HARVARD
Kenneth C. Griffin

GRADUATE SCHOOL
OF ARTS AND SCIENCES

THESIS ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Biostatistics

have examined a dissertation entitled

"Statistical methods for transcription factor footprinting in 3D genome assays"

presented by Corriene E. Sept

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature Martin Aryee
Martin Aryee (May 1, 2025 13:33 EDT)

Typed name: Prof. Martin Aryee

Signature Anders Sejr Hansen
Anders Sejr Hansen (May 2, 2025 18:43 EDT)

Typed name: Prof. Anders Sejr Hansen

Signature Rafael Irizarry
Rafael Irizarry (May 4, 2025 12:07 EDT)

Typed name: Prof. Rafael Irizarry

Signature _____

Typed name:

Signature _____

Typed name:

Date: April 30, 2025

Statistical methods for transcription factor footprinting in 3D genome assays

A DISSERTATION PRESENTED

BY

CORRIENE E. SEPT

TO

THE DEPARTMENT OF BIOSTATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN THE SUBJECT OF

BIOSTATISTICS

HARVARD UNIVERSITY

CAMBRIDGE, MASSACHUSETTS

APRIL 2025

© 2025 CORRIENE E. SEPT
ALL RIGHTS RESERVED.

Statistical methods for transcription factor footprinting in 3D genome assays

ABSTRACT

Chromatin loops are drivers of gene regulation, bringing distal enhancers into close proximity with their target genes. While a subset of these long-range contacts are mediated by the architectural factors CTCF and cohesin, the mechanisms underlying the majority of these regulatory contacts remains unclear. While transcription factors and other proteins are known to be contributors to DNA looping, no methods currently exist for simultaneously profiling 3D genome structure and the DNA-binding proteins involved. This presents a significant limitation for evaluating transcription factor impact on 3D contacts and constructing high-resolution protein occupancy maps of regulatory regions. This dissertation seeks to bridge this gap by (1) developing statistical methods that quantitatively assess DNA-binding protein occupancy in 3D genome assays and (2) leveraging these new methods to (i) investigate the relationship between protein-binding and genome architecture and (ii) provide high-resolution maps of DNA-binding proteins at enhancers and promoters.

Chapter 1 proposes a statistical method that determines CTCF binding in CTCF MNase HiChIP assays. MNase, beyond its use in profiling genome structure, has also been used to determine transcription factor and nucleosome occupancy at high-resolution in assays like CUT&RUN and MNase-seq due to its endo-exonuclease activity, where it cuts regions unprotected by proteins and chews back the fragments until it reaches protein-protected DNA.

This enables inference of both protein size and location, since nucleosomes protect more than twice the DNA of a transcription factor, thus yielding DNA fragments of significantly longer length. We leverage short, transcription factor (TF) protected fragments to pinpoint locations of CTCF binding at base-pair resolution. We then use TF-protected fragments at CTCF binding sites to implement a novel fragment-level view of CTCF-mediated chromatin looping dynamics. With this approach, we determine that fully extruded chromatin loops between convergent CTCF-bound sites are rare genome-wide and that, in addition to CTCF, active regulatory elements hinder cohesin-mediated loop extrusion. This supports a model by which the partially extruded chromatin loop can enable distal enhancer-promoter contacts.

Chapter 2 expands upon the method developed in Chapter 1 by broadly mapping locations of DNA-binding transcription factors in Micro-C. Unlike Chapter 1, Chapter 2 does not rely on a ChIP step to profile one protein's occupancy. Not being limited to just investigating one protein, CTCF, facilitates additional novel insights into protein occupancy and their relation to 3D contacts and transcription. We find that expression level is tightly linked with the size of the nucleosome depleted region at the TSS and the presence of a large TF complex at the promoter, immediately upstream of the TSS, with unexpressed genes exhibiting a TSS obstructed by a nucleosome and a lack of TF binding. Furthermore, the TF-sized proteins upstream of the TSS at expressed genes appear to facilitate long-range, cohesin-independent looping contacts, which may partially explain why gene expression is largely maintained when cohesin is degraded. Further investigation into whether specific TFs may enable these long-range cohesin-independent contacts identified transcription factor motif families such as the KLF/SP and NF-Y motif families as likely candidates for cohesin-independent looping factors.

Chapter 3 applies the approach developed in Chapter 2 to gain a high-resolution view of the *MYC* oncogene and its distal cell-type specific enhancers. This analysis identifies *MYC* distal enhancers as regions highly occupied by TF-enhancer assemblies, which depend on RNA for their coalescence. This chapter reveals the cooperation between TF binding and RNA required for gene regulation and genome structure, and presents a framework for developing high-resolution maps of regulatory architecture.

Contents

Title page i
Copyright ii
Abstract iii
Table of Contents vi
Author List viii
List of Figures ix
List of Tables xi
Acknowledgements xii
Dedication xiv

1 High-resolution CTCF footprinting reveals impact of chromatin state on cohesin extrusion

1.1 Introduction 1
1.2 Results 4
1.3 Discussion 21
1.4 Methods 25
1.5 Data availability 37
1.6 Code availability 38
1.7 Acknowledgements 38
1.8 Author contributions 38

2 DNA-binding protein footprints reveal cohesin independent long-range looping factors

2.1 Introduction 39
2.2 Results 42
2.3 Discussion 54

3 Regulatory architecture of transcriptional assemblies at nanoscale resolution

3.1 Introduction 57

3.2 Results 59

3.3 Discussion 69

3.4 Acknowledgements 72

3.5 Author contributions 73

3.6 Data and materials availability 73

Appendix A Supplemental material for Chapter 1 74

Appendix B Supplemental material for Chapter 2 85

Appendix C Supplemental material for Chapter 3 89

References 115

Author List

The following authors contributed to chapters in this dissertation:

- Chapter 1: Y. Esther Tak, Viraat Goel, Mital S. Bhakta, Christian G. Cerda-Smith, Haley M. Hutchinson, Marco Blanchette, Christine E. Eyler, Sarah E. Johnstone, J. Keith Joung, Anders S. Hansen & Martin J. Aryee

- Chapter 2: Kalon Overholt, Adrienne Vancura, Viraat Goel, Sarah Johnstone, Anders S. Hansen, Richard A. Young, & Martin J. Aryee

- Chapter 3: Kalon J. Overholt, Adrienne Vancura, Daniel Stoyko, Ido Sagi, Viraat Y. Goel, Susana W. Hawken, Aleksandra Galitsyna, David R. Pease, Shannon Moreno, Alejandra Ramos, Alfonso Gomez-Gonzalez, Hailey Erb, Emily Plyler, Tong Ihn Lee, John Easton, Brian J. Abraham, Anders S. Hansen, Martin J. Aryee, Richard A. Young

List of Figures

1.1 MNase CTCF HiChIP data contains short (~ <80 bp) CTCF-protected fragments and longer (~ >120 bp) nucleosome-protected fragments 6
1.2 True CTCF binding sites have a bimodal strand-specific distribution centered on the CTCF motif 8
1.3 CTCF binding sites identified by *CAMEL* with single basepair resolution in MNase K562 CTCF HiChIP data 11
1.4 Cohesin footprint observed in MNase HiChIP and experimentally confirmed in mESC RCMC 13
1.5 Cohesin extrudes further through quiescent regions than active regions 17
1.6 Schematic of supported model whereby single promoter-proximal CTCF sites enable an enrichment of enhancer-promoter contacts 24

2.1 Promoter-proximal transcription factors enable long-range cohesin-independent contacts at the TSS of expressed genes 45

2.2 Footprinting transcription factors in MicroC enables base-pair resolution of locus-specific transcription factor binding and aggregate examination of protein behavior 49

2.3 Cohesin independent looping factors maintain enhancer-promoter contacts upon cohesin depletion 51

3.1 Regulatory elements dominate the assembly architecture of the 2.8 Mb *MYC* locus 60

3.2 TF footprinting reveals TF nano-assemblies in regulatory elements at base-pair resolution 63

3.3 RNA-associated nano-assemblies underlie regulatory architecture 66

3.4 Biopolymer associations in transcriptional assemblies 70

A.1 Comparison of K562 CTCF MNase HiChIP to K562 Intact Hi-C 74

A.2 Percent observed ligations increases with read length 75

A.3 Comparison of K562 CTCF MNase HiChIP replicates 76

A.4 Companion figure to Figure 1.4: CTCF and cohesin footprints 77

A.5 Filtering to short fragments enables identification of CTCF-mediated contacts 78

A.6 Companion figure to Figure 1.5: Cohesin extrudes further through quiescent regions than active regions 80

A.7 mESC *Fbn2* CBS:RNAPII contact lost upon depleting cohesin 82

A.8 The probability of observing a high CAMEL statistic under the null hypothesis is higher at low read counts 83

A.9 Nucleosomes are preferentially positioned inside the loop 84

B.1 Fragment length density histogram (chr1 fragments) 85

B.2 Increased expression is associated with a larger NDR and TF-sized footprint enrichment at the TSS.....86

B.3 High density of short fragments reveals transcription factor binding events.....87

B.4 Evaluating significance of short fragments at transcription factor motifs 88

C.1 Region Capture Micro-C (RCMC) reveals the 3D chromatin structure of the 2.8 Mb MYC TAD.....102

C.2 Microcompartment domains in the 2.8 Mb MYC TAD. 103

C.3 Regulatory elements preferentially engage in long-range interactions.. 104

C.4 Transcription factor footprinting and benchmarking against DNase I and ATAC-seq footprinting.. 105

C.5 ChIP-seq data and DNA binding motifs support TF assignments 107

C.6 TF footprints at MYC enhancers and MYC promoter 108

C.7 TF ChIP-seq signal density differs inside and outside of enhancer-associated RCMC loop regions.....109

C.8 Nano-assembly model of an enhancer.....110

C.9 Nano-assemblies facilitate interpretation of disease-associated genetic variation.....111

C.10 RNA-dependent features of chromatin architecture and TF assemblies.....113

List of Tables

C.1 RCMC loop regions assigned to enhancer and promoter elements defined by ENCODE .. 114
C.2 Footprint assignment using TF DNA-binding motifs alone 114
C.3 Nascent RNA expression at studied enhancers 114
C.4 TFs with RNA-binding protein characteristics 114

Acknowledgements

They say it's more important to pick an advisor based on personality and mentoring style than scientific topic, but I was fortunate enough to have a dissertation advisor who has both been fantastic to work with and also guided me towards dissertation work that I've found both interesting and meaningful. Thank you Martin for your unfailing optimism about solving problems I once thought were unsolvable, and your critical thinking skills that have pushed me to be a better scientist and writer throughout my PhD. Beyond your scientific abilities, I also appreciate how you always encourage taking vacations and not working all the time – although this advice has become increasingly hard to follow as I near my dissertation defense date.

I am also very thankful for the guidance and support provided by my committee members, Anders Hansen and Rafael Irizarry. Anders, your encyclopedia-like knowledge of all papers in and adjacent to the 3D genomics field is truly inspiring. You always know what new interesting directions to look in and your feedback on my projects has undoubtedly made them more scientifically rigorous and interesting. Rafa – thank you for challenging me to improve my statistical methods development, my papers have without a doubt become more statistically rigorous because of your feedback. Thank you also for all the time and energy you put into improving PhD student lives and working conditions as the Dana Farber Data Science chair – in particular the dedication to keeping the students in offices with windows and the snack cabinet full. It does not go unnoticed how much you care about the students here.

To my cohort members, with whom I've enjoyed commiserating over the years – our zoom calls and virtual game nights the first year of grad school, when we were completely online due to the Covid-19 pandemic, kept me sane during a period of great uncertainty. In particular, to my roommates Max, Lin, and Ellen – I've enjoyed living together in our little apartment for all of grad school. I've been lucky to make such great friends and roommates, and living with 1/3 of the cohort in one apartment has been great fun. Jenna – our New Hampshire ski expeditions and shared enthusiasm for books and Taylor Swift have been a joy over the course of the PhD – I am fortunate to be your friend.

To my siblings – we didn't always get along growing up, but you all have become some of my best friends in our old age. Family gatherings would be far more boring without you. To my parents - thank you for relentlessly supporting my PhD despite my poor attempts to explain what exactly I do and why anyone should care about it. Your continued optimism that everything will work out fine has been a steadying force. To my partner Jakob – our little vacations, late-night chats, and 5+ daily phone calls every time I have a thought have kept me sane even during the most tumultuous, stressful points of the PhD. Thanks for always being willing to listen to me complain even about the most trivial things, and for steadying me with your kindness and positivity.

Lastly, I would like to thank Trader Joe's cold brew concentrate for getting me through this final year of the PhD. I look forward to drinking less coffee and sleeping more in the (hopefully) not too distant future.

Dedicated to my family.

Chapter 1

High-resolution CTCF footprinting reveals impact of chromatin state on cohesin extrusion

Abstract

Cohesin-mediated DNA loop extrusion enables gene regulation by distal enhancers through the establishment of chromosome structure and long-range enhancer-promoter interactions. The best characterized cohesin-related structures, such as topologically associating domains (TADs) anchored at convergent CTCF binding sites, represent static conformations. Consequently, loop extrusion dynamics remain poorly understood. To better characterize static and dynamically extruding chromatin loop structures, we use MNase-based 3D genome assays to simultaneously determine CTCF and cohesin localization as well as the 3D contacts they mediate. Here we present CTCF Analyzer (with) Multinomial Estimation (*CAMEL)*, a tool that identifies CTCF footprints at near base-pair resolution in CTCF MNase HiChiP. We also use Region Capture Micro-C to identify a CTCF-adjacent footprint that is attributed to cohesin occupancy. We leverage this substantial advance in resolution to determine that the fully extruded (CTCF-CTCF loop) state is rare genome-wide with locus-specific variation from ~1-10%. We further investigate the impact of chromatin state on loop extrusion dynamics and find that active regulatory elements impede cohesin extrusion. These findings support a model of topological regulation whereby the transient, partially extruded state facilitates enhancer-promoter contacts that can regulate transcription.

1.1 Introduction

The three-dimensional organization of the genome within the nucleus is a key regulator of gene expression. The cohesin complex, through its loop extrusion ability that brings distant genomic regions into close physical proximity, is a key determinant of this spatial arrangement[1]. Cohesin-mediated loop extrusion is a modulator of transcriptional programs through its ability to both facilitate and block interactions between enhancers and promoters. As a consequence, disruptions to cohesin-mediated loop extrusion can lead to misregulation of gene expression and are implicated in various genetic diseases[2–5].

CCCTC-binding factor (CTCF) can act as an extrusion barrier through its ability to bind and stabilize cohesin on DNA[6], serving to preferentially localize and anchor one or both ends of cohesin loops. Through its interaction with cohesin, CTCF can act to promote or inhibit gene expression in a context dependent manner. For example, many topologically associated domains (TADs) are bounded by convergently oriented CTCF sites and insulate the genes contained within from activation by enhancers located outside the TAD[7–9]. On the other hand, promoter-proximal CTCF binding sites can enable gene activation by distal enhancers[10–12], although the mechanisms involved are not well understood.

While the best understood cohesin-related structures, including TADs and enhancer-promoter loops, represent static conformations, loop extrusion has also been shown to be a highly dynamic process with an extrusion rate of ~1kb/s[13]. Recent live cell-imaging studies of loop extrusion

dynamics at two TADs bounded by convergent CTCF sites found that only a small fraction (~3-30%) of time is spent in the static "fully extruded" state where the loop is anchored by CTCF at both ends[14,15]. It remains unclear whether these two TADs are representative of loop extrusion dynamics across the human genome, what chromatin factors influence extrusion, and what role these actively extruding structures play in transcription.

To explore the genome-wide dynamic behavior of CTCF-anchored loops, we developed an analysis approach for 3D chromatin proximity ligation assays that enables simultaneous assessment of 3D contacts and fragment-level identification of CTCF/cohesin occupancy. We chose to use MNase digestion-based assays (MNase HiChIP[16–19], Region Capture Micro-C[12,20–22]) as the endo-exonuclease activity of MNase allows "footprinting" the location of bound factors[23–27]. We show that by analyzing proximity ligation events involving CTCF binding site locations, it is possible to use precise fragment size characteristics to determine at a single molecule level whether a given DNA molecule was bound by a nucleosome or CTCF/cohesin. Analysis of the CTCF/cohesin bound molecules then enables us to quantify the fully extruded (CTCF-CTCF loop) frequency at individual CTCF sites and find that it varies from ~1-10%. Further, we show that extruded loop length differs substantially depending on chromatin state, ranging from ~140 kb in active chromatin regions to ~250 kb in quiescent regions. These data advance our ability to quantify loop dynamics from topological data and support a role for partially extruded loops in gene regulation.

1.2 Results

MNase Digestion Fragment Length Distinguishes CTCF and nucleosome bound DNA

We used Micrococcal nuclease (MNase) HiChIP[16–19] for CTCF to profile 3D architecture in K562 cells, generating 150 bp reads with over 380 million unique pairwise contacts across four replicates. Briefly, following cell fixation with DSG and formaldehyde, chromatin is digested by MNase, immunoprecipitated to enrich for CTCF-bound DNA, and free ends are then ligated. After reverse-crosslinking, the resulting ligation products are sequenced from both ends and the mapping locations of the paired reads can be used to infer chromosomal locations of the physically interacting loci. We confirmed that the majority (70%) of loops identified by FitHiChIP[28] overlap those identified in a publicly available, high-coverage K562 Intact Hi-C data-set[29] (1.2 billion paired contacts) (see Methods, Figure A.1).

We explored whether the MNase HiChIP assay could simultaneously be used to infer 3D contacts and the identity of proteins bound at interacting loci. MNase has previously been used to identify "footprints" of bound factors[23–26], as it has both endonuclease activity that selectively cleaves naked DNA not shielded by bound proteins, and exonuclease activity that subsequently trims back the remaining unprotected DNA fragment ends. We sought to estimate the protected DNA fragment length, as we hypothesized that this length would let us infer characteristics of a bound protein. This is possible in cases where an MNase digestion fragment is shorter than the read length such that the read comprises two separate segments and the ligation junction is directly observed (Figure 1.1A, Figure A.2). Exact fragment lengths could thus be determined for

fragments with length less than 150 bp as the 150 bp read length results in censoring of longer fragments.

As expected, due to the high abundance of histones in chromatin (Figure 1.1B), the predominant fragment length is approximately 150 bp or longer, suggestive of cuts between nucleosomes[30] (Figure 1.1C). We also noted a distribution of shorter fragment lengths, with 20% of fragments representing lengths shorter than 120 bp (Figure 1.1D). Restricting to fragments that overlap CTCF motifs shows a strong enrichment of these short fragments (Figure 1.1C), with <80 bp fragments having a 10-fold higher overlap frequency with CTCF motifs than >120 bp fragments (Figure 1.1D).

A fragment pileup metaplot centered on CTCF motif loci (Figure 1.1E) shows a strong enrichment of short fragments centered on the CTCF motif sequence, and a concomitant depletion of long fragments over the motif (Figure 1.1F). Long fragments, in contrast, show peaks with a strong ~200 bp periodicity adjacent to the CTCF binding site (Figure 1.1F), consistent with the ability of CTCF to precisely position a series of nucleosomes adjacent to its binding site[31]. Note that while long (>120 bp) fragments are depleted at CTCF binding sites, they still represent a significant fraction of reads at these sites (Figure 1.1C, F). This likely reflects that CTCF motif loci without a bound CTCF are frequently instead occupied by histones[31,32], and even CTCF motifs with very strong CTCF ChIP-seq signal show only partial occupancy by CTCF across a cell population[33].

In summary, long (~150bp) fragments likely correspond to nucleosome-protected DNA whereas shorter fragments at CTCF binding sites likely arise from CTCF-protected DNA. This can be explained by the different sizes of CTCF and histone octamers, which translate into different lengths of DNA protected from MNase exonuclease activity[31].

