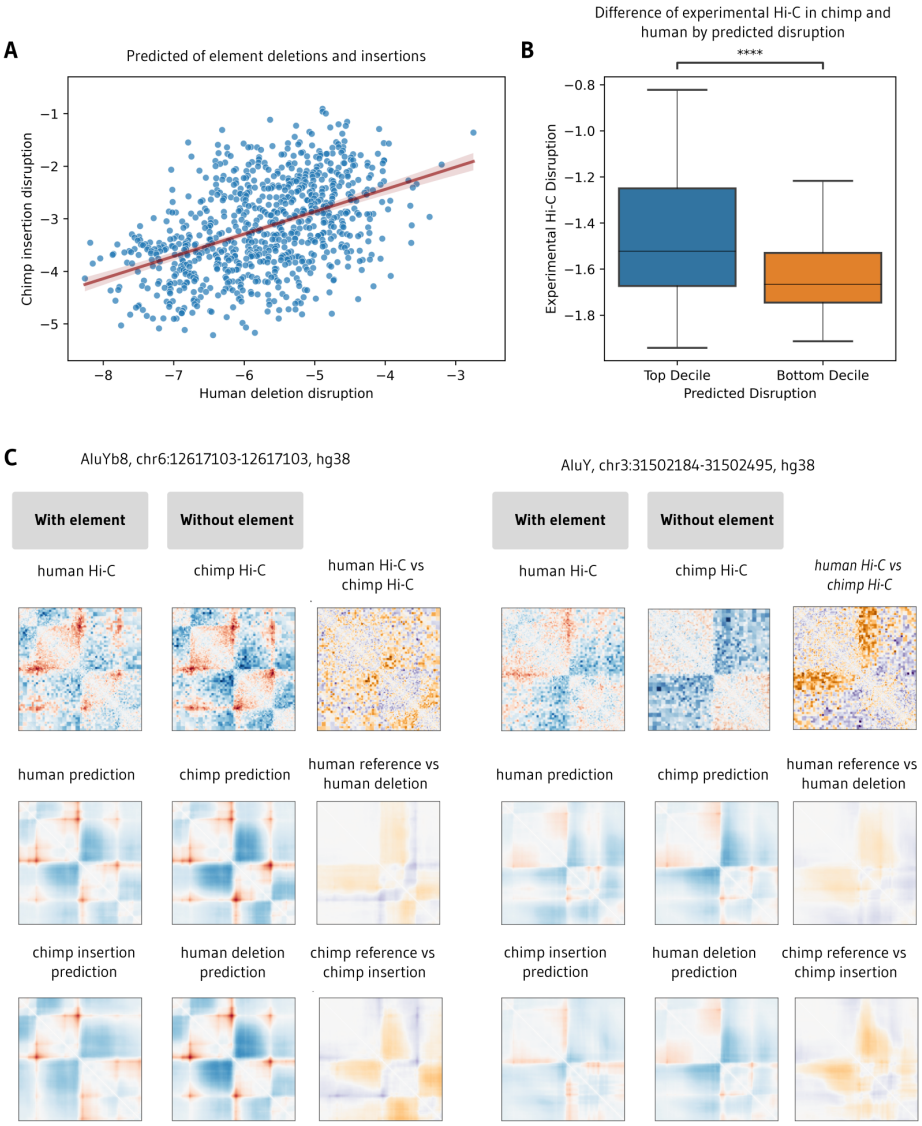
Supplemental Figure 2.5. Disruption score comparison across HFFc6 and hESC.

a. Disruption scores for whole-genome tiled 5kb deletions and a subset of repetitive element deletions are strongly correlated in HFF and hESC (Spearman's correlation of 0.962 for 5kb deletions and 0.967 for repetitive element deletions). **b.** 5kb deletion overlapping A compartments produce higher disruption scores than fragments overlapping B compartments in HFF and hESC. **c.** 5kb deletions overlapping distal-enhancers produce the strongest disruption scores in HFF and hESC. **d.** Deletion of All elements, SVA elements, and snRNA elements produce the largest disruption scores upon deletion in HFF and hESC. Other small RNAs were not included, as they were not sampled at a high enough frequency in hESC. **e.** Volcano plot of differentially expressed genes in HFF and hESC given 8 total RNA-Seq experiments. **f.** Regions ranging from 300 bp upstream to 300 bp downstream the TSS were deleted across all genes in HFF and hESC. Differentially expressed genes produced higher disruption scores than lowly expressed genes as well as genes that were not transcribed in either cell type.



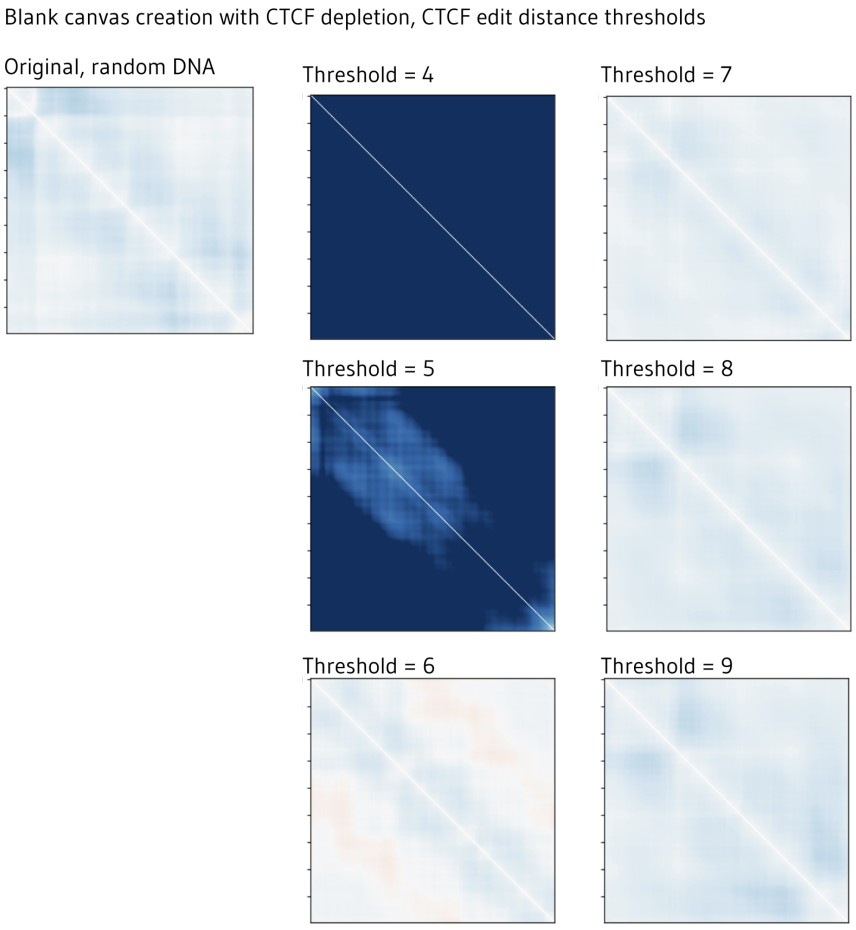
Supplemental Figure 2.6. Disruption and Mappability.

Comparison of multi-read mappability at chr1:100Mb-200Mb and disruption scores. **a.** Average mappability by repetitive element family. **b.** Average mappability by repetitive element type. **c.** Average mappability of 5-kb deleted genome windows. **d-f.** Average mappability of repetitive element types within the Alu, tRNA, and L1 LINE families.



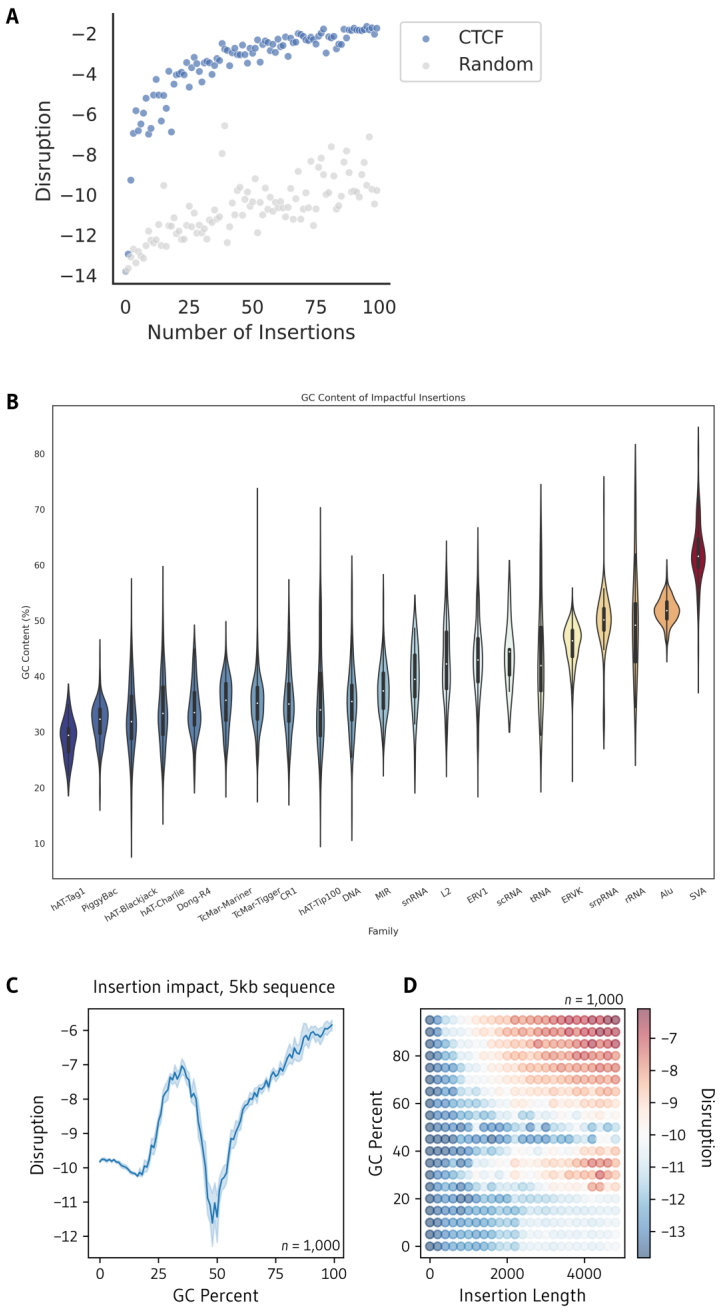
Supplemental Figure 2.7. Investigation of recent human-specific transposable elements.

862 human-specific mobile elements were mapped to the chimp panTro6 genome to investigate the impact of genome folding with and without transposable elements. The vast majority of these elements were Alu elements (813 elements), although several SVA, L1 LINE, and LTR elements were included. We compared the differences in experimental Hi-C at these regions, as well as predicting the folding of the human genome with the element deleted and the chimp genome with the element inserted. **a.** The predicted effect of element insertions into the chimp genome correlates with the predicted effect of element deletions in the human genome, indicating that certain elements are more disruptive when both inserted and deleted and their impact is consistent (Spearman's correlation = 0.456). **b.** Comparing the most disruptive elements upon deletion to the least disruptive elements upon deletion (top decile, disruption > -4.48 and bottom decile, disruption < -7.01, respectively), we find that the predicted most disruptive elements produce significantly higher differences between experimental human and chimp Hi-C than the least disruptive elements (two-sided Mann-Whitney-Wilcoxon test, p-value =6.160e-05). **c.** Examples of Alu elements that produce a weakening of boundaries (**left**) and additional stripes (**right**). Elements produce consistent differences in contact when present vs absent, as seen when comparing human to chimp experimental Hi-C, predicted human reference to a deletion in human, and predicted chimp reference to an insertion in chimp.



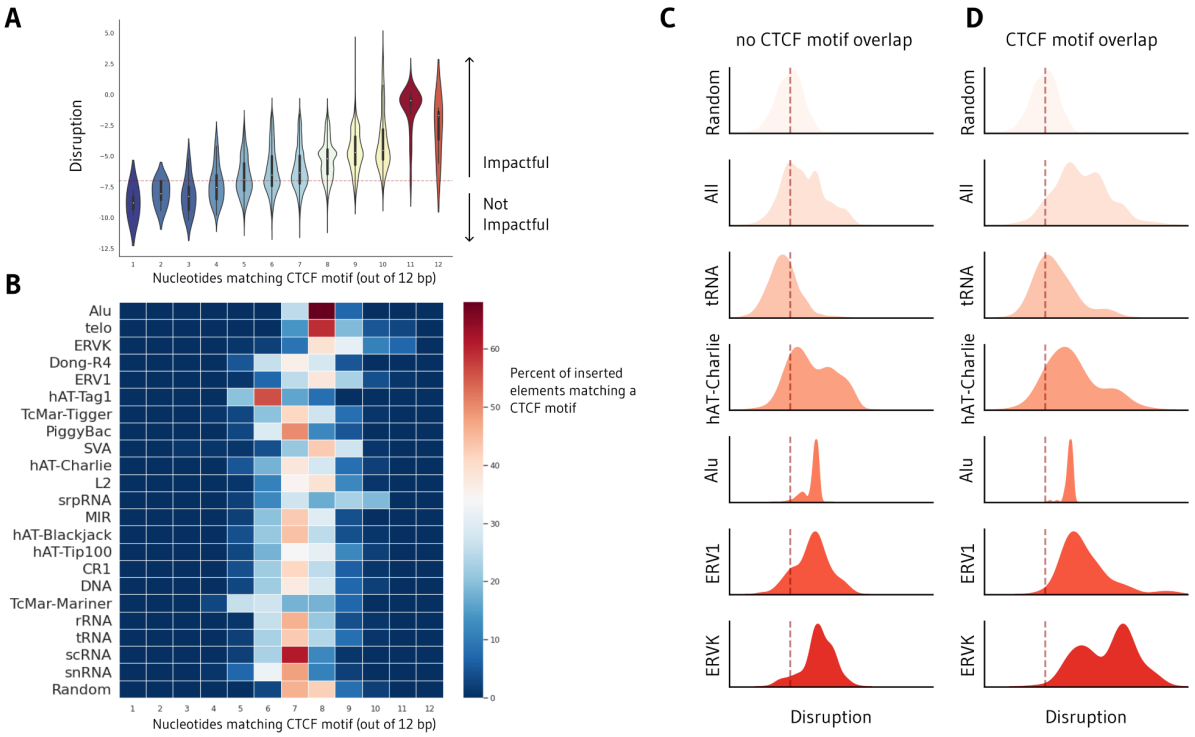
Supplemental Figure 2.8. Edit distance thresholds for blank canvas map creation.

We create a blank map to insert elements by predicting genome folding of random DNA sequence. The original map, by chance, contains spurious structure, so we deplete the sequence of any subsequence within a given edit distance of CTCF. An aggressive threshold (e.g. 4) does not produce a biologically plausible sequence, while a permissive threshold (e.g. 9) leaves structure. We select a threshold of 7.



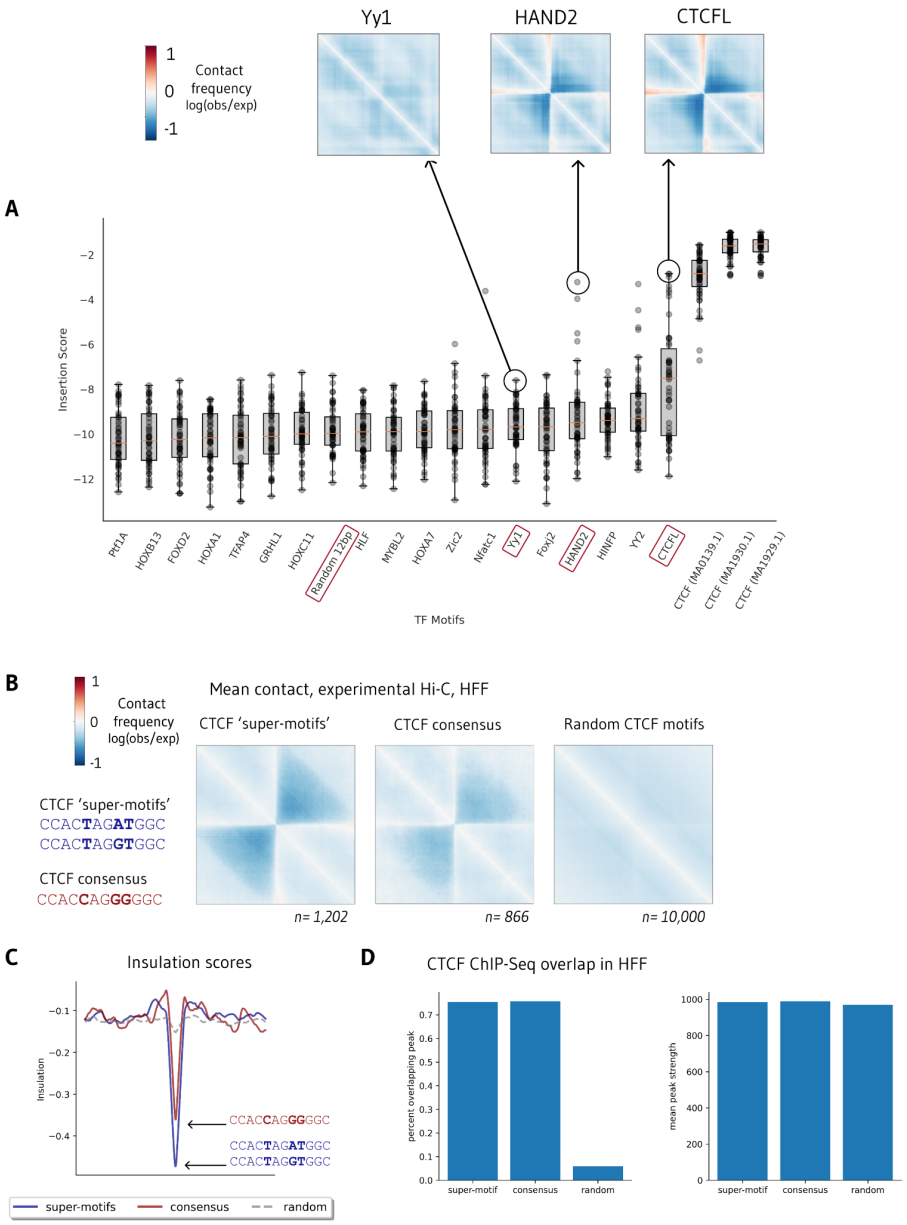
Supplemental Figure 2.9. Motif insertion strength and GC content.

a. Impact of increasing the number of CTCF and random 12-bp sequence insertions into a blank map. Insertions are separated by 100 bp randomly generated DNA sequence. **b.** GC content of disruptive repetitive elements with a MSE greater than -5 upon insertion into a blank map. **c.** Disruption caused by insertion of randomly generated 5-kb DNA sequences with GC percentages ranging from 0% to 100%. **d**. Disruption produced by random insertions into a blank map ranging from GC percentages from 0% to 100% and lengths from 1 bp to 5 kb.



Supplemental Figure 2.10. Insertion strength and CTCF.

a. Number of nucleotides of inserted repetitive elements matching the consensus CTCF motif versus element disruption score. **b.** Elements were scanned one nucleotide at a time to calculate the edit distance of all 12-bp subsequences to CTCF. 12 indicates that the element contains a perfect CTCF motif match. 1 indicates the element contains no subsequences matching the CTCF motif. Motifs more similar to CTCF are higher scoring. Number of nucleotides of inserted repetitive elements matching the CTCF motif, by family (bottom). Only elements with a disruption score above -7 (red threshold) are shown below. **c-d.** Disruption scores across all repetitive element insertions into a blank, CTCF-depleted map, striated by overlap of the original element with an annotated JASPAR CTCF motif.



Supplemental Figure 2.11.Transcription factor motif insertions.

a. The top 20 highest scoring motifs, along with YY1, inserted with a randomized spacing between motifs between 1bp and 100bp across 50 trials. A randomly generated 12bp sequence is included as a control. **b-c.** To experimentally support CTCF super-motifs, we searched the 761,299 annotated CTCF motif locations in the JASPAR 2022 database for the 12-bp core CTCF consensus sequence (CCAC**C**AG**GG**GGC) and the two core sequences found in the top ten highest-scoring synthetic CTCF sequences (CCAC**T**AG**AT**GGC, CCAC**T**AG**GT**GGC), resulting in 866 consensus matches and 1,202 synthetic "super-motif" matches. Comparing centered, averaged Micro-C contact frequencies (log observed/expected) in HFFc6 at super-motif locations compared to consensus motif locations, we observe that super-motifs produce a stronger boundary with greater insulation than consensus sites and randomly selected CTCF sites. **d.** Super-motif sites are as likely to overlap CTCF ChIP-Seq as consensus sites. Sites that do overlap CTCF ChIP-Seq have an equally strong peak strength under super-motif sites and consensus sites.

Supplementary Table 2.1: Data table

Filename	n	Type	Context	Description
5kb_deletions.csv	562,744	deletion	original	Tiled 5kb deletions across the genome.
random_sized_deletions_100kb.csv	39,207	deletion	original	Random deletions between 1 and 100kb from chromosome 1.
random_sized_deletions_1kb.csv	2,000	deletion	original	Random deletions between 1 and 1kb from chromosome 17.
RepeatMasker_deletions.csv	1,164,108	deletion	original	Deletions from the RepeatMasker database.
CTCF_motif_insertions.csv	991	insertion	blank	Sampled CTCF JASPAR motifs inserted into blank map.
CTCF_PWM_sampling_insertions.csv	326,177	insertion	blank	Sampled CTCF PWM inserted into blank map.
random_20bp_insertions.csv	236,119	insertion	blank	Random 20bp sequences inserted into blank map.
CTCF_consensus_motif_insertions.csv	1000	insertion	blank	CTCF consensus sequence inserted into blank map (different random sequence spacing).
random_201bp_insertions.csv	1000	insertion	blank	Random DNA sequence inserted into blank map.
all_motif_insertions.csv	842	insertion	blank	JASPAR motif insertions into blank map.
RepeatMasker_insertions.csv	14514	insertion	blank	RepeatMasker element insertions into blank map.
tss_mutagenesis.csv	1,073,334	mutagenesis	original	Random single nucleotide mutagenesis around the TSS.

3

Chapter 3: Comparing chromatin contact maps at scale: methods and insights

3.1 Abstract

Comparing chromatin contact maps is an essential step in quantifying how three-dimensional (3D) genome organization shapes development, evolution, and disease. However, no gold standard exists for comparing contact maps, and even simple methods often disagree. In this study, we propose novel comparison methods and evaluate them alongside existing approaches using genome-wide Hi-C data and 22,500 *in silico* predicted contact maps. We also quantify the robustness of methods to common sources of biological and technical variation, such as boundary size and noise. We find that simple difference-based methods such as mean squared error are suitable for initial screening, but biologically informed methods are necessary to identify why maps diverge and propose specific functional hypotheses. We provide a reference guide, codebase, and benchmark for rapidly comparing chromatin contact maps at scale to enable biological insights into the 3D organization of the genome.

3.2 Introduction

The same genomic locus can adopt different three-dimensional (3D) conformations in different cells, species, and disease states, which can impact gene regulation, cell identity, and replication timing (**Fig. 3.1A**),(Yang *et al.*, 2017; Kragesteen *et al.*, 2018; Spielmann, Lupiáñez and Mundlos, 2018; Eres *et al.*, 2019; Gorkin *et al.*,

2019; Galan *et al.*, 2020; Hoencamp *et al.*, 2021). Chromosome-conformation capture methods (3C, 4C, 5C, Hi-C, Micro-C)(Dekker *et al.*, 2002; Lieberman-Aiden *et al.*, 2009; Dixon, Gorkin and Ren, 2016; Hsieh *et al.*, 2020; Krietenstein *et al.*, 2020) measure how the genome folds across scales, including chromosomal territories, topologically associating domains (TADs), enhancer-promoter loops, and architectural stripes (Dixon, Gorkin and Ren, 2016; Fudenberg *et al.*, 2016; Vian *et al.*, 2018; Kraft *et al.*, 2019). In recent years, single-cell and deep learning techniques accelerated the study of chromatin conformation across an expanding range of biological contexts (Nagano *et al.*, 2013, 2017; Tan *et al.*, 2018, 2021; Fudenberg, Kelley and Pollard, 2020; Schwessinger *et al.*, 2020; Zhang, Zhou and Ma, 2022a).

There are many ways to compare chromatin conformation maps, but no gold standard exists. Existing approaches rank differences between pairs of maps(Yan *et al.*, 2017; Yang *et al.*, 2017; Stansfield *et al.*, 2018; Yardımcı *et al.*, 2019; Galan *et al.*, 2020; Yang, Chung and Kim, 2022), test reproducibility between replicates and modalities (Yan *et al.*, 2017; Yang *et al.*, 2017; Yardımcı *et al.*, 2019; Boninsegna *et al.*, 2022), identify tissue specific contacts(Yang, Chung and Kim, 2022), and highlight differential chromatin interactions(Stansfield *et al.*, 2018; Galan *et al.*, 2020). Some scores are designed to identify global differences like boundaries and contact intensities (**Fig. 3.1B, left and center**), while others target focal changes like enhancer stripes (**Fig. 3.1B, right**). To rank thousands of loci with diverse folding patterns, one must consider how scoring metrics prioritize different map features and respond to technical artifacts.

Here, we develop a unifying framework to guide strategies for comparing contact maps for new use cases. We introduce three novel methods—eigenvector difference, contact decay probability difference, and triangle track comparison—and benchmark these along with representative methods from the literature to evaluate 11 total approaches (**Fig. 3.1C**). We quantify how methods differentially rank pairs of contact maps across experimental Hi-C data, 22,500 *in silico* sequence insertions and deletions, and simulated contact maps that capture both biological and technical variation. Our analyses identify when methods diverge and when they

are consistent, which methods are redundant or complementary, and where methods commonly fail. The new methods we introduce have relatively high concordance with existing metrics while providing rich information about biological mechanisms. We summarize our recommendations and release a library of open-source code for scoring differences between contact frequency maps to enable scientists to choose and apply the right method for their research question.

3.3 Results

3.3.1 Diverse strategies for scoring pairs of contact maps

When scoring differences between pairs of contact maps, it is common to apply *basic* methods that consider entire 2D contact matrices (e.g., mean squared error (Yang *et al.*, 2017; Fudenberg, Kelley and Pollard, 2020; Galan *et al.*, 2020)) or *feature-informed* methods that sum differences in specific structures (e.g., loops(Rao *et al.*, 2014)). These methods represent two extremes. Basic methods are global summary statistics that can overlook small differences that are most biologically interesting. In contrast, feature-informed algorithms specifically target elements such as TADs, stripes, and loops, but are agnostic to overall contact change and may emphasize artifactual differences. As a compromise between these extremes, we extend statistics previously developed to quantify compartments (eigenvectors/PCA(Nichols and Corces, 2021)), boundaries (directionality index(Dixon *et al.*, 2012), insulation(Crane *et al.*, 2015)), and contact decay(Lieberman-Aiden *et al.*, 2009) in individual maps to instead score differences between pairs of maps. We also propose a new method, called triangle score, which calculates average contact frequencies across all submatrices in a larger contact matrix. These new *map-informed* methods (Supplemental Text) transform 2D contact matrices into 1D tracks

that capture features relevant to genome folding, and then score them using Spearman's correlation or mean squared error (MSE). The intermediate 1D track allows for the interpretation of which regions contribute most.

To comprehensively characterize the behavior of the basic, map-informed, and feature-informed scoring approaches, we implemented 11 representative methods in open-source code (**Fig. 3.1C, Supplementary Table 3.1, Supplemental Text**): MSE, Spearman's rank correlation coefficient (ρ), structural similarity (SSIM), stratum-adjusted correlation coefficient (SCC), eigenvector difference, directionality difference, insulation difference, contact probability decay difference, triangle score, the HiCCUPS loop caller(Rao *et al.*, 2014), and the cooltools TAD caller(Crane *et al.*, 2015; Open2C, Abdennur, Abraham, *et al.*, 2022). We evaluated how these methods perform across diverse settings. We first applied the methods to

Micro-C from human foreskin fibroblasts (HFF) and embryonic stem cells (ESC) to develop biological intuition about the type of map differences each method captures. We then evaluated their performance using a mass screen of *in silico* genetic perturbations. Finally, simulations isolated the effects of specific kinds of technical and biological variation. This three-part benchmark focuses on how methods rank map pairs, rather than the statistical significance of specific differences; stricter or looser significance thresholds can be applied to any score. In sum, we explored and quantified the behavior of scoring methods to learn when they are discordant with each other.

3.3.2 Beware! Map comparison methods produce discordant rankings

Spearman's correlation, Pearson's correlation, and mean squared error are most commonly used to score two maps(Imakaev *et al.*, 2012; Rao *et al.*, 2014; Dixon *et al.*, 2015), as they are computationally efficient and require

no feature selection. We compared their behavior using Micro-C contact maps from HFF and ESC cells across all 7,840 1-Mb windows of the human genome (Methods). These basic methods prioritized markedly different regions (**Fig. 3.2, r^2 = 0.0002, Supplemental Fig. 3.1**)(Krietenstein *et al.*, 2020), often for reasons unrelated to underlying biology. For example, a pair of maps with visible structural rearrangements but a low range of contact frequencies was prioritized by correlation, but not by MSE, as the absolute difference between them is small (**Fig. 3.2A**). Conversely, two maps with similar overall structure but different contact frequency ranges produce a large MSE even though they are very strongly correlated with each other (**Fig. 3.2D**). These inconsistencies occur because Spearman's correlation is agnostic to intensity changes, while MSE is sensitive to intensity. Basic methods were not designed to identify specific chromatin features, and therefore may not always be biologically interpretable on their own. They often disagree.

3.3.3 Map-informed methods highlight changes in genome structure

The *map-informed* methods we created or extended have never been benchmarked. To gain intuition about their behavior, we used our comparison across experimental Micro-C maps in HFF and ESCs to evaluate how these methods behave on contact maps containing three common changes linked to disruption in gene regulation: a boundary change, a stripe change, and a loop change (**Fig. 3.3A i, red boxes**). Triangle score, directionality index, insulation difference, and eigenvector difference all correctly identified large contact changes across the three examples (**Fig. 3.3A ii-v**). Eigenvector difference in particular showed a strong separation between tracks at the emergence of a new boundary and the strengthening of an existing boundary (**Fig. 3.3A iii**). Compared to other approaches, directionality index performed best in identifying focal changes, like the loss of loops (**Fig. 3.3A iv**), while eigenvector difference and insulation difference instead prioritized global changes in contact. Finally, eigenvector difference and contact decay were sensitive to overall contrast difference.

We observed a divergence in the contact decay tracks across the first pair where a map gains distal contact **(Fig. 3.3 vi)**. In sum, the design of these methods highlight different features in the tracks, from overall structural differences and average contact, to sharp changes in contrast.

3.3.4 Feature-informed methods prioritize changes to interacting chromatin regions

To evaluate comparison approaches based on TAD and loop calling methods, we chose two regions with differential structure between ESC and HFF maps (**Fig. 3.3B i**) and tuned the parameters of the cooltools TAD caller (Crane *et al.*, 2015; Open2C, Abdennur, Abraham, *et al.*, 2022) and the HiCCUPS loop caller (Rao *et al.*, 2014) (**Supplemental Fig. 3.2**) (Forcato *et al.*, 2017; Zufferey *et al.*, 2018). As expected, the TAD caller correctly identified all three TAD boundaries visible in ESCs, including one that is lacking in HFF (**Fig. 3.3B ii**). Similarly, the loop caller identified a loop that is unique to ESCs (**Fig. 3.3B iii**). While these feature-informed approaches are biologically interpretable, they tend to be slower, address only one element at a time, and require additional parameter selection (**Table 3.1, Supplemental Table 3.1**). These methods also require a significance cutoff for initial feature calls, which may result in missed features of low signal. Additionally, most maps contain fewer than ten called features in a 1-Mb window, creating a small range of possible scores. Therefore, caution should be exercised when using these scores at a large scale, especially in maps without strong TADs or loops, where they can produce artificial results.

3.3.5 In silico perturbation enables evaluation of contact map comparison methods at scale

Although differences between cell-types exist, 3D genome organization is often highly conserved(Dixon *et al.*, 2012; Yang *et al.*, 2017; McArthur and Capra, 2021). To evaluate the performance of map comparison methods across a wider variety of possible changes in chromatin structure, we used an *in silico* approach to generate pairs

of 1-Mb maps across the genome with a variety of perturbations. We applied Akita (Fudenberg, Kelley and Pollard, 2020), a convolutional neural network that predicts genome folding from sequence alone, to generate contact frequency maps from sequences with and without a genetic perturbation likely to disrupt genome folding **(Fig. 3.4A**). We designed three types of perturbations: CTCF canonical motif insertions (Castro-Mondragon *et al.*, 2022), endogenous CTCF motif deletions, and random 100 base pair deletions (Methods). In total, we produced 22,500 unique contact frequency map pairs on which to test all three types of methods. To enable large-scale evaluation, we applied the 11 methods and transformed their scores such that higher values indicate greater disruption of 3D organization and smaller values indicate more similar organization (**Methods, Supplemental Fig. 3.3**).

We quantified the similarities and differences between methods by comparing the scores for all 22,500 *in silico* perturbations across all possible pairs of methods. We found that TAD- and loop-based scores are most different from the rest, as they only detect a specific type of change (**Fig. 3.4B**). Correlation-based measures (i.e., Spearman's correlation, SSIM, and correlation of contact decay) cluster together distinct from MSE-based methods (i.e., MSE, triangle (MSE), insulation (MSE)). This result aligns with our initial observation that Spearman's correlation and MSE often do not agree, especially across their top-scoring variants (**Fig. 3.2, Supplemental Fig. 3.4, Supplemental Fig. 3.5**). Principal component analysis (PCA) on the disruption scores shows similar clustering (**Fig. 3.4C**).

We next simultaneously clustered the perturbed map pairs and scores across methods to identify groups of perturbations that differentiate them (**Fig. 3.4D**). While all correlation-based methods exhibit similar behavior, insulation (corr), SSIM, and DI (corr) produce scores which are more uniformly distributed and less extreme across perturbations, highlighting the necessity of appropriate normalization when comparing across methods (**Fig. 3.4E i and vi**). We also find that perturbations created by CTCF insertion group together, as they are often the most disruptive of 3D organization. However, we observed substantial sub-structure within

the cluster, reflecting differences in the behavior of scores on these maps. For example, cluster *i* is highly scored by all methods, and a representative perturbation example shows a variety of changes: gained loops, lost stripes, and boundary changes. The magnitude of changes in this set likely contributes to the universally high scores. Clusters *iv* and *v* are primarily composed of CTCF insertions, where scores are similar across most methods, but higher only for MSE-based methods. Profile *v* is the most dissimilar. Here, the representative map pair has minimal structural differences but extreme contrast, suggesting that this cluster is defined by examples of high dynamic range that are over-prioritized by MSE-based methods (**Fig. 3.4E v**).

We further compared methods by quantifying how well the top-ranked maps agree across methods. Some methods have high overlap (**Supplemental Fig. 3.5, Supplemental Fig. 3.6**). For example, 85% of map pairs are ranked in the top 5th percentile for both SCC and Spearman's correlation, indicating some general agreement in the methods. However, many methods have minimal overlap, suggesting they prioritize different features. For example, only 32% of the top 5th percentile of maps ranked by insulation (MSE) and SSIM are shared. Finally, we applied methods to map pairs selected to represent a range of effect sizes and confirmed all methods are sensitive to large changes and insensitive to small changes (**Supplemental Fig. 3.7**).

3.3.6 Simulation studies quantify method sensitivity

Our *in silico* screen produced a diversity of structural alterations, often affecting multiple aspects of the map. For instance, a CTCF site insertion can both create a new TAD boundary and alter overall contact intensity. To disentangle how each method responds to changes in particular map features, we generated simulated maps and synthetically altered a single variable at a time. We then measured the sensitivity of each score to each specific change. As a template, we created a contact frequency map with two CTCF motifs forming a TAD and used this canvas to simulate both biologically meaningful changes (e.g., change in TAD size, substructure, or

intensity) and technical artifacts (e.g., change in noise or resolution) (**Methods**). For each change, we gradually increased the strength of the perturbation across 100 maps and subsequently applied scoring methods (**Supplemental Fig. 3.8**).

Each method responded differently across the simulated changes (**Fig. 3.5**). Steeper curves represent high sensitivity to the perturbation, while flatter curves represent less sensitivity. We find that basic methods are most sensitive to technical variations, such as increased noise and decreased resolution, while map-informed methods are most robust (**Fig. 3.5A-B**). As expected, correlation-based methods are unaffected by changes in contrast and intensity, while MSE-based methods are highly sensitive (**Fig. 3.5C-D**). All methods except eigenvector difference identify TAD size and sub-structure changes. However, some prioritize certain types of organizational changes over others (**Fig. 3.5E-F**). For example, insulation difference and triangle profile are sensitive to boundary changes, while directionality index highlights new boundaries but is less effective in identifying changes to existing boundaries. We synthesized these results along with findings from *in silico* perturbations in the Guidelines to provide recommendations based on the intended application.

3.3.7 Guidelines

Our study assessed the effectiveness of 11 existing and new methods for comparing 3D genome contact maps (**Supplementary Table 3.1**). Although there were similarities between the top-scoring variants of most methods, our results indicate that they differ substantially in their sensitivity to biological and technical variation (**Supplemental Fig. 3.6**). We summarize these findings and guidelines in Table 3.1.

All of the methods can identify structural changes, such as changes to domain size or the addition of substructure, but to varying degrees. Of the basic methods, MSE and SCC more readily identify subtle organizational changes. Among the map-informed methods, insulation difference and triangle difference are the

most effective at identifying changes in both existing and new domain boundaries. Directionality index highlights new boundaries or substructures but less readily identifies changes in existing boundaries. Eigenvector difference and contact probability decay are the least sensitive to small-scale organizational changes, but prioritize larger-scale changes in the overall structure of the map. These statistics have been deployed primarily for identifying differences at the scale of compartments and whole chromosomes, so it is not surprising that they are not sensitive to map differences within 1-Mb windows.

In general, the new map-informed methods we proposed are concordant with basic methods and each other, especially when comparing the top 5% of scores genome-wide (30%-80% of examples are shared; **Supplemental Fig. 3.6**). Triangle difference stands out among our newly implemented methods as highly concordant with other methods and able to detect a variety of map differences, but it is also the slowest (**Supplementary Table 3.1**). Insulation difference is faster and also fairly concordant with other methods. Most top 5% map pairs called by other methods are also high-scoring with loop calling, but not TAD calling. Loop calling also identifies many additional map pairs that are not in the top 5% of other methods.

Correlation-based methods are insensitive to changes in contrast and intensity, while MSE-based methods are highly sensitive to these changes. In contrast, map-informed methods summarize maps across a feature track and are therefore more robust to these changes. One notable exception is insulation difference, which is more sensitive to resolution changes that obscure domain boundaries. Some map changes, such as contrast or intensity, may either be biologically meaningful or a consequence of technical variability, depending on the scenario. The basic methods, especially MSE, SCC, and SSIM, are particularly sensitive to technical variation such as increased noise and decreased resolution. SSIM falls in between. We also note that MSE is by far the fastest approach (**Supplementary Table 3.1**). All others require less than 10 seconds per thousand calls, aside from eigenvector difference and triangle track, which can be accelerated by decreasing the resolution of the maps prior to comparison.

We recommend using multiple methods in tandem. We find that there is no "one size fits all" metric that best identifies every feature of interest in a chromatin contact map. Researchers should consider the intended application and the types of changes that are meaningful when selecting the most effective and relevant metrics. We recommend first applying basic methods as an initial screen to identify the most disrupted maps, especially when evaluating large datasets. Using both correlation- and MSE-based scores will help mitigate biases of each. We next suggest applying a map-informed method, such as triangle or insulation difference, to a subset of disruptive perturbations to gain insight into the types of changes present. Finally, feature-informed methods can be used to explore TAD and loop gains/losses and to develop mechanistic hypotheses.

3.3.8 Code

Our codebase is publicly available to enable researchers to easily test and apply all 11 approaches to their own research questions. The code is written in Python and is accompanied by documentation and tutorials to help users get started. The methods have flexible hyperparameters and can be run simultaneously on one dataset, making it easier to compare the results of different approaches and select the most appropriate methods. To aid in interpretation of the methods, we also provide guidance on how to visualize map-informed and feature-informed approaches across contact matrices. Overall, our codebase provides a valuable resource for researchers who wish to apply multiple methods to their own datasets and rank pairs of maps based on their differences.

3.4 Discussion

In this study, we evaluated and compared the behavior of 11 methods for quantifying differences between pairs of 3D contact maps, including many methods that have not been previously used for this application. We

introduced insulation difference, eigenvector difference, and contact decay difference, as well as the new triangle comparison method, which is robust to noise while capturing structural differences between maps. We found that the choice of scoring function can have a significant impact on the conclusions drawn from the data, and therefore suggest that multiple comparison metrics should be used when seeking biological insights into the function of the 3D genome.

Several limitations should be considered when evaluating our results. While we consider a range of experimental, predicted, and simulated maps, our findings may not apply to other experimental conditions, such as single-cell contact matrices or other scenarios in which maps have a high level of noise and/or sparsity. Additionally, some of the methods we evaluated have variables that can be tuned to optimize performance in a given context (**Supplementary Text**, **Supplementary Table 3.1**). We only tested one TAD caller and one loop caller to examine their general utility (Forcato *et al.*, 2017; Zufferey *et al.*, 2018). Finally, we did not directly address the problem of identifying a threshold beyond which the differences should be considered biologically or statistically significant. One could apply previously proposed (Xu *et al.*, 2016; Stansfield *et al.*, 2018; Galan *et al.*, 2020) and novel thresholding methods to the ranks computed with scoring methods to define a significant set of map pairs.

Our work provides useful guidelines for scoring contact maps that will enable further discovery into the mechanisms of the 3D genome. We provide a codebase of methods for flexible and fast scoring across contact maps under a unified framework. The experiments we performed as a part of this study, such as the *in silico* deletion and insertion of thousands of CTCF motifs genome-wide, provide a useful dataset for evaluating diverse biological questions or utility as controls for the level of 3D genome variation expected based on CTCF and random perturbations. We anticipate that incorporating methods with stronger biological interpretability, like those evaluated here, may further improve machine learning methods for predicting contact maps. Overall,

by developing novel and more robust scoring functions, our study provides a foundation for analyzing contact maps at scale.

3.5 Methods

3.5.1 Datasets

Experimental maps

Maps of 3D chromatin contact are represented as 2D matrices of pairwise interaction frequencies. Regions of maps with high values indicate genomic loci with a high frequency of interaction in physical space, on average. Following experimental Hi-C, maps begin as raw read counts, which are subsequently balanced and normalized to reflect log(observed/expected) contact frequencies (Lyu, Liu and Wu, 2020).

Experimental data considered in this study from HFF and ESCs were preprocessed as training datasets for the Akita model(Fudenberg, Kelley and Pollard, 2020; Krietenstein *et al.*, 2020). Specifically, these high-quality Micro-C datasets were normalized with genome-wide iterative correction (ICE), adaptively coarse-grained, normalized for distance-dependent decrease in contact frequency, log clipped to (-2,2), linearly interpolated to fill missing bins, and convolved with a 2D Gaussian filter for smoothing. Processing maps ensures consistency across the experimental data and computational predictions since we do not evaluate raw experimental read counts.

Predicted maps

To effectively compare contact maps at scale, we generated a dataset of thousands of maps predicted from *in silico* CTCF motive insertions, CTCF motif deletions, random 100 bp sequence insertions, and random 100 bp

sequence deletions. These alterations were passed into Akita(Fudenberg, Kelley and Pollard, 2020), a model predicting genome folding from sequence, to generate pairs of maps with structural rearrangements. We first curated sequences for insertion. CTCF motif sequences were randomly selected from annotated CTCF sites in the reference genome from the hg38 build of the JASPAR database(Castro-Mondragon *et al.*, 2022). Random 100 bp fragments were also selected from chromosome 1 for insertion. Both the CTCF and random sequences were inserted into the center of 1-Mb of DNA with start locations randomly selected from chromosome 1. Akita requires a fixed input of 2^{20} bp. Additional sequence was trimmed from the 3' end, such that the final sequence remained 1-Mb. To curate deletions, we again selected random CTCF sites from JASPAR, pulled the surrounding 1-Mb of DNA, removed the motif sequence, and pulled in additional sequence from the 3' end such that the entire sequence remained 1-Mb in length. The same strategy was applied to randomly selected 100 bp fragments for deletion. All generated 1-Mb genomic query sequences were filtered to exclude overlap with ENCODE blacklisted regions(Amemiya, Kundaje and Boyle, 2019). For each perturbation, both the original genomic sequence and the perturbed sequence were provided to Akita, resulting in two predicted 448x448 contact maps where the resolution of each pixel is 2048 bp representing a total length of ~1 Mb (2^{20}) of DNA sequence (Fudenberg, Kelley and Pollard, 2020). This dataset consists of 7,500 matched contact maps for each category of perturbation for 30,000 total map pairs. Random 100 bp insertions were generally excluded from analysis, as they had almost no effect.

Simulated maps

To generate simulated maps, we initially generated predicted maps with Akita from random DNA sequence. Predicted maps still showed minimal structure from randomly occurring CTCF-like motifs. Sequence matches to the forward and reverse canonical CTCF motif(Castro-Mondragon *et al.*, 2022) were therefore shuffled to

produce a predicted blank canvas map devoid of all higher-order folding patterns. Structure was reintroduced to simulated maps by inserting forward and reverse CTCF motifs ¼ and ¾ through the random DNA sequence, producing TAD-like boundaries. We tuned simulated parameters as described below.

- *Noise:* Gaussian noise was added to the maps with a standard deviation ranging from 0 (no added noise) to 0.2.
- *Resolution:* The original 448x448 map was downsampled ranging from a resolution of 2,048 bp (original resolution) to 50,972 bp.
- *Contrast:* Pixel intensities of the contact map were multiplied by a scalar ranging from 1 (no increase in contrast) to 2.
- *Intensity:* A scalar value ranging from 0 (no addition) to 0.2 was added to all pixels in the contact map.
- *Size:* The size of the substructure within the map was increased by resizing the original map by a scalar and trimming the matrix back down to the original dimensions. Map sizes were increased by a factor of 1 (no resize) to 1.1.
- *Substructure:* An additional map was created by introducing CTCF halfway into the random sequence to produce an additional boundary. The original map was combined with the substructure map with a multiplier ranging from 0 (no added structure) to 1 (total added structure).