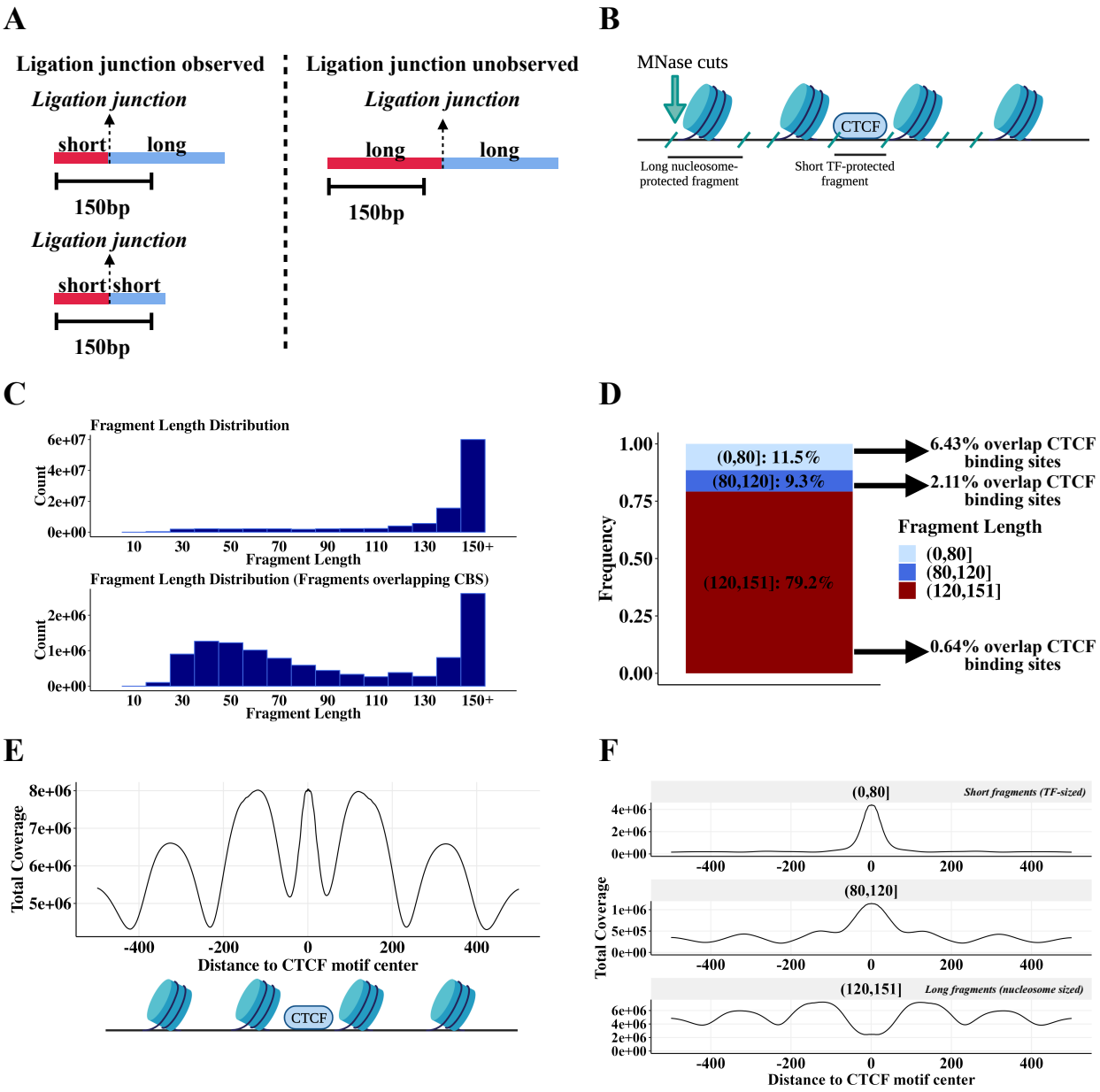


Figure 1.1 MNase CTCF HiChIP data contains short (~ <80 bp) CTCF-protected fragments and longer (~ >120 bp) nucleosome-protected fragments. **A** Schematic illustrating relationship

between short fragments and observed ligations. **B** Schematic illustrating how the fragment length results from MNase cutting around bound proteins of different sizes. Created in BioRender. Aryee, M. (2025) https://BioRender.com/g98u240 **C** Fragment length distribution for all fragments (top plot) and fragments overlapping occupied CTCF motifs (lower plot). Occupied CTCF motifs are defined here as CTCF motifs within 30 bp of a CTCF ChIP-seq peak summit. **D** Bar plot quantifying the frequency of different fragment lengths genome-wide and how often each fragment length group overlaps an occupied CTCF motif. Occupied CTCF motifs are defined here as CTCF motifs within 30 bp of a CTCF ChIP-seq peak summit. **E** Fragment coverage metaplot +/- 500 bp around CTCF binding sites. Schematic below the coverage metaplot illustrates the likely proteins producing these peaks. Schematic created in BioRender. Aryee, M. (2025) https://BioRender.com/g98u240 **F** Plot **(E)** stratified by fragment length.

Sub nucleosome-sized protected fragments enable precise CTCF binding site localization

In order to characterize CTCF-mediated chromatin interactions, we first set out to map CTCF binding sites with high resolution. We examined the MNase CTCF HiChIP fragment distribution around individual CTCF motifs and found a strong enrichment for short fragments (see methods) in the 150bp windows centered on motifs (Figure 1.2A, B). Further, a strand-specific analysis shows a bimodal read start position distribution centered on the CBS, with read 5' location peaks observed upstream (positive strand) and downstream (negative strand) of the CBS (Figure 1.2C, D). We refer to these regions as quadrants 2 and 4 (Q2 and Q4) respectively (Figure 1.2D, E). This suggests that CTCF binding can be detected by an enrichment of reads in Q2 and Q4 relative to reads in Q1 and Q3 (Figure 1.2D, E). This pattern was observed consistently across four replicates (Figure A.3A). At sites without protein binding, MNase can cut at any location resulting in no specific quadrant enrichment (Figure 1.2E). As a result, we attempted to determine binding events by testing if there are significantly more reads in Q2 and Q4 than Q1 and Q3 (Figure 1.2F).

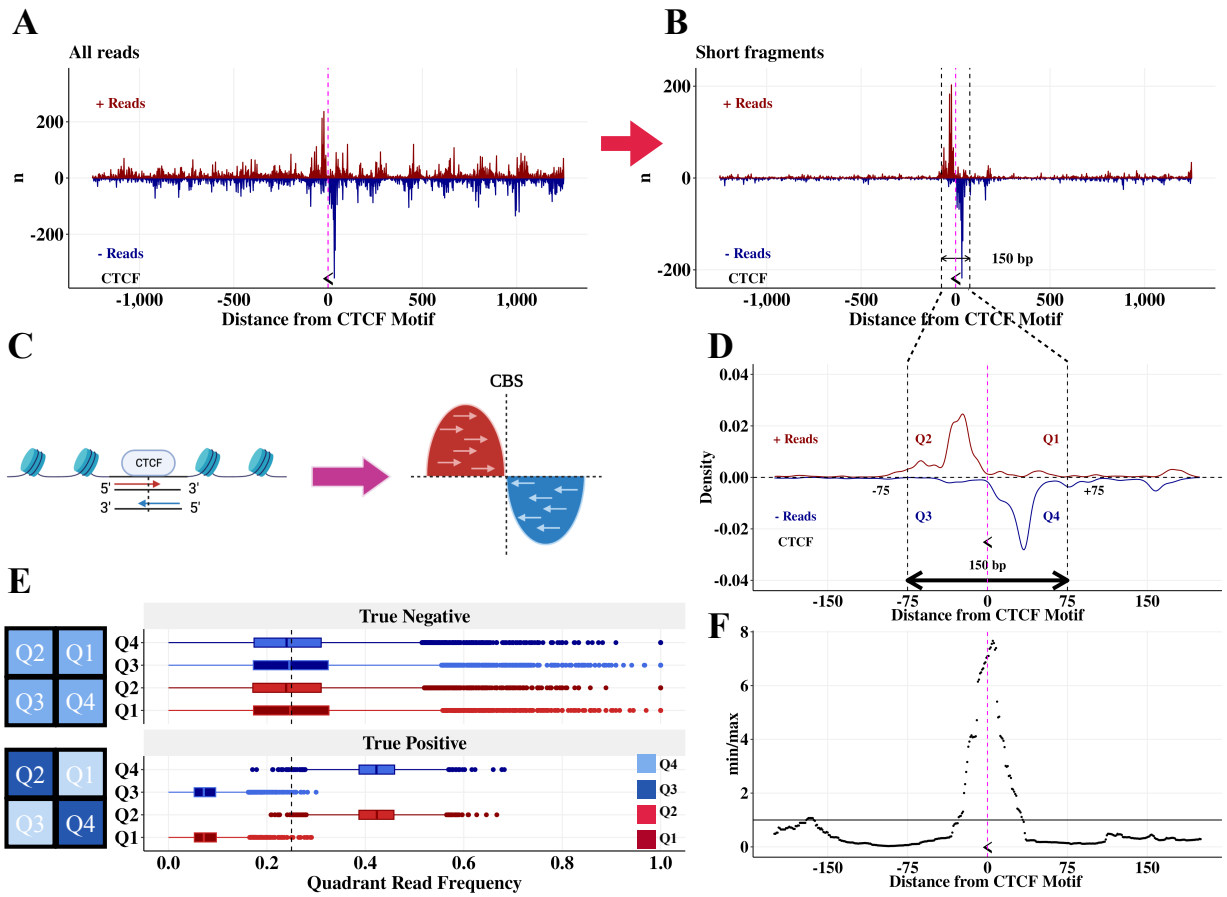


Figure 1.2 True CTCF binding sites have a bimodal strand-specific distribution centered on the CTCF motif. **A** Unfiltered reads +/- 1250 bp around a CTCF binding site located on the negative strand (chr1: 30,779,763 - 30,779,781). The midpoint of the CTCF motif is marked with the symbol " < ", representing that it is on the negative strand, and a pink line. **B** Plot **(A)** filtered to observed ligations (equivalently, short fragments.) **C** Schematic demonstrating the bimodal read pile-up around a CTCF binding site. Created in BioRender. Aryee, M. (2025) https://BioRender.com/g98u240 **D** Plot **(B)** as a density plot and zoomed in on the CTCF motif, with quadrant annotations. **E** Distributions of reads in quadrants for true negative and true positive CTCF binding sites in DNA loop anchors. True positives are defined as CTCF motifs that are the only CTCF motif in a loop anchor and within 30 bp of a CTCF ChIP-seq peak. True negatives are areas of the loop anchors with one CTCF motif that are at least 200 bp from the CTCF motif. 3,984,150 fragments across 4,523 loop anchors each containing one CTCF binding site are used to make these boxplots. Boxplots are made with ggplot2::geom_boxplot() and show

the 25% quantile (lower bound of box), median (center line), and 75% quantile (upper bound of box). Boxplot whiskers subtract (lower whisker) or add (upper whisker) 1.5 * IQR from the lower bound (lower whisker) or upper bound (upper whisker) of the box. Schematics of the quadrant read pile-up patterns are shown next to the corresponding true positive and true negative boxplots. **F** *CAMEL* statistic ($\hat{\alpha} = \frac{\min(n_2,n_4)}{\max(n_1,n_3)}$) for plot **(D)** peaks at the CTCF motif.

Statistically we assess each genomic location by considering each adjacent read in the 75bp upstream and 75bp downstream region as an independent draw from a multinomial distribution with four categories corresponding to the four quadrants. Under the null hypothesis, each read has equal probability of belonging to any of the four quadrants $Q_i, i \in \{1,2,3,4\}$. Because CTCF binding induces a strong preferential read pile-up in *both* quadrants 2 and 4 (Figure 1.2C, D, E), we test for an enrichment of reads in Q2 and Q4 compared to Q1 and Q3 by estimating the *CAMEL* statistic $\hat{\alpha} = \frac{\min(n_2,n_4)}{\max(n_1,n_3)}$, where n_i is the number of reads in Q_i. We then test if $\hat{\alpha}$ is significantly greater than 1. Note that min and max are used to enforce that both quadrants 2 and 4 must have more reads than both quadrants 1 and 3; using the average enables spurious read pile-ups that occur in quadrant 2 or 4 (but not both) to be called as false positive CTCF binding events. Nominal p-values are computed using an empirical null distribution generated through multinomial sampling (see methods).

We sought to evaluate this CTCF binding site identification approach using CTCF motif locations[34], CTCF ChIP-seq peaks[29], and DNA loop anchors identified by FitHiChIP at 2.5kb resolution[28]. We first defined a high stringency true positive set of CTCF binding sites as CTCF motifs in loop anchors that are located within 30 bp of a CTCF ChIP-seq peak summit. To avoid ambiguity due to multiple closely spaced motifs, we further selected only those motifs that are

unique within a 2.5kb loop anchor. Using this true positive set, we observe that the *CAMEL* statistic, $\log2(\hat{\alpha}) = \log2(\frac{\min(n_2,n_4)}{\max(n_1,n_3)})$ has signal greater than 0 (equivalently, $\hat{\alpha} > 1$) almost exclusively within 20 bp of the CTCF motif center and centered on 0 bp from the CTCF motif center (Figure 1.3A, Figure A.3B). Using this same set of true positive sites (true negatives are the regions of the loop anchors >200 bp from a CTCF motif), we achieve > 90% precision and > 90% recall at a nominal p-value threshold of 1e-05, and maintain high recall and precision at all p-value thresholds < 1e-05 (Figure 1.3B, Figure A.3C). This high level of recall and precision is achieved because of the very different *CAMEL* statistic distributions for true positives and true negatives (Figure 1.3C).

Running *CAMEL* binding site detection genome-wide, we observe that almost all peaks (93%) are within 20 bp of a CTCF motif center, with a median distance of 5 bp (Figure 1.3D). Defining accuracy as motif occurrence within 20 bp of a peak summit, we find that *CAMEL* maintains ~95% motif occurrence (Figure 1.3E). Further, applying the motif discovery tool STREME[35] to 30 bp sequences centered on *CAMEL* peak summits produces a motif sequence that exactly matches the JASPAR CTCF motif (Figure 1.3F), supporting *CAMEL*'s ability to identify true CTCF binding sites.

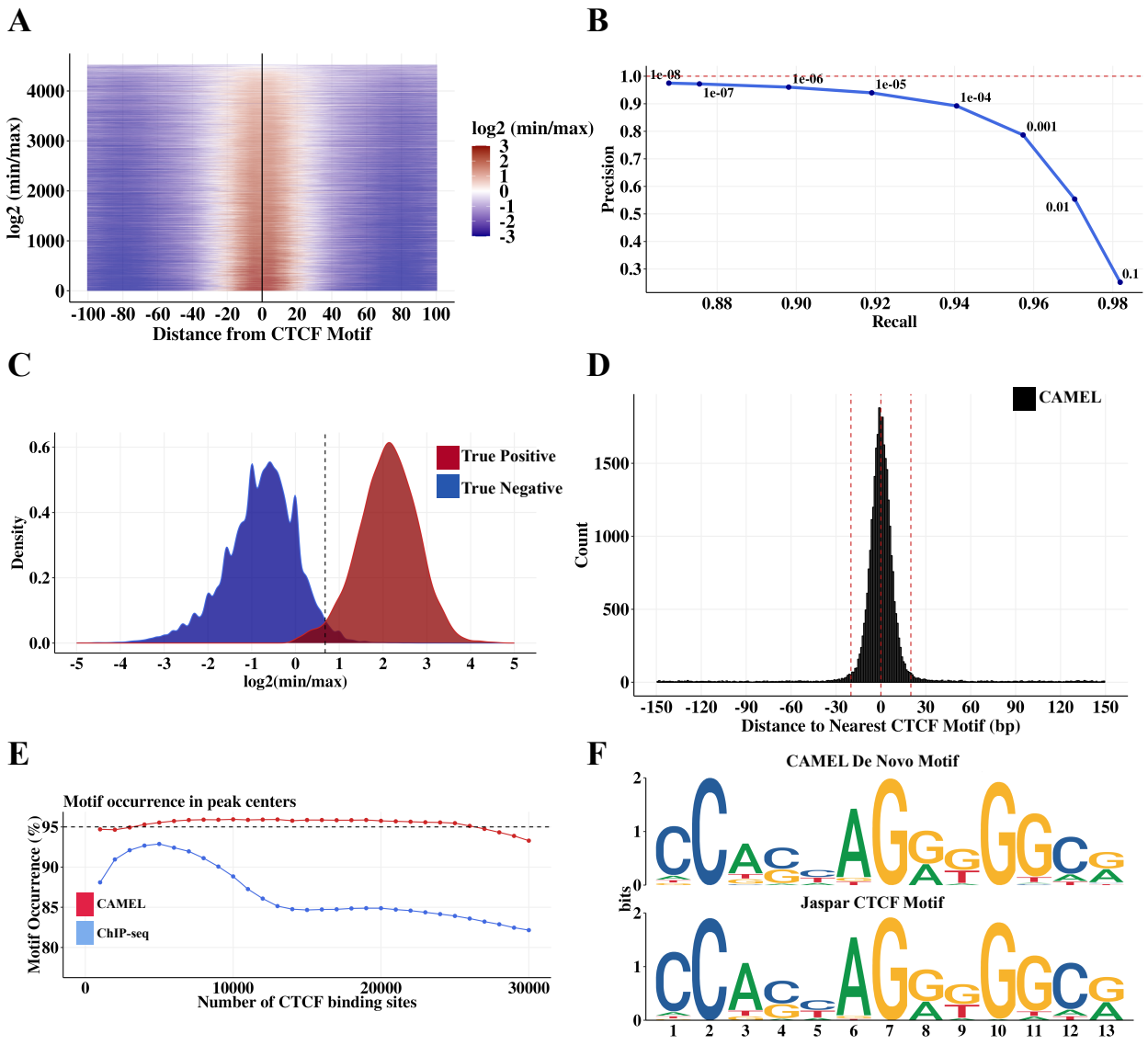


Figure 1.3 CTCF binding sites identified by *CAMEL* with single basepair resolution in MNase K562 CTCF HiChIP data. **A** Heatmap of log2(min/max) as a function of distance between *CAMEL* peak center and CTCF motif center within loop anchors. Only CTCF motifs that are unique within a loop anchor and within 30 bp of a CTCF ChIP-seq peak are used. **B** Precision recall curve for true negative and true positive CTCF binding sites in DNA loop anchors. True positives are defined as in **(A)**. True negatives are areas of the loop anchors in **(A)** that are at least 200 bp from the one CTCF motif. Precision is calculated as TP / (TP + FP), recall is calculated as TP / (TP + FN). **C** *CAMEL* statistic $\log2(\hat{\alpha}) = \log2(\frac{\min(n_2,n_4)}{\max(n_1,n_3)})$ density plots using the same set of true positives and true negatives as **(B). D** Histogram with 1 bp bin size depicting *CAMEL* resolution for all peaks genome-wide (not just in loop anchors). **E** Motif occurrence in ChIP-seq and *CAMEL* peak centers genome-wide. Motif occurrence is calculated as % peak

centers within 20 bp of CTCF motif. Only peak centers within 150 bp of a CTCF motif are used for this figure. **F** 30 bp sequences centered on genome-wide *CAMEL* peak centers produce a de novo motif (top) that matches the core JASPAR CTCF motif (bottom). De novo motif is made using STREME[35].