Visualizations of these changes can be found in **Supplemental Fig. 3.8.**

3.5.1 Benchmarking Methods

Adapting new methods

Triangle profile is a novel scoring method. Directionality index(Dixon *et al.*, 2012), PCA(Nichols and Corces, 2021), insulation(Crane *et al.*, 2015), and contact decay(Lieberman-Aiden *et al.*, 2009) are established methods for analysis on individual Hi-C maps, but have not previously been used to to score pairs of maps. For map-motivated and feature-motivated methods, it is possible to plot the scoring method results along the length of the map, or on the map itself, as seen in **Figure 3.3**. The common behavior across maps with a small change, a large change, and no change is illustrated in **Supplemental Fig. 3.7**.

Comparing contact maps

We applied all comparison metrics to pairs of experimental, predicted, and synthetic maps. For details regarding how each metric is computed, see **Supplemental Text**. Any missing values were masked prior to evaluation and not considered by the comparison metrics. Scoring method implementations can be found within scoring.py in the codebase. MSE, Spearman's rank correlation coefficient, and Pearson correlation coefficient were applied to map-informed methods to collapse two 2D tracks into a scalar value. Pearson correlation behaved almost identically to Spearman's rank correlation, and therefore was excluded from analysis (**Supplemental Fig. 3.1)**. For computationally intensive methods, we reduced the resolution of the input from 2kb to 10kb to speed evaluation time across thousands of comparisons.

To ensure that scores across approaches are comparable, we flip some methods such that higher values indicate greater disruption and smaller values indicate more similar maps. For methods like correlations, we use 1 - correlation such that a perfect correlation (1) is flipped to mean no difference (0). For all the results, we provide raw scores and normalized scores so that it is easier to interpret how a raw score for one method

compares to a raw score of another method. We additionally scale all values by the mean score of all random 100 bp deletions using Akita, which we find to have minimal impact (**Supplemental Fig. 3.3**). For example, a raw MSE of 0.0065 and a raw 1 - pearson correlation of 0.036 both correspond to the same normalized score of 2. That is, a disruption of that magnitude corresponds to 2 times the average disruption of a 100 bp deletion.

For loop and TAD callers, we quantify the ratio of changed (e.g. added or lost) features (TADs or loops) to extend these approaches and generate a single score for each pair of maps.

Method parameters

The following methods required no adjustable input parameters: mean squared error, Spearman's rank correlation coefficient, and pearson correlation coefficient, SSIM, SCC, contact decay, eigenvector, and triangle correlation. We describe tunable parameters choices for the remaining methods below. We did not optimize tunable parameter choices but instead selected default choices from existing approaches. Results from alternative parameter selection are demonstrated in **Supplemental Fig. 3.2** and **Supplemental Fig. 3.9**.

Insulation:

window_size=10: size of the diamond-shaped window considered

Directionality index:

window_resolution=10000: resolution of sliding window in bp

replace_ends: replaces ends of DI track with 0s

buffer=50: how far from the track ends to replace with 0

Loop difference:

p=2: the width of the interaction region surrounding the peak

width=5: the size to get the donut filter

ther=1.1: the threshold for the ratio of center windows to the donut filter and lower left filter

ther_H=1.1: the threshold for the ratio of center windows to the horizontal filter

ther_V=1.1: the threshold for the ratio of center windows to the vertical filter

radius=5: the upper bound of distance of two loop points considered as same

TAD difference:

window_size=5: size of the diamond-shaped window

ther=0.2: the threshold for TAD boundaries

radius=5: the upper bound of distance of two TADs considered as same

3.6 Figures

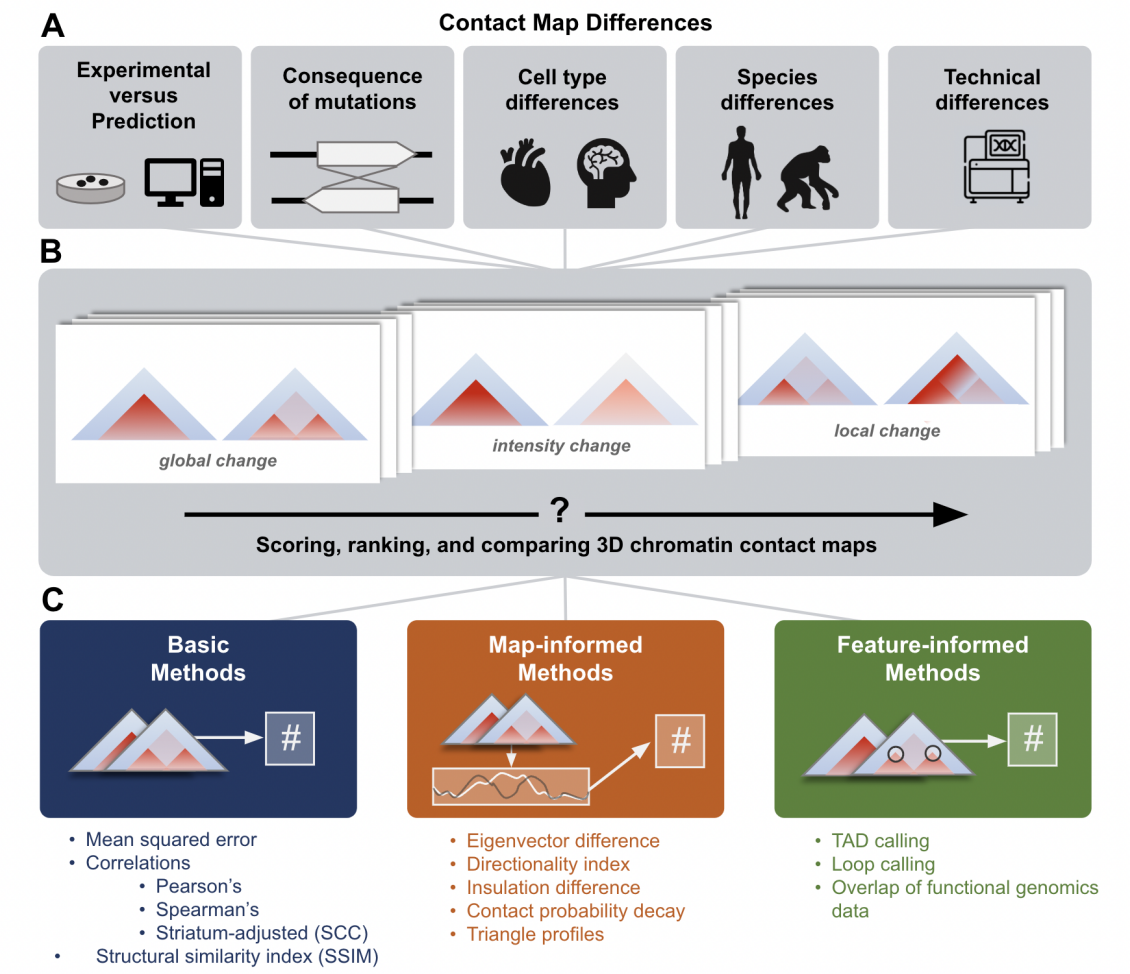


Figure 3.1. Approaches for comparing 3D chromatin contact maps.

(A) 3D genome comparisons drive insights into many domains of chromatin biology. Differences observed between maps may reflect consequences of mutations, cell type differences, species differences, or technical biases. **(B)** 3D contact maps exhibit a range of functionally meaningful differences, e.g., in global folding patterns, contact intensity, or small, focal changes to part of the map. **(C)** We define three categories of comparison methods and evaluate 11 representative methods. Basic methods (left) compare the contact intensities at each contact bin across two maps with simple measures such as mean squared error or correlations. Map-informed methods (middle) transform the 2D contact maps into 1D tracks that describe qualities like the directionality index or insulation score. These tracks were compared to obtain a score. Feature-informed methods (right) are designed to identify relevant elements (e.g., from functional genomics data) or structures (e.g., TADs or loops).

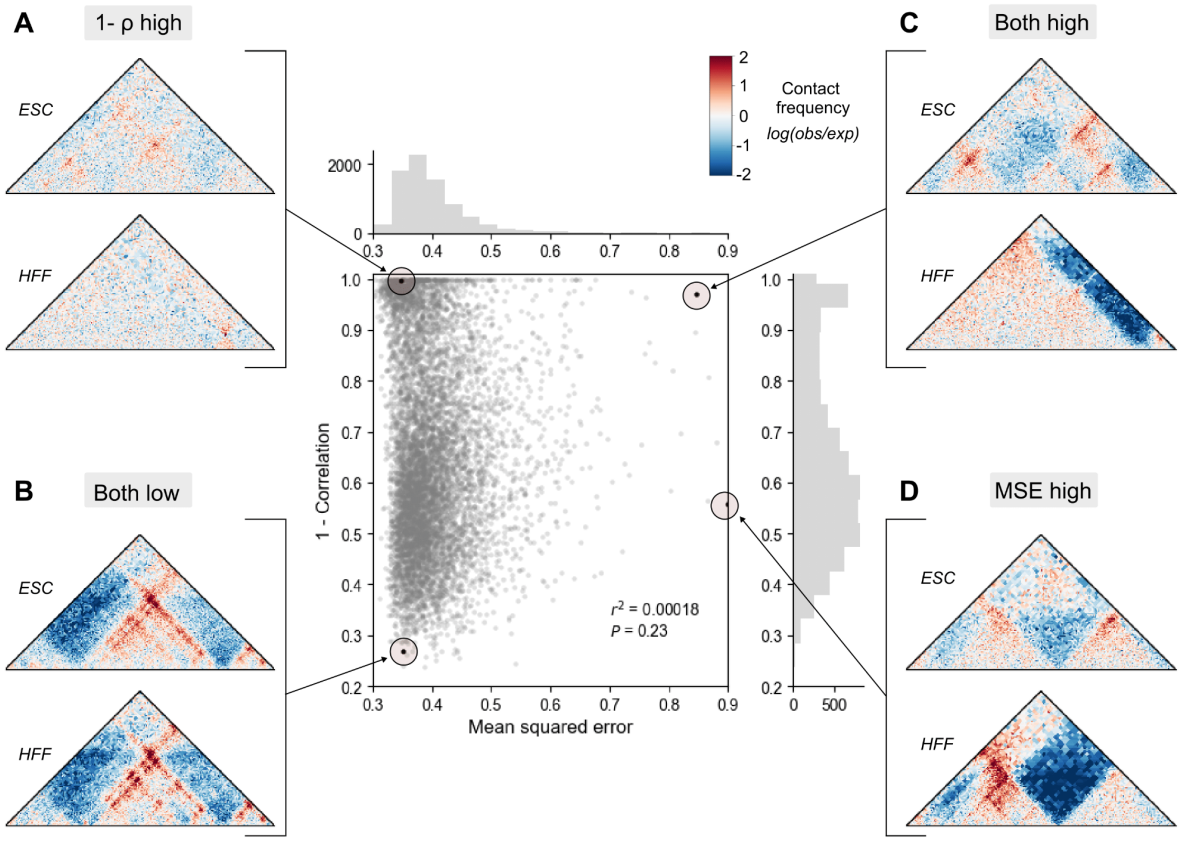


Figure 3.2. Basic methods to compare contact frequency maps rank map pairs differently.

Mean squared error (MSE) and Spearman's correlation (ρ) were calculated across the genome on experimental contact maps from embryonic stem cell (ESC) and human foreskin fibroblast (HFF) (n = 7840). Each point represents a comparison score between a pair of contact maps. We highlight examples where (**A**) only correlation ranks highly, (**B**) both methods agree the maps are similar, (**C**) both methods agree the maps are different, and (**D**) only MSE ranks highly.

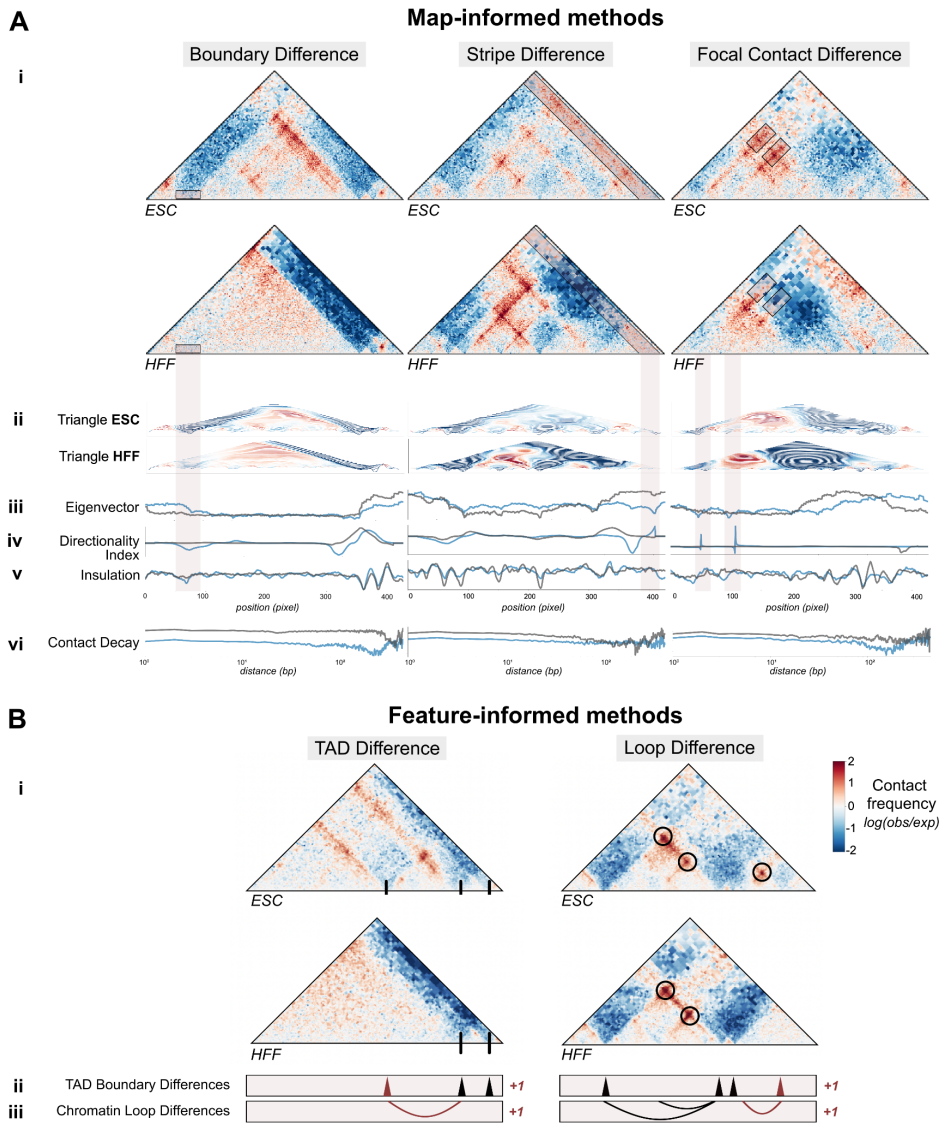


Figure 3.3: Map-informed and feature-informed methods capture differences in TAD boundaries, stripes, and loops.

A. ***i.*** Examples of regions where contact frequency maps differ between HFF and ESCs across three structural changes: a lost TAD boundary (left panel), a lost stripe (middle panel), and lost loops (right panel), as marked by red boxes. ***ii-vi***. Tracks corresponding to each map-informed disruption score method are shown below for ESCs (blue) and HFF (gray). Tracks for methods in (ii. - v.) correspond to the coordinates of the contact maps, while contact decay in (v.) is plotted across genomic distance. **B.** ***i.*** Two loci in HFF and ESC with a boundary and loop change (GRCh38 chr3:137129984-138178560 and GRCh38 chr3:138702848-139751424, respectively). ***ii.*** Applying a TAD boundary caller identifies a boundary change between cell types ***iii.*** Comparing chromatin loops identifies a genomic region with differential looping.

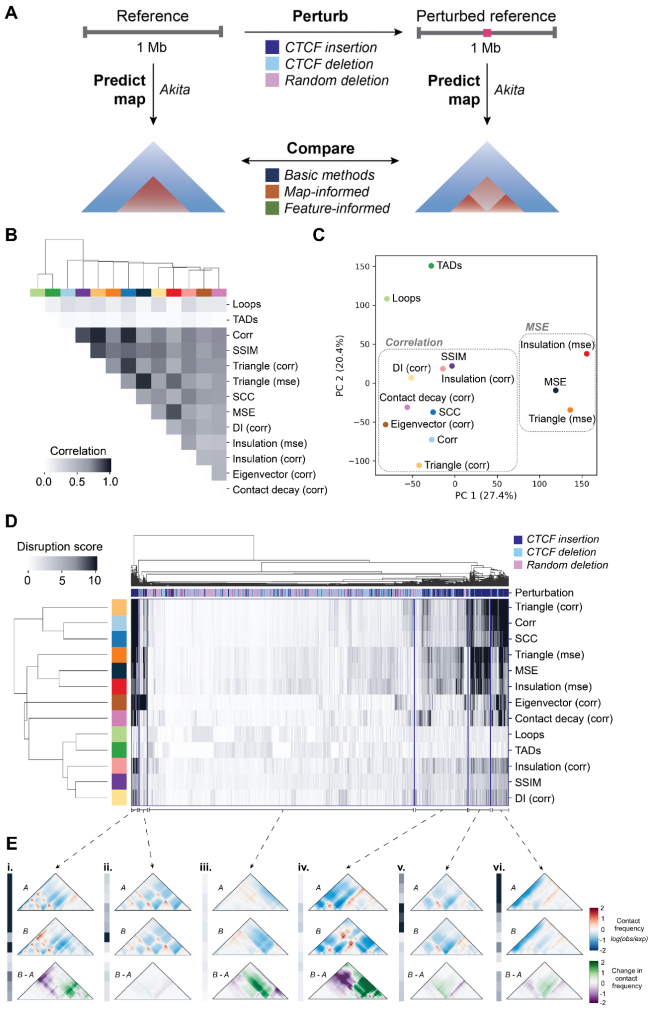


Figure 3.4. Comparison of disruption score methods.

(**A**) Schematic describing the strategy for comparing *in silico* perturbed contact maps. Random ~1 Mb windows of the human genome (GRCh38) are selected and input into Akita to predict chromatin contacts (left). The same window is also perturbed with a CTCF motif insertion, deletion, or random 100 base pair deletion. The resulting sequence is also input into Akita to predict chromatin contacts of this perturbed reference sequence (right). The perturbed and unperturbed maps were compared by applying the 11 basic, map-informed, and feature-informed methods. (**B**) Correlation matrix of the methods tested, where cells are shaded according to how well their scores correlated across perturbations. Concordance of the top-scoring perturbations (**Supp. Fig. 3.6**) also shows agreement between corr and SSIM, while highlighting that loops and triangle (corr) are quite concordant with other methods when considering only the top scores. Colors across the top of the heatmap identify the individual methods. (**C**) Principal component analysis of disruption scores of each method from perturbed map pairs. (**D**) Heatmap of normalized disruption scores across all methods and perturbations. The colored key along the top of the heatmap indicates whether the perturbation was a random deletion (pink), a CTCF insertion (navy), or a CTCF deletion (light blue). Method colors are the same as in (C). Four broad trends in disruption score patterns across methods are marked with brackets. (**E**) Representative example map pairs chosen from the groups identified in D: i. high scores across 5 methods; ii. low across all methods except for eigenvector (corr); iii. low scores across all methods; iv. low scores across methods but higher for MSE-based scores; v. high scores only for MSE-based scores; vi. high scores for correlation-based scores: triangle (corr), corr, and SCC.

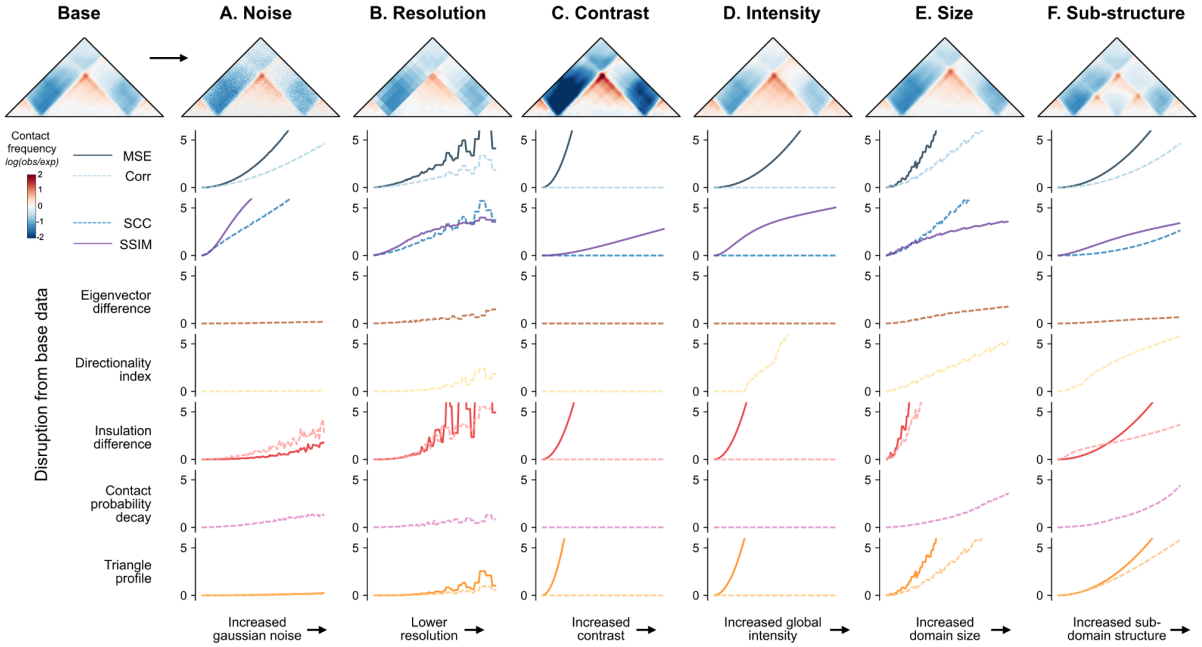


Figure 3.5. Simulated contact frequency maps with controlled perturbations estimate disruption score method sensitivities.

Normalized disruption scores are plotted for a simulated contact frequency map containing a TAD across 6 types of perturbations, plotted on the x-axis. Each perturbation was added at 100 different degrees. The images shown correspond to the final degree–the maximum perturbation added. Line plots show disruption scores from comparing the original map (top left corner) to each perturbed map. Maps corresponding to the incremental increases in perturbation are shown alongside the changed scores in **Supp. Fig. 3.8**. (**A**) Noise is added by introducing random values drawn from a Gaussian distribution to the maps; (**B**) Resolution is lowered by increasing bin size; (**C**) Contrast is applied by increasing the range of the signal; (**D**) Intensity is increased globally by adding a constant to all values; (**E**) Size is increased by slightly enlarging the domain width; (**E**) A sub-structure is added by gradually incorporating a new boundary at the center of the existing TAD.

Table 3.1: Strengths, weaknesses, and suggested applications of disruption score methods.

Trends and patterns across disruption scores summarized from statistical comparisons (**Fig. 3.4**), simulations (**Fig. 3.5**), and manual parsing of the most highly disruptive perturbations for each method (**Supp. Fig. 3.5**). While this summary is not exhaustive of all possible outcomes, it provides qualitative guidelines for users to make informed decisions when selecting a comparison method based on the scale and application of their research. We use green checks to indicate advantages and red X's to indicate disadvantages for each method category: basic methods (blue), map-informed methods (orange), and feature-informed methods (green). Double signs represent strong patterns, while no sign indicates no pattern, and NA denotes that the method was not tested.

		Technical sensitivities				Application-based features				Specs	
	Comparison method	Resistant to noise	Resistant to decreased resolution	Biased towards ___ contrast maps: High	Biased towards ___ contrast maps: Low	Sensitive to small structural changes	Sensitive to signal change at boundaries	Sensitive to loss or gain of boundaries	Sensitive to changes in TAD substructures	Needs tuning	Time & memory
	MSE			✗			✓	✓✓	✓		✓
	Corr		✓		✗		✓	✓✓			✓
	SCC				✗		✓	✓✓			✓
	SSIM					✓	✓	✓✓	✓		✓
	Eigenvector (corr)*	✓✓	✓✓			✗	✓✓	✓	✗	✓	
	DI (corr)	✓✓	✓✓			✓	✓	✓✓	✓	✓	✓
	Insulation (mse)	✓		✗		✓	✓✓	✓	✓	✓	
	Insulation (corr)	✓			✗		✓✓	✓		✓	
	Contact decay (corr)**	✓	✓✓		✗			✓✓	✗		
	Triangle (mse)*	✓	✓	✗		✓	✓✓	✓✓	✓		✗✗
	Triangle (corr)*	✓✓	✓✓		✗		✓	✓✓			✗✗
	Loop calling	*NA*	*NA*			✗	✓	✗	✓	✓	
	TAD calling	*NA*	*NA*		✗	✗	✗	✓	✓	✓	

3.7 Supplemental Notes

3.71 Basic methods

Mean Squared Error

The mean squared error (MSE) measures the average squared difference between two flattened contact matrices, such that

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

Because MSE is a measure of absolute difference, it consistently prioritizes the greatest changes in intensity between contact maps. MSE has been widely adopted across machine learning as a loss function for consistent performance and ease of use (Fudenberg, Kelley and Pollard, 2020; Schwessinger *et al.*, 2020; R. Yang *et al.*, 2021; Tan *et al.*, 2023) . Large changes between maps score highly, while visually smaller or localized changes produce lower MSE values. However, maps with differences in read count or normalization intensity will produce high MSE, despite little change in structure. For this reason, technical artifacts may dominate top map rankings scored by MSE. MSE will also deprioritize maps with large structural changes and low overall contact intensity. 2D map features will not be individually captured since the matrices are collapsed to 1D vectors.

Spearman's Rank Correlation Coefficient

The Spearman's rank correlation coefficient (ρ) assesses the correlation between the intensity of corresponding pixels in two maps by quantifying how well the relationship between the corresponding pixels can be described using a monotonic function. If the rank of intensity of all pixels in two contact maps are the same, the correlation is 1. If there is no relationship between the rank of pixel intensity between maps, the correlation is 0. Spearman's Rank Correlation coefficient can be described as:

$$\rho = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)},$$

where the number of points in the data set is represented by n, and d^2 is the squared difference in the ranks of a single coordinate y_i between the two maps, which is summed over all points.

Correlation coefficients have been used extensively to compare contact maps(Lieberman-Aiden *et al.*, 2009; Dixon *et al.*, 2015; Fudenberg, Kelley and Pollard, 2020; Tan *et al.*, 2023). Large-scale structural changes have high scores with correlation, because the ranks of each pixel in the maps are very different. This approach works well even when the contact intensity is low because magnitude of the values is not considered when converting to rank. However, Spearman's correlation is low even when the contact intensity is negligible (e.g. at an extreme, random noise will generate a very low correlation). The method does not pick up on small or focal changes in intensity, nor does it prioritize large-scale changes in intensity that do not change the map structure– the rank will stay the same even if the magnitude of the values change. Because matrices are flattened before calculating correlation, correlation also ignores the physical relationships between pixels of the map.

Stratum-adjusted Correlation Coefficient (SCC)

Contact frequency in Hi-C maps is known to exhibit a distance-dependent decay. The high similarity of the dependence pattern might bias the correlation between Hi-C maps, thus causing high, spurious correlations. SCC addresses this distance-dependence effect by stratifying Hi-C data based on genomic distance, calculating a Pearson correlation coefficient for each stratum and aggregating the weighted stratum-specific correlation coefficients with weights derived from the generalized Cochran–Mantel–Haenszel (CMH) statistic(Yang *et al.*, 2017). SCC values range from -1 to 1 and share a similar interpretation as standard correlations. The equation to calculate SCC can be written as:

$$\rho_s = \sum_k w_k p_k$$

SCC was first implemented for Hi-C map comparison by Yang et al. in the R package HiCRep (Yang *et al.*, 2017). By including distance-aware weights, SCC is able to measure the overall reproducibility of the Hi-C

matrices better than standard correlations and is resistant to decreased resolution. However, SCC is less likely to identify small changes in TAD substructures compared to some other methods surveyed here (see **Table 1**).

Structural similarity index measure (SSIM)

Structural similarity index quantifies the perceived change in structural information of two images by incorporating three terms:

Luminescence: Contrast: Structure:

$$l(x,y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \qquad c(x,y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \qquad s(x,y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3}$$

Where the integrated SSIM score is equal to:

$$\mathrm{SSIM}(x,y) = \left[l(x,y)^\alpha \cdot c(x,y)^\beta \cdot s(x,y)^\gamma\right]$$

SSIM is well-suited for identifying structural changes, and unlike correlation and MSE measures, is not biased by map contrast values. For this reason it has been incorporated into Hi-C map comparison methods previously(Galan *et al.*, 2020). However, SSIM is sometimes very sensitive to small changes relative to larger-scale changes that may appear more pronounced to the human eye. SSIM is also sensitive to the order of the input. It should be applied to the matrix as a whole (not vector-by-vector) as it is designed to account for neighboring values. NaN values must be interpolated or masked to zero.

3.7.2 Map-informed methods

Eigenvector difference

This method is inspired by genomic compartments, which are called by calculating the first eigenvector from Hi-C contact maps and assigning each genomic region to its sign (Lieberman-Aiden *et al.*, 2009). Similarly, eigenvector difference is calculated from the first eigenvector that corresponds to each contact frequency map, creating a vector annotated at each bin for both maps. These vectors are then compared using spearman's rank correlation. Because the components can have different signs that are arbitrarily assigned, MSE is not used for this method as it is sensitive to these signs and would result in falsely high scores when the maps are assigned opposite signs.

Directionality Index (DI)

The Directionality Index (DI) is a measure of contact frequency bias towards either upstream sequence or downstream sequence at some DNA locus. An inflection of DI values from negative to positive and vice versa indicates a potential chromatin boundary, where DI can be calculated by:

$$DI = \frac{B-A}{|B-A|} * \left(\frac{(A-E)^2}{E} + \frac{(B-E)^2}{E}\right)$$

Where A is the number of reads (or average normalized frequency value) that map from a given locus to upstream bins, B is that value for downstream bins, and E is the expectation under the null hypothesis, equal to $\frac{(A+B)}{2}$.

DI was first proposed by Dixon et al. in 2012 (Dixon *et al.*, 2012). It depends on two parameters: the size of the focal bin whose relative upstream and downstream contact frequency is being compared, and the size of the upstream and downstream bins (40 kb and 2 Mb in the original publication). To create a composite DI disruption score for a variant, DI is calculated for each locus in the region of both maps and compared using MSE or correlation. This composite score is subject to the caveats of the chosen comparison method.

Insulation score

Also known as the ratio score or boundary score, this method seeks to identify TAD boundary-like regions by comparing the frequency of within-region contacts upstream (A) and downstream (B) of some point X to the inter-region contacts between regions A and B (Crane *et al.*, 2015; Gong *et al.*, 2018). The higher the ratio is in a given region, the more likely this region is to be a TAD boundary. The metric is calculated as follows:

$$Insulation = \frac{max(mean(A), mean(B))}{mean(X)},$$

where X quantifies the frequency of local contacts within the central region spanning 20 kilobases, and A and B represent contact frequency in the regions upstream and downstream of X, respectively, spanning 200 kilobases each.

Correlation or MSE can be applied to the insulation tracks of two conditions for a scalar disruption score. The magnitude of the insulation score is dependent on differences in contact intensity between the two maps at each bin, and therefore is sensitive to global change in contact frequency. Variants in regions of DNA with wider ranges of contact intensity (high contrast) may have inflated insulation scores relative to other regions. This

method can potentially be improved by adjusting the following parameters: the size of central window X, i.e. the region for which the insulation score is being calculated (default: 20kb), and the size of upstream/downstream windows A and B (default: 200kb)

Contact probability decay

Contact decay, or the P(S) curve, measures chromatin interaction as a function of genomic distance (Nagano *et al.*, 2017; Zhou *et al.*, 2019). Interaction frequency across the contact map is ranked by genomic distance between all pairs of contact, resulting in a track of distance vs interaction frequency. As distance increases, the probability of contact between loci decreases. Decay curves may be calculated at a given resolution such that the chromosome is divided into $n = L/r$ bins, where L is the chromosome length and r is resolution. Across an $n \times n$ contact map, the contact frequency of each entry $A_{i,j}$ is ordered by the distance between loci, i-j. A steeper decay in contact frequency indicates a greater distance between further loci, while a shallow contact decay suggests more interaction between distant loci. Contact decay measures a global signal of relative interaction increase or decrease, but will not be sensitive to local structural changes to contact matrices.

Triangle Method

Basic methods (correlations, MSE) all ignore the physical relationships between pixels of the map when they are flattened into vectors. They are therefore over-simplified characterizations of the relationships between maps. This method tries to leverage our understanding that contacts are represented by different subsets of triangles within the larger map to address this gap. The triangle-based method compares the average contact intensity within all sub-triangles of two contact maps. In comparison to MSE and correlations, the flattened representation of the map is the average contact intensity of all sub-triangles instead of just each pixel on its own. The flattened representations can then be compared with either MSE or a correlation method.

The performance of this method depends on the correlation or MSE used over the sub-triangles (see their individual pros and cons). The advantage over those basic methods is that triangle comparison is more feature-informed to capture relevant contact relationships. Because there are so many more smaller triangles than larger triangles, this method likely prioritizes more local changes; however, one could weight the triangles or subset to only the larger or smaller sub-triangles to prioritize only larger or smaller scale interactions. One caveat is that this method is significantly slower than other methods, but speed can be improved by creating lower resolution maps before computing.

3.7.3 Feature-informed methods

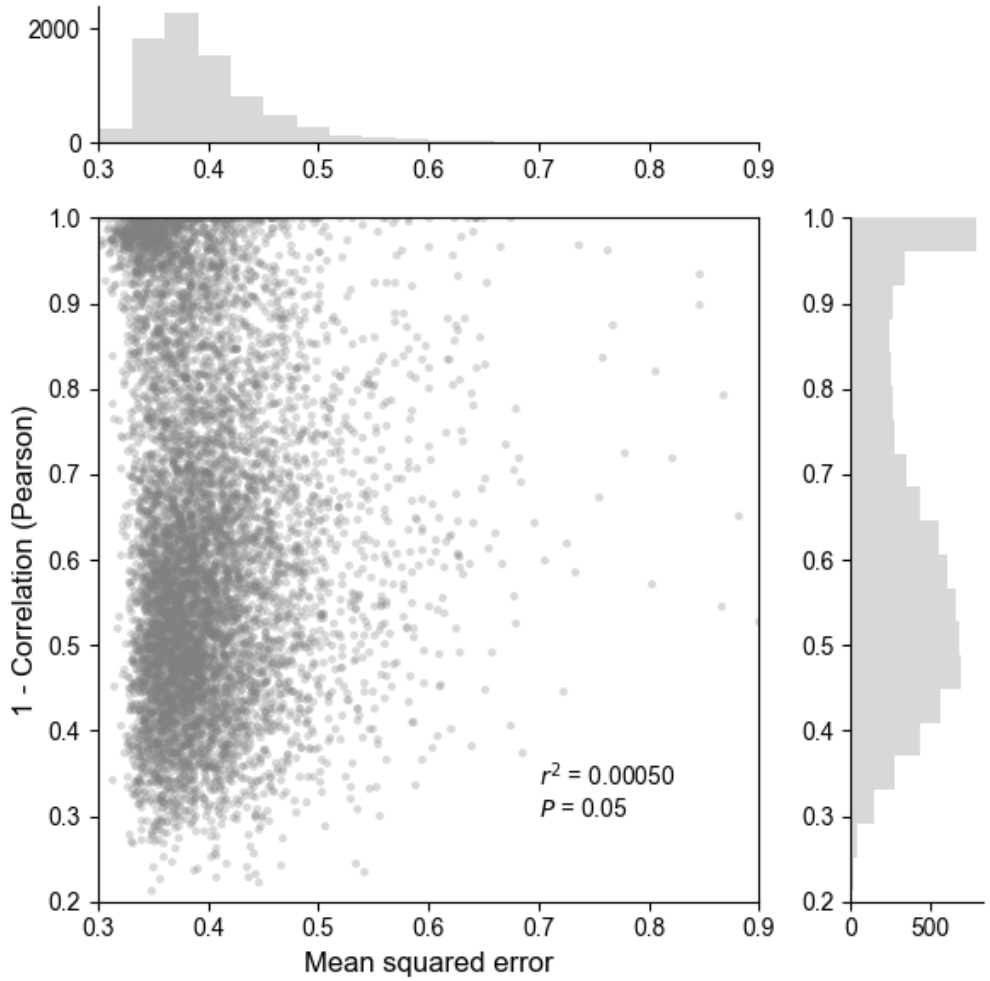
TADs

TAD boundaries are called by finding the local minima of the insulation profile, which is calculated using a diamond-shaped window-based method proposed by Crane et al. (Crane *et al.*, 2015). Specifically, a square (a W x W diamond-shaped window) is slid along each diagonal bin of the matrix and the averaged contact frequency within each window is calculated and called as insulation score. Bins with a low insulation score indicate a high insulatory effect, thus the bins reaching the local minima are identified as candidate TAD boundaries. The boundary strength is calculated for each local minima using peak prominence and candidates with strength above a threshold are referred to as TAD boundaries. The scores for the bins at the end of the diagonal and within the window size are not calculated. The overlap, gain, and loss of TAD boundaries between two Hi-C matrices are reported to show their consistency and changes. The boundary locations within a set resolution r are considered the same. This method could be further improved by changing the following parameters: window size (w), threshold of boundary strength, and upper bound of distance when two TAD boundaries are considered the one.

Loops

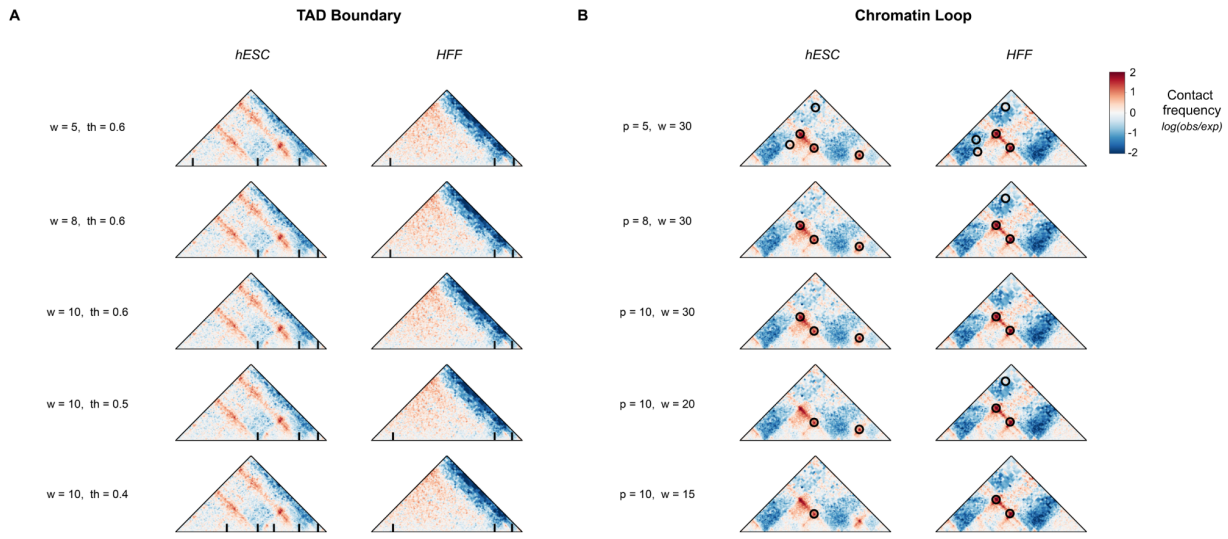
Chromatin loops are the positions where a pair of loci showing closer proximity compared to loci lying between them, corresponding to pixels with higher contact frequency than the ones in their neighborhood. We identify loops by comparing regions with their local background, as in HiCCUPS(Rao *et al.*, 2014). Specifically, for each bin in the upper triangle window of the matrix, we first check whether it is a local maximum (across neighborhood window size w) and then calculate the mean signal of center window (window size p) surrounding the bin as well as the mean signal in a donut-shape neighborhood, a lower-left neighborhood, vertical and horizontal neighborhoods around the pixel. The bins enriched above its neighborhood with ratios of mean signals of the center window to the neighborhoods higher than certain thresholds are considered as candidate loops. The bins at the corners are not considered. Loops that are the same, gained, lost between two Hi-C matrices are identified. The loops that are located within a window of size r of one another are treated as the same. This method can potentially be improved by adjusting the following parameters: the center window size (p), window size (w), threshold of the ratio of center window to donut and lower-left filter, threshold of the ratio of center window to vertical filter, threshold of the ratio of center window to horizontal filter, and the upper bound of bin distance where two loops are considered as same one.

3.8 Supplemental Figures



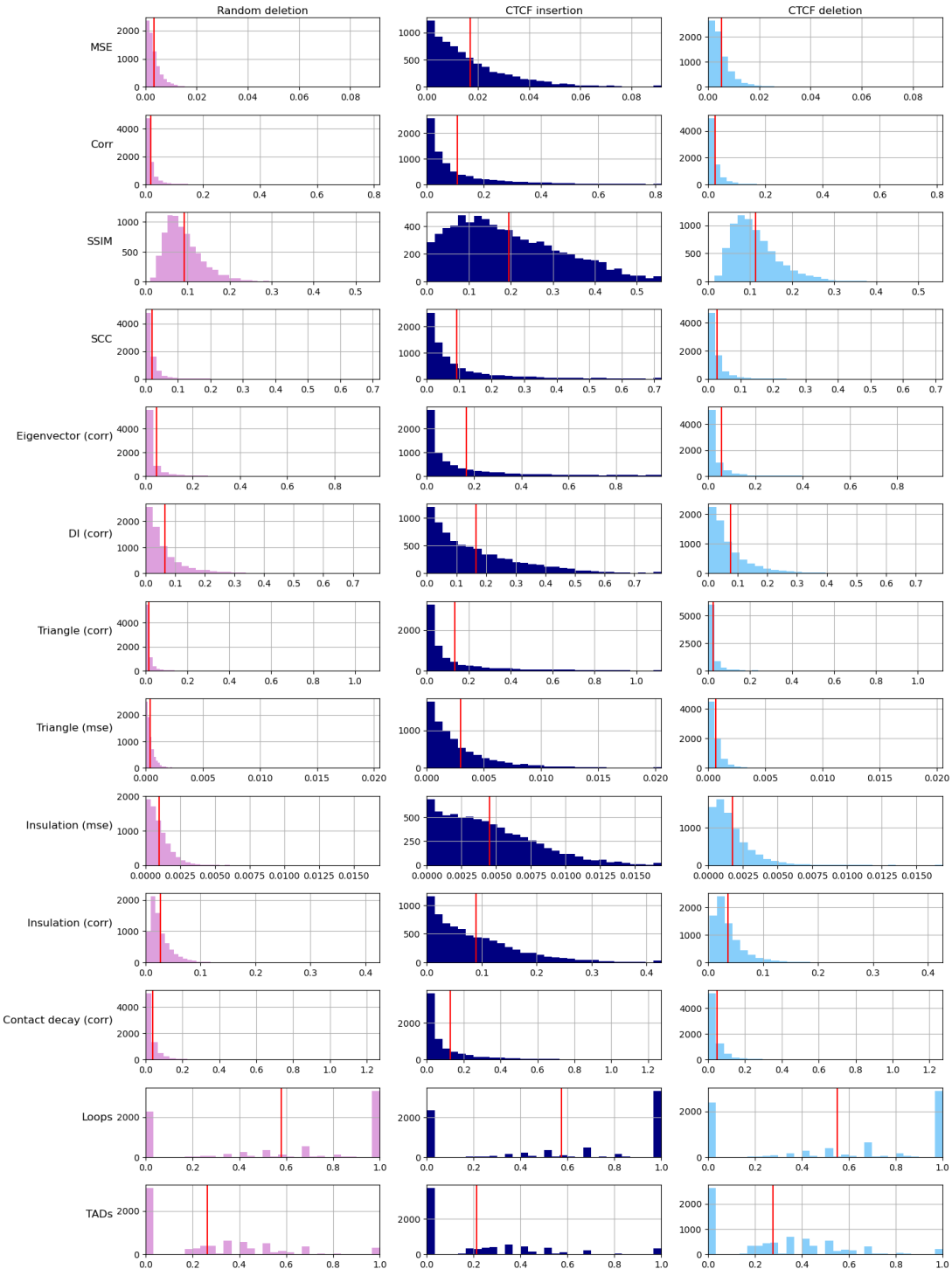
Supplemental Figure 3.1. Pearson correlation versus mean squared error comparisons of contact maps.