Short, TF-scale fragments are enriched for long-range interactions

We next examined the length characteristics of MNase HiChIP fragments overlapping individual CTCF motifs, in an attempt to infer the identity of the proteins occupying each locus. For all motifs with non-zero coverage, we observed long, 150+ bp fragments, as shown for two representative motifs in Figure 1.4A. We hypothesized that these fragments represent cells with a nucleosome located at the motif locus. In addition, particularly for CTCF motifs with an overlapping CTCF ChIP-seq peak, we also observed short, sub-nucleosome sized (<120 bp) fragments (Figure 1.4A, left). The fraction of reads overlapping a CAMEL-identified CBS (+) that are short (<120bp) strongly correlates with the strength of its CTCF ChIP-seq signal (Figure 1.4B).

Given the association between short fragment frequency and binding of the architectural protein CTCF, we asked whether short and long fragments might have different long-range interaction behavior. We stratified the 2.3 million HiChIP contacts anchored at CAMEL-identified CBS (+) by the size of the motif-overlapping fragment and examined the distance to the fragments' interaction partner. This reveals that interactions that span linear distances greater than 10kb are significantly enriched (2.2-fold, $p < 10^{-10}$ (chi-square)) for short TF-scale fragments (Figure A.4A). Consistent with this, HiChIP contact maps created using only <120bp fragments more focally highlight CTCF-mediated contacts (Figure A.5).

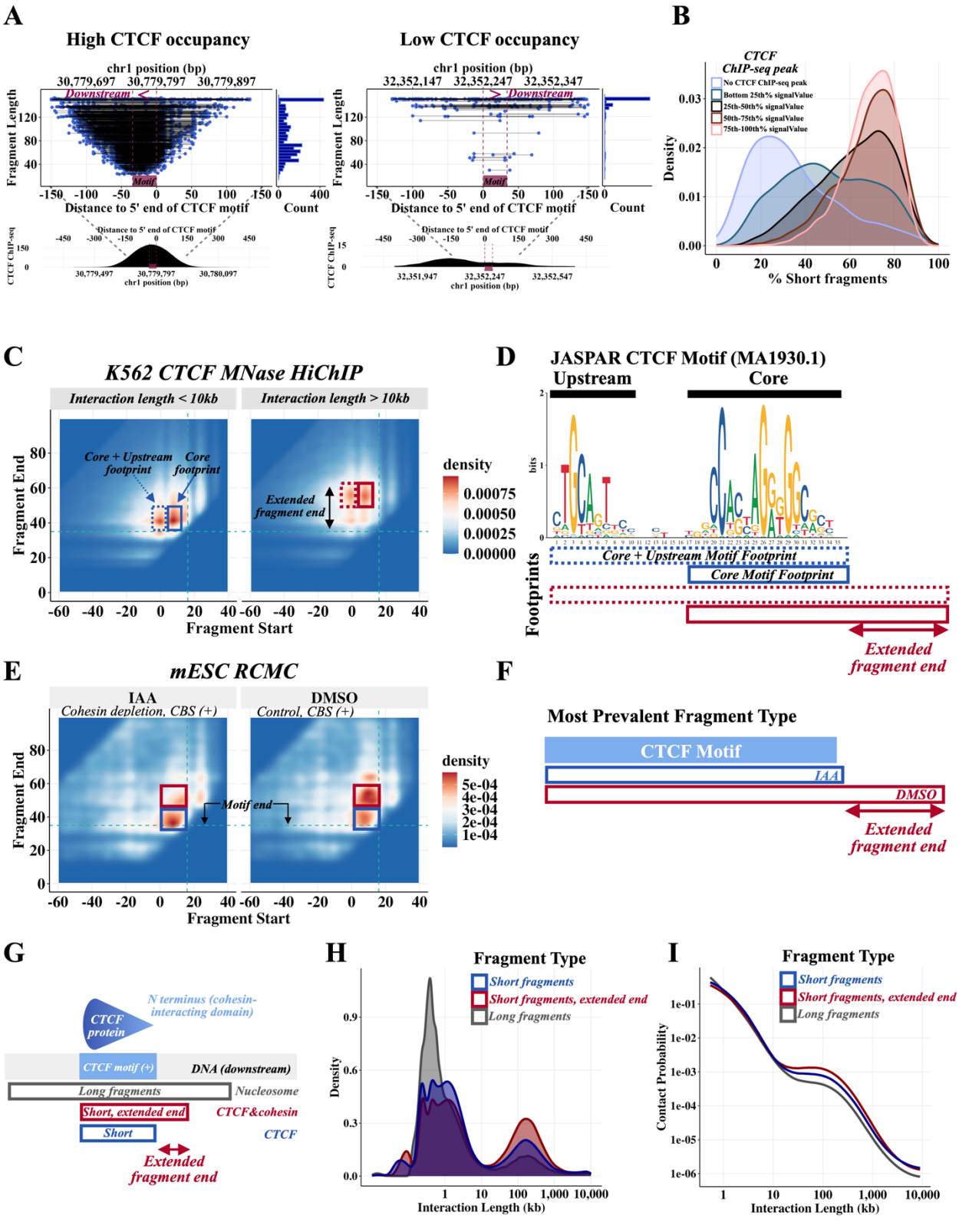

A
High CTCF occupancy
Low CTCF occupancy
chr1 position (bp)
Fragment Length
Downstream
Motif
Distance to 5' end of CTCF motif
Count
CTCF ChIP-seq
B
CTCF ChIP-seq peak
No CTCF ChIP-seq peak
Bottom 25th% signalValue
25th-50th% signalValue
50th-75th% signalValue
75th-100th% signalValue
Density
% Short fragments
C
K562 CTCF MNase HiChIP
Interaction length < 10kb
Interaction length > 10kb
Core + Upstream footprint
Core footprint
Extended fragment end
density
Fragment End
Fragment Start
D
JASPAR CTCF Motif (MA1930.1)
Upstream
Core
bits
Core + Upstream Motif Footprint
Core Motif Footprint
Footprints
Extended fragment end
E
mESC RCMC
IAA
Cohesin depletion, CBS (+)
DMSO
Control, CBS (+)
Motif end
F
Most Prevalent Fragment Type
CTCF Motif
IAA
DMSO
Extended fragment end
G
N terminus (cohesin-interacting domain)
CTCF protein
CTCF motif (+)
DNA (downstream)
Long fragments
Nucleosome
Short, extended end
CTCF&cohesin
Short
CTCF
Extended fragment end
H
Fragment Type
Short fragments
Short fragments, extended end
Long fragments
Density
Interaction Length (kb)
I
Fragment Type
Short fragments
Short fragments, extended end
Long fragments
Contact Probability
Interaction Length (kb)

Figure 1.4 Cohesin footprint observed in MNase HiChIP and experimentally confirmed in mESC RCMC. **A** High and low CTCF occupancy motifs. For each occupancy level, CTCF ChIP-seq and all fragments overlapping the CTCF motif are depicted, along with the corresponding fragment length histogram. **B** The fraction of reads overlapping a CBS (+) that are short (<120bp) strongly correlates with the strength of its CTCF ChIP-seq signal. The fraction of short reads (<120bp) was calculated for each CBS, plotted in a density, and colored by absence of CTCF ChIP-seq peak or CTCF ChIP-seq signalValue quartile. **C** K562 CTCF MNase HiChIP 2D density for short (<120bp) left fragments overlapping CAMEL-identified CBS (+) (N = 10,906 CBS). The fragment start and fragment end are aligned relative to the start of the 35bp CTCF motif (MA1930.1), so that "0" in the plots corresponds to the start of the 35bp CTCF motif. **D** Annotated 35bp JASPAR CTCF motif (MA1930.1). **E** mESC RCMC 2D density for short (<120bp) left fragments overlapping CTCF motifs within 30bp of a mESC CTCF ChIP-seq peak (N = 65 CBS). The fragment start and fragment end are aligned relative to the start of the 35bp CTCF motif (MA1930.1), so that "0" in the plots corresponds to the start of the 35bp CTCF motif. **F** Schematic illustrating the most prevalent fragment type observed in the mESC RCMC perturbations shown in **(E)**. **G** Schematic showing the three classified types of fragments relative to the location of the CTCF motif. The interaction partners of these three fragment types are graphed in **(H, I)**. **H** Short TF-sized fragments (<120bp) with an extended fragment end have a noticeably larger bump in density of long range interactions (>10kb) compared to TF-sized short fragments without an extended fragment end and nucleosome-sized long fragments (>120bp). **I** P(S) curve for fragments depicted in **(H)**.

CTCF and Cohesin occupancy footprints

Focusing on the short, CBS-overlapping fragments (< 120bp), we found that they exhibit a bimodal interaction length distribution (Figure A.4A), with 39% involving an interaction partner >10kb away. We sought to understand factors that might explain this long- vs short-range interaction behavior through a closer examination of the fragment characteristics. To this end we created aggregate 2D density "footprint plots" showing fragment positions relative to CTCF motifs at CAMEL-identified CTCF binding sites (+) by plotting fragment start positions (x-axis) against fragment end positions (y-axis) (Figure 1.4C), such that each (fragment start, fragment end) combination represents a single point. Fragment start and fragment end have been aligned to the start of the 35bp JASPAR CTCF motif (MA1930.1) such that 0 represents the start of the

motif. Examining first only those fragments with short-range (<10kb) interactions (Figure 1.4C left) we noted that two fragment footprint types were most prominent. The footprints correspond to two different fragment types, with a shared end position but differing starts. To interpret these footprints we overlaid them onto the CTCF motif (MA1930.1). The motif contains a core 19bp region that is present at the majority of CTCF binding sites genome-wide, and an upstream 10bp region that is present at a smaller subset of CTCF sites[36,37] (Figure 1.4D, Figure A.4B). Strikingly, we find that the boundaries of the footprints correspond almost precisely to two different binding events: one that protects the 19bp core region only and another binding event that protects the full 35bp core + upstream region (Figure 1.4C, D).

We next examined footprint plots for the short TF-scale fragments (<120) with long-range (>10kb) interaction partners (Figure 1.4C right). These fragments also comprise two distinct types with the same start positions as the short-range interaction fragments, corresponding to the beginning of either the core or upstream CTCF motif region. Interestingly however, both the Core and Core+Upstream fragment types have an extended downstream end that stretches ~15bp beyond the CTCF footprint (Figure 1.4C, D, Figure A.4C), suggesting the possible presence of an additional protein protecting this region. CTCF binds its motif in an orientation that presents the N-terminus towards the downstream end, and recent work has shown that regions within this N-terminus interact directly with the loop-extruding cohesin complex[6,38]. Based on this and the link between the CTCF-downstream footprint and long-range interactions, we hypothesized that the footprint indicates cohesin's presence.

To investigate whether the CTCF downstream extended footprint reflects cohesin occupancy we turned to a degron strategy that achieves 97% cohesin depletion 3 hours after auxin (IAA) induction[20]. Goel *et al.* have previously used this system with Region Capture Micro-C[20], a procedure involving a similar MNase digestion to MNase HiChIP, to profile 3MB of DNA containing 65 CTCF binding sites at the *Fbn2*, *Klf1*, and *Ppm1g* loci in mouse embryonic stem cells. We resequenced these RCMC libraries from the IAA-induced cohesin degron and DMSO control conditions[20] using 150bp paired end reads to allow accurate assessment of fragment length. Examining fragment footprint plots for short TF-scale fragments we observe a 15bp extended footprint downstream of the CTCF motif in the DMSO control condition that is greatly attenuated upon cohesin depletion by IAA (Figure 1.4E, F - red box).

Taken together, these findings suggest that base-pair resolution analysis of MNase digestion fragment start and end positions at CTCF binding sites can classify individual fragments as likely bound by a nucleosome, CTCF alone, or CTCF together with cohesin. Given cohesin's role in mediating DNA loops, we hypothesized that cohesin-associated DNA would be more likely to participate in long-range chromatin interactions. Using the proxy of a CTCF-downstream cohesin footprint to identify likely cohesin-protected DNA fragments, we indeed observe that these fragments have increased propensity for long-range interactions in the 10kb-500kb range compared to DNA bound to CTCF alone (Figure 1.4G-I).

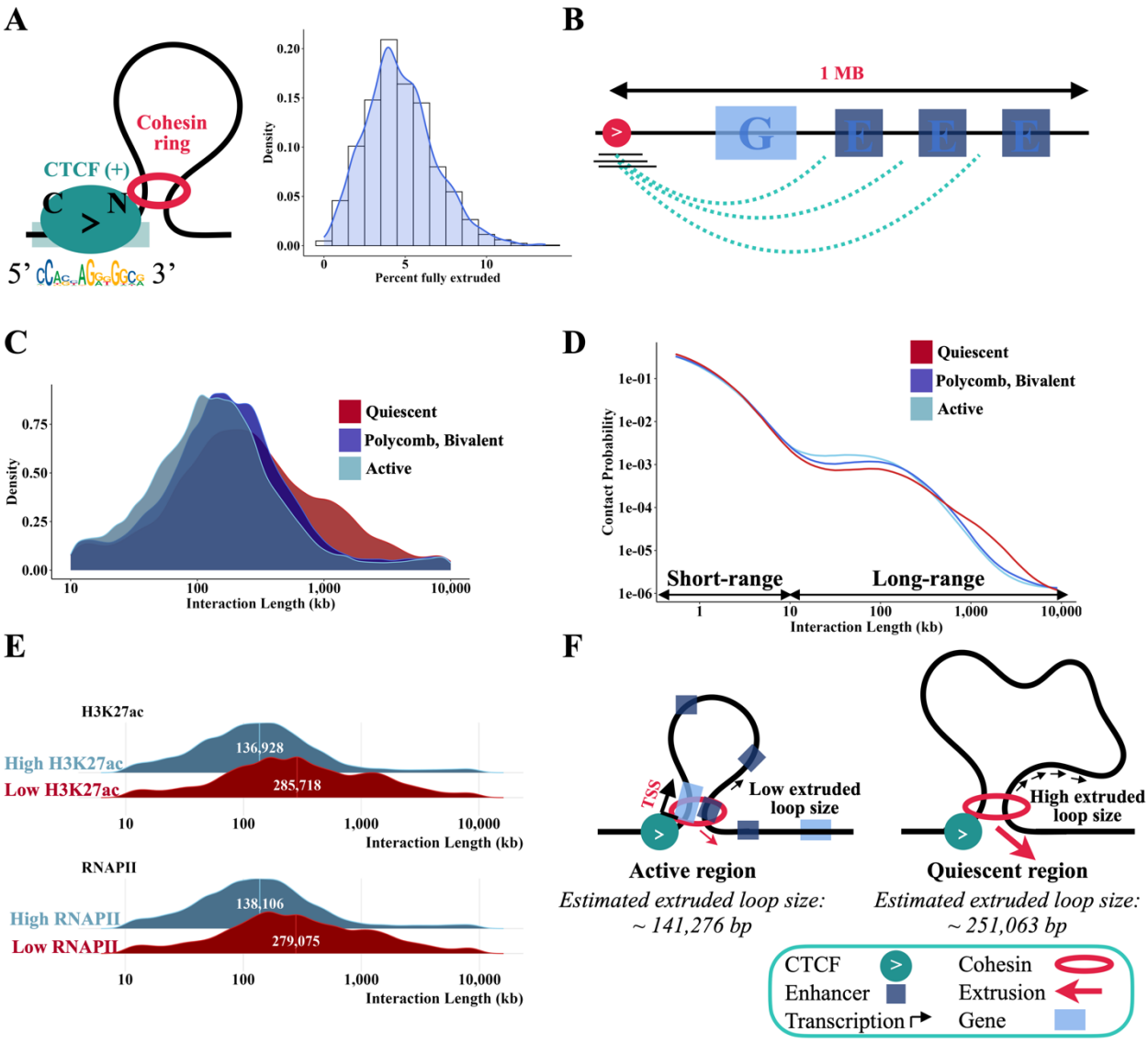


Figure 1.5 Cohesin extrudes further through quiescent regions than active regions. **A** Most CTCF-mediated looping contacts do not reflect the fully extruded state. Estimate is obtained using left fragments that overlap *CAMEL* identified CBS (+) and have an interaction length greater than 10kb with start and end at least 5 bp from motif start and end, length < 120, and extended fragment end. For each CBS with at least 50 long-range TF-protected fragments overlapping the motif, % convergent is calculated as the number of interaction partners overlapping CTCF (-) motifs / total number of fragments at motif. Because this estimate is conditional on CTCF binding at the anchor, we divide estimates by two to account for the ~50% occupancy of CTCF[33]. **B** Depiction of how 1 MB regions downstream of CBS were annotated using ChromHMM. Density **(C)** and P(S) curves **(D)** for chromatin state clusters shown in **(Figure A.6A)**, filtered to the top 20%. Chromatin annotations making up each cluster are added together and quantiles are obtained to determine fragments in the top 20% of active chromatin,

quiescent chromatin, and bivalent / polycomb chromatin. Left fragments that overlap *CAMEL* identified CBS (+) and have an interaction length greater than 10kb with start and end at least 5 bp from motif start and end, length < 120, and extended fragment end are used. **E** Ridge plots for the bottom 10% quantile ("Low") and top 10% quantile ("High") of H3K27ac bp and number of RNAPII binding sites. ChIP-seq from ENCODE was used to annotate 1 MB downstream of left fragments overlapping CBS (+) for this figure. Left fragments that overlap *CAMEL* identified CBS (+) and have an interaction length greater than 10kb with start and end at least 5 bp from motif start and end, length < 120, and extended fragment end are used. Plots are labeled with the average log10 interaction length. **F** Diagram illustrating differences in extrusion rates between active and quiescent chromatin states, with numbers obtained from Figure A.6B.

Active chromatin obstructs cohesin-mediated loop extrusion

Using the techniques described above, MNase proximity ligation assays enable us to simultaneously locate CBS at high resolution, identify footprints of bound proteins, and interrogate specific chromatin contacts at the single molecule level. We next used these data to identify cohesin-protected DNA fragments and characterize cohesin extrusion dynamics in a range of genomic contexts.

We first sought to determine the genome-wide frequency of fully extruded (CTCF-CTCF) loops, which have been previously estimated using locus-specific live-cell imaging studies[14,15]. We used K562 MNase HiChIP data to estimate the fully extruded rate for each CAMEL-identified CTCF binding event by selecting short fragments overlapping the CBS with an extended fragment end (Figure 1.4C, Figure A.4C) and determining the fraction of interaction partners that overlap a downstream convergent CTCF motif (see methods). We find wide CBS to CBS variability with fully extruded rate estimates ranging from ~1-10% (Figure 1.5A) and a genome-wide average frequency of 5%. These findings are compatible with an imaging study of the 505kb *Fbn2* TAD

that estimated a fully extruded loop frequency of 3%-6%[14], and suggest that the majority of CTCF-anchored chromatin contacts genome-wide are in the 'extruding' state, rather than joining two CTCF sites. We note that these fully extruded loop estimates are contingent on at least one end of a long-range interaction being anchored by CTCF, and does not include long-range interactions that are not anchored by CTCF on either end (the fully unlooped state).

We next sought to use our data to examine how cohesin extrusion is impacted by chromatin context. Since HiChIP libraries are a snapshot of millions of cells, we can estimate dynamic extrusion parameters such as the average loop size extruded by cohesin[39] from interaction length distributions. To determine the impact of chromatin state on cohesin extrusion, we first annotated the 1 MB region downstream of each *CAMEL*-identified CBS with ChromHMM chromatin states[40] (Figure 1.5B, see methods) to characterize the DNA through which a cohesin anchored at the CBS would extrude. We collapsed ChromHMM states into three main categories: active, polycomb/bivalent or quiescent (Figure A.6A), and computed the fraction of downstream chromatin in each category. Each of the three chromatin categories was represented by the 20% of regions with the highest fraction of DNA in this state. We used CTCF/cohesin-protected fragments (Figure 1.4C, Figure A.4C) overlapping the three sets of CBS and estimated extruded loop sizes using fragment-level interaction lengths.