Mean squared error (MSE) and Pearson correlation coefficient calculated across the genome on experimental contact maps from embryonic stem cell (ESC) and human foreskin fibroblast (HFF). Each point represents a comparison between maps from HFF and ESC cell types (n = 7840 windows). There is a weak relationship between the Pearson correlation and MSE (r^2 = 0.0005, P = 0.05)



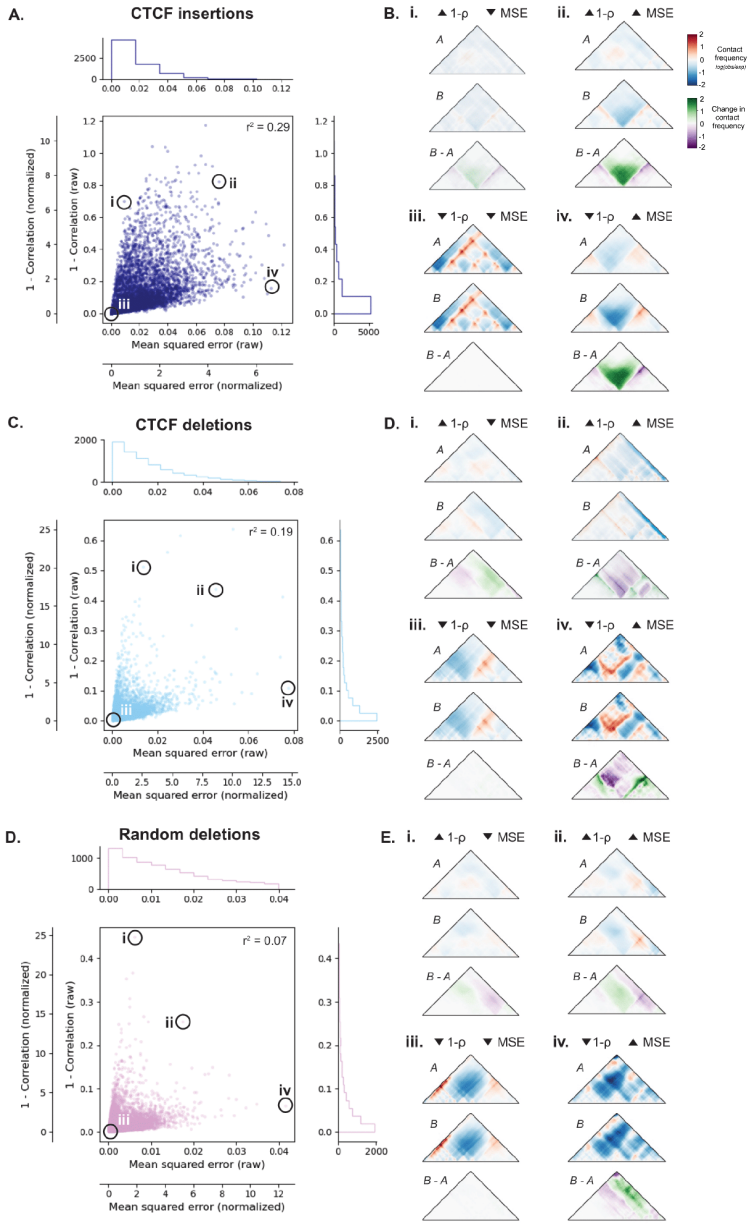
Supplemental Figure 3.2. Sensitivity of TAD and loop caller on parameter shifts.

(**A**) TAD boundaries (highlighted with black bar) called with different sizes of diamond-shaped window (*w*) and thresholds of insulation scores (*th*). (**B**) Chromatin loops (highlighted with black circle) identified using different sizes of center window (*p*) and donut filter (*w*). Example maps used here are the same as in **Fig. 3.3Bi**.



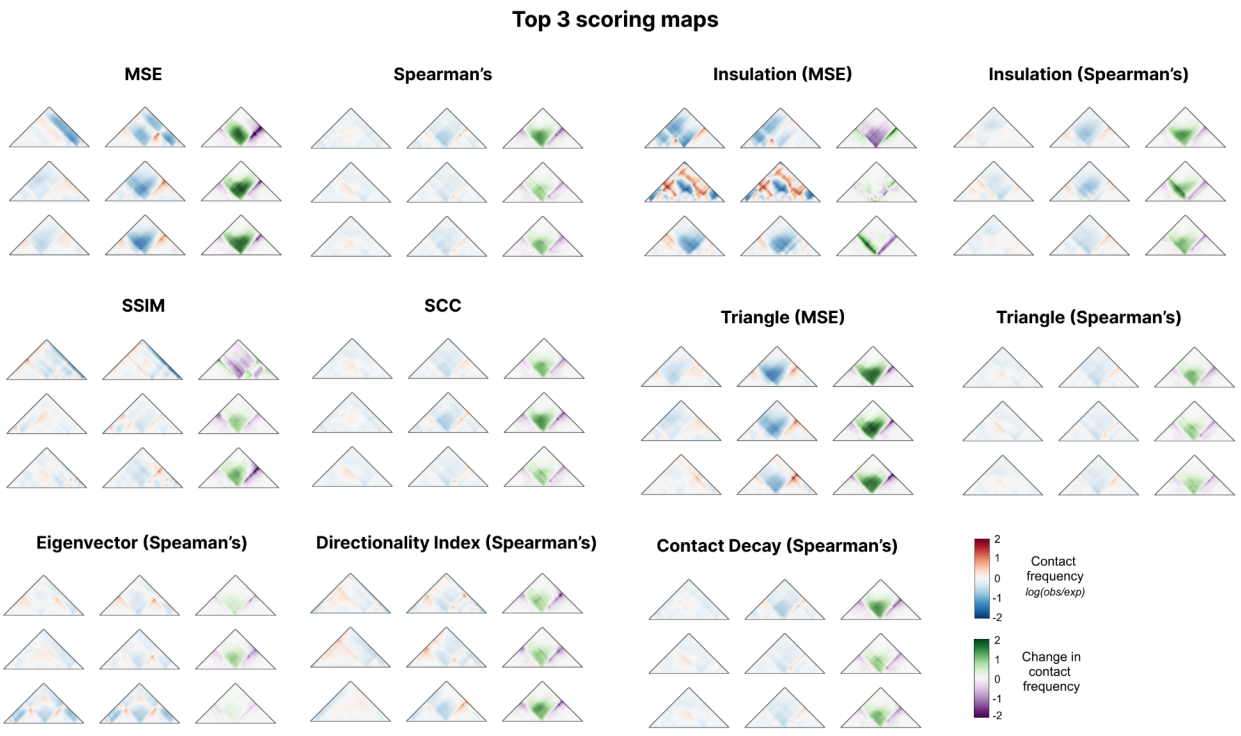
Supplemental Figure 3.3. Score distributions of random deletions, CTCF deletions, and CTCF insertions.

Each disruption score method (rows) produces a different range and mean (red line) across scores produced. Histograms show the raw scores comparing maps produced by 7500 random 100 bp deletions (left), 7500 CTCF insertions (middle), and 7500 CTCF deletions (right). To enable comparisons between the different scores, the main text figures report scores standardized to the mean disruption produced by a random 100 bp deletion. Thus, both an MSE-based disruption score and correlation-based disruption score describe that the maps are twice as different as the average 100 bp deletion.



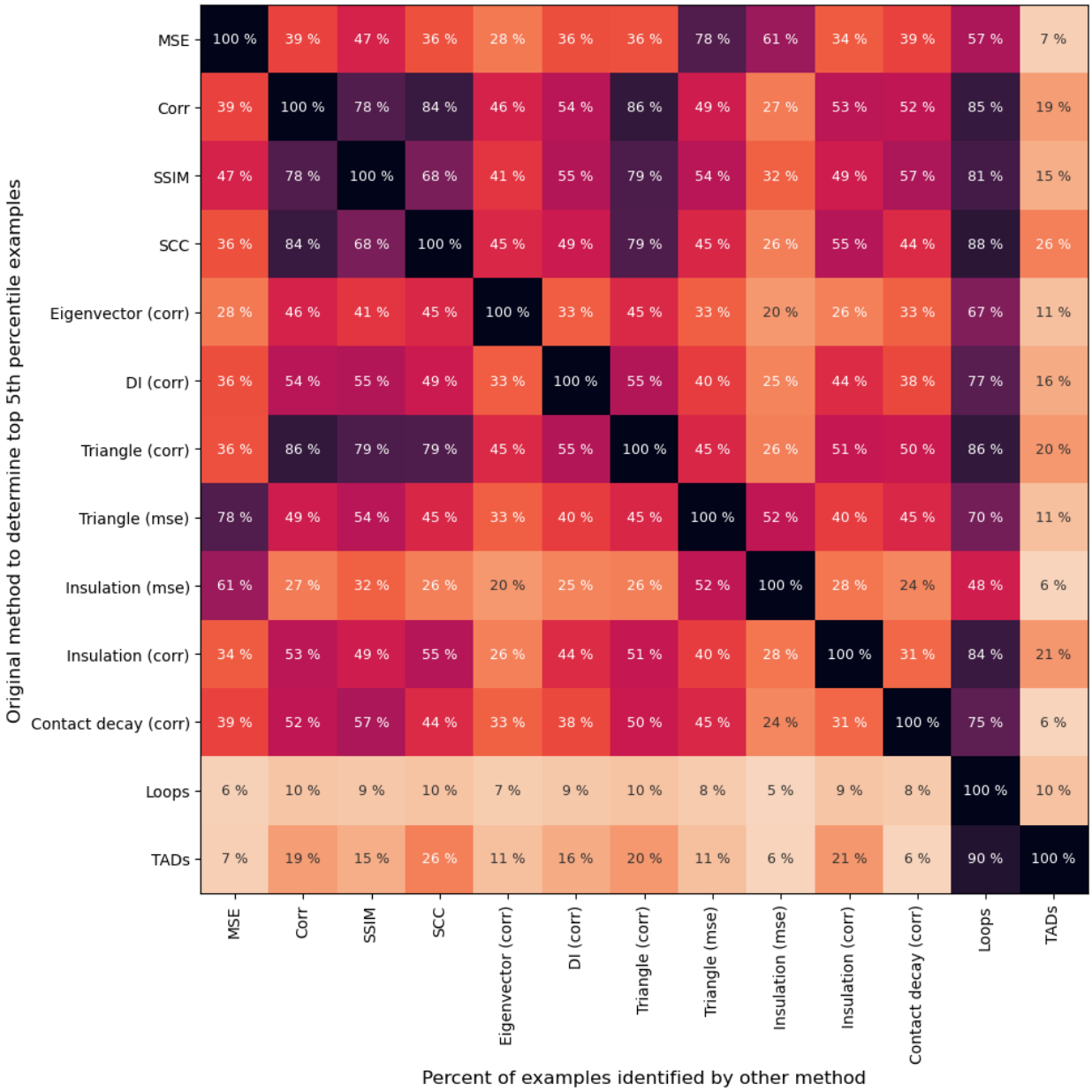
Supplemental Figure 3.4. Basic methods to compare contact frequency maps rank maps differently on *in silico* perturbations.

Mean squared error (MSE) vs Spearman's correlation (ρ) scores across an *in silico* screen of 7,500 map pairs with and without CTCF insertions (A-B), CTCF deletions (C-D) and random 100 bp deletions (E-F), similar to **Fig. 3.2.** We plot 1 - ρ such that higher values for both methods reflect increasing differences between maps. MSE versus Spearman's correlation are plotted where each point represents a comparison between a reference and perturbed map (**Fig. 3.4a**). Normalized scores are divided by the mean of the distribution of random deletions. Across all three perturbations, there is a weak relationship between the two disruption scores (r^2 = 0.29, 0.19, and 0.07 for CTCF insertions, CTCF deletions and random deletions, respectively). The relationship is strongest for CTCF insertions, for which scores are highest, followed by CTCF deletions, which have the next highest scores. Yet perturbations with at least one high score are not always concordantly scored. Examples of extreme scores for each perturbation are shown in B, D, and F in panels i through iv, illustrating that perturbations with high MSE and low 1 - ρ are consistently maps with high contrast, while low MSE and high 1 - ρ perturbations are maps with overall low contrast.



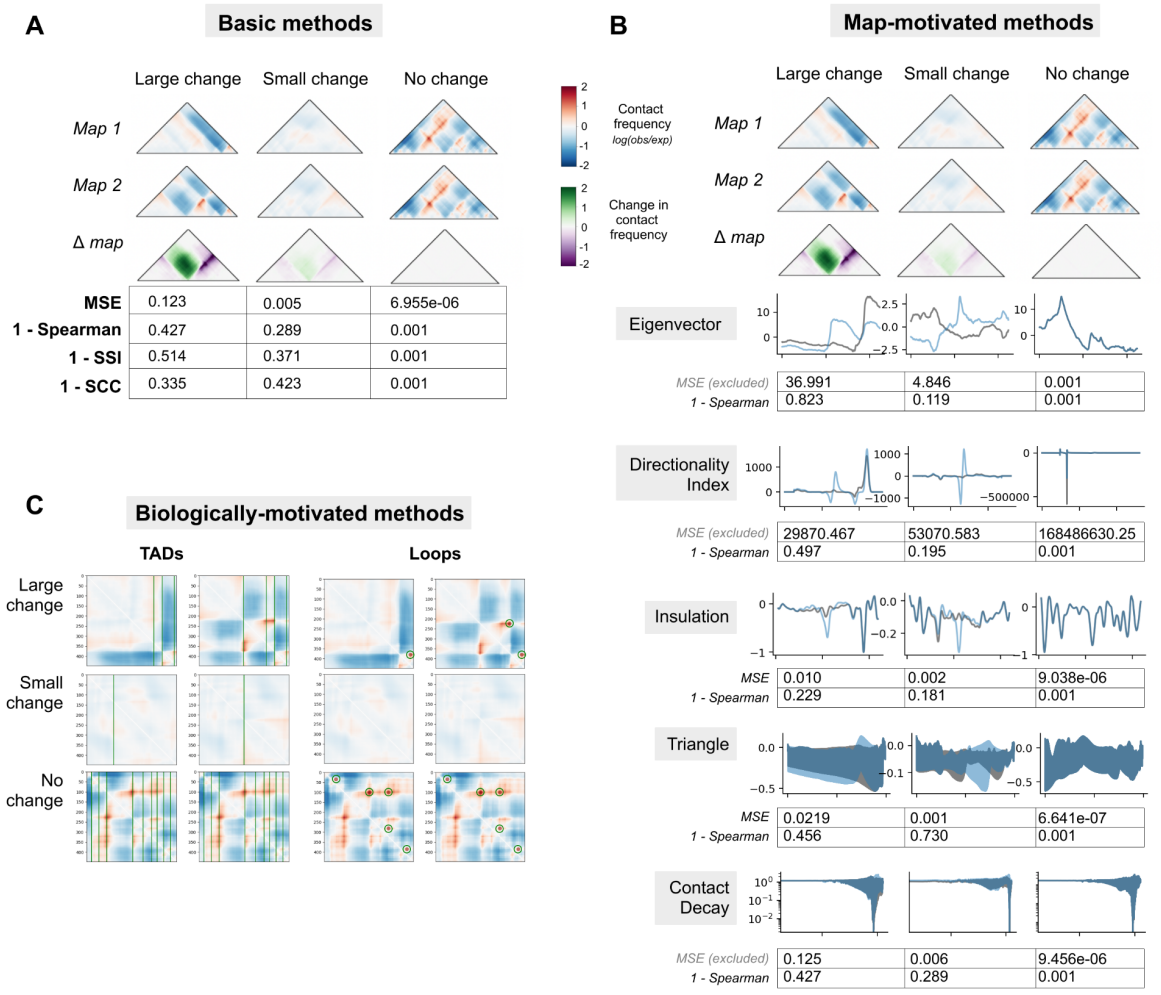
Supplemental Figure 3.5. The three most disruptive map pairs of each scoring method.

For each example row, the unperturbed map is shown on the left, the perturbed map is shown in the middle, and the difference between the two maps is shown on the right. The top three disruptive maps were chosen across the *in silico* screen.



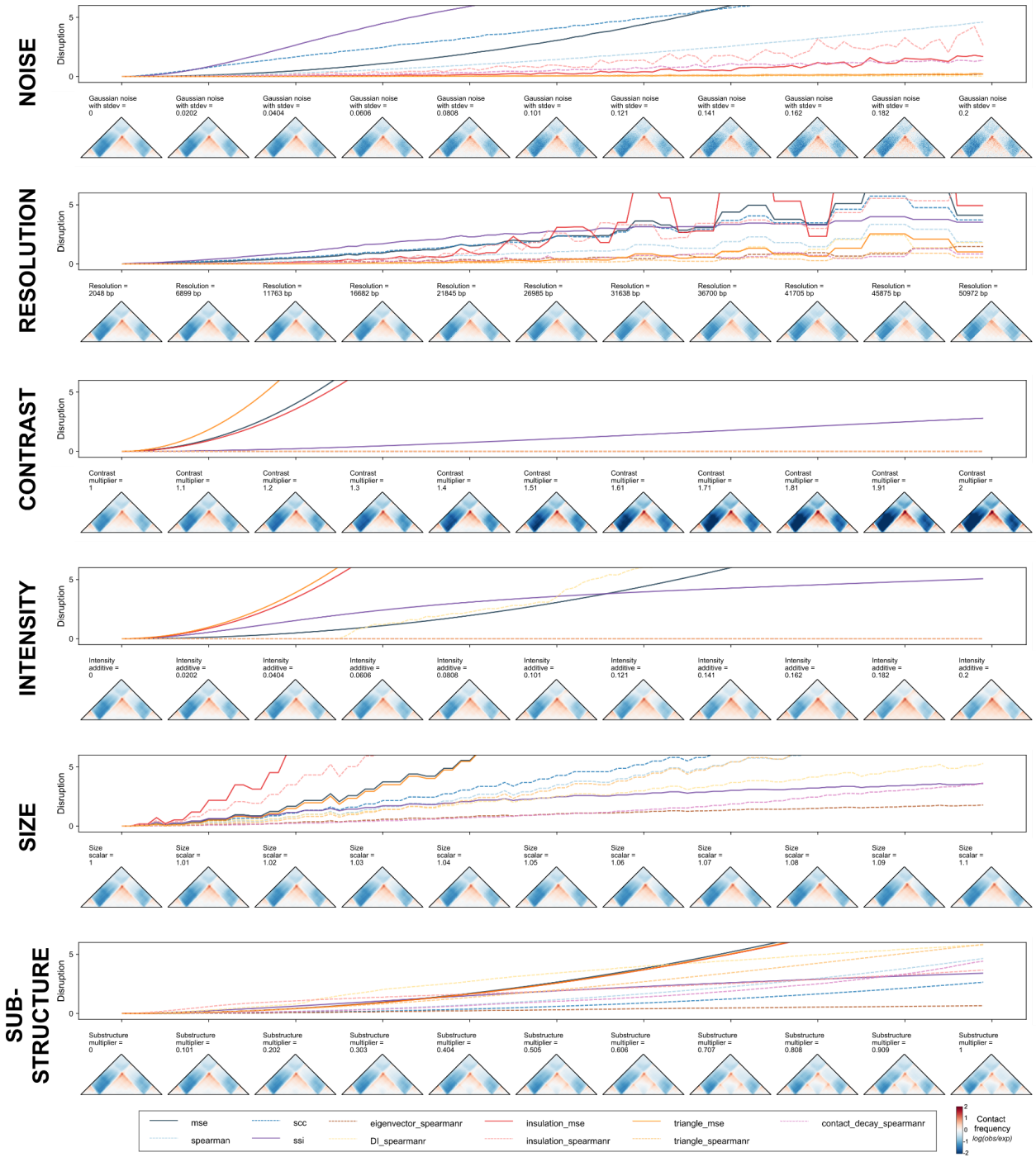
Supplemental Figure 3.6. Overlap of the most disruptive map pairs identified by each scoring method.

Each cell in the heatmap represents the percentage of map pairs that are above the 5% cutoff for the method in row and above the 5% cutoff for the method in the column. Darker colors indicate higher concordance for the top scoring loci. The heatmap is symmetric except for Loops and TADs. The imbalance of these two methods is caused by multiple map pairs that have scores equal to the 5th percentile, which results from methods producing low counts of discrete values.



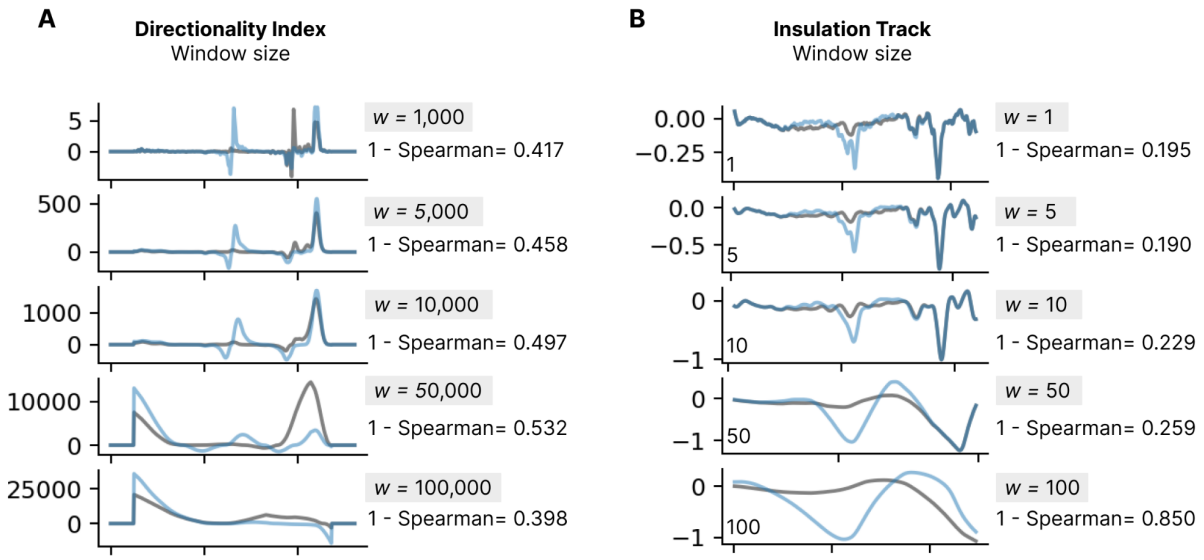
Supplemental Figure 3.7. Scoring metrics on contact map pairs with large, small, and minimal changes.

(**A**) Basic method scoring results across three example loci with a large, small, and minimal change upon CTCF motif insertion. (**B**) Map-motivated scoring results across three example loci. Raw tracks are shown for each measurement and the MSE and Spearman's correlation between the tracks are shown below. (**C**) Feature-informed scoring examples across three example loci with a no change, a minimal change, and a large change to folding.



Supplemental Figure 3.8. Changes of disruption scores with gradual increases in perturbations.

Each subpanel shows the changes of disruption scores (top row) and contact maps (bottom row) against the incremental changes in a technical or biological variation. The colors of the scoring metric are the same as seen in **Fig. 3.5**.



Supplemental Figure 3.9. Sensitivity of directionality index and insulation tracks on parameter shifts.

Directionality index (**A**) and insulation (**B**) tracks across a range of input window size choices, as well as the resulting Spearman's correlation between the two tracks. A window size of 10 Mb was used for both approaches to produce the *in silico* scoring results in the **Results** section.