Interestingly, we find that cohesin loop sizes are ~1.75 fold greater in quiescent regions (~250kb) than active regions (~140kb), corresponding to a difference in average extruded loop size of ~110kb, $p < 10^{-10}$ (Figure 1.5C, Figure A.6B). The P(s) curve, a plot of interaction frequency vs.

distance, confirms a depletion of the longest-range interactions in active regions (Figure 1.5D). This ~250kb estimate for loop size in quiescent regions is consistent with a live cell imaging study of the *Fbn2* locus in the absence of transcription that estimated an average loop size of 300kb[14]. As quiescent regions are characterized by low TF binding and low transcription[41], we hypothesized that the substantial difference in extruded loop size might relate to gene activity and enhancer density obstructing loop extrusion. Consistent with this, we found that higher levels of H3K27ac and RNA Pol II binding in the 1MB region downstream of the CBS strongly correlate with lower average extruded loop size (Figure 1.5E). In addition, a focal enrichment of contacts between CBS and downstream H3K27ac-marked enhancers (Figure A.6C) suggests an ability of these regulatory elements to stall extrusion.

We used the mESC RCMC cohesin degron system to assess the cohesin dependency of contact enrichment between CTCF binding sites and regulatory elements. The captured region contains the *Fbn2* gene and its encompassing convergently oriented CTCF binding sites (Figure A.7A). A pseudo 4C plot showing interactions anchored at the *Fbn2* upstream CTCF viewpoint confirms the presence of the fully extruded 505kb TAD-defining CTCF-CTCF loop and also reveals a shorter range contact enrichment between the CTCF anchor and the 230kb downstream RNAPII-marked *Fbn2* transcriptional start site (TSS) (Figure A.7B). Consistent with this, the TSS is also marked by cohesin (RAD21 & SMC1A) ChIP-seq peaks. As expected, cohesin depletion significantly weakens the fully extruded CTCF-CTCF TAD contacts. Importantly, the loss of cohesin also attenuates the 3D contacts between the CTCF anchor and the TSS, suggesting that these contacts are also cohesin-dependent. These findings

further suggest that regulatory elements are able to impede cohesin extrusion, and are consistent with recent reports of RNAPII/cohesin interactions[42,43].

We performed two additional sets of analyses to confirm that the estimated differences in loop extrusion length between chromatin states are not confounded by background, non cohesin-related, interactions. First, since each genomic region has locus-specific genetic and epigenetic architecture, we fit a linear mixed effects model to account for the locus-to-locus variability in interaction patterns (see methods). Specifically, we compute the 'cohesin effect' on loop length, defined as the average increase in interaction length for CTCF/cohesin bound fragments while adjusting for the interaction length of nucleosome bound fragments at each individual CBS. Controlling for the background interaction frequency due to the local sequence and epigenetic state context of a region in this way confirms that cohesin-associated loops are significantly shorter in active chromatin (see Methods and Figure A.6D). Second, we examined whether differences in CTCF binding density between active and quiescent chromatin might account for the observed differences in average extruded loop size. We repeated the loop length analyses after excluding contacts between convergent CTCF motifs and confirmed that the relationship between chromatin state on loop length remains largely unaltered (Figure A.6E, F).

1.3 Discussion

While chromatin loops as visualized on proximity ligation contact maps are often viewed as static, it is evident from recent single locus live cell imaging studies that these structures can be very dynamic. Here we provide a genome-wide view of extrusion dynamics by using MNase-based proximity ligation data to simultaneously infer 3D genome structure and the presence of

the architectural factors CTCF and cohesin on individual DNA molecules. The dual endo- and exo-nuclease activity of MNase degrades unprotected DNA up to the edge of bound proteins resulting in fragment sizes that are characteristic of either nucleosome or CTCF/cohesin binding. Our findings demonstrate that an analysis of precise fragment start and end positions enables the classification of individual DNA molecules participating in 3D interactions based on their association with either nucleosomes or CTCF. Further, a subset of CTCF-bound molecules have an additional extended footprint immediately downstream of the CTCF binding site that can be attributed to the co-presence of a cohesin complex. This is consistent with a co-crystal structure of CTCF and the SA2-SCC2 cohesin subunits that shows cohesin docking at the N-terminus of the CTCF protein, positioned at the downstream edge of the motif.

By leveraging this 'footprinting' property of MNase together with a strand-specific analysis of digestion fragment sizes, we show that MNase CTCF HiChIP can be used to identify CTCF binding sites with 5bp precision in K562 cells, on-par with the precision of the dedicated localization assay CUT&RUN[25]. This precision allows the discrimination of bound and unbound CTCF motifs located in close proximity. Since loop anchors often contain multiple CTCF motifs (Figure A.6G), not all of which are bound, distinguishing between occupied and unoccupied CTCF motifs at high resolution is vital for determining the CBS mediating a chromatin loop. We have implemented this approach in an accessible software package, *CAMEL* (CTCF Analyzer with Multinomial Estimation, https://github.com/aryeelab/cohesin_extrusion_reproducibility).

We use the ability to identify individual 3D contacts anchored by CTCF and/or cohesin to make several inferences about cohesin extrusion dynamics genome-wide. We estimate the frequency

with which a CTCF bound locus forms a loop with a downstream CTCF site and find that it varies considerably from CBS to CBS, with a genome-wide range from ~1-10%. This is consistent with estimates from two recent live-cell imaging studies that found that CTCF-mediated loops predominantly exist in the partially extruded state at two imaged TADs[14,15]. We note a limitation of our study in that the use of HiChIP may lead to an over-estimate of the fully extruded rate.

We next explored the impact of chromatin state on loop extrusion and observed an approximately 2-fold increase in extruded loop size comparing quiescent chromatin to active chromatin, an effect that correlates with differences in H3K27ac and RNAPII levels between these chromatin states. Consistent with this, we also observe punctate contact enrichment between CTCF bound sites and downstream H3K27ac-marked enhancers and RNAPII-marked transcriptional start sites. Analysis of a cohesin degron system revealed that these contacts are attenuated upon cohesin depletion. This is in line with previous ChIP-seq studies showing co-localization of RNAPII and cohesin in the absence of CTCF[44], despite evidence against preferential loading of cohesin at regulatory elements[42]. These findings are consistent with a model of stalled extrusion at regulatory elements and previous studies suggesting that gene and enhancer activity obstruct cohesin activity[45–50]. A limitation of our study, however, is that our data do not allow us to exclude alternate reasons such as differences in extrusion speed or cohesin unloading activity between active and less active chromatin regions.

This dynamic CTCF-mediated enhancer-promoter contact model is consistent with recent evidence that cohesin-mediated chromatin loops are dynamic[14,15] and that promoter proximal CTCF binding sites facilitate contacts with distal enhancers[10–12,50,51]. Our findings are also compatible with the "kiss and kick" model of gene regulation where enhancer-promoter contacts are temporarily maintained while RNAPII is paused at transcriptional start sites, and are subsequently lost during elongation[52].

In summary, these findings suggest that the concept of chromatin loops should be expanded beyond a static structure to include an actively extruding state, where loops anchored by a single CTCF may play a key role in enabling the 'scanning' of DNA for regulatory elements (Figure 1.6).

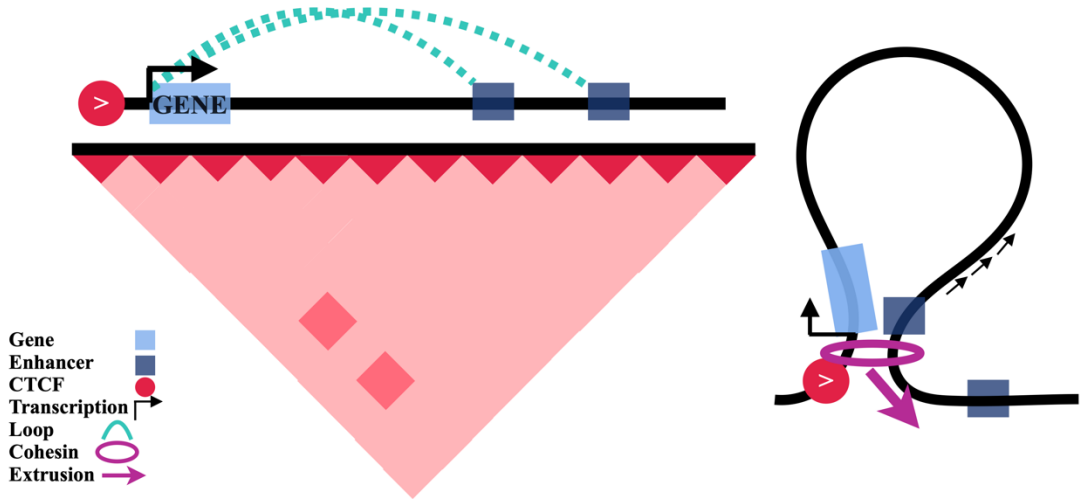


Figure 1.6 Schematic of supported model whereby single promoter-proximal CTCF sites enable an enrichment of enhancer-promoter contacts.

1.4 Methods

Data generation

K562 (ATCC CCL-243) CTCF MNase HiChIP

Four MNase K562 CTCF HiChIP (150 bp paired-end) libraries were generated using the Cantata Bio / Dovetail Genomics MNase HiChIP kit[17–19]. CTCF MNase HiChIP was performed as described in the Dovetail HiChIP MNase Kit protocol v.2.0. Briefly, 5 million K562 cells per sample were crosslinked with 3mM DSG and 1% formaldehyde and digested with 1ul MNase ("YET" samples) or 2ul MNase ("GW" samples) in 100ul of 1X nuclease digestion buffer. Cells were lysed with 1X RIPA containing 0.1% SDS, and CTCF ChIP was performed using 1500ng of chromatin (40-70% mononucleosomes) and 500 ng of CTCF antibody (Cell Signaling, cat #: 3418). Protein A/G beads pull-down, proximity ligation, and library preparation were done according to the protocol. Libraries were sequenced to a read depth of ~172 million paired end reads per sample on the Illumina Nextseq 2000 platform.

mESC (RRID: CVCL_J962) RCMC

WT, 3 hour DMSO, and 3 hour IAA libraries (two replicates of each) at the *Fbn2* (chr18:57,899,000-58,901,000), *Ppm1g* (chr5:31,100,000-32,225,000), and *Klf1* (chr8:84,120,000-85,130,000) loci from Goel *et al.*[20] were re-sequenced with 150bp paired-end reads. Note that the provided coordinates of the *Fbn2*, *Ppm1g*, and *Klf1* captured loci use the mm10 reference genome.

Data Pre-processing

Four replicates of K562 MNase CTCF HiChIP and two replicates each of mESC RCMC WT, mESC RCMC DMSO, and mESC RCMC IAA were aligned to the reference genome (hg38 and mm10 respectively) using the BWA-MEM algorithm[53] (with *-5SP -T0*). Ligation events were then recorded using pairtools parse v. 0.3.0[54] (with *--min-mapq 40 --walks-policy 5unique --max-inter-align-gap 30 --add-columns pos5,pos3,dist_to_5,dist_to_3,read_len*). Parsed reads were then sorted (pairtools sort) and PCR duplicates were removed (pairtools dedup). Final pairs and bam files were then generated using *pairtools split --nproc-in 8 --nproc-out 8 --output-pairs ${filename}.mapped.pairs --output-sam -|samtools view -bS -@16 | samtools sort -@16 -o ${filename}.mapped.PT.bam;samtools index ${filename}.mapped.PT.bam* We used a Terra pipeline to keep data processing consistent.

Replicates were merged using pairtools merge and for the RCMC samples we then filtered to both fragments of an interaction pair in the captured regions using pairtools select (captured regions (mm10): *Fbn2* (chr18:57,899,000-58,901,000), *Ppm1g* (chr5:31,100,000-32,225,000), and *Klf1* (chr8:84,120,000-85,130,000)). For example, *pairtools select '(chrom1=="chr18") and (chrom2=="chr18") and (57899000 <= pos1 <= 58901000) and (57899000 <= pos2 <= 58901000)' RCMC_WT.mapped.pairs > RCMC_WT_Fbn2.mapped.pairs*.

CTCF MNase HiChIP loop calls were then made from the combined replicate pairs file using FitHiChIP Peak to Peak[28] with 2.5kb loop anchor bin size. Balanced contact maps were created using the merged replicate genome-wide pairs file (K562 CTCF MNase HiChIP) and the merged

replicate captured region pairs files (mESC RCMC). Filtering to both fragments in the captured window for RCMC data processing is the approach taken by the RCMC authors to avoid problems with ICE balancing[20]. To create .mcool files, we compressed the combined replicate pairs files (bgzip), indexed the compressed pairs files (pairix), created cool files (cooler cload pairix), and then obtained ICE-balanced mcool files (cooler zoomify --balance). This data processing approach was developed starting with guidelines from https://hichip.readthedocs.io/en/latest/before_you_begin.html.

The fragment start and fragment end of each interaction partner are obtained by taking the minimum of the 5' and 3' ends of a fragment (this is the fragment start) and the maximum of the 5' and 3' ends of a fragment (this is the fragment end). Including *--add-columns pos5,pos3* in the pairtools parse command adds the 5' and 3' ends of both fragments to the pairs file and is crucial for identifying fragment starts, fragment ends, and fragment lengths for both fragments in an interaction pair. Reproducible code is available at https://github.com/aryeelab/cohesin_extrusion_reproducibility.

Fragment lengths can be determined for fragments with length less than 150 bp; the 150 bp read length results in censoring of fragments longer than 150 bp. Since at least ~25 bp are required to align a sequence to the reference genome, we are able to identify both mapping locations for a ligation event when the longer fragment is <=125bp. We use this as a proxy for identifying short fragments in Figures 1.2 and 1.3. The fraction of informative, sub-nucleosome fragments decreases with shorter sequencing read length (Figure A.2).

Visualizing contact maps in R

Contact maps were visualized on the hg38 (human) and mm10 (mouse) reference genomes using HiContacts::plotMatrix after importing a given resolution of the mcool file using HiCExperiment::import. Publicly available bigwig files (ChIP-seq and ATAC-seq) were imported using rtracklayer::import, visualized with ggplot, and aligned with contact maps using cowplot::plot_grid. Downloaded mESC bigwig files were CrossMapped[55] from mm9 and mm39 to the mm10 reference genome. Downloaded K562 bigwig files were already aligned to hg38, and thus did not require CrossMapping.

Downloaded mESC files include: mESC CTCF ChIP-seq (GSM3508478), mESC ATAC-seq (GSE98390), mESC H3K27ac ChIP-seq (GSM2417096), mESC RNAPII ChIP-seq (GSM6809981), mESC RAD21 ChIP-seq (GSE137272), and mESC SMC1A ChIP-seq (GSE137272). Downloaded K562 files include: CTCF ChIP-seq (ENCFF736NYC.bed, ENCFF168IFW.bigWig), H3K27ac ChIP-seq (ENCFF544LXB.bed), RNAPII ChIP-seq (ENCFF355MNE.bed, ENCFF914WIS.bigWig). We obtained CTCF motifs from the CTCF R package[34], which provides CTCF motifs predicted by FIMO[56]. The CTCF annotation hub we use is AH104729 (hg38.MA0139.1), and the mESC CTCF annotation hub is AH104755 (mm10.MA0139.1).

CTCF MNase HiChIP comparison to Intact Hi-C

Publicly available K562 Intact Hi-C was downloaded from GEO (GSE237898) and compared to our MNase K562 CTCF HiChIP data. We first compared CTCF-CTCF loop calls obtained with FitHiChIP in our K562 CTCF MNase HiChIP experiment to the K562 Intact Hi-C loop calls, and found that 70% of our loop calls were also identified in the Intact Hi-C dataset (named "GSE237898_ENCFF256ZMD_loops_GRCh38.bedpe.gz" in the GEO link GSE237898). We next examine an example individual CTCF-mediated TAD (included in both datasets' loop calls) and compare balanced contact maps for CTCF MNase HiChIP and Intact Hi-C. We observe similar structures in both assays (Figure A.1A).

To more closely examine the data at high resolution, we zoom in on the 10kb x 10kb window centered on the CTCF-CTCF loop marked in Figure A.1A and plot the individual ligated fragment pairs (Figure A.1B). We see a strong enrichment directly at the intersection of the two CTCF sites in both datasets. While the same pattern is observed in both datasets, it is clear through the marginal density plot that chipping for CTCF in the CTCF MNase HiChIP experiment concentrates the signal at CTCF binding sites as expected (observed as a narrowing of the marginal density).

We next made a metaplot of Intact Hi-C contacts aggregated over chromosome one CTCF HiChIP loops (Figure A.1C). We observe, as expected, a strong enrichment between loop anchors and within the enclosed chromatin loops. These analyses suggest that the overall 3D genome structure related to CTCF binding sites that we observe in CTCF MNase HiChIP is recapitulated in Intact Hi-C data.

CAMEL nominal p-value estimation

To evaluate the significance of $\hat{\alpha}$ at a particular total read count $N = \sum_{i=1}^{4} n_i$, we simulated 100 million samples under the null hypothesis that each fragment is equally likely to occur in any of the four quadrants (Figure 1.2E). This was done at each total read count ranging from 5 to 500. P-values at read counts beyond 500 are very similar to those at 500, so 500+ read counts are treated as bins with 500 total read count (Figure A.8). The empirical CDF of the 100 million $\log2(\hat{\alpha})$ at a given total read count was then computed and used to evaluate the probability of observing a value more extreme than $\log2(\hat{\alpha})$ under the null hypothesis. The empirical CDF was evaluated at a sequence of possible $\log2(\hat{\alpha})$ between 0 and 5 at step sizes of 0.01(this corresponds to $\hat{\alpha} \in [1,32]$.) This approach produces the same p-values as using $\hat{\alpha}$ instead of $\log2(\hat{\alpha})$, but using the log enables smaller step size at large values of $\hat{\alpha}$. After acquiring the grid of p-values for each $\hat{\alpha}$ at a given read count N, we match the observed $\hat{\alpha}$ at a read count of N with the corresponding p-value from the table. Because this approach only requires quadrant-specific read counts to match with the given table of p-values, it is very computationally efficient. Furthermore, by using the multinomial framework we place no distributional assumptions on the reads within each quadrant.

Precision/Recall Curves (K562 CTCF HiChIP)

The precision recall curve is calculated for true negative and true positive CTCF binding sites in DNA loop anchors. True positives are defined as CTCF motifs that are unique within a FitHiChIP loop anchor and within 30 bp of a CTCF ChIP-seq peak. True negatives are areas of the FitHiChIP 2.5kb loop anchors (included in the true positive set) that are at least 200 bp from

the one CTCF motif. Precision is calculated as TP / (TP + FP), recall is calculated as TP / (TP + FN). TP refers to true positive, FP refers to false positive, and FN refers to false negative respectively.

Identification of significant motifs

We use CTCF motifs identified as significant ($p < 1e-05$) by *CAMEL* as the set of CTCF binding sites. This p-value threshold was chosen based on the precision recall curve (Figure 1.3B), and corresponds to the p-value at which 5% FDR is obtained.

Short (<120bp) fragment contact maps

We created HiChIP contact maps after filtering interactions to those involving <120bp fragments on both sides. Filtering to the left fragment < 120bp produces a narrower peak centered on the CBS (+) (Figure A.5B), and filtering to the right fragment < 120bp produces a narrower peak centered on the CBS (-) (Figure A.5B).

Comparison of replicates (K562 CTCF HiChIP)

We compared our four K562 CTCF MNase HiChIP replicates (Figure A.3) and found a similar distribution of reads around a CBS (Figure 1.2D, Figure A.3A). We next confirmed that the distribution of the CAMEL statistic is maintained across all four replicates at true positive CBS, defined as CTCF motifs that are unique within a loop anchor and within 30 bp of a CTCF ChIP-

seq peak (Figure 1.3A, Figure A.3B). Finally, we determined that high precision is maintained across all four replicates using precision-recall curves (Figure 1.3B, Figure A.3C).