Supplementary Table 3.1: Method summary table

Method (abbreviation)	Description	Parameters	Time per 1k calls (s)
Mean squared error (MSE)	Measures the average squared difference between two flattened contact matrices.	NA	0.24
Spearman's correlation coefficient (corr)	Assesses the correlation between the intensity of corresponding pixels in two maps.	NA	5.31
Stratum-adjusted correlation coefficient (SCC)	Addresses bias due distance-dependence patterns in Hi-C data by stratifying based on their genomic distance, calculating a Pearson correlation coefficient for each stratum and aggregating the weighted stratum-specific correlation coefficients with weights derived from the generalized Cochran–Mantel–Haenszel statistic.	NA	2.96
Structural similarity index measure (SSIM)	Quantifies the perceived change in structural information of two images by incorporating three terms that correspond to structure, luminescence, and contrast	NA	2.37
Eigenvector difference (eigenvector)	Calculated from the first eigenvector that corresponds to each contact frequency map. Eigenvector tracks are compared between maps using spearman's rank correlation.	NA	41.51
Directionality index difference (DI)	A measure of contact frequency bias towards either upstream sequence or downstream sequence at a locus. An inflection of DI values from negative to positive and vice versa indicates a potential chromatin boundary.	(1) size of the focal bin where upstream and downstream contact frequency is compared (2). size of upstream and downstream bins	3.98
Insulation difference (insulation)	Identifies TAD boundary-like regions by comparing the frequency of within-region contacts upstream and downstream of some point to the inter-region contacts between regions A and B	(1). size of central window (region where insulation score is calculated, x_size) (2). size of upstream/downstream windows used (a_size)	10.98
Contact decay difference (contact decay)	Measures chromatin interaction as a function of genomic distance. Interaction frequency across the contact map is ranked by genomic distance between all pairs of contact, resulting in a track of distance vs interaction frequency. As distance increases, the probability of contact between loci decreases.	(1). mean vs median	7.17
Triangle profile (triangle)	Tries to leverage our understanding that contacts are represented by different subsets of triangles within the larger map. Compares the average contact intensity within all sub-triangles of two contact maps.	NA	4284.15
TAD calling (TADs)	TAD boundaries are called by finding the local minima of the insulation profile, which is calculated using a diamond-shaped window-based method. Specifically, a square is slid along each diagonal bin of the matrix and the averaged contact frequency within each window is calculated and called as insulation score.	(1) window size (w) (2). threshold of boundary strength (3). upper bound of distance when two TAD boundaries are considered one	8.16
Loop calling (Loops)	Loops are identified by comparing with their local background. For each bin in the upper triangle window of the matrix, we first check whether it is a local maximum and then calculate the mean signals of center window surrounding the bin and also the mean signals in a donut-shape neighborhood, a lower-left neighborhood, vertical and horizontal neighborhoods around the pixel.	(1). center window size (p) (2). window size (w) (3). threshold of the ratio of center window to donut and lower-left filter (4). threshold of the ratio of center window to vertical filter (5). threshold of the ratio of center window to horizontal filter (6). upper bound of bin distance where two loops are considered as same one.	104.18

4

ChromaFactor: deconvolution of single-molecule chromatin organization with non-negative matrix factorization

4.1 Abstract

The investigation of chromatin organization in single cells holds great promise for identifying causal relationships between genome structure and function. However, analysis of single-molecule data is hampered by extreme yet inherent heterogeneity, making it challenging to determine the contributions of individual chromatin fibers to bulk trends. To address this challenge, we propose ChromaFactor, a novel computational approach based on non-negative matrix factorization that deconvolves single-molecule chromatin organization datasets into their most salient primary components. ChromaFactor provides the ability to identify trends accounting for the maximum variance in the dataset while simultaneously describing the contribution of individual molecules to each component. Applying our approach to two single-molecule imaging datasets across different genomic scales, we find that these primary components demonstrate significant correlation with key functional phenotypes, including active transcription, enhancer-promoter distance, and genomic compartment. ChromaFactor offers a robust tool for understanding the complex interplay between chromatin structure and function on individual DNA molecules, pinpointing which subpopulations drive functional changes and fostering new insights into cellular heterogeneity and its implications for bulk genomic phenomena.

4.2 Introduction

Chromatin is intrinsically dynamic, and its behavior across time restricts and permits the precise regulatory landscape controlling gene expression (Misteli, 2020; Hafner and Boettiger, 2022). Recent single-cell technologies such as single-cell Hi-C (Nagano *et al.*, 2017; Ramani *et al.*, 2017) and chromatin microscopy techniques (Cardozo Gizzi *et al.*, 2019; Mateo *et al.*, 2019; Liu *et al.*, 2020; Su *et al.*, 2020; Takei *et al.*, 2021) now offer unique insight into genome folding, allowing us to directly observe chromatin folding as well as functional readouts in individual cells to disentangle their mechanistic relationship.

Linking chromatin conformation to function in single cells presents several key challenges: 1) Single cell data is extremely sparse (Zhou, Zhang and Ma, 2021). Current single-cell technology often yields incomplete information, such as missing values or misallocated genomic coordinates. 2) Single-cell measurements capture snapshots, whereas chromatin function may result from a dynamic behavior as it moves across time. Phenomena observed in bulk experiments, such as Hi-C, may be artifacts of averaging across cell populations and patterns seen in bulk may not exist in single cells (Hafner *et al.*, 2022). 3) Capturing chromatin folding and phenotypic measurements like nascent transcription in the same cells has only recently become possible, but temporal offsets between folding and function could introduce uncertainty. 4) The bulk trends we observe in aggregate may be driven by a small fraction of cells, and identifying this subset amidst heterogeneous single-cell chromatin measurements presents a complex challenge. Connecting chromatin behavior to function in individual cells remains intractable given these technical barriers.

Several computational methods have been developed in response to emerging single-cell imaging and high-throughput sequencing techniques to measure chromatin conformation. Topic modeling(Kim *et al.*,

2020), random-walk methods(Zhou *et al.*, 2019) and recent deep learning approaches(Zhang, Zhou and Ma, 2022a; Zheng, Shen and Keleş, 2022) effectively cluster cells into subpopulations. Recent methods also offer rich annotations in single cells, including A/B compartments (Zhang, Zhou and Ma, 2022a, 2022b), subcompartments(Xiong, Zhang and Ma, 2023), topologically associating domains (TADs)(Zhang, Zhou and Ma, 2022a, 2022b), and chromatin loops(Yu *et al.*, 2021). Rajpurkar et al. first applied a convolutional neural network to directly predict nascent transcription from chromatin folding(Rajpurkar *et al.*, 2021) and Zhan et al.(Zhan *et al.*, 2023) propose an effective deep-learning-based dimensionality reduction method to cluster conformations. However, these works did not yet connect the behavior of individual cells to populations of similar conformations that are transcriptionally on or off. We build on these works to relate the behavior of individual cells to bulk trends as well as mechanistically link chromatin behavior to transcription.

We introduce ChromaFactor, a non-negative matrix factorization (NMF) technique to decompose single-cell datasets into interpretable components and identify key subpopulations driving cellular phenotypes. Non-negative matrix factorization (NMF) offers an ideal approach to analyze such complex data due to its inherent capacity to reduce high-dimensional data into a lower-dimensional, interpretable format(Lee and Seung, 1999). NMF has a robust legacy in genomics as it allows for the deconvolution of composite signals into a set of additive components and can therefore discern patterns and structures in noisy, large-scale data. Notably, it has been used on bulk Hi-C data for TAD calling(Lee and Roy, 2021) and has found applications in other emergent single-cell modalities, including single-cell RNA-Seq(Kotliar *et al.*, 2019) and spatial transcriptomics datasets(Townes and Engelhardt, 2023). By applying NMF to single-cell genome folding datasets, we can identify significant components or *'templates'* that account for the majority of cellular variation. Linking these templates to matched functional readouts describes how differences in cell populations correspond to differences in phenotypes. ChromaFactor deconvolves single-cell chromatin organization datasets into their

most meaningful primary components, providing new insights into the interplay between chromatin structure and function. Here, we apply ChromaFactor to two single-cell imaging datasets and link templates to nascent transcription. This tool may also be applied to any set of ordered matrices in single cells.

4.3 Results

NMF to decompose single-cell genome conformation datasets

We were motivated to develop ChromaFactor by the disconnect between meaningful signal observed bulk cell populations and the extreme heterogeneity of single-molecule examples. One such dataset, Mateo et al. 2019(Mateo *et al.*, 2019), profiles local chromatin conformation at the bithorax complex (BX-C) in *Drosophila* embryos and additionally includes matched nascent transcription in the same cells (n = 19,103). To discover how cell populations vary, we often take the difference between the average contact maps under two conditions. We observed a pronounced boundary in cells within the 30 kb region actively transcribing the Abd-A gene, as compared to non-transcribing cells (**Fig. 4.1a**). However, these patterns are nearly impossible to discern in individual cells (**Fig. 4.1b**). Given the heterogeneity and dynamic behavior of chromatin in single cells, identifying which cells contribute to overall trends is complex. Are these contact patterns visible at the single-cell level, or are they composite effects resulting from population-wide averaging? To bridge the gap between single-cells and bulk averages and identify which cells contribute most, we propose applying NMF to single-molecule chromatin conformation datasets with ChromaFactor.

In this approach, NMF decomposes a non-negative distance matrix into two lower-rank non-negative matrices, such that their product approximates the original matrix. ChromaFactor decomposes an *n* by *n* by *m* count matrix, where *n* is the number of genomic loci profiled and *m* is the number of cells, into an *n* by *n* by *k* component matrix W, where *k* is a specified number of components, and a *k* by *m* weight matrix, H (**Fig. 4.1c**). The matrix W represents the basis vectors, which we call *templates*, as they resemble patterns observed across the cell population. The method accepts both distance matrices from single-molecule imaging experiments and contact matrices from single-cell Hi-C. Here, the number of components (*k*) was selected to balance template interpretability and reconstruction error. The matrix H represents the weight matrix, signifying contributions of cells to components such that the data for each molecule is approximated as the weighted average of the components plus noise. To estimate W and H, matrices are randomly initialized and updated to minimize the reconstruction error between their product and the single-cell dataset.

Relating single cells to bulk trends with ChromaFactor

When ChromaFactor is applied to the Mateo et al. dataset with twenty components, we find that several templates resemble chromatin boundaries (**Methods**, **Fig. 4.1d, Supp. Fig. 4.1, Supp. Fig. 4.2**). To visualize the relationships within the single-cell dataset, we apply UMAP on the weight matrix, H (**Fig. 4.2a**), and label cells by the component with the largest weight. Investigating individual examples, we find that single cells can resemble these template patterns. To illustrate, we show the 3D coordinates and distance maps of three cells, along with their component contributions from weight matrix H (**Fig. 4.2b-d**). Since every cell is an additive combination of the templates, we can multiply each cell's weights by the component matrix to reconstruct single-cell examples, which are less noisy than the original cell measurements (**Fig. 4.2e**). Notably, these cells closely resemble the component with the highest contribution, indicating that templates can be

representative of cell subpopulations (**Fig. 4.2f**). Indeed, considering all cells in the same group, we find that median contact resembles the closest template (**Fig. 4.2g**)

Templates are significantly correlated with active transcription

Templates may capture subsets of cells and cellular patterns. Which components, if any, correspond to biological phenotypes? To investigate if templates are correlated with downstream biological function, we train a random forest model to predict nascent transcription of nearby genes from the weight matrix H alone (**Methods, Fig. 4.3a**). In the Mateo et al. dataset, three genes were profiled in the same cells that were imaged, producing matched chromatin organization and transcription data(Mateo *et al.*, 2019). Predictive performance would indicate that the components capture salient information about transcriptional state and may serve as a proxy for the raw input distance matrices themselves.

Different random forest models were separately trained to predict transcription in the 17 measured gene isoforms. We find that the weight matrix can modestly predict transcription across several genes, including Abd-a, Ubx, and Abd-b on balanced datasets of transcribing and non-transcribing cells (**Fig. 4.3b**). Indeed, the performance of the random forest parallels the performance of a random forest trained directly on the distance matrices, achieving an accuracy of 67.4% and 65.3%, respectively. Examining the feature importance of the Abd-a trained model, we find that components 0, 1, 5, and 14 are particularly influential for the model's prediction (**Fig. 4.3c**).

We can alternatively address which components are preferentially upweighted by transcribed cells by evaluating component weights separately for transcribing and non-transcribing cells. Twelve of the twenty components have significantly different weights between Abd-A transcribing and non-transcribing cells (two-sided Mann Whitney U, p-value < 0.001, **Fig**. **4.3d**). This effect is most extreme in components 0, 1, 5, and 14, the same components identified as salient by the random forest models. Visually, these components show a separation of chromatin into two distinct compartments across three separate points across the locus (components 0, 1, 5), as well as a sharp decrease in contact at the center of the locus (component 14). Significant templates suggest how contact differs upon active transcription to favor stricter subcompartmentalization within the genomic locus.

Subpopulations of transcribing cells drive contact patterns observed in aggregate

Our aim is to understand not just how contact differs, but which subpopulations of cells drive the changes we observe in bulk (**Fig. 4.1a**). We consider transcribing cells with the top 50% of weights in components 0, 1, 5, and 14, which we call 'high contribution' cells, and contrast them with non-transcribing cells in the bottom 50% of component contributions ('low contribution' cells). These cells make up only 12.7% of the total cell population, but their component weights are the most predictive of transcriptional state. These high and low contribution cells occupy different areas of the UMAP plot seen previously– low contribution cells are more likely to be mixes of components and high contribution cells are more likely to favor one component, suggesting that biologically consequential cells resemble templates (**Fig. 4.3e**).

These subpopulations of cells differ not just in chromatin conformation and transcriptional state, but also in their local behavior of regulatory elements. The distance between Abd-A and all proximate enhancers at

the same locus is notably smaller in high contribution cells as compared to low contribution cells (**Fig. 4.3f**). Enhancers are closer to the gene promoter in the subset of active cells identified by ChromaFactor. Moreover, examining the median contact of high and low contribution cells, we observe contact patterns far more pronounced than those observed in bulk (**Fig. 4.3g**). Cells with high component weights possess stronger boundary separation as well as a stripe of contact centered at the location of Abd-A, which is absent in low-contribution, transcriptionally-off cells. In this case, the cell population identified by ChromaFactor exhibits a more potent and unified profile of compartmentalized chromatin driving smaller enhancer-promoter distances when compared with all transcribing cells.

In sum, template analysis at this locus paints a holistic portrait of higher genomic sequestration between loci, reducing the distance between enhancers and promoters, thereby increasing the likelihood of transcription. This trend, although suggested at the level of the bulk population, is strongly driven by a small subpopulation of single cells. The remaining population is extremely heterogeneous across transcribing and non-transcribing cells such that their contact effectively cancels out.

Application of ChromaFactor to holocarboxylase synthetase (HLCS) locus highlights local and compartment-level chromatin shifts upon transcription

To demonstrate the efficacy of ChromaFactor across genomic scales, we next apply NMF to a 10 Mb locus in human IMR90 cells with 40 kb resolution(Su *et al.*, 2020). We profile a population of 7,590 cells derived from genome-wide profiling of chromatin conformation and nascent transcription with microscopy. The median genomic distance between cells actively transcribing and not transcribing the HLCS gene reveals no visually discernible change in contact (**Fig. 4.4a**). However, after subtracting one contact matrix from the other to

examine the difference in contact between populations, we observe a weakened boundary directly upstream of the HLCS locus.

We decompose the single-cell imaging dataset into twenty components with ChromaFactor and identify components with significantly different weights in transcribing cells (**Fig. 4.4b, Supp. Fig. 4.3**). Of the twenty components, five are significantly different in transcribed cells (two-sided Mann-Whitney-U, p-value < 0.05), exhibiting a diverse range in contact differences across templates. Component weights are elevated in transcribing cells in all components but component 11, where a decrease in contact is observed at the precise location of the boundary loss observed in bulk. Curiously, component 14 highlights a sharp change in contact at two particular genomic loci, one of which is the location of the HLCS gene itself. The method has no knowledge of transcriptional state nor gene location, indicating that change in contact at this locus may nonetheless contribute to a significant amount of variation across the cell population.

We find that two significant components, 0 and 11, correspond to a steep shift in the directionality index, a measure of contact frequency bias towards either upstream sequence or downstream sequence, at the location of the HLCS locus (**Fig. 4.4c**). Components 3 and 13 display a sharp increase in insulation at the location of a compartment switch from A to B in IMR90 cells (**Fig. 4.4d-e**). In sum, template analysis at this locus suggests that the change of transcription correlates with a reorganization of chromatin directly at the site of active transcription, at boundary shifts directly upstream of the locus, and within broader compartments downstream of the locus.

4.4 Discussion

This study presented ChromaFactor, a novel application leveraging Non-negative Matrix Factorization (NMF) for dissecting single-cell chromatin conformation datasets. ChromaFactor uncovers nuanced layers of genome conformation dynamics and their correlation with transcriptional states, which would otherwise be obscured in bulk analyses. Correlations between template patterns and active transcription suggest that these templates are not merely reflections of cellular heterogeneity, but could be mechanistically linked to transcriptional regulation. Our application of ChromaFactor to the Mateo et al. and Su et al. datasets leads us to two intriguing biological interpretations: 1) bulk behavior can sometimes be observed in individual cells, and 2) only a small minority of cells in the population drive population-wide signal. This reasoning is not possible at the bulk level, where trends may be artifacts of averaging, nor at the single-cell level, where it is unknown which snapshots capture relevant signal.

Looking ahead, the application of ChromaFactor to a wider array of cell types, genomic phenomena, and sc-HiC will help us to understand if these conclusions hold across biological contexts. ChromaFactor's ability to isolate functional portions of single-cell chromatin datasets could clarify the structural dynamics underpinning cell-type-specific gene regulation and the development of cellular heterogeneity. Additionally, integrating ChromaFactor with multi-omics approaches, such as single-cell RNA-Seq, ATAC-Seq, or CUT&Tag, could help resolve the interplay between chromatin structure, epigenetic modifications, and gene expression. ChromaFactor's application to pathological states or CTCF degradation experiments presents another exciting area of future exploration.

While the application of ChromaFactor is promising, it is important to note its current limitations. The number of templates, *k,* is manually selected. The approach also relies heavily on the quality of the single-cell chromatin datasets and the resolution at which they are produced. The inherent noise and technical artifacts present in these datasets can influence the deconvolution process and the interpretation of the resulting components. Extremely noisy datasets with consistent dropout locations will produce templates capturing dropout. Further improvements in single-cell chromatin imaging and sequencing technologies will likely enhance the accuracy and interpretability of ChromaFactor's outputs.

Finally, a deeper investigation of the biological interpretation of components is warranted. While we have shown that these components correlate with transcription and other genomic features, the exact mechanisms through which these templates influence cellular behaviors remain largely unknown. Additional studies are needed to mechanistically link these structural templates with specific functional outputs and to explore their potential role in modulating regulatory response. In sum, this study introduces ChromaFactor as a promising tool for decoding single-cell chromatin conformation data. It provides a more granular view of the dynamic nature of genome architecture and its role in gene regulation, thereby broadening our perspective on the intricate interplay of genome organization and function.

Code Availability

This package as well as notebooks to reconstruct the figures can be found at the following repository: https://github.com/lgunsalus/ChromaFactor.

4.5 Methods

Datasets and processing

Mateo et al. dataset

The Mateo et al. microscopy dataset contains 3D genomic coordinates and transcriptional activity for single molecules spanning the Drosophila Bithorax complex (BX-C) locus(Mateo *et al.*, 2019), which can be found at the following repository: https://zenodo.org/records/4741214. We employed the data preprocessing procedure from Rajpurkar et al.(Rajpurkar *et al.*, 2021) to handle missing values, which can be found at the following repository:

https://github.com/aparna-arr/DeepLearningChromatinStructure/tree/master/DataPreprocessing.

Cells with over 80% of coordinates missing were excluded from our analysis. For the remaining cells, missing coordinates were imputed by linear interpolation between adjacent loci using the scipy.interpolate.interp1d function(Virtanen *et al.*, 2020). Maps were normalized by dividing by the maximum distance observed as followed prior to NMF.

Su. et al. dataset

The Su et al. dataset(Su *et al.*, 2020) comprises genome-wide chromatin folding data from single molecules in human IMR90 fibroblasts imaged using DNA FISH. Data can be downloaded from the following repository: https://zenodo.org/record/3928890. In particular, we analyze paired coordinate and transcription data in 'genomic-scale-with transcription and nuclear bodies.tsv'. Custom Python code was written to extract specific genomic regions from the raw dataset, transform coordinate data into distance matrices, and identify cells transcribing the HLCS (ENSG00000159267) gene for downstream analysis, and can be found in the provided

repo with an example of processing. Maps were considered with a 40k resolution. Cells with more than 25% missing coordinates within the 10 Mb HLCS genomic locus were excluded. Any remaining missing values were imputed by linear interpolation using numpy.interp(Harris *et al.*, 2020).

Non-negative matrix factorization (NMF) and ChromaFactor application

We applied non-negative matrix factorization (NMF) using the ChannelReducer wrapper in the Lucid NMF library(*lucid: A collection of infrastructure and tools for research in neural network interpretability*, no date) built on top of the scikit-learn implementation(Pedregosa *et al.*, 2011), which unravels the 3D input matrices into 2D vectors suitable for NMF. We set the number of components to k=20 to balance interpretability of templates and reconstruction error (**Supplementary Figure 4.1**). Default scikit-learn NMF parameters were used: NNDSVD initialization, a coordinate descent solver, Frobenius loss, tolerance of 1e-4, maximum 200 iterations, and an element-wise L2 regularization penalty. Additional code is provided to process both datasets and plot 2D distance matrices and 3D coordinates.

Random forest

We trained random forest classifiers using RandomForestClassifier in scikit-learn(Pedregosa *et al.*, 2011) to predict transcriptional activity from NMF component weights. Models were trained separately for each gene with binary on/off transcription labels. Balanced datasets were created for each gene with equal transcribing and non-transcribing cells. Data was split 70/30 into train and test sets. All other random forest parameters were left as scikit-learn defaults.

Genomic features annotations

Gene annotations and enhancer locations were derived from the original publications for each dataset. Enhancer and gene locations were provided from Mateo et al(Mateo *et al.*, 2019). Compartment annotations were used from Rao et al.(Rao *et al.*, 2014) (4DNFIHM89EGL). Directionality index and insulation tracks were produced from scoring code provided in Gunsalus et al(Gunsalus *et al.*, 2023). UMAP dimensionality reduction for visualization used the scikit-learn implementation with n_neighbors=5 and remaining default parameters.

Statistical Analysis

Differences in contact patterns and NMF component weights between transcribing and non-transcribing cells were evaluated using the non-parametric two-sided Mann-Whitney U statistical test. P-values less than 0.05 were considered statistically significant.

4.6 Figures

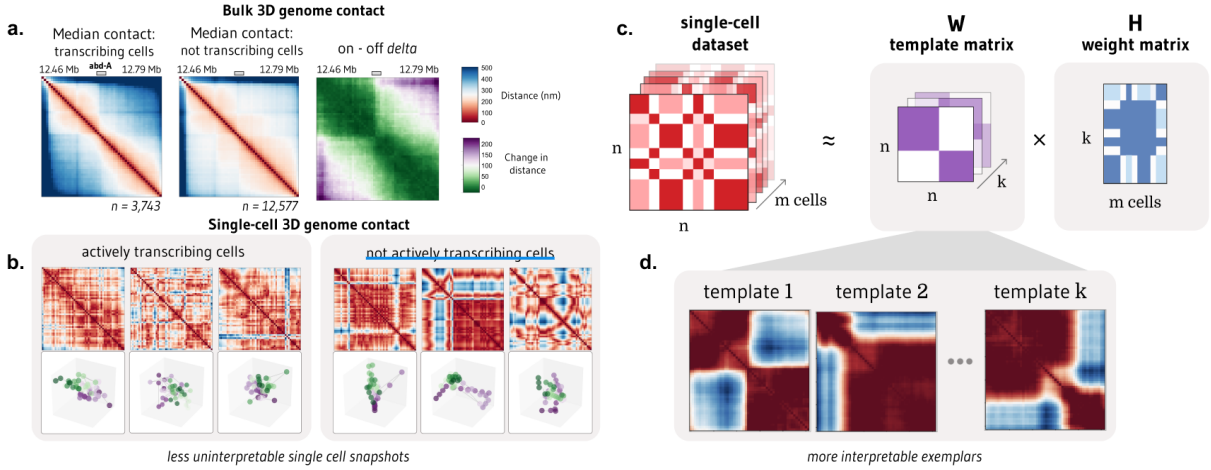


Figure 4.1. NMF provides interpretable decomposition of single-molecule chromatin conformation datasets.
a. Matrices representing the median all-by-all Euclidean difference (nm) between genomic loci in single molecules at a 30 kb locus in *Drosophila melanogaster* actively transcribing (left) and not transcribing (middle) the Abd-A gene from Mateo et al. (n = 16,320 molecules). The rightmost panel shows the difference in distance matrices, indicating two domains with elevated local interactions and reduced distal interactions in populations transcribing Abd-A. **b.** Bulk trends in contact change are challenging to observe in single cells actively transcribing (left) and not transcribing (right) Abd-A. **c.** Non-negative matrix factorization (NMF) decomposes a dataset of single-cell distance matrices into a *template* matrix with interpretable chromatin domain boundaries and a *contribution* matrix describing the weight of each template to each cell. **d.** Three templates produced when NMF is applied to distance matrices at the Abd-A locus.

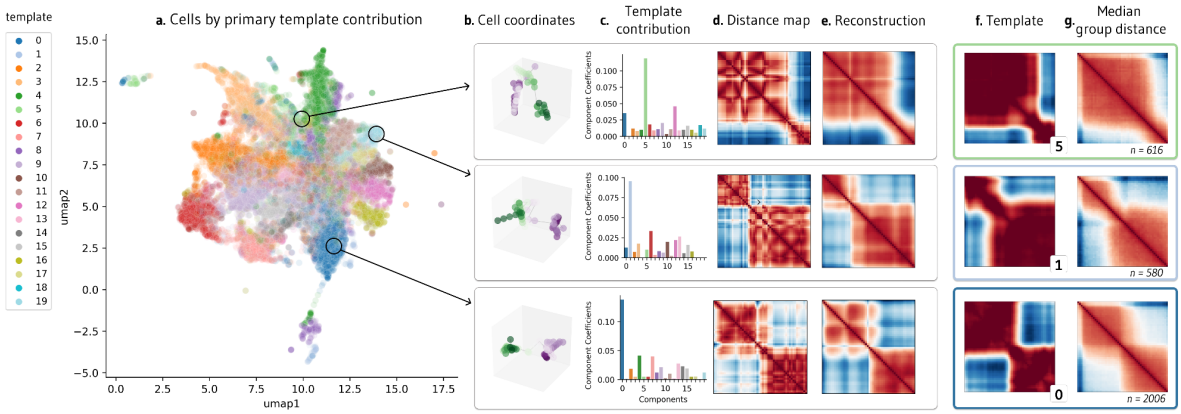


Figure 4.2. Visualization of NMF outputs and their relationship to single-cell behavior.
a. UMAP visualization of contribution matrix, colored by the template with the predominant contribution in each cell. **b.** Depiction of cell coordinates from selected individual cells. **c.** Component contributions for each cell, emphasizing high weight for templates 5, 1, and 0. **d.** Distance matrices corresponding to each cell**. e**. Denoised reconstructions of the distance matrices, created by multiplying the template contribution of a cell shown in (c) by the template matrix. **f.** NMF templates 5, 1, and 0, which had the highest weight contributions for the three individual cells in (c). **g**. Median contact distance across all cells in the dataset with the highest weight contributions to templates 5, 1, and 0, respectively.

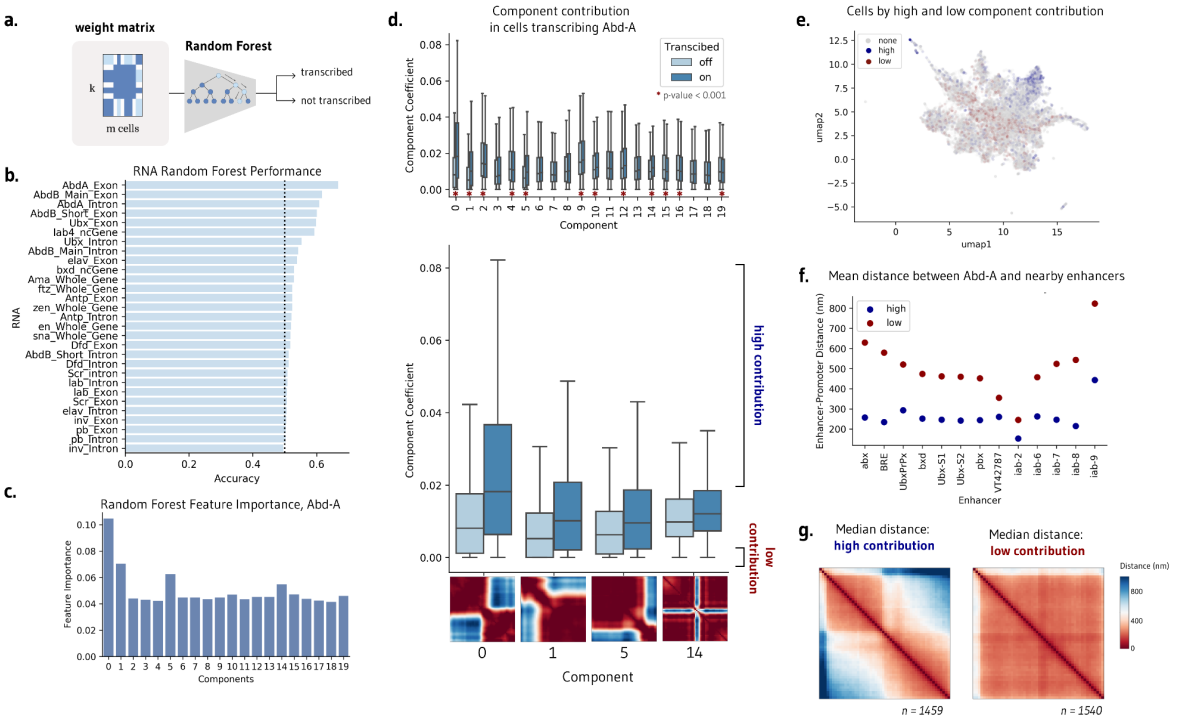


Figure 4.3. NMF templates are significantly correlated with transcription.
a. Application of random forest models to predict cell transcription from the contribution matrix alone. **b.** A random forest can modestly predict transcription in abd-A, Abd-B, and Ubx, demonstrating that the components capture salient information for transcription. **c.** Random forest feature importance highlights templates 0, 1, and 5 as most important for predicting transcription**. d**. Several components, including 0, 1, 5, and 14, have significantly different component contribution weights in transcribing and not transcribing cells. **e.** UMAP visualization of component contribution matrix, colored by cells with a high contribution of components 0, 1, 5, and 14 (blue) and a low contribution of these components (red). **f**. Mean distance between abd-A and nearby enhancers at the same locus across the subset of cells with high and low component contributions. **g**. Median contact of cells with high and low component contributions, encompassing the subset of cells which may be responsible for changes in contact observed in bulk.

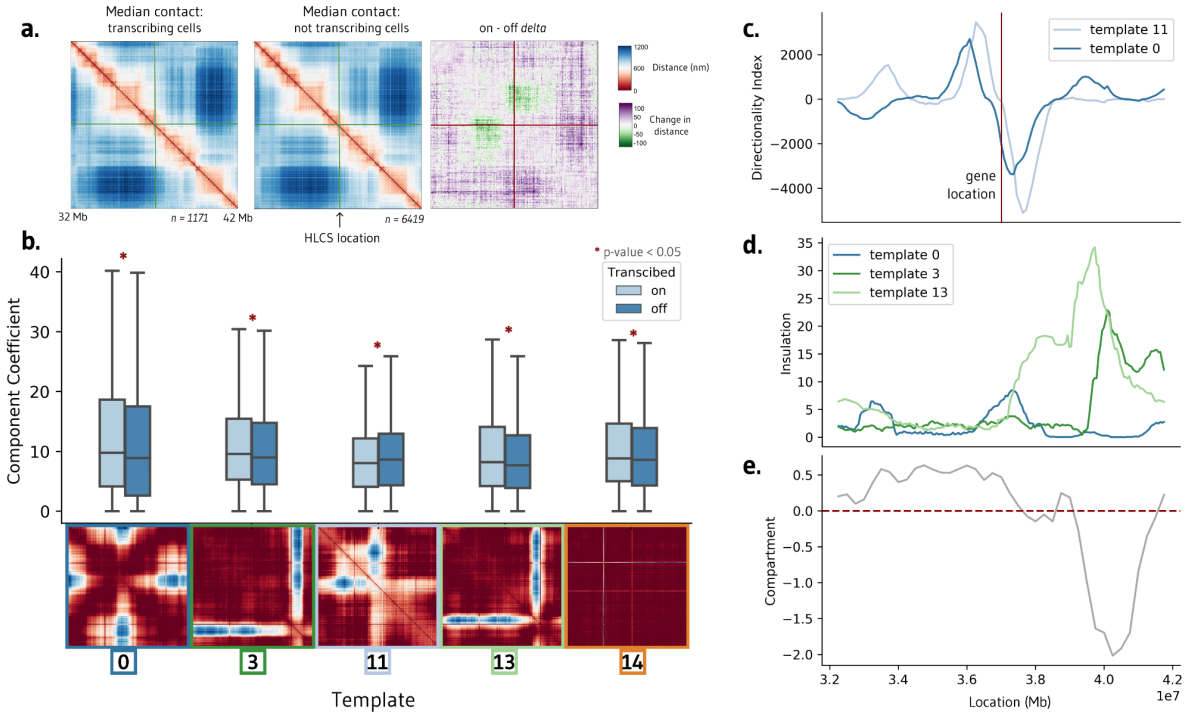
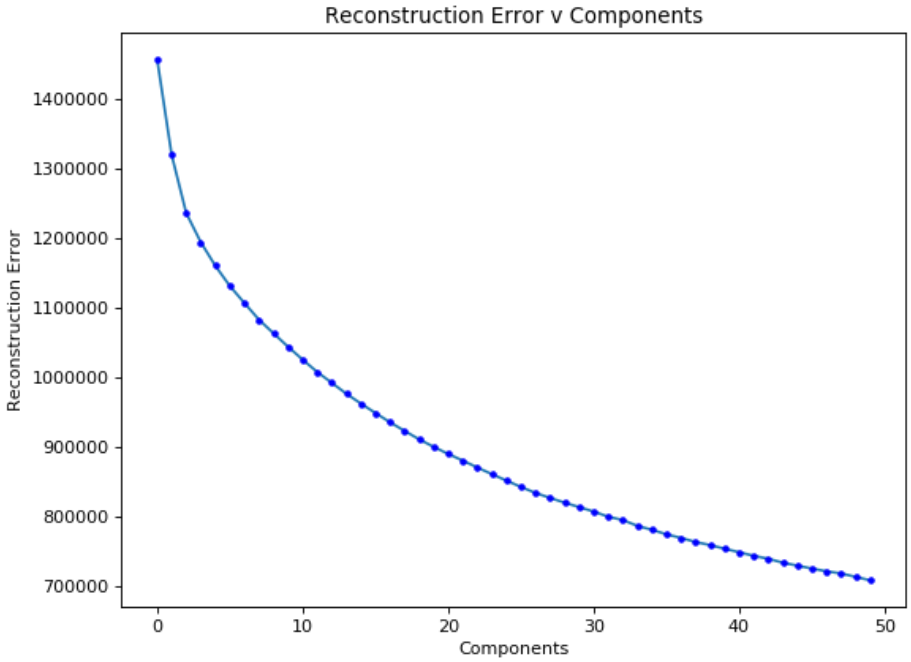
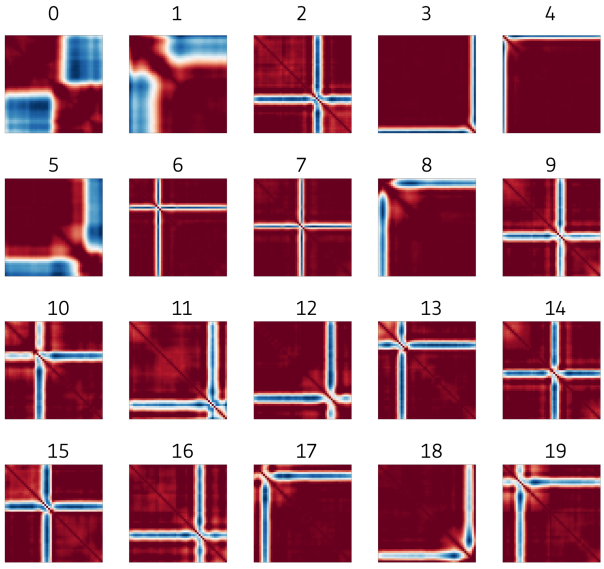


Figure 4.4. Interpretable templates at the HCLS locus in IMR-90 cells.
a. Average chromatin contacts in cells actively transcribing (left) and non-transcribing (middle) the HCLS gene within the surrounding 10 Mb region. The right panel highlights the contrast in contact patterns, emphasizing a stronger boundary in actively transcribing cells. **b.** Templates generated using Non-Negative Matrix Factorization (NMF) on this cell population. Among the 20 components, five exhibit significant differences between cells transcribing and not transcribing the HCLS gene. **c.** Directionality index of templates 0 and 11 correspond with the location of the transcribed HLCS gene. **d-e**. Insulation scores of templates 3 and 4 align with a shift in compartment in IMR-90 cells.

4.7 Supplementary Figures

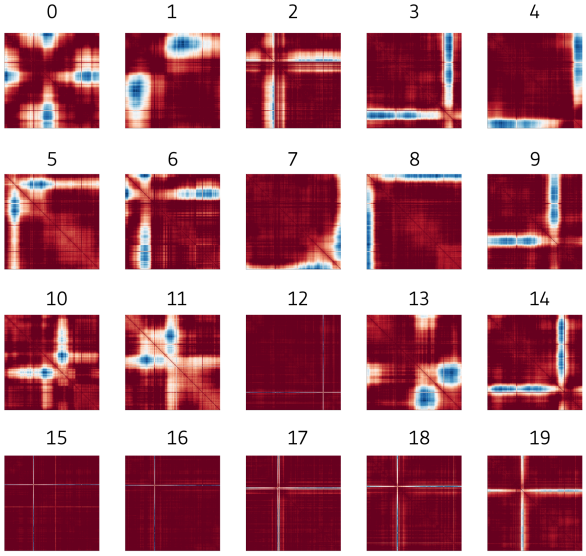


Supplemental Figure 4.1. Reconstruction error across number of components, k.
Reconstruction error, measured by the difference between the original single-cell chromatin folding dataset and the NMF reconstructed approximation, across different values of *k* components. Adding more components reduces the reconstruction error as NMF can better capture patterns in the data at the expense of interpretability. An elbow is visible around k=15-20 where adding additional components leads to diminishing returns in error reduction. We selected k=20 components for our analysis to balance reconstruction accuracy and interpretability.



Supplemental Figure 4.2. Components at BX-C locus.

All 20 components generated by applying NMF across the cells at this locus from the Mateo et al. dataset.



Supplemental Figure 4.3. Components at HLCS locus.

All 20 components generated by applying NMF across the cells at this locus from the Su et al. dataset.

5

Conclusions and Outlook

Conclusions and Future Outlook

This dissertation puts forth computational strategies to advance our understanding of the principles governing three-dimensional chromatin structure and their implications for gene regulation. In Chapter 2, I performed large-scale *in silico* mutagenesis to systematically identify DNA sequences that encode the folding of the human genome. Chapter 3 tackled challenges in quantitatively comparing differences between Hi-C maps by benchmarking methods using thousands of perturbations. Finally, Chapter 4 introduced a technique to deconvolve single-cell heterogeneity in chromatin folding and link cell subsets to average genomic patterns. Together, these aims provide new computational approaches to gain fundamental insights into the form and function of the dynamically folded genome. In this concluding chapter, I reflect on key lessons, limitations, and opportunities for future work related to the presented studies.

Broader Impacts and Future Directions

Stepping back, this dissertation makes contributions on a specific technical level as well as a broader conceptual level. The presented studies introduce novel computational techniques directly advancing capabilities in several domains: identifying DNA sequence elements governing folding, quantifying differences in chromatin maps, and unraveling single-cell heterogeneity. The particular methods put forth in each chapter address limitations of existing approaches to make progress on longstanding questions in decoding the grammar of genome topology.

However, the reach of these contributions also extends beyond the specific aims targeted. This work broadly highlights the potential of emerging computational methods, especially machine learning, to accelerate discovery in genomics. Powerful deep learning models can now predict diverse molecular phenotypes directly from raw DNA sequence, including transcription factor binding, gene expression, epigenetic state, and 3D chromatin structure (Kelley, Snoek and Rinn, 2016; Fudenberg, Kelley and Pollard, 2020; Avsec, Agarwal, *et al.*, 2021; Avsec, Weilert, *et al.*, 2021). As we demonstrate in Chapter 2, these models enable completely new experimental paradigms. Rather than costly and time-consuming *in vivo* assays, we can computationally test millions of *in silico* sequence perturbations at will. The creative application of machine learning for biological hypothesis generation is an area open for future development.

Looking forward, several promising directions emerge. Here I highlight a few key opportunities motivated by this dissertation at the frontiers of deciphering principles of genome folding, connecting chromatin architecture to function, and further leveraging the potential of machine learning.

Relating Heterogeneous Chromatin Structure to Gene Regulation in Single Cells

Chapter 4 developed an approach to link single-cell variability in chromatin folding to average principles and phenotypes. However, significant future work remains to unravel the complex relationship between 3D architecture and transcription at the single-cell level. One avenue is applying unsupervised techniques like graph neural networks to segment single-cell chromatin folding measurements and then correlate communities with expression states. Alternatively, multi-modal neural networks could explicitly predict gene regulation from heterogeneous chromatin structure annotations (Rajpurkar *et al.*, 2021).

Capturing paired measurements of conformation and transcription in individual cells is an ongoing experimental challenge. Emerging combinatorial barcoding schemes enable parallel capture of spatial transcriptomics and DNA contacts (Wang *et al.*, 2018). Expanding these approaches to additional modalities like enhancer contacts could provide ideal training data. We anticipate joint profiling and modeling will enable new traction in disentangling mechanisms linking form and function within single cells.

Inference of Causal Models from Population and Single-Cell Data

Critically, correlation does not imply causation. Observing links between chromatin folding and transcription provides circumstantial evidence but cannot conclusively determine that structure drives expression or vice versa. Recent work to infer causality from population chromatin data shows promise in nominating putative drivers of compartmentalization and domain formation (Gayoso *et al.*, 2023). Extending causal network models to single-cell measurements could help resolve longstanding questions about the sequence of events potentiating gene activation. Do enhancer-promoter contacts induce transcription, or does recruitment of polymerases stabilize looping? Do shifts in compartmentalization alter folding, or does altered folding drive compartment changes? To address these questions with computational techniques, we will require large-scale single-cell chromatin perturbation experiments as training data. Nonetheless, I am especially excited for the future of molecular biology technology development, which may be the only conclusive avenue to understand causality in chromatin biology. The ability to image chromatin in live cells over time may finally resolve long-standing unknowns about how enhancers and promoters interact to facilitate transcription.

Designing Novel Regulatory Sequences with Generative Models

Chapter 2 nominated DNA elements governing folding, but further work is needed to refine and expand this vocabulary. Recently, generative machine learning models have shown promise in designing novel sequences

optimized for desired molecular functions. These models learn patterns from training data and can sample new examples adhering to the implicit rules.

Applying similar generative strategies could enable designing synthetic DNA sequences to precisely shape chromatin architecture. For instance, variational autoencoders trained on TF binding motifs can generate new motifs with tuned strengths. Adapting this concept, we envision creating sequences optimized to recapitulate specific folding behavior. Reinforcement learning presents another avenue to iteratively refine sequences to match target conformations. Generative modeling could provide a rapid design cycle complementing in vivo testing.

In summary, future studies integrating diverse data modalities, capitalizing on single-cell approaches, and inferring causal mechanisms will help refine our understanding of the sequence encoding, dynamics, and consequences of higher-order genome topology. Complementary computational and experimental techniques should be combined iteratively to unravel multi-scale organizational principles. The work presented in this dissertation provides both specific methods and a conceptual template to advance the next stage of insights into the form, function, and variability underpinning the folded genome.

References

Akgol Oksuz, B. *et al.* (2021) 'Systematic evaluation of chromosome conformation capture assays', *Nature methods*, 18(9), pp. 1046–1055.

de Almeida, B.P. *et al.* (2022) 'DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers', *Nature genetics*, 54(5), pp. 613–624.

Amemiya, H.M., Kundaje, A. and Boyle, A.P. (2019) 'The ENCODE Blacklist: Identification of Problematic Regions of the Genome', *Scientific reports*, 9(1), p. 9354.

Avsec, Ž., Weilert, M., *et al.* (2021) 'Base-resolution models of transcription-factor binding reveal soft motif syntax', *Nature genetics*, 53(3), pp. 354–366.

Avsec, Ž., Agarwal, V., *et al.* (2021) 'Effective gene expression prediction from sequence by integrating long-range interactions', *Nature methods*, 18(10), pp. 1196–1203.

Ay, F. and Noble, W.S. (2015) 'Analysis methods for studying the 3D architecture of the genome', *Genome biology*, 16, p. 183.

Bailey, T.L. *et al.* (2015) 'The MEME Suite', *Nucleic acids research*, 43(W1), pp. W39–49.

Barutcu, A.R. *et al.* (2018) 'A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus', *Nature communications*, 9(1), p. 1444.

Bembom, O. (no date) *seqlogo: Sequence logos for DNA sequence alignments*, *R package version 1.48.0.* Available at: https://bioconductor.org/packages/release/bioc/html/seqLogo.html.

Bintu, B. *et al.* (2018) 'Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells', *Science*, 362(6413). Available at: https://doi.org/10.1126/science.aau1783.

Boninsegna, L. *et al.* (2022) 'Integrative genome modeling platform reveals essentiality of rare contact events in 3D genome organizations', *Nature methods*, 19(8), pp. 938–949.

Bourque, G. *et al.* (2008) 'Evolution of the mammalian transcription factor binding repertoire via transposable elements', *Genome research*, 18(11), pp. 1752–1762.

Buenrostro, J.D. *et al.* (2018) 'Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation', *Cell*, 173(6), pp. 1535–1548.e16.

Busslinger, G.A. *et al.* (2017) 'Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl', *Nature*, 544(7651), pp. 503–507.

Cao, Y. *et al.* (2019) 'Widespread roles of enhancer-like transposable elements in cell identity and long-range genomic interactions', *Genome research*, 29(1), pp. 40–52.

Cardozo Gizzi, A.M. *et al.* (2019) 'Microscopy-Based Chromosome Conformation Capture Enables Simultaneous Visualization of Genome Organization and Transcription in Intact Organisms', *Molecular cell*, 74(1), pp. 212–222.e5.