CTCF ChIP-seq signal comparison to number of short fragments

The fraction of reads overlapping a CBS (+) that are short (<120bp) strongly correlates with the strength of its CTCF ChIP-seq signal. CAMEL-identified CBS (+) were overlapped with CTCF ChIP-seq peaks (ENCFF736NYC.bed). CTCF ChIP-seq signalValue was cut into quartiles and CAMEL-identified CBS (+) more than 200bp from a CTCF ChIP-seq peak were labeled as not having a CTCF ChIP-seq peak. The fraction of short reads (<120bp) was then calculated for each CBS, plotted in a density, and colored by absence of CTCF ChIP-seq peak or CTCF ChIP-seq signalValue quartile (Figure 1.4B).

Estimating cohesin footprints

To estimate the cohesin footprints (Figure 1.4C-F), we made 2D kernel density plots (bivariate normal kernel) with short (< 120bp) fragments overlapping CBS (+). Specifically, we took left fragments in an interaction pair overlapping CAMEL-identified CTCF binding sites (+) (MNase HiChIP, n > 10,000) or CTCF motifs within 30bp of a CTCF ChIP-seq peak (mESC RCMC, n = 65). We then aligned the fragment start and fragment end relative to the start of the 35bp CTCF motif (MA1930.1), so that “0” in the plots corresponds to the start of the 35bp CTCF motif. (We chose this coordinate system to enable easy pinpointing of the fragment end relative to the end of

the motif.) In this way we can then aggregate fragment starts and fragment ends across CBS (+) to observe the overall pattern of fragment starts and fragment ends.

The advantage of using a 2D kernel density for this aggregation is it avoids binning fragment starts and ends into arbitrary boxes (which might misrepresent the underlying signal), but instead obtains aggregated estimates of the individually-smoothed fragment starts and ends evaluated on a grid. Furthermore, density plots have the additional added advantage of automatically normalizing, which enables easy comparison of the relative prevalence of fragment start and fragment end combinations across different conditions or perturbations.

The degree of smoothing along the x (fragment start) and y (fragment end) axes is determined by the bandwidth parameter. The bandwidth parameter re-weights the distance from a specific fragment's start and end to each point (fragment start, fragment end) on the grid to determine the contribution of a fragment to the (fragment start, fragment end) point at which the 2D kernel density is being evaluated. As the bandwidth increases, the 2D kernel density estimate becomes more smoothed (fragments further away from the evaluation point contribute more.) The specific base command we are running to generate these 2D density plots (in R) is *ggplot2::stat_density_2d*, which leverages the *MASS::kde2d* function to obtain 2D kernel density estimates.

The 2D density estimation (as opposed to a 1D view) is key to obtain an accurate depiction of the fragment ends conditional on the fragment starts. Because CTCF has both a 19bp and 35bp

motif, viewing the aggregated and smoothed fragment start, fragment end combinations is crucial to simultaneously viewing both the extended fragment end (due to cohesin) and the extended fragment start (due to the addition of the upstream motif with the core 19bp motif).

Estimating the fully extruded state

K562 CTCF HiChIP

We estimated a genome-wide range for the fully extruded state by obtaining CTCF/cohesin-protected upstream fragments (left fragment in interaction pair) overlapping CBS (+) and estimating the fraction of interaction partners overlapping a downstream convergent negative strand CTCF motif. CBS (+) were required to have at least 50 CTCF/cohesin-protected upstream fragments overlapping the motif to enable sufficient sample size for the motif-specific percent convergent calculation. We then accounted for CTCF occupancy (estimated as ~50%)[33] by dividing this estimate by two. The point estimate (5%) is the number of interaction partners overlapping a downstream convergent negative strand CTCF motif genome-wide / the total number of fragments genome-wide, and the range (1-10%) are the 1st and 99th percentile of the CBS-level CTCF-CTCF chromatin loop estimate. We confirm that these interactions are not predominantly due to random contacts by comparing to an estimate obtained using trans contacts (Figure A.6H). Note that while some fully extruded loop interactions involve short-short fragment ligations, the majority involve one immediately CTCF-adjacent long fragment.

mESC RCMC

We also explored the use of RCMC for estimating fully extruded (CTCF-CTCF) loop fraction using the 150bp mESC RCMC libraries (re-sequenced from Goel *et al.*[20]) covering the *Fbn2* TAD previously studied by live cell imaging in Gabriele *et al*[14]. We obtain short (<120bp) left fragments overlapping the *Fbn2* upstream TAD-anchor CBS (+) (chr18: 57,976,797-57,976,815) and estimate the frequency of interaction partners within 1kb-5kb of the CBS (-) located 505kb downstream (chr18: 58,481,866-58,481,884) at the other end of the TAD. The estimates we obtain (1-5%) are consistent with Gabriele *et al*[14] and the range reflects different window sizes: 1% of fragments are within 1kb of the CBS (-) and 5% of fragments are within 5kb of the CBS (-). Note that these estimates have been divided by two to adjust for the estimated ~50% occupancy of CTCF[33].

Evaluating extruded loop size dependent on chromatin state

We used upstream fragments overlapping CAMEL-identified CTCF binding sites (+) for this analysis. 1 MB regions downstream of the CBS (+) were annotated using ChromHMM[40] to quantify the percentage of bp assigned to each of the 15 chromatin states. To simplify annotation, we grouped the 15 chromatin states into three categories (quiescent, polycomb/bivalent, and active) based on their correlation (Figure A.6A). Regions were clustered using Ward's hierarchical clustering method[57] (Figure A.6A.). For extrusion dynamics analyses (Figure 1.5C, D, E), each of the three chromatin categories was represented by the 20% of regions with the highest fraction of DNA in this state. Extruded loop size was then estimated as the average log10 interaction length for each annotation. Only long range TF-protected fragments (start and end at

least 5bp from motif start and end, length < 120, extended fragment end (fragment end at least 48bp from the 35bp motif start), and interaction length > 10kb) were included in this estimate.

Similarly, high/low H3K27ac corresponds to the top 10% and bottom 10% of the number of basepairs covered by H3K27ac ChIP-seq peaks in the 1 MB regions downstream of CBS (+). High/low RNAPII corresponds to the top 10% and bottom 10% of the number of RNAPII ChIP-seq peaks located in the 1 MB regions downstream of CBS (+). Extruded loop size estimates were obtained in the same way for these annotated regions; long range TF-protected fragments with an extended fragment end were used to estimate the average log10 interaction length.

Directionality of CBS-adjacent nucleosome position signal

Interestingly, the strength of the nucleosome positioning signal is related to the orientation of the DNA contact. Stratifying nucleosome-bound fragments based on whether they are the upstream or downstream long-range (>10kb) fragment in a pair (effectively single-cell left or right loop anchor) produces a differential nucleosome signal inside and outside the loop (Figure A.9). For both upstream and downstream nucleosome-bound fragments, the nucleosome closest to the CTCF binding site and inside the loop exhibits a substantially stronger signal than the closest nucleosome outside the loop. HiChIP ligations are unlikely to fully account for this signal as a previous study using MNase-seq also showed a directional nucleosome preference around CBS (see Figure 1A)[31].

1.5 Data availability

The RCMC and HiChIP raw and processed data generated in this study have been deposited in the NCBI Gene Expression Omnibus (GEO) database under accession code GSE285087 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE285087].

Publicly available K562 ChIP-seq RAD21 BED file (Accession ID: ENCFF330SHG), CTCF BED file (Accession ID: ENCFF736NYC), CTCF bigWig signal value (Accession ID: ENCFF168IFW), RNAPII BED file (Accession ID: ENCFF355MNE), and H3K27ac BED file (Accession ID: ENCFF544LXB) were downloaded from ENCODE, and CTCF motifs were obtained from the R package *CTCF*[34] (annotation record: AH104729 (hg38.MA0139.1), documentation: https://bioconductor.org/packages/release/data/annotation/vignettes/CTCF/inst/doc/CTCF.html). K562 Intact Hi-C was downloaded from GEO (GSE237898). All K562 files downloaded were already aligned to hg38.

Publicly available mESC CTCF (GSM3508478), ATAC-seq (GSE98390), H3K27ac (GSM2417096), RNAPII (GSM6809981), RAD21 (GSE137272), and SMC1A (GSE137272) were downloaded. We use mm10 coordinates for all mESC figures in this paper; files downloaded in other mouse genomes were converted to mm10 using CrossMap. CTCF motifs were obtained from the R package *CTCF*[34] (annotation record: AH104755 (mm10.MA0139.1), documentation: https://bioconductor.org/packages/release/data/annotation/vignettes/CTCF/inst/doc/CTCF.html).

1.6 Code availability

Preprocessing, analysis and figure code used in this paper are available at

https://github.com/aryeelab/cohesin_extrusion_reproducibility

Data figures in this paper were made in R v.4.1.2 using ggplot.

1.7 Acknowledgements

This work was supported by the National Institutes of Health grants RM1HG009490 (MJA, JKJ, CS), R35GM118158 (JKJ), T32GM135117 (CS), and a Career Development Award from the American Society of Gene & Cell Therapy (YET). This work is supported by the Novo Nordisk Foundation NNF21SA0072102 (ASH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the American Society of Gene & Cell Therapy. Dovetail Genomics / Cantata Bio supported data generation costs.

1.8 Author Contributions

C.S. and M.A. conceived the study and developed analysis methods. C.S. performed all bioinformatics analyses. E.T. and M.B. performed the HiChIP experiments. M.B., C.C., H.H., C.E., K.J., E.T., and S.J. contributed to data interpretation. V.G. and A.S.H. provided RCMC libraries and assisted in interpreting the role of cohesin in the study. M.A. supervised the overall work.

Chapter 2

DNA-binding protein footprints reveal cohesin independent long-range looping factors

Abstract

Chromatin loops enable gene regulation by distal enhancers through maintaining structural domains and mediating focal long-range contacts between enhancers and promoters. Changes in long-range enhancer-promoter (EP) contacts and concomitant changes in gene expression have been associated with diverse developmental processes and diseases. While a subset of DNA loops are facilitated by the well-characterized architectural factors CTCF and cohesin, the majority of loops across the genome are resistant to loss of both of these proteins. The underlying mechanisms remain unclear. Here, we develop a novel MicroC transcription factor (TF) footprinting method that enables base-pair resolution identification of DNA-binding events. We apply this approach to examine protein-mediated DNA-DNA contacts with and without cohesin. We find that transcription start site (TSS) contacts are largely maintained upon cohesin loss by short fragments likely representing DNA-binding TFs. We further examine protein-mediated contacts of individual TFs and find that specific non-CTCF transcription factors are able to mediate long-range contacts regardless of cohesin. This suggests a model by which transcription factors located at enhancers and promoters are able to maintain cohesin-independent long-range functional contacts.

2.1 Introduction

Long-range enhancer-promoter contacts enable cell-type specific regulation of gene expression[58–61]. Disruptions in enhancer-promoter chromatin loops can drive pathogenic gene expression changes, resulting in aberrant phenotypes such as limb malformation[2,60] and cancer[3,5,11]. Chromatin loop and topologically associating domain (TAD) structure are maintained by CCCTC-binding factor (CTCF) through its interactions with the looping protein complex, cohesin[1,62]. CTCF and cohesin both enable contacts between distal regions of the genome through their role in chromatin looping and prevent contacts between chromatin located in separate TADs[63,64].

Despite the critical role these proteins play in maintaining and preventing specific enhancer-promoter contacts, degrading cohesin disrupts genome organization but results in gene expression changes at only a handful of genes[65]. A recent study described another layer of genome organization, termed microcompartments, that are cohesin and transcription independent[20]. Microcompartments predominantly occur between promoter-promoter (P-P) and enhancer-promoter (E-P) pairs[20]. Given their resistance to cohesin loss and role in facilitating functional contacts, microcompartments may explain why gene expression persists when cohesin is absent. However, since microcompartments are unaffected by cohesin and transcription, it is unclear what gives rise to their formation.

Chromatin loops and TADs in humans and mice are mediated by the proteins CTCF[63,64] and to a lesser extent YY1[66], which both depend on cohesin to facilitate long-range contacts. Other

organisms such as *Drosophila* have several looping proteins[67], but if there are additional proteins that can enable distal contacts in mice and humans they are unknown. We hypothesized that cohesin-independent E-P and P-P contacts in humans and mice may arise from protein-mediated 3D genome organization by currently unidentified architectural proteins.

To investigate these potential cohesin-independent looping proteins, we developed a statistical method to assess transcription factor occupancy in the MNase-based 3D genome assays Micro-C[68] and Region Capture Micro-C[20]. This method enables protein and long-range contact identification in the same assay without requiring ChIP protein enrichment. Our approach facilitates simultaneous examination of nucleosome and transcription factor spacing at specific loci and in aggregate, and confirms the known strand-specific spacing of nucleosomes around transcription start sites (TSS) while also profiling an enrichment of smaller proteins located slightly upstream of the TSS. This group of TSS-adjacent smaller proteins appears to enable long-range contacts in a cohesin-independent manner. Upon further inspection, we find that a subset of transcription factors including the NF-Y protein complex and KLF/SP transcription factor family are particularly strongly associated with cohesin-independent long-range contacts. This study suggests a potential mechanism by which genome structure (through cohesin-independent transcription factors) orchestrates enhancer-promoter contacts, partially resolving a paradox that has plagued the topology field – why genome structure matters if depletion of cohesin in cells does not largely alter gene expression.

2.2 Results

Short, TF-sized fragments maintain long-range TSS contacts with and without cohesin

We used Micro-C, a chromosome conformation capture (3C) proximity ligation method that leverages MNase-digestion to obtain high resolution chromatin structure[68], to profile genome architecture in mouse embryonic stem cells (mESC) with and without cohesin depletion. MNase has been similarly used to map transcription factor[25,69,70] and nucleosome occupancy[30] due to its endo-exonuclease ability to both cut unprotected chromatin and chew back the DNA until it reaches protein-protected regions, enabling precise identification of protein footprints. Previously, we developed a method that determines CTCF binding in CTCF MNase HiChIP assays, which combines a CTCF ChIP step with the Micro-C protocol to enrich for DNA fragments occupied by CTCF[70]. This approach was limited to identifying the ChIPed protein, CTCF. We sought to expand this method to Micro-C to broadly profile proteins and examine their long-range contacts.

MNase digestion facilitates precise maps of protein occupancy due to its ability to cut and chew back DNA until it reaches a protein. As a result, fragment length and size can be used to infer protein size and location. As expected, nucleosome size (150bp) fragments make up the majority of our observed fragments, but we notably observe a group of smaller fragments as well (Figure B.1). Because nucleosomes protect ~147bp[25,71] of DNA whereas transcription factors protect <120bp of DNA[25], we hypothesized that this group of smaller fragments represents TF-protected DNA, consistent with the relationship between protein size and fragment length in MNase

HiChIP[70]. Fragment length could then be used to distinguish nucleosome-occupied DNA from TF-occupied DNA. Graphing fragment length vs fragment position where each fragment is represented as a segment should size-stratify nucleosomes from TFs. We can then represent each fragment as a point with (x,y) coordinates (middle of fragment, fragment length) and smooth these fragments with a 2D kernel density to obtain regions of high fragment density, likely corresponding to bound proteins (Figure 2.1A). Specifically, we use an axis-aligned bivariate normal kernel and bandwidth (10,10) such that fragments within two standard deviations of the evaluation point fall within 20bp with respect to fragment length and position. Use of the Gaussian kernel ensures that fragments closer to the evaluation point are weighted higher than fragments further away from the evaluation point.

Aggregating fragments across transcription start sites, we observe an enrichment of 150 bp-sized fragments, likely representing nucleosomes, located adjacent to the TSS as well as an enrichment of shorter fragments at the TSS (Figure 2.1B). Stratifying these transcription start sites by gene strand and fragment length, we confirm the strand-specific preferential spacing of nucleosomes around transcription start sites, with the +1 nucleosomes exhibiting the strongest signal. This suggests that we can use fragment density to infer higher protein occupancy. Notably, we also observe an enrichment of short fragments located slightly upstream of the TSS, encompassing a region that is roughly 200bp upstream and 50bp downstream of the TSS. This promoter region furthermore has a peak of ChIP-seq supported transcription factor binding sites (TFBS) located in the same range, suggesting that this enrichment of TSS-proximal short fragments may represent high transcription factor occupancy slightly upstream of the TSS (Figure 2.1C).

Interestingly, if we separate out expressed from unexpressed genes and re-align the fragment coordinates so that x-coordinates reflect a fragment's upstream or downstream location relative to the TSS and gene strand, we observe that expressed genes exhibit strong nucleosome positioning downstream of the TSS and an enrichment of short fragments slightly upstream of the TSS located in the nucleosome depleted region (NDR). Unexpressed genes however have a very different protein distribution, with what appears to be a nucleosome obstructing the TSS and no short fragment enrichment. The TSS-obstructing nucleosome is notably not symmetrically positioned on the TSS, but is instead shifted slightly downstream (Figure 2.1D, Figure B.2).

In addition to the protein occupancy information that can be gleaned from a 1D application of this data, using Micro-C to profile protein footprints has the added advantage of providing interaction partners for each fragment, allowing protein-specific identification of 3D contacts. Graphing the median interaction length of fragments at each position relative to the TSS allows for aggregate distillation of contact patterns across transcription start sites. Using this approach, we find that the group of TSS-proximal short fragments located at expressed genes aligns with a peak in median interaction length that persists when cohesin is degraded, suggesting that transcription-factor sized TSS-proximal proteins may facilitate long-range contacts independently of cohesin (Figure 2.1E).