Castro-Mondragon, J.A. *et al.* (2022) 'JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles', *Nucleic acids research*, 50(D1), pp. D165–D173.

Chen, H. *et al.* (2012) 'Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome', *PloS one*, 7(7), p. e41374.

Choudhary, M.N. *et al.* (2020) 'Co-opted transposons help perpetuate conserved higher-order chromosomal structures', *Genome biology*, 21(1), p. 16.

Choudhary, M.N.K. *et al.* (2022) 'Widespread contribution of transposable elements to the rewiring of mammalian 3D genomes and gene regulation', *bioRxiv*. Available at: https://doi.org/10.1101/2022.02.01.475239.

Crane, E. *et al.* (2015) 'Condensin-driven remodelling of X chromosome topology during dosage compensation', *Nature*, 523(7559), pp. 240–244.

Cuartero, S. *et al.* (2018) 'Control of inducible gene expression links cohesin to hematopoietic progenitor self-renewal and differentiation', *Nature immunology*, 19(9), pp. 932–941.

Dekker, J. *et al.* (2002) 'Capturing Chromosome Conformation', *Science*, 295(5558), pp. 1306–1311.

Dekker, J. and Mirny, L. (2016) 'The 3D Genome as Moderator of Chromosomal Communication', *Cell*, 164(6), pp. 1110–1121.

Deng, W. *et al.* (2012) 'Controlling long-range genomic interactions at a native locus by targeted tethering of a

looping factor', *Cell*, 149(6), pp. 1233–1244.

Diehl, A.G., Ouyang, N. and Boyle, A.P. (2020) 'Transposable elements contribute to cell and species-specific chromatin looping and gene regulation in mammalian genomes', *Nature communications*, 11(1), p. 1796.

Dixon, J.R. *et al.* (2012) 'Topological domains in mammalian genomes identified by analysis of chromatin interactions', *Nature*, 485(7398), pp. 376–380.

Dixon, J.R. *et al.* (2015) 'Chromatin architecture reorganization during stem cell differentiation', *Nature*, 518(7539), pp. 331–336.

Dixon, J.R., Gorkin, D.U. and Ren, B. (2016) 'Chromatin Domains: The Unit of Chromosome Organization', *Molecular cell*, 62(5), pp. 668–680.

ENCODE Project Consortium *et al.* (2020) 'Expanded encyclopaedias of DNA elements in the human and mouse genomes', *Nature*, 583(7818), pp. 699–710.

Eres, I.E. *et al.* (2019) 'Reorganization of 3D genome structure may contribute to gene regulatory evolution in primates', *PLoS genetics*, 15(7), p. e1008278.

Ferrari, R. *et al.* (2020) 'TFIIIC Binding to Alu Elements Controls Gene Expression via Chromatin Looping and Histone Acetylation', *Molecular cell*, 77(3), pp. 475–487.e11.

Forcato, M. *et al.* (2017) 'Comparison of computational methods for Hi-C data analysis', *Nature methods*, 14(7), pp. 679–685.

Frankish, A. *et al.* (2021) 'GENCODE 2021', *Nucleic acids research*, 49(D1), pp. D916–D923.

Fudenberg, G. *et al.* (2016) 'Formation of Chromosomal Domains by Loop Extrusion', *Cell reports*, 15(9), pp. 2038–2049.

Fudenberg, G. *et al.* (2017) 'Emerging Evidence of Chromosome Folding by Loop Extrusion', *Cold Spring Harbor symposia on quantitative biology*, 82, pp. 45–55.

Fudenberg, G., Kelley, D.R. and Pollard, K.S. (2020) 'Predicting 3D genome folding from DNA sequence with Akita', *Nature methods*, 17(11), pp. 1111–1117.

Fudenberg, G. and Pollard, K.S. (2019) 'Chromatin features constrain structural variation across evolutionary

timescales', *Proceedings of the National Academy of Sciences of the United States of America*, 116(6), pp. 2175–2180.

Galan, S. *et al.* (2020) 'CHESS enables quantitative comparison of chromatin contact data and automatic feature extraction', *Nature genetics*, 52(11), pp. 1247–1255.

Gayoso, A. *et al.* (2023) 'Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells', *Nature methods* [Preprint]. Available at: https://doi.org/10.1038/s41592-023-01994-w.

Giorgetti, L. *et al.* (2016) 'Structural organization of the inactive X chromosome in the mouse', *Nature*, 535(7613), pp. 575–579.

Gong, Y. *et al.* (2018) 'Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries', *Nature communications*, 9(1), p. 542.

Gorkin, D.U. *et al.* (2019) 'Common DNA sequence variation influences 3-dimensional conformation of the human genome', *Genome biology*, 20(1), p. 255.

Gunsalus, L. *et al.* (2023) 'Comparing chromatin contact maps at scale: methods and insights', *Research square* [Preprint]. Available at: https://doi.org/10.21203/rs.3.rs-2842981/v1.

Guo, Y. *et al.* (2015) 'CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function', *Cell*, 162(4), pp. 900–910.

Hafner, A. *et al.* (2022) 'Loop stacking organizes genome folding from TADs to chromosomes', *bioRxiv*. Available at: https://doi.org/10.1101/2022.07.13.499982.

Hafner, A. and Boettiger, A. (2022) 'The spatial organization of transcriptional control', *Nature reviews. Genetics* [Preprint]. Available at: https://doi.org/10.1038/s41576-022-00526-0.

Hark, A.T. *et al.* (2000) 'CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus', *Nature*, 405(6785), pp. 486–489.

Harris, C.R. *et al.* (2020) 'Array programming with NumPy', *Nature*, 585(7825), pp. 357–362.

Hoencamp, C. *et al.* (2021) '3D genomics across the tree of life reveals condensin II as a determinant of architecture type', *Science*, 372(6545), pp. 984–989.

Hsieh, T.-H.S. *et al.* (2020) 'Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding', *Molecular cell*, 78(3), pp. 539–553.e8.

Hsieh, T.-H.S. *et al.* (2022) 'Enhancer-promoter interactions and transcription are largely maintained upon acute loss of CTCF, cohesin, WAPL or YY1', *Nature genetics*, 54(12), pp. 1919–1932.

Huda, A., Mariño-Ramírez, L. and Jordan, I.K. (2010) 'Epigenetic histone modifications of human transposable elements: genome defense versus exaptation', *Mobile DNA*, 1(1), p. 2.

Imakaev, M. *et al.* (2012) 'Iterative correction of Hi-C data reveals hallmarks of chromosome organization', *Nature methods*, 9(10), pp. 999–1003.

Jia, Z. *et al.* (2020) 'Tandem CTCF sites function as insulators to balance spatial chromatin contacts and topological enhancer-promoter selection', *Genome biology*, 21(1), p. 75.

Karimzadeh, M. *et al.* (2018) 'Umap and Bismap: quantifying genome and methylome mappability', *Nucleic acids research*, 46(20), p. e120.

Kelley, D.R. *et al.* (2018) 'Sequential regulatory activity prediction across chromosomes with convolutional neural networks', *Genome research*, 28(5), pp. 739–750.

Kelley, D.R., Snoek, J. and Rinn, J.L. (2016) 'Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks', *Genome research*, 26(7), pp. 990–999.

Kentepozidou, E. *et al.* (2020) 'Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains', *Genome biology*, 21(1), p. 5.

Kent, W.J. *et al.* (2002) 'The human genome browser at UCSC', *Genome research*, 12(6), pp. 996–1006.

Keough, K.C. *et al.* (2023) 'Three-dimensional genome rewiring in loci with human accelerated regions', *Science*, 380(6643), p. eabm1696.

Khan, A. (2021) *pyJASPAR: a Pythonic interface to JASPAR transcription factor motifs*. Available at: https://doi.org/10.5281/zenodo.4509415.

Kim, H.-J. *et al.* (2020) 'Capturing cell type-specific chromatin compartment patterns by applying topic modeling to single-cell Hi-C data', *PLoS computational biology*, 16(9), p. e1008173.

Kim, T.H. *et al.* (2007) 'Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome', *Cell*, 128(6), pp. 1231–1245.

Kotliar, D. *et al.* (2019) 'Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq', *eLife*, 8. Available at: https://doi.org/10.7554/eLife.43803.

Kraft, K. *et al.* (2019) 'Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations', *Nature cell biology*, 21(3), pp. 305–310.

Kragesteen, B.K. *et al.* (2018) 'Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis', *Nature genetics*, 50(10), pp. 1463–1473.

Krietenstein, N. *et al.* (2020) 'Ultrastructural Details of Mammalian Chromosome Architecture', *Molecular cell*, 78(3), pp. 554–565.e7.

Kubo, N. *et al.* (2021) 'Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation', *Nature structural & molecular biology*, 28(2), pp. 152–161.

Kunarso, G. *et al.* (2010) 'Transposable elements have rewired the core regulatory network of human embryonic stem cells', *Nature genetics*, 42(7), pp. 631–634.

Lee, D.D. and Seung, H.S. (1999) 'Learning the parts of objects by non-negative matrix factorization', *Nature*, 401(6755), pp. 788–791.

Lee, D.-I. and Roy, S. (2021) 'GRiNCH: simultaneous smoothing and detection of topological units of genome organization from sparse chromatin contact count matrices with matrix factorization', *Genome biology*, 22(1), p. 164.

Lieberman-Aiden, E. *et al.* (2009) 'Comprehensive mapping of long-range interactions reveals folding principles of the human genome', *Science*, 326(5950), pp. 289–293.

Liu, M. *et al.* (2020) 'Multiplexed imaging of nucleome architectures in single cells of mammalian tissue', *Nature communications*, 11(1), p. 2907.

lucid: A collection of infrastructure and tools for research in neural network interpretability (no date). Github. Available at: https://github.com/tensorflow/lucid (Accessed: 29 June 2023).

Lu, J.Y. *et al.* (2021) 'Homotypic clustering of L1 and B1/Alu repeats compartmentalizes the 3D genome', *Cell*

research, 31(6), pp. 613–630.

Lupiáñez, D.G. *et al.* (2015) 'Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions', *Cell*, 161(5), pp. 1012–1025.

Lyu, H., Liu, E. and Wu, Z. (2020) 'Comparison of normalization methods for Hi-C data', *BioTechniques*, 68(2), pp. 56–64.

Mateo, L.J. *et al.* (2019) 'Visualizing DNA folding and RNA in embryos at single-cell resolution', *Nature*, 568(7750), pp. 49–54.

McArthur, E. *et al.* (2022) 'Reconstructing the 3D genome organization of Neanderthals reveals that chromatin folding shaped phenotypic and sequence divergence', *bioRxiv*. Available at: https://doi.org/10.1101/2022.02.07.479462.

McArthur, E. and Capra, J.A. (2021) 'Topologically associating domain boundaries that are stable across diverse cell types are evolutionarily constrained and enriched for heritability', *American journal of human genetics*, 108(2), pp. 269–283.

Merkenschlager, M. and Nora, E.P. (2016) 'CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation', *Annual review of genomics and human genetics*, 17, pp. 17–43.

Miga, K.H. *et al.* (2014) 'Centromere reference models for human chromosomes X and Y satellite arrays', *Genome research*, 24(4), pp. 697–707.

Misteli, T. (2020) 'The Self-Organizing Genome: Principles of Genome Architecture and Function', *Cell*, 183(1), pp. 28–45.

Morgan, S.L. *et al.* (2017) 'Manipulation of nuclear architecture through CRISPR-mediated chromosomal looping', *Nature communications*, 8, p. 15993.

Nagano, T. *et al.* (2013) 'Single-cell Hi-C reveals cell-to-cell variability in chromosome structure', *Nature*, 502(7469), pp. 59–64.

Nagano, T. *et al.* (2017) 'Cell-cycle dynamics of chromosomal organization at single-cell resolution', *Nature*, 547(7661), pp. 61–67.

Naughton, C. *et al.* (2013) 'Transcription forms and remodels supercoiling domains unfolding large-scale

chromatin structures', *Nature structural & molecular biology*, 20(3), pp. 387–395.

Naumova, N. *et al.* (2013) 'Organization of the mitotic chromosome', *Science*, 342(6161), pp. 948–953.

Nichols, M.H. and Corces, V.G. (2021) 'Principles of 3D compartmentalization of the human genome', *Cell reports*, 35(13), p. 109330.

Nora, E.P. *et al.* (2012) 'Spatial partitioning of the regulatory landscape of the X-inactivation centre', *Nature*, 485(7398), pp. 381–385.

Nora, E.P. *et al.* (2017) 'Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization', *Cell*, 169(5), pp. 930–944.e22.

Nurk, S. *et al.* (2022) 'The complete sequence of a human genome', *Science*, 376(6588), pp. 44–53.

Open2C, Abdennur, N., Fudenberg, G., *et al.* (2022) 'Bioframe: Operations on Genomic Intervals in Pandas Dataframes', *bioRxiv*. Available at: https://doi.org/10.1101/2022.02.16.480748.

Open2C, Abdennur, N., Abraham, S., *et al.* (2022) 'Cooltools: enabling high-resolution Hi-C analysis in Python', *bioRxiv*. Available at: https://doi.org/10.1101/2022.10.31.514564.

Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine Learning in Python', *Journal of machine learning research: JMLR*, 12(85), pp. 2825–2830.

Rajpurkar, A.R. *et al.* (2021) 'Deep learning connects DNA traces to transcription to reveal predictive features beyond enhancer–promoter contact', *Nature communications*, 12(1), pp. 1–15.

Ramani, V. *et al.* (2017) 'Massively multiplex single-cell Hi-C', *Nature methods*, 14(3), pp. 263–266.

Rao, S.S.P. *et al.* (2014) 'A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping', *Cell*, 159(7), pp. 1665–1680.

Rao, S.S.P. *et al.* (2017) 'Cohesin Loss Eliminates All Loop Domains', *Cell*, 171(2), pp. 305–320.e24.

Raviram, R. *et al.* (2018) 'Analysis of 3D genomic interactions identifies candidate host genes that transposable elements potentially regulate', *Genome biology*, 19(1), p. 216.

Rege, M. *et al.* (2018) 'LADL: Light-activated dynamic looping for endogenous gene expression control', *bioRxiv*. Available at: https://doi.org/10.1101/349340.

Schmidt, D. *et al.* (2012) 'Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages', *Cell*, 148(1-2), pp. 335–348.

Schmitt, A.D. *et al.* (2016) 'A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome', *Cell reports*, 17(8), pp. 2042–2059.

Schwarzer, W. *et al.* (2017) 'Two independent modes of chromatin organization revealed by cohesin removal', *Nature*, 551(7678), pp. 51–56.

Schwessinger, R. *et al.* (2020) 'DeepC: predicting 3D genome folding using megabase-scale transfer learning', *Nature methods*, 17(11), pp. 1118–1124.

Sherman, M.D. (no date) *seqlogo: Python port of the R Bioconductor `seqLogo` package*. Github. Available at: https://github.com/betteridiot/seqlogo (Accessed: 9 May 2022).

Shrikumar, A., Greenside, P. and Kundaje, A. (2017) 'Learning Important Features Through Propagating Activation Differences', *arXiv [cs.CV]*. Available at: http://arxiv.org/abs/1704.02685.

Slotkin, R.K. and Martienssen, R. (2007) 'Transposable elements and the epigenetic regulation of the genome', *Nature reviews. Genetics*, 8(4), pp. 272–285.

Smit, AFA, Hubley, R & Green, P. (no date) *RepeatMasker Open-4.0*. Available at: http://www.repeatmasker.org (Accessed: 2013-2015).

Spielmann, M., Lupiáñez, D.G. and Mundlos, S. (2018) 'Structural variation in the 3D genome', *Nature reviews. Genetics*, 19(7), pp. 453–467.

Stansfield, J.C. *et al.* (2018) 'HiCcompare: an R-package for joint normalization and comparison of HI-C datasets', *BMC bioinformatics*, 19(1). Available at: https://doi.org/10.1186/s12859-018-2288-x.

van Steensel, B. and Furlong, E.E.M. (2019) 'The role of transcription in shaping the spatial organization of the genome', *Nature reviews. Molecular cell biology*, 20(6), pp. 327–337.

Su, J.-H. *et al.* (2020) 'Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin', *Cell* [Preprint]. Available at: https://doi.org/10.1016/j.cell.2020.07.032.

Su, M. *et al.* (2014) 'Evolution of Alu elements toward enhancers', *Cell reports*, 7(2), pp. 376–385.

Takei, Y. *et al.* (2021) 'Integrated spatial genomics reveals global architecture of single nuclei', *Nature*, 590(7845), pp. 344–350.

Tan, J. *et al.* (2022) 'Cell type-specific prediction of 3D chromatin architecture', *bioRxiv*. Available at: https://doi.org/10.1101/2022.03.05.483136.

Tan, J. *et al.* (2023) 'Cell-type-specific prediction of 3D chromatin organization enables high-throughput in silico genetic screening', *Nature biotechnology* [Preprint]. Available at: https://doi.org/10.1038/s41587-022-01612-8.

Tan, L. *et al.* (2018) 'Three-dimensional genome structures of single diploid human cells', *Science*, 361(6405), pp. 924–928.

Tan, L. *et al.* (2021) 'Changes in genome architecture and transcriptional dynamics progress independently of sensory experience during post-natal brain development', *Cell*, 184(3), pp. 741–758.e17.

Taskiran, I.I. *et al.* (2022) 'Cell type directed design of synthetic enhancers', *bioRxiv*. Available at: https://doi.org/10.1101/2022.07.26.501466.

Townes, F.W. and Engelhardt, B.E. (2023) 'Nonnegative spatial factorization applied to spatial genomics', *Nature methods*, 20(2), pp. 229–238.

Trigiante, G., Blanes Ruiz, N. and Cerase, A. (2021) 'Emerging roles of repetitive and repeat-containing RNA in nuclear and chromatin organization and gene expression', *Frontiers in cell and developmental biology*, 9, p. 735527.

Van Bortle, K. *et al.* (2014) 'Insulator function and topological domain border strength scale with architectural protein occupancy', *Genome biology*, 15(6), p. R82.

Vian, L. *et al.* (2018) 'The Energetics and Physiological Impact of Cohesin Extrusion', *Cell*, 175(1), pp. 292–294.

Virtanen, P. *et al.* (2020) 'SciPy 1.0: fundamental algorithms for scientific computing in Python', *Nature methods*, 17(3), pp. 261–272.

Wang, J. *et al.* (2015) 'MIR retrotransposon sequences provide insulators to the human genome', *Proceedings of*

the National Academy of Sciences of the United States of America, 112(32), pp. E4428–37.

Wang, T. *et al.* (2007) 'Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53', *Proceedings of the National Academy of Sciences of the United States of America*, 104(47), pp. 18613–18618.

Wang, X. *et al.* (2018) 'Three-dimensional intact-tissue sequencing of single-cell transcriptional states', *Science*, 361(6400). Available at: https://doi.org/10.1126/science.aat5691.

de Wit, E. *et al.* (2015) 'CTCF Binding Polarity Determines Chromatin Looping', *Molecular cell*, 60(4), pp. 676–684.

Xiong, K., Zhang, R. and Ma, J. (2023) 'scGHOST: Identifying single-cell 3D genome subcompartments', *bioRxiv : the preprint server for biology* [Preprint]. Available at: https://doi.org/10.1101/2023.05.24.542032.

Xu, Z. *et al.* (2016) 'A hidden Markov random field-based Bayesian method for the detection of long-range chromosomal interactions in Hi-C data', *Bioinformatics* , 32(5), pp. 650–656.

Yang, D., Chung, T. and Kim, D. (2022) 'DeepLUCIA: predicting tissue-specific chromatin loops using Deep Learning-based Universal Chromatin Interaction Annotator', *Bioinformatics* , 38(14), pp. 3501–3512.

Yang, M. and Ma, J. (2022) 'Machine Learning Methods for Exploring Sequence Determinants of 3D Genome Organization', *Journal of molecular biology*, 434(15), p. 167666.

Yang, R. *et al.* (2021) 'Epiphany: predicting Hi-C contact maps from 1D epigenomic signals', *bioRxiv*. Available at: https://doi.org/10.1101/2021.12.02.470663.

Yang, R. *et al.* (2021) 'Epiphany: predicting hi-c contact maps from 1d epigenomic signals', *bioRxiv* [Preprint]. Available at: https://www.biorxiv.org/content/10.1101/2021.12.02.470663.abstract.

Yang, T. *et al.* (2017) 'HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient', *Genome research*, 27(11), pp. 1939–1949.

Yang, Y. *et al.* (2022) 'Concert: Genome-wide prediction of sequence elements that modulate DNA replication timing', *bioRxiv*. Available at: https://doi.org/10.1101/2022.04.21.488684.

Yan, K.-K. *et al.* (2017) 'HiC-spector: a matrix library for spectral and reproducibility analysis of Hi-C contact maps', *Bioinformatics* , 33(14), pp. 2199–2201.

Yardımcı, G.G. *et al.* (2019) 'Measuring the reproducibility and quality of Hi-C data', *Genome biology*, 20(1), p. 57.

Yu, M. *et al.* (2021) 'SnapHiC: a computational pipeline to identify chromatin loops from single-cell Hi-C data', *Nature methods*, 18(9), pp. 1056–1059.

Zhang, R., Zhou, T. and Ma, J. (2022a) 'Multiscale and integrative single-cell Hi-C analysis with Higashi', *Nature biotechnology*, 40(2), pp. 254–261.

Zhang, R., Zhou, T. and Ma, J. (2022b) 'Ultrafast and interpretable single-cell 3D genome analysis with Fast-Higashi', *Cell systems*, 13(10), pp. 798–807.e6.

Zhang, S. *et al.* (2021) 'RNA polymerase II is required for spatial chromatin reorganization following exit from mitosis', *Science advances*, 7(43), p. eabg8205.

Zhang, Y. *et al.* (2019) 'Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells', *Nature genetics*, 51(9), pp. 1380–1388.

Zhan, Y. *et al.* (2023) 'Conformational analysis of chromosome structures reveals vital role of chromosome morphology in gene function', *bioRxiv*. Available at: https://doi.org/10.1101/2023.02.18.528138.

Zheng, Y., Shen, S. and Keleş, S. (2022) 'Normalization and de-noising of single-cell Hi-C data with BandNorm and scVI-3D', *Genome biology*, 23(1), p. 222.

Zhou, J. *et al.* (2019) 'Robust single-cell Hi-C clustering by convolution- and random-walk–based imputation', *Proceedings of the National Academy of Sciences*, 116(28), pp. 14011–14018.

Zhou, J. (2021) 'Sequence-based modeling of genome 3D architecture from kilobase to chromosome-scale', *bioRxiv*. Available at: https://doi.org/10.1101/2021.05.19.444847.

Zhou, J. (2022) 'Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale', *Nature Genetics*, pp. 725–734. Available at: https://doi.org/10.1038/s41588-022-01065-4.

Zhou, T., Zhang, R. and Ma, J. (2021) 'The 3D Genome Structure of Single Cells', *Annual review of biomedical data science*, 4, pp. 21–41.

Zufferey, M. *et al.* (2018) 'Comparison of computational methods for the identification of topologically associating domains', *Genome biology*, 19(1), p. 217.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Laura Gunsalus
5C3CBB38457E446...
Author Signature

12/13/2023
Date