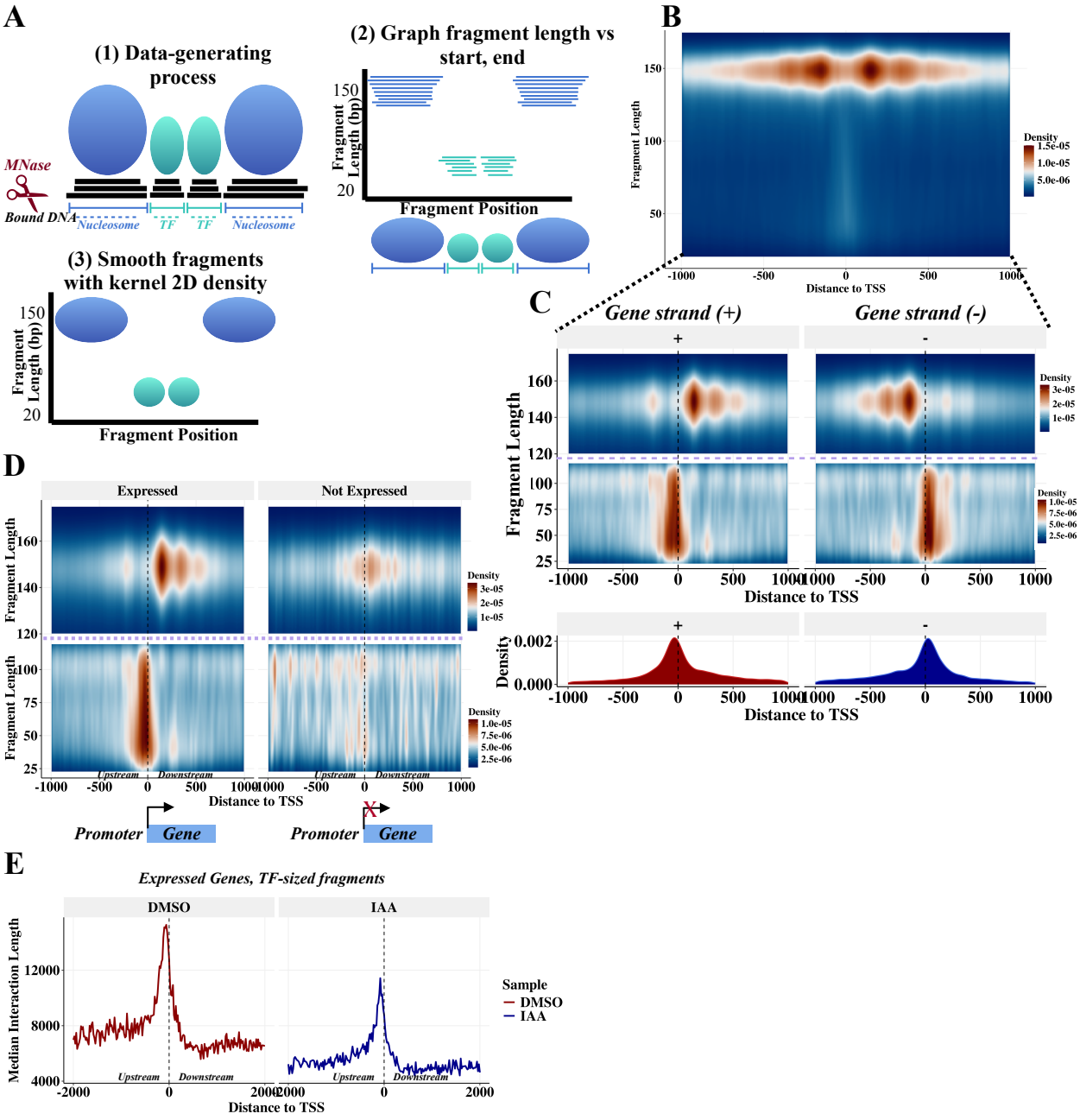


Figure 2.1 Promoter-proximal transcription factors enable long-range cohesin-independent contacts at the TSS of expressed genes. **A** Schematic demonstrating how MNase digestion of protein-bound DNA enables graphical viewing of size and position stratified high density fragment regions, likely corresponding to bound proteins. **B** Fragment length vs fragment position (middle of fragment - TSS position) 2D density for all genes on chromosome one. **C** Plot **(B)** stratified by fragment length and gene strand with aligned UniBind[72] TFBS density track below. **D** Plot **(B)** stratified by fragment length and gene expression (expressed vs not expressed). X-coordinates were re-aligned to reflect fragment position in the upstream or downstream position relative to the TSS and gene strand. (Fragments at genes on the positive

strand coordinates are the same as **(C)**, fragments at genes on the negative strand had their coordinate sign flipped.) Annotation below of expressed and unexpressed genes. **E** Plot **(D)** short fragments (<120bp) at expressed genes median interaction length at binned distance to TSS positions.

Short fragments enable precise TF footprinting

To determine if specific transcription factors can be associated with these cohesin-independent contacts, we developed a statistical method to identify transcription factor binding events using Micro-C data. We noticed that similar to CTCF MNase HiChIP[70], MNase digestion in Micro-C produces a high density of short fragments at CTCF binding sites (CBS) and high density of long fragments adjacent to CBS, representing CTCF binding to its motif and positioning adjacent nucleosomes (Figure B.3A). Similarly, in a locus-specific view of a 1000bp wide CTCF ChIP-seq peak, we observe high density of short fragments at occupied motifs and adjacent spacing of positioned nucleosomes, enabling precise identification of the motifs bound (Figure B.3B).

Expanding beyond looking at just CTCF bound motifs, we can use fragment length and position to examine proteins occupying enhancer-promoter chromatin loops. The *Esrrb* gene, which plays a key role in maintaining pluripotency in mESC[73], contacts a downstream super-enhancer in a cohesin-independent manner (Figure 2.2A). Looking at the *Esrrb* (+) TSS (Figure 2.2B top), we observe a strongly positioned +1 nucleosome (downstream of the TSS) and strong TF binding directly upstream of the TSS, consistent with general trends observed in Figure 2.1C, D. We further observe transcription factor binding inside the gene body that aligns with UniBind[72] true binding sites, although this signal is less strong than the +1 nucleosome and TSS-proximal footprints. Zooming in on one of *Esrrb*'s enhancers located inside the super-enhancer, we

similarly observe transcription factor binding between positioned nucleosomes that align with UniBind[72] TFBS, suggesting that we can use regions of high short fragment density to map bound transcription factors (Figure 2.2B bottom).

Examining the distribution of short and long fragments across all UniBind[72] true binding sites as a function of distance to the motif, we see that in general areas of transcription factor binding are characterized by nucleosome depleted regions (dips in long fragments) and peaks of short fragments at the motif (Figure 2.2C). This suggests that we can use short fragments located at the motif to ascertain TF binding. Assuming that TF binding is a relatively rare event, we can obtain a background distribution for expected number of short fragments at a motif by fitting a negative binomial distribution to short fragments counts across the genome (Figure B.4A, B). Maximum likelihood estimation was used to estimate the parameters. Using a negative binomial instead of a poisson accounts for the large observed variance in short fragment counts (Figure B.4B).

Comparing the distribution of short fragment counts genome-wide to the fitted negative binomial probability distribution function (pdf), we see that the model is correctly specified (Figure B.4B). Similarly, comparing the distribution of short fragment counts at unbound motifs of different transcription factors to the fitted negative binomial pdf, we observe that unbound motifs align with the fitted null distribution (Figure B.4C). Bound motifs[72] (as identified by UniBind) diverge from the fitted negative binomial pdf and short fragment distribution at unbound motifs (Figure B.4C, D). Obtaining p-values for bound and unbound motifs using the fitted negative binomial distribution produces unbound motifs with predominantly large p-values and bound motifs with

predominantly small p-values, suggesting that we can distinguish between bound and unbound motif sequences (Figure B.4E). We use Benjamini-Hochberg to control the false discovery rate (FDR), as p-values in regions not bound by transcription factors follow a uniform distribution. Transcription-factor specific precision-recall curves further confirm that we can identify bound motifs (Figure 2.2D).

Using true positive ChIP-seq supported bound motifs[72] also identified as bound by our method, we can examine transcription-factor specific protein behavior, specifically with regards to nucleosome positioning around bound sites (Figure 2.2E, F). As expected, CTCF[31] and REST[74] position multiple nucleosomes around their binding sites. Notably, other transcription factors such as KLF4, NRF1, NF-YB, and MYC also exhibit this nucleosome-positioning ability, with the size of the NDR varying by protein. MYC in particular has the smallest NDR, whereas NF-YB has the largest (Figure 2.2F). This may partially be a reflection of the larger observed size of NF-YB's transcription factor footprint (Figure 2.2E), likely due to the larger space occupied by the NF-Y protein complex that NF-YB forms with additional subunits NF-YA and NF-YC[75,76].

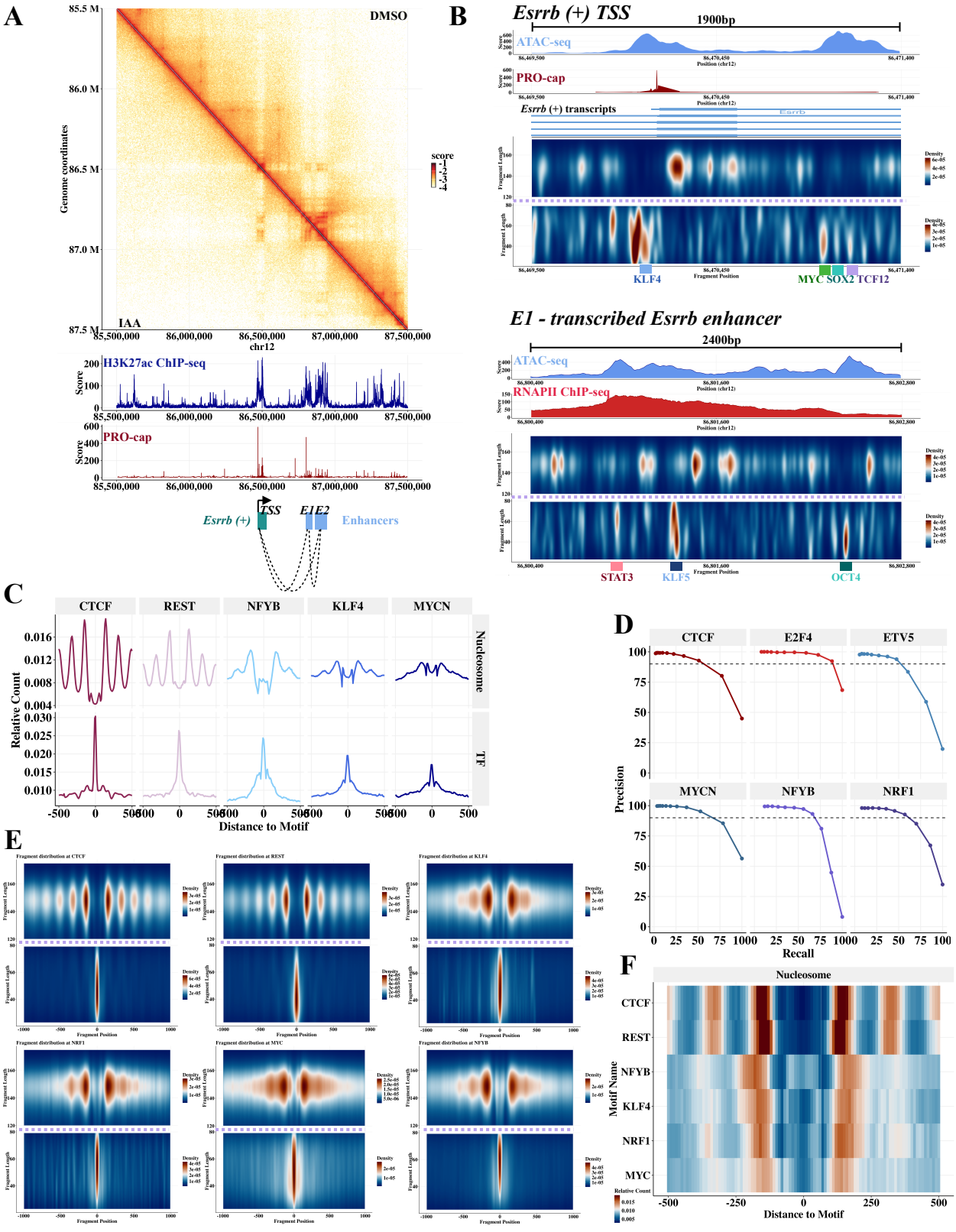


Figure 2.2 Footprinting transcription factors in MicroC enables base-pair resolution of locus-specific transcription factor binding and aggregate examination of protein behavior. **A** Balanced

contact map at *Esrrb* locus (4kb resolution) with H3K27ac ChIP-seq, PRO-cap, and *Esrrb* TSS + enhancer element annotations below. DMSO (control) perturbation is above the diagonal, IAA (cohesin depletion) perturbation is below the diagonal. **B** *Esrrb* TSS and *Esrrb* enhancer element 2D density plots with annotations from publicly available ATAC-seq, PRO-cap, and RNAPII ChIP-seq as well as TFBS locations[72]. **C** Transcription factor specific relative frequency of nucleosome fragments (>120bp) and TF fragments (<80bp) as a function of distance to TFBS[72]. **D** Precision-recall curves with true positive set as UniBind TFBS[72] and true negative set as transcription factor motifs at least 200bp away from a TFBS. **E** Aggregate plots of true positive TFBS[72] also identified by our method. **F** Plot **(E)** relative count of nucleosome-sized fragments surrounding TFBS reveals varying NDR size by protein.

Cohesin-independent transcription factors likely maintain enhancer-promoter contacts when cohesin is degraded

Examining re-sequenced Region Capture Micro-C (150bp reads)[20,70] at the *Klf1*-locus microcompartment identified by Goel *et al.*[20], we confirm that the microcompartment structure is maintained when cohesin is degraded (Figure 2.3A). Subsetting the pairs file to fragments where both fragments are nucleosome-sized (>120bp, "long fragments") or at least one fragment is TF-sized (<120bp, "short fragments"), it is clear that the short fragment contact map makes the microcompartment structure more distinct (Figure 2.3B) whereas the long fragment contact map makes the microcompartment structure fade (Figure 2.3C). These TF-sized fragments that enrich the contact map for long-range contacts are unlikely to represent CTCF binding events, as the cohesin depletion (below diagonal) contact map also shows this effect (Figure 2.3B).

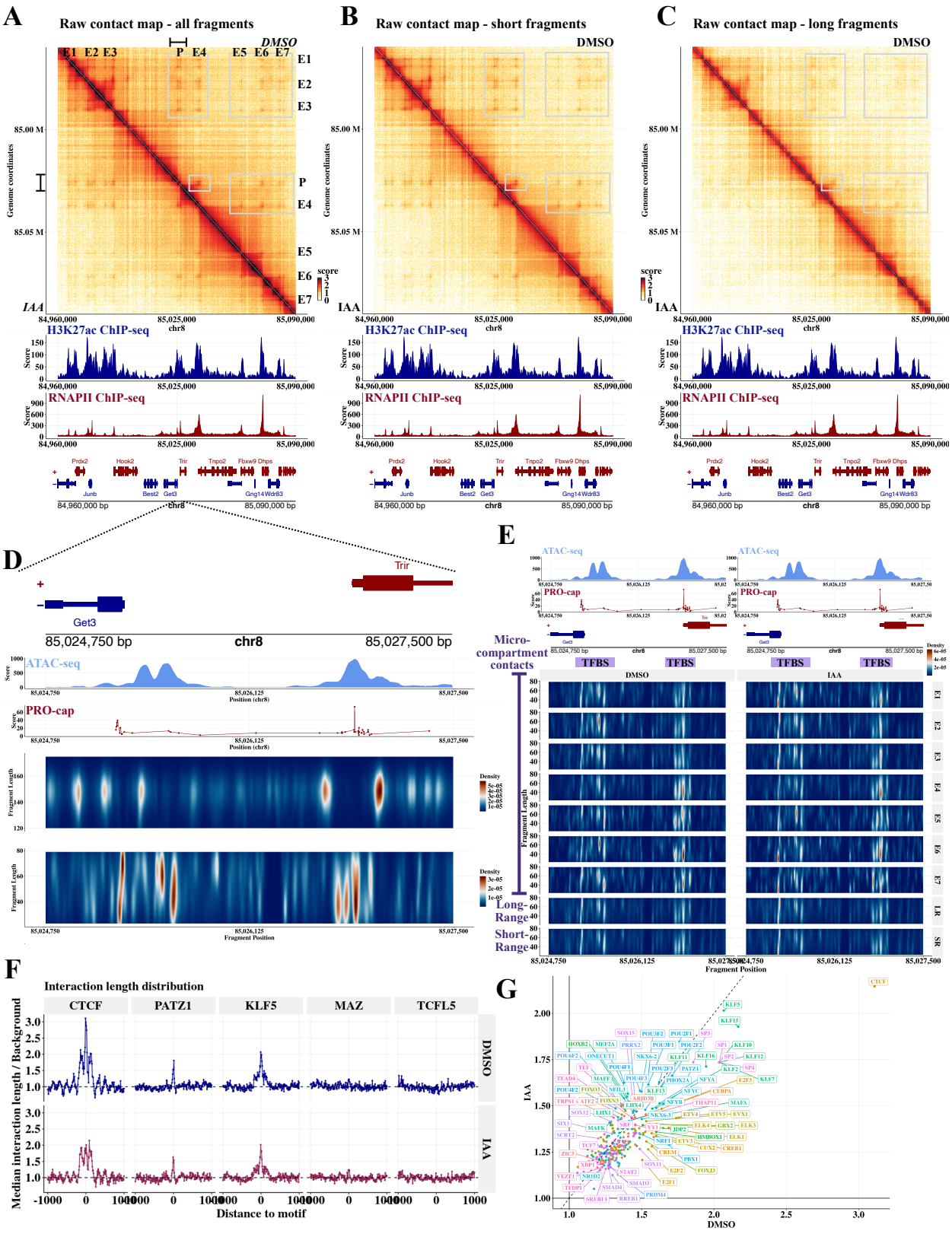

A
Raw contact map - all fragments
DMSO
IAA
E1 E2 E3 P E4 E5 E6 E7
Genome coordinates
85.00 M
85.05 M
score
H3K27ac ChIP-seq
RNAPII ChIP-seq
chr8
84,960,000
85,025,000
85,090,000
B
Raw contact map - short fragments
C
Raw contact map - long fragments
D
Get3
Trir
85,024,750 bp
chr8
85,027,500 bp
ATAC-seq
PRO-cap
Fragment Length
Fragment Position
E
ATAC-seq
PRO-cap
Micro-compartment contacts
TFBS
DMSO
IAA
Long-Range
Short-Range
F
Interaction length distribution
CTCF
PATZ1
KLF5
MAZ
TCFL5
Median interaction length / Background
Distance to motif
G
DMSO
IAA

Figure 2.3 Cohesin independent looping factors maintain enhancer-promoter contacts upon cohesin depletion. **A** mESC RCMC Contact map (log10 scale, counts) at 250bp resolution for all fragments. DMSO (control) is above the diagonal, IAA (cohesin depletion) is below the diagonal. Enhancer elements (E1:E7) and *Get3* / *Trir* promoter (P) are labeled and publicly available gene annotations, H3K27ac ChIP-seq, and RNAPII ChIP-seq are included. **B** Plot **(A)** filtered to short fragments (at least one fragment in a pair < 120bp). **C** Plot **(A)** filtered to long fragments (both fragments in a pair > 120bp). **D** mESC RCMC 2D density plot of *Get3 (-)* and *Trir (+)* promoter region with gene, ATAC-seq, and PRO-cap annotations. **E** Plot **(D)** faceted by perturbation and interaction partner location - E1:E7 are the enhancer regions annotated in Plot **(A),** "LR" (long-range) are fragments with an interaction partner at least 10kb away that is not in E1:E7, and "SR" (short-range) are fragments with an interaction partner less than 10kb away. **F** mESC Micro-C median interaction length at each position/ median interaction length in the +/- 1kb region (background) as a function of distance to the motif, faceted by TF motif and perturbation. **G** Plot **(F)** interaction length / background value at the motif in IAA (cohesin depletion) vs DMSO (control) perturbations for mESC expressed transcription factors. y = x dotted line included to make comparison between samples easier.

We posited that this relationship between TF-sized fragments and long-range cohesin-independent contacts observed both in Micro-C at the TSS (Figure 2.1E) and in Region Capture Micro-C at microcompartment contacts (Figure 2.3B) may indicate cohesin-independent looping factor ability of unknown transcription factors. Zooming in on the highly expressed *Get3* (-) and *Trir* (+) TSSs (marked by brackets and the letter "P" in Figure 2.3A), which have several cohesin-independent microcompartment contacts with other functional areas in this region, we notice the characteristic nucleosome positioning adjacent to and NDR at both TSSs (Figure 2.3D). The nucleosome depleted areas contain high density regions of short fragments, representing TF binding. Interestingly, when faceting the TF-sized (<80bp) fragments by location of their interaction partner, a pattern emerges where fragments with interaction partners in microcompartment regions disproportionately emanate from TSS-proximal TFBS in comparison to fragments with short-range or long-range (not microcompartment) contacts (Figure 2.3E). This

pattern does not go away when cohesin is depleted, suggesting a potential role of TFBS in maintaining cohesin-independent microcompartments.

To determine if specific transcription factor families had particularly strong associations with long-range interactions, we aggregated fragments around accessible sequences with strong matches to a particular transcription factor's motif and plotted the median interaction length at each position / the background median interaction length to evaluate whether there is a peak above background at the motif (Figure 2.3F). As expected, CTCF has a large peak in interaction length at its motif that is attenuated when cohesin is depleted. Surprisingly, other transcription factors like PATZ1 and KLF5 also had peaks in interaction length at their motifs but these peaks stayed consistent when cohesin was degraded, suggesting that transcription factors in the KLF/SP/PATZ1 motif family may maintain cohesin-independent interactions. Notably, not all transcription factors had a peak at their motif - the MAZ and TCFL5 motifs for example contained no enrichment at the motif for long-range interactions in DMSO or IAA perturbations (Figure 2.3F).

To ascertain whether other transcription factors may also have a peak in interaction length at their motifs, we plotted median interaction length at the motif / background in IAA (cohesin depletion) and DMSO (control) conditions on a scatterplot for expressed transcription factors (Figure 2.3G). CTCF is far off the y=x line on this graph, suggesting that it mediates long-range interactions but is highly cohesin dependent. Notably, the KLF/SP motif family and NF-Y motif

family are close to the diagonal, suggesting proteins in those motif families may be able to mediate long-range, cohesin-independent interactions.

2.3 Discussion

The only known looping factors in mice and humans, CTCF and YY1, mediate chromatin looping through their interaction with cohesin via the loop extrusion model[1,64,66]. However, while degrading cohesin reduces contacts between CTCF binding sites, many long-range enhancer promoter contacts persist and gene expression stays largely the same[65]. It is unclear by what mechanism these microcompartment contacts are maintained, since they do not depend on cohesin or transcription[20].

While these EP contacts may not depend on transcription itself, we find that the region upstream of the TSS associated with high amounts of TF binding appears to enable long-range interactions even without cohesin (Figure 2.1E). This area of high TF occupancy and nucleosome depletion is associated with gene expression (Figure 2.1D, Figure B.2). This suggests that it may not be transcription that maintains these interactions, but instead the transcriptional machinery that predates transcription. The triptolide treatment that was used to determine that microcompartments do not depend on transcription[20] would not prevent formation of the high density TFBS and NDR at the TSS, but would instead prevent the downstream process of transcription initiation.

We next sought to determine if specific transcription factor motif families are associated with longer interactions independently of cohesin. To accomplish this goal, we developed a statistical method that identifies transcription factor binding sites in Micro-C. We then dove deeper into the *Klf1* RCMC microcompartment locus, which has a highly nested structure that is resistant to cohesin loss[20]. Interestingly, we observed that filtering the contact map to long fragments caused the microcompartment structure to visibly fade, whereas filtering the contact map to short fragments enhanced the microcompartment structure (Figure 2.3A-C). Zooming in on the *Get3* and *Trir* TSSs, which act as a cohesin-independent hub of functional contacts in this microcompartment structure, we noticed that their TSS-proximal TFBS are associated with cohesin-independent microcompartment contacts with long-range enhancer elements (Figure 2.3E). This further provided support that transcription factors may mediate cohesin-independent EP contacts.

To test if specific motif families facilitate these longer interaction lengths, we investigated the long-range contact enrichment at motifs relative to adjacent regions in DMSO (control) and IAA (cohesin depletion) perturbations, and noticed that some motif families appear to be better at mediating long-range interactions than others. CTCF ranked the highest on the list, but notably lost its ability to facilitate long-range interactions when cohesin was degraded. Other motif families however, such as the KLF/SP motif family and NF-Y motif family, seem to have longer interactions at their motifs that are maintained when cohesin is lost (Figure 2.3G). Our proposed KLF/SP and NF-Y looping abilities are consistent with recent papers that have found KLF4 depletion reduces EP looping[77] and that the NF-Y complex promotes chromatin looping in plants[78,79].

This suggests a model by which cohesin-independent transcription factor families may maintain long-range EP contacts without cohesin, partially accounting for the lack of major gene expression changes when cohesin is degraded by maintaining TSS contacts with distal enhancers. It is still unclear how these transcription factor motif families may maintain these contacts - whether this is through formation of protein complexes and condensates, consistent with a phase separation model of EP contact maintenance[80,81] or some other mechanism. Furthermore, it cannot be ruled out that the relevant EP contact maintaining transcription factor families may vary between cell types, or that it may be the aggregate behavior of many TFs coming together rather than the individual behavior of a single TF that enables these cohesin-independent distal contacts. However, it is at least clear from this study that looping factors other than the frequently studied CTCF architectural protein likely contribute to fine-scale genome architecture without relying on cohesin.

Chapter 3

Regulatory architecture of transcriptional assemblies at nanoscale resolution

Abstract

Actively transcribed genetic loci coalesce into biopolymer-rich assemblies containing genes and their regulatory elements. To gain insights into the mechanisms that structure these assemblies, we developed a method to define transcription factor (TF)-enhancer assemblies at base-pair resolution and used it to study an active oncogene locus. We find that TF-enhancer assemblies consist of clusters of TFs that co-occur with nascent RNA molecules. These TF-enhancer assemblies interact with the oncogene promoter and depend on RNA for their stability. The approach described here reveals how TFs populate enhancers at base-pair resolution, indicates that TFs and nascent RNA molecules together govern the relationship between gene regulation and genome structure, and provides a foundation for high-resolution maps of regulatory architecture across the genome.

3.1 Introduction

Genome architecture is highly influenced by gene regulation[82–85]. At thousands of actively transcribed protein-coding genes, regulatory elements assemble large numbers of transcription factors, cofactors and chromatin regulators that coalesce and drive various structural features of these loci. Similar structuring events occur at thousands of actively transcribed noncoding loci[86,87]. We do not yet have high-resolution models of these assemblies, which limits understanding of gene regulation and genome architecture.

Foundational models of transcriptional regulation involve the assembly of protein complexes at the enhancer and promoter elements of genes[88–90]. Enhancers are DNA elements that influence the level of transcription of target genes, often from considerable genomic distances, and are bound by collections of transcription factors (TFs) together with cofactors and RNA polymerase II (Pol II)[83,91–96]. These enhancer-bound assemblies modulate transcription by interacting with additional transcription apparatus at gene promoters[52,97–103]. Over one million candidate enhancers have been identified in the human genome[104] and the importance of these elements is emphasized by the observation that most disease-associated genome variation occurs in these elements[104–108]. There are remarkably few models that depict precisely how TFs are assembled on human regulatory elements, although some studies have made pioneering attempts in this direction[27,109–112]. Furthermore, TF-enhancer models have yet to incorporate the recent insight that most mammalian TFs have functionally important interactions with both DNA and RNA[113].

Many genes are regulated by multiple enhancers[114–116], and multiple enhancer-bound complexes can coalesce into transcriptional compartments containing DNA, RNA and transcription apparatus[116–123]. This linkage between gene regulation and local genome architecture was highlighted recently by results from high resolution DNA contact mapping. Region Capture Micro-C (RCMC), which enables mapping of 3D contacts among genomic loci at high resolution, and protein occupancy data suggest that local genome architecture is likely a product of the coalescence of apparatus at enhancers and promoters[20,22,85]. Here, we develop an approach that extends RCMC by integrating 3D architectural chromatin mapping with high-resolution TF footprinting. This approach enables us to deduce the contact structure and footprint landscape of an oncogene-associated transcriptional assembly and learn how individual TFs populate regulatory elements at base-pair resolution.

3.2 Results

Regulatory elements dominate the assembly architecture of an active locus

We sought to create a map of a human genetic locus at unprecedented resolution using a method that captures both nanoscale 3D chromatin architecture (Figure 3.1A) and the footprints of DNA binding factors. This method is based on RCMC, a technique that generates exceptional read depth for DNA fragments that are protected from micrococcal nuclease (MNase) by proteins that are bound to DNA (Figure C.1A)[20,22]. RCMC has been used previously to identify microcompartments, which are 3D genome structures that appear as "grid-of-dots" patterns in contact maps, and whose formation is thought to occur through coalescence of transcription apparatus assembled at enhancers and promoters. Thus, RCMC should reveal the 3D genome structure within the transcriptional assemblies that occur at an active locus. We selected the *MYC* locus for study because the *MYC* oncogene is regulated by clusters of enhancers within a "gene

desert"[11], offering the opportunity to deduce the 3D architecture of regulatory elements for a single gene at exceptionally high resolution, and to link this architecture to assemblies of TFs throughout the locus.

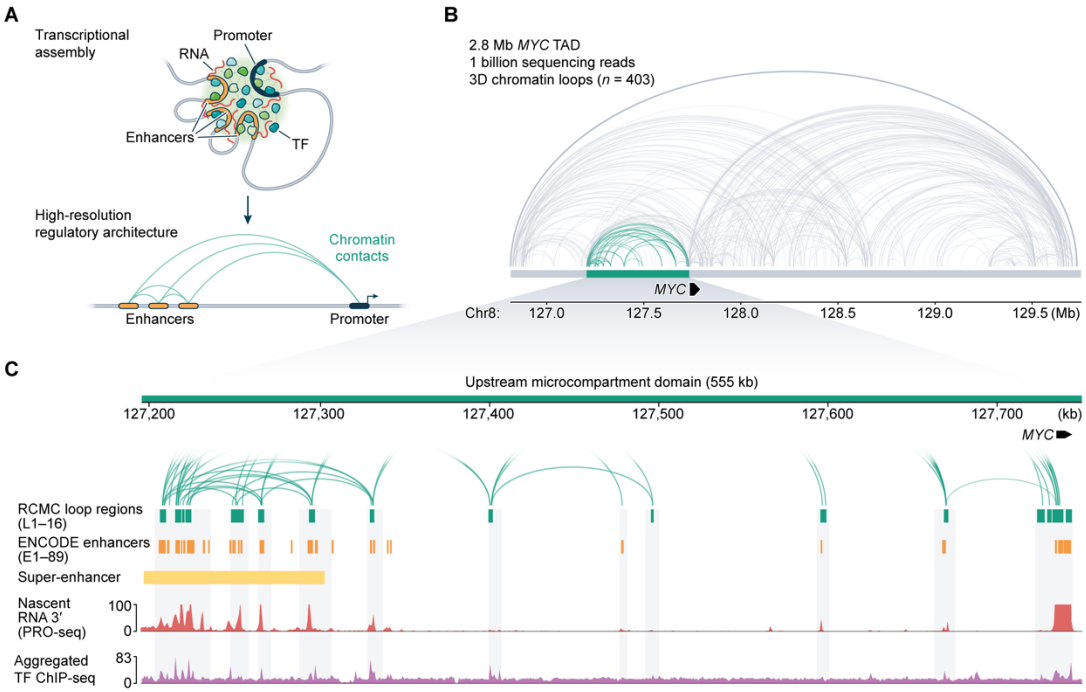


Figure 3.1 Regulatory elements dominate the assembly architecture of the 2.8 Mb *MYC* locus. A Compartments formed by assemblies of transcription factors and transcription apparatus can be probed by Region Capture Micro-C (RCMC) to produce high-resolution contact information for the regulatory architecture of a locus. **B** RCMC data from the 2.8 Mb topologically associated domain (TAD) containing the *MYC* oncogene in HCT 116 cells shows 3D chromatin contacts that produce a highly nested structure of DNA loops (green and gray arcs). A microcompartment domain upstream of the *MYC* gene is illustrated by the green bar and green arcs. Statistically significant RCMC loops are shown ($p<0.01$). **C** RCMC loop regions largely coincide with annotated regulatory elements and nascent RNA data. Shown are annotations of RCMC loops (green arcs), loop regions (green bars), ENCODE enhancers (orange bars), a super-enhancer (yellow bar), nascent RNA signal from PRO-seq reads of RNA 3' ends (red track), together with the aggregated signal from 18 TF ChIP-seq datasets (purple track).

We performed RCMC on the 2.8 megabase (Mb) topologically associating domain (TAD) harboring the *MYC* oncogene in human HCT 116 colorectal cancer cells, and obtained over one

billion sequencing reads that resulted in the identification of 403 DNA loops in the locus (Figure 3.1B, C.1B). The map of these loops revealed that multiple nested compartments surround the *MYC* gene (Figure 3.1B, C.1B, C.2A). In all tumor cells driven by *MYC* expression, large clusters of enhancers, called super-enhancers, occur within the 2.8 Mb locus[11,107,124]. In HCT 116 cells, the RCMC data shows that the *MYC* promoter occurs in a 555 kb microcompartment domain together with a super-enhancer upstream of the gene (Figure 3.1C). The RCMC loop regions, defined as regions containing multiple significant loop anchors, highly overlapped regions identified as enhancers by the ENCODE consortium ($P<0.0001$; binomial test) (Figures 3.1C, C.1C, C.2B, Table C.1). TFs and nascent RNA species were almost universally associated with these enhancer elements (Figure 3.1C), consistent with evidence that enhancers are generally actively transcribed[125–127]. The constituent enhancers in the super-enhancer displayed contacts with one another, as well as with the *MYC* promoter, consistent with evidence that the transcription apparatus bound to these regulatory elements can coalesce to form a transcriptional condensate[120,123]. Chromatin lacking these regulatory elements showed little evidence of long-range interactions or nascent RNA, but was engaged in short-range interactions (Figure C.3), which may be due to some degree of folding into the assembly. We conclude that transcriptionally active assemblies at the *MYC* locus are dominated by contacts among the components of enhancer and promoter elements, which include TFs and nascent RNAs.

TF assemblies at regulatory elements

We next sought to identify TFs that occupy regulatory elements within the *MYC* transcriptional assembly at base-pair resolution. To do this, we used a combination of protein footprint data derived from RCMC, together with chromatin immunoprecipitation sequencing (ChIP-seq) data and DNA sequence motif data for TFs (Figures 3.2A, C.4A, C.4B). The protein footprint data is a

product of protein-mediated protection of short lengths of DNA from MNase digestion. Binding of one or more TFs is expected to protect ~20-80 bp fragments of DNA[24,70,128,129], and when multiple overlapping protected DNA fragments are sequenced, the local enrichment of fragment density produces a recognizable footprint in 2D fragment density maps. Boundaries for the footprints were identified by "segmenting" these maps (Methods). The TFs that occupy specific footprint sites can be predicted from TF ChIP-seq data, and their precise sites of binding can be inferred from known TF binding site motifs. This allows us to produce models for the TFs assembled at enhancers, which we term TF-enhancer assemblies. These models are obtained from populations of cells, so likely reflect a composite view of the more frequent and stable molecular interactions.

TF footprints, determined as described above, occurred within nearly all ENCODE enhancers in the 555 kb compartment containing the *MYC* promoter (146/157), but also extended beyond these elements (Figure 3.2B, 3.2C, C.5, C.6). For example, 5 footprints were found to span the ENCODE enhancer E36, and 6 additional footprints occur within the larger 1 kb segment within RCMC loop region L5 (Figure 3.2C, Table C.2). The 5 footprints within the enhancer had evidence for binding by more TFs than those outside the enhancer based on ChIP-seq data and presence of motifs (Figure 3.2C). Throughout this locus, footprints in enhancer-associated RCMC loop regions were bound by larger numbers of TFs than footprints in surrounding DNA (Figure C.7). It thus appears that DNA sequences surrounding enhancers can serve as binding sites for some TFs, but not at a threshold density that produce enhancer-associated signals or DNase hypersensitivity, both metrics used by ENCODE to identify enhancers[130].

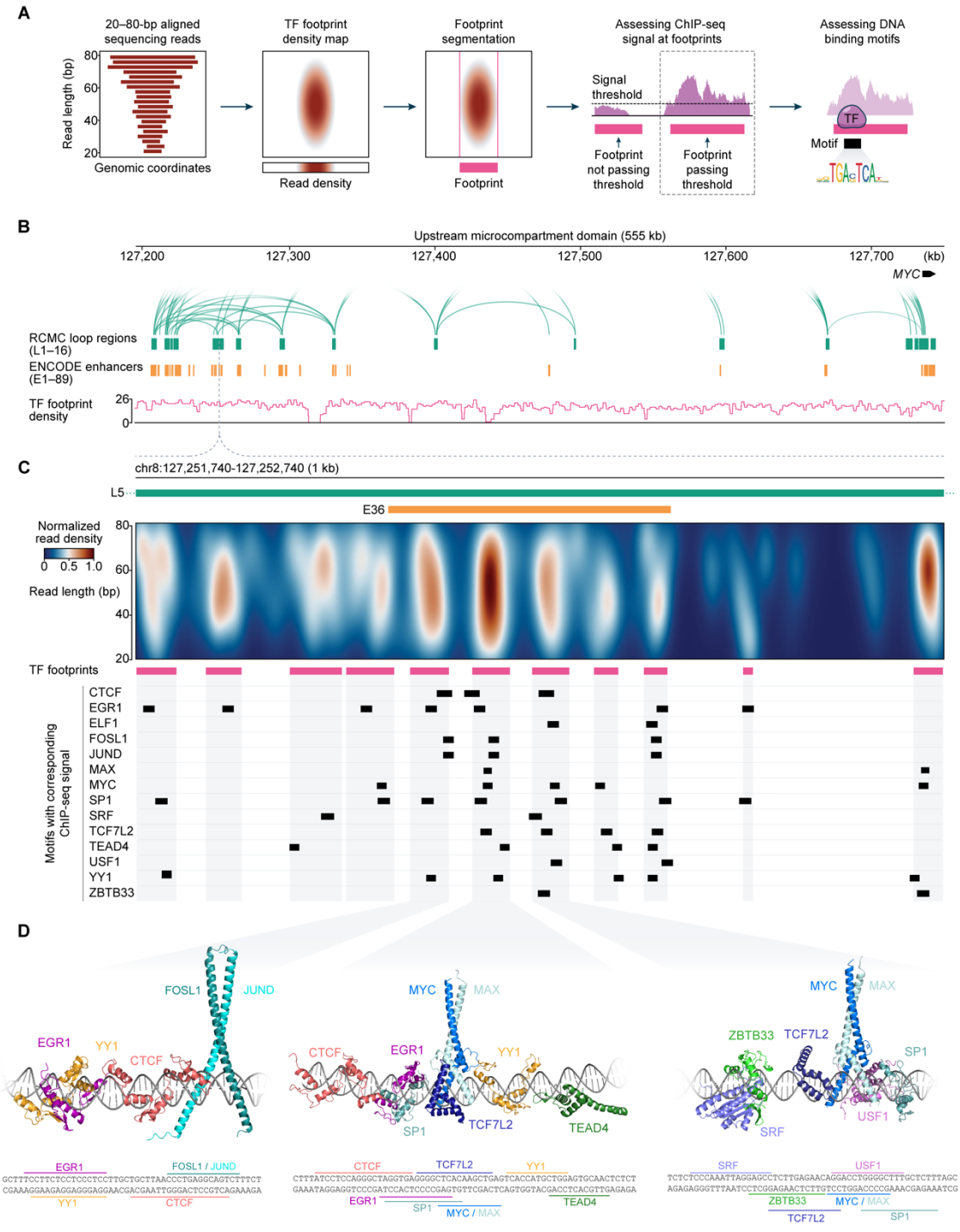


Figure 3.2 TF footprinting reveals TF nano-assemblies in regulatory elements at base-pair resolution. A Schematic of a method for footprinting transcription factors (TFs) using short (20-80 bp) RCMC sequencing reads. TF ChIP-seq data and TF DNA binding motifs were used to assign specific TFs to footprints and to sites within these footprints (see Methods). **B** Contact map

of the upstream microcompartment domain containing *MYC* enhancers and promoter, together with TF footprint density averaged over 2 kb windows, showing that protected segments occur throughout the locus. **C** TF binding locations deduced for footprints within the RCMC loop region L5 (green bar) of the upstream microcompartment domain, centered on the enhancer E36 (orange bar). The x-axis of the footprint density map represents the coordinate of a protected fragment center and the y-axis indicates the length of the protected DNA fragments. The relative density of 20-80 bp fragments is indicated by the color scale. A binary representation of detected footprints (pink) sits below the footprint density map. TF DNA binding motifs (black) within footprints are shown for TFs whose binding at the footprint was confirmed by ChIP-seq and where motif sequences were determined to meet a significance threshold (see Methods). The density of TF occupancy is generally higher at footprints within annotated enhancer elements. As illustrated here, TFs in the 5 footprints that occur within the E36 enhancer would be defined as an enhancer TF nano-assembly. **D** Structural models for TFs in three of the five E36 enhancer footprints. Footprints are large enough to structurally accommodate multiple TFs, as shown. The models consist of a subset of the TFs listed in **(C)**, docked on their motifs using available co-crystal structures (for EGR1, YY1, MYC, MAX, TEAD4, SRF, ZBTB33, and USF1), published docked crystal structures (for SP1), or TF-DNA structures generated by AlphaFold3 (for FOSL1, JUND, CTCF, and TCF7L2). Some TFs were not displayed in the model due to steric constraints. Intrinsically disordered regions (IDRs) comprise substantial portions of the total amino acid content of TFs but are not displayed in these models. DNA sequences and locations of TF DNA binding motifs are shown beneath each assembly.

The average enhancer in the microcompartment domain containing the *MYC* promoter (Figure 3.2B) is 273 bp, the average TF-enhancer assembly has 3 footprints of 49 bp, and each footprint is separated by approximately 60 bp. The structures of the DNA-binding domains of many of these TFs have been solved, which enables modeling of their binding within footprinted regions (Figure 3.2D, Figure C.8). The results of this analysis indicate that TF-enhancer assemblies consist of multiple clusters of TFs (TF nano-assemblies) that span these regulatory elements and produce initial models of TF occupancy across enhancers at base-pair resolution.

TF-enhancer assemblies facilitate variant-to-function study

Models of TF-enhancer assemblies such as those described here should provide a foundation to investigate whether disease-associated sequence variation affects binding by specific TFs, and thus suggest mechanistic hypotheses linking genetic variants to biomolecular function. The vast majority of disease-associated human genetic variation has been reported to occur in proximity to enhancers[105–108,131]. Although thousands of noncoding variants have been identified, the effects of variation on TF interactions with enhancers have been determined for only a small minority, and thus the mechanistic effects of variation have only occasionally been ascertained[132–139].

To explore the possibility that the TF-enhancer assembly models generated here might provide such insights, we interrogated disease-associated single nucleotide polymorphisms (SNPs) (Figure C.9A) and focused on those associated with colon cancer in the *MYC* locus[140]. The highest scoring risk SNP (rsID: rs6983267) occurs in a TCF7L2-occupied footprint in HCT 116 colorectal cancer cells within a larger TF nano-assembly (Figure C.9B); this SNP was previously shown to increase the binding affinity of TCF7L2 and because this TF is bound by β-catenin, could serve to enhance *MYC* expression due to dysregulated β-catenin signaling[138] (Figure C.9C). We identified 19 variants that are high scoring cancer-associated SNPs in this locus (rsIDs: rs6983267, rs7013278, and rs4733767) or SNPs in high linkage disequilibrium (r^2>0.8) with these three top scoring SNPs, and found that 7 of these (37%) occurred within TF nano-assembly footprints (Figure C.9D). Because these assemblies show evidence of contacts with the larger assembly that forms at the *MYC* locus, it is reasonable to postulate that these variants influence regulation of the assembly. Conventional methods, including DNase footprinting, ATAC-seq, and ENCODE cCREs, were far

less successful at identifying these sites (Figure C.9B, C.9D). These observations suggest that defining TF nano-assemblies as described here will prove to be valuable for elucidating the mechanistic impact of disease-associated variation in regulatory elements.

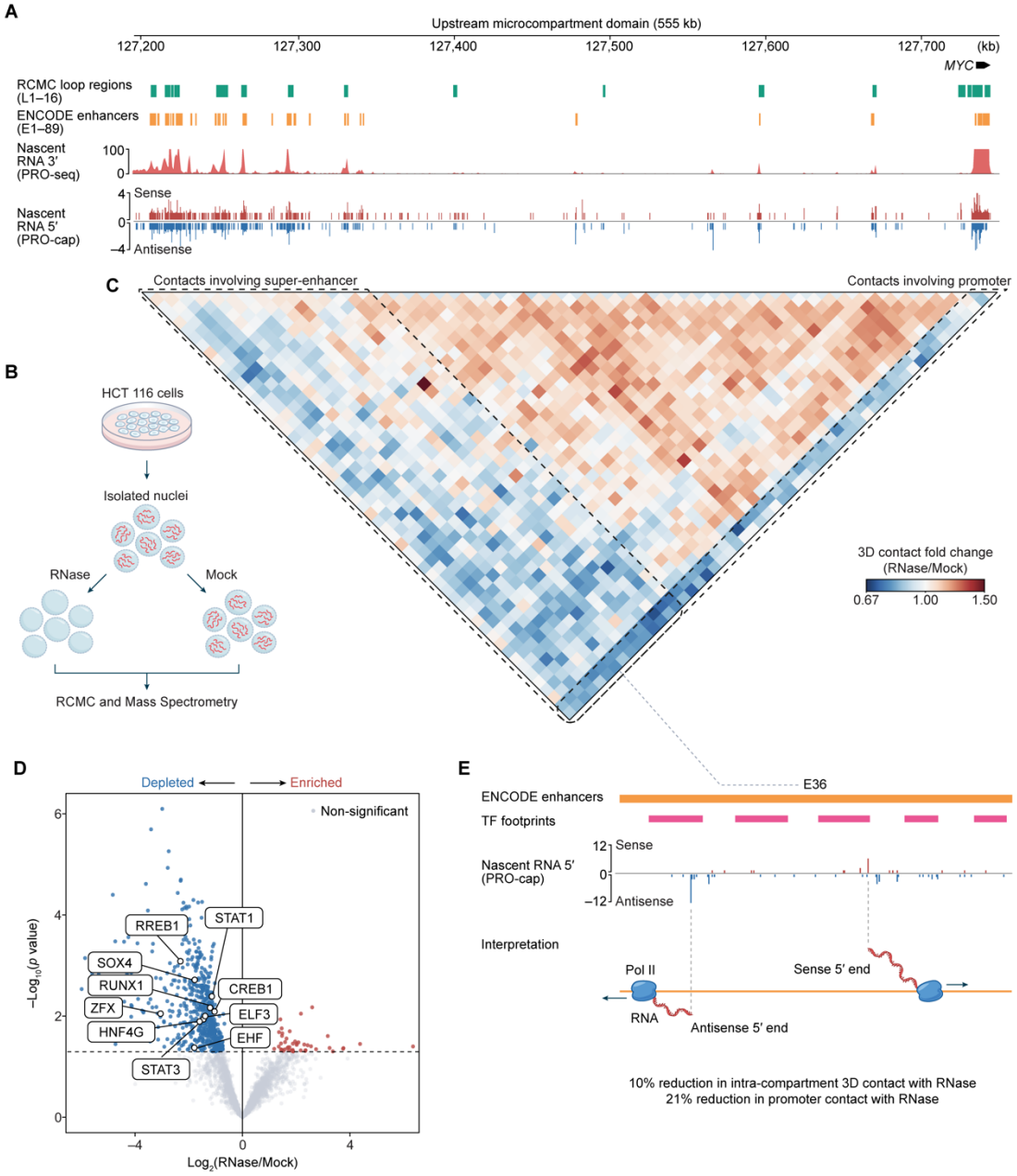


Figure 3.3 RNA-associated nano-assemblies underlie regulatory architecture. A Association between nascent RNA and 3D contact signals. Tracks display, from top to bottom, RCMC loop regions (green), ENCODE enhancers (orange), 3' ends of nascent RNA (PRO-seq, light red), 5' ends of nascent RNA (PRO-cap, red plus strands, dark blue minus strands). **B** Experimental

schematic for assessing the dependence of 3D chromatin architecture and TF association with chromatin on nuclear RNA by treating isolated nuclei with an RNase A/T1 cocktail ("RNase") or the enzyme carrying buffer alone ("Mock"). **C** RNA-dependent chromatin architecture revealed from RNase treatment. Difference between RNase and Mock-treated RCMC contact maps for the upstream *MYC* microcompartment domain. Color bar represents the fold change between RNase and Mock reads (without ICE balancing), where blue corresponds to diminished contact signal after RNase treatment and red corresponds to increased signal. Bin size = 10 kb. **D** Volcano plot of protein abundance on isolated chromatin detected by mass spectrometry (MS) following RNase treatment relative to Mock treatment. The abundance of each protein detected by MS is normalized to the abundance of histone H4 in each technical replicate (n=4). Negative values indicate a depletion of a protein from the chromatin fraction, while positive values indicate an enrichment of a protein in the chromatin fraction. Unadjusted p-values were calculated using an unpaired Student's t-test (dashed line at p=0.05). **E** Nascent RNA signals at ENCODE enhancer E36. Top to bottom: enhancer boundaries (orange), TF footprints (pink), 5' ends of nascent RNA (PRO-cap, red plus strands, dark blue minus strands), and graphical interpretation of bidirectional enhancer transcription. Reduction in 3D contact is calculated for the 10 kb bin including enhancer E36 across the microcompartment domain shown in (C) (intra-compartment fold change median = 0.897, standard deviation = 0.065) and for the *MYC* promoter (promoter fold change = 0.785).

RNA-associated nano-assemblies underlie regulatory architecture

Enhancer RNA molecules are likely to be integral components of TF-enhancer assemblies. Most enhancers are transcribed, and enhancer transcription appears to be among the best indicators of an enhancer's ability to activate genes[125–127,141–144]. Most TFs are RNA-binding proteins and RNA binding can stimulate assembly of TFs on enhancers (Henninger et al., 2021) and increase the time they are associated with these elements[113,145]. Analysis of nascent RNA transcription data revealed that nascent RNA molecules are present at nearly all enhancer sites in the larger *MYC* assembly (Figure 3.3A, Table C.3), suggesting the hypothesis that RNA may be necessary both for optimal TF occupancy and for optimal assembly of the *MYC* compartment.

The observation that regulatory elements are densely occupied by TF assemblies, are prominent sites of nascent RNA transcription, and that TFs are RNA-binding proteins (Figure 3.3A) led us to hypothesize that depletion of RNA would result in depletion of 3D chromatin interactions in cells. To test this hypothesis, we conducted RCMC after depletion of nuclear RNA. Isolated nuclei from HCT 116 cells were treated with a cocktail of RNase A and RNase T1 ("RNase") or buffer alone ("Mock") for 30 minutes (Figure 3.3B, C.10A). RNase treatment diminished long-range RCMC contact signals between regulatory sites, with the most prominent effects occurring at transcribed enhancers and promoters (Figure 3.3C, C.10B). These results support the model that RNA molecules promote both optimal stability of TF-enhancer assemblies and spatial interactions with other such assemblies.

Given evidence that TF-DNA occupancy is enhanced by RNA[113,145], we next tested the hypothesis that RNA contributes to the occupancy of TFs in cells. RNase- and Mock-treated nuclei were subjected to mass spectrometry (Figure 3.3B). Although both preparations contained equivalent amounts of core histones, nuclei treated with RNase showed substantially reduced levels of TFs on chromatin (Figure 3.3D). The TFs that showed the most profound reduction with RNase treatment were among those previously shown to bind to RNA (Figure C.10C, C.10D, Table C.4). These results are consistent with the model that RNA contributes to the stability of TF nano-assemblies in cells.

Examination of nascent transcripts and the effects of RNA depletion at a single enhancer reinforce these interpretations (Figure 3.3E). PRO-seq data shows that the E36 enhancer is bidirectionally

transcribed, consistent with previous studies that indicate bidirectional transcription at both enhancer and promoter elements[143,146,147]. The RCMC data indicates that E36 interacts with most other enhancers in the locus and that these interactions are diminished when RNA is depleted (Figure 3.3C, 3.3E).

3.3 Discussion

Models of genome organization generally highlight evidence that gene regulation and genome architecture are linked[83,88], yet we lack precise, high-resolution information about the molecules that contribute to this linkage. Here, we describe a method that captures the 3D chromatin contact structure of a locus and, at nanoscale resolution, assemblies of enhancer-bound TFs that dominate the contact architecture of the locus. Our results suggest that TF-enhancer assemblies consist of multiple small clusters of TFs. Each of the TF-enhancer assemblies can form contacts with the other TF-enhancer assemblies within the locus, and all of these can also interact with the target gene promoter. These TF-enhancer assemblies occur where RNA is transcribed and appear to depend on RNA for their stability.

The models of TF-enhancer assemblies described here suggest that enhancers often consist of multiple small clusters of TFs, called here TF nano-assemblies, each separated on average by ~60 bp from the other (Figure 3.4). A previous model proposed for an enhancer TF assembly in mammalian cells suggests that TFs can fully occupy a 50 bp span of DNA, suggesting the possibility that the population of TFs at an enhancer might form one large uninterrupted and cooperative assembly[110,111]. The distribution of footprints we observe in enhancers suggest that human enhancers more closely resemble those derived from *Drosophila* studies[148–151], where TFs

or collections of TFs bind multiple sites that are separated by some distance, rather than continuously occupying the span of enhancer DNA sequence. The clustering of TFs in a nano-assembly might be a consequence of local interactions among the TFs with DNA, RNA and protein cofactor molecules, and we speculate that the intervals of unbound DNA sequences may have evolved to provide highly dynamic structures, where TFs rapidly bind and dissociate, and where there is some degree of biochemical and biophysical flexibility that would not be available to one large uninterrupted and cooperative assembly. Additionally, the spaces between highly bound sequences may be occupied by the intrinsically disordered regions (IDRs) of TFs, which harbor functional effector domains and often comprise the majority of a TF by mass. Clustering of TFs on DNA with semi-regular spacing may contribute to TF-nucleated condensate formation[152] or impart DNA with block copolymer-like properties that favor the formation of higher-order assemblies[153,154].

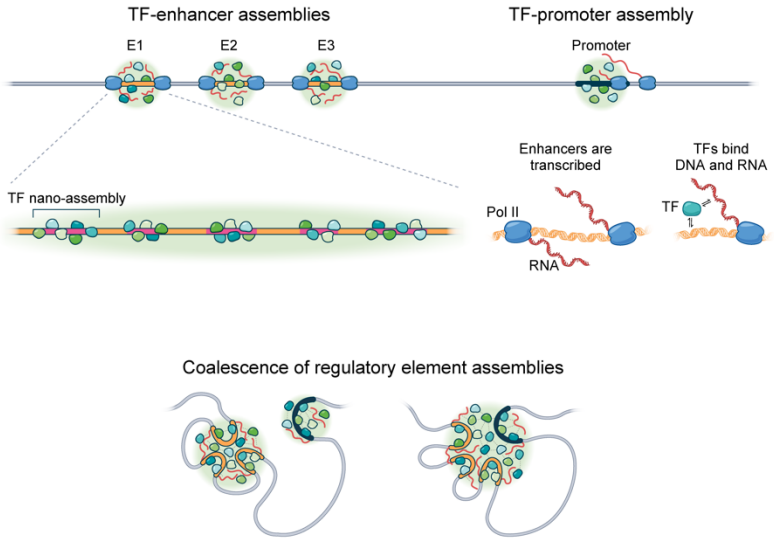


Figure 3.4 Biopolymer associations in transcriptional assemblies. Model of TF-enhancer assemblies (E1, E2, E3) and a promoter assembly (top). Clusters of TFs form nano-assemblies, where local RNAs facilitate TF occupancy on enhancer DNA (middle). These TF-DNA-RNA clusters can coalesce into larger assemblies of transcriptional components at an active locus (bottom).

TFs appear to use both DNA and RNA binding to fully occupy enhancer assemblies (Figure 3.4). RNA is actively transcribed at enhancers, as evidenced by nascent RNA data, and most TFs can bind RNA through a common arginine-rich motif (ARM) or via portions of their disordered effector domains[113,125–127]. RNA tethered to weak enhancers can enhance recruitment of TFs[145] and TF molecules with mutant ARMs appear to spend less time at regulatory sites enhancers[113]. These observations suggest that RNA contributes to the formation and/or stability of TF assemblies and is supported by evidence that RNA depletion causes loss of TFs from chromatin and loss of the dominant contacts within locus architecture (Figure 3.3).

The RCMC contact map indicates that TF-enhancer assemblies can interact with each other and with the *MYC* promoter, producing an assembly that may encompass all of the regulatory elements (Figure 3.4). It would be expected, however, that subsets of the TF-enhancer assemblies interact with one another and with the promoter assemblies in a dynamic fashion, as suggested by recent studies[98,99]. In these regulatory element assemblies, individual TFs dynamically interact with DNA, RNA and various protein cofactors, and future studies will be needed to more fully understand the dynamic relationships among these molecules.

The approach described here can be used to identify TF-enhancer nano-assemblies at high resolution and deduce the regulatory architecture of many loci in the genome. Doing so at loci where disease-associated variation occurs in enhancers should enhance our ability to interpret the impact of disease-associated variation in these regulatory elements and may lead to new therapeutic hypotheses.

3.4 Acknowledgements

We would like to thank Phillip Sharp, Jesse Platt, Hao-Che Wang, Ming Zheng, Catherine Van Dongen, Michael Gallagher, Nancy Hannett, Bjorn Stolte, Jon Henninger, Alessandra Dall'Agnese, Tomer Barzilay, Stephen Jiang, Anna Dvaladze, Ifediora Nwakuche, Charalampos Lazaris, Eric Olson, Clarice Hong, Miles Huseyin, Shdema Filler-Hayut, An Hoang, Olivia Corradin, Peter Mikhael, and Domitilla Del Vecchio for assistance, advice, and helpful discussions. We thank Caitlin Rausch for assistance with graphical design.

We also thank the MIT Koch Institute Genomics Core, MIT BioMicroCenter, and the St. Jude Computational Biology Genomics Laboratory, and Fabian Schulte, Brooke Linnehan, and other members of the Whitehead Mass Spectrometry Core for valuable technical assistance.

Funding: This work was supported by National Institutes of Health grant GM144283 (R.A.Y.), National Institutes of Health grant CA155258 (R.A.Y.), National Institutes of Health grant R33CA257878 (A.S.H.), National Institutes of Health grant R01EB035127 (A.S.H.), National Institutes of Health grant R01CA300848 (A.S.H.), National Science Foundation grant PHY2044895 (R.A.Y.), the St. Jude Children's Research Hospital Transcription Collaborative (R.A.Y., J.E., B.J.A.), the Novo Nordisk Foundation grant NNF21SA0072102 (A.S.H.) the NSF Graduate Research Fellowship grant 1745302 (K.J.O.), and the Hope Funds for Cancer Research Postdoctoral Fellowship 2024 (A.V.).

3.5 Author Contributions

Conceptualization: KJO, AV, CES, SWH, MJA, RAY

Methodology: KJO, AV, CES, DS, VYG, SWH, AG, DRP, SM, HE, BJA, ASH, MJA, RAY

Investigation: KJO, AV, CES, DS, IS, VYG, SWH, AG, DRP, SM, AR, AG, HE, EP, JE, BJA, ASH, MJA, RAY

Visualization: KJO, AV, CES, IS, AG, MJA, TIL, RAY

Funding acquisition: KO, AV, JE, BJA, MJA, RAY

Project administration: MJA, RAY

Supervision: MJA, RAY

Writing – original draft: KJO, AV, TIL, RAY

Writing – review & editing: KJO, AV, CES, DS, IS, VYG, SWH, DRP, SM, AG, HE, TIL, JE, BJA, ASH, MJA, RAY

3.6 Data and materials availability

All code used in the study will be made available via a GitHub repository. All data generated in the paper is available upon request directed to the corresponding author(s).

Appendix A

Supplemental material for Chapter 1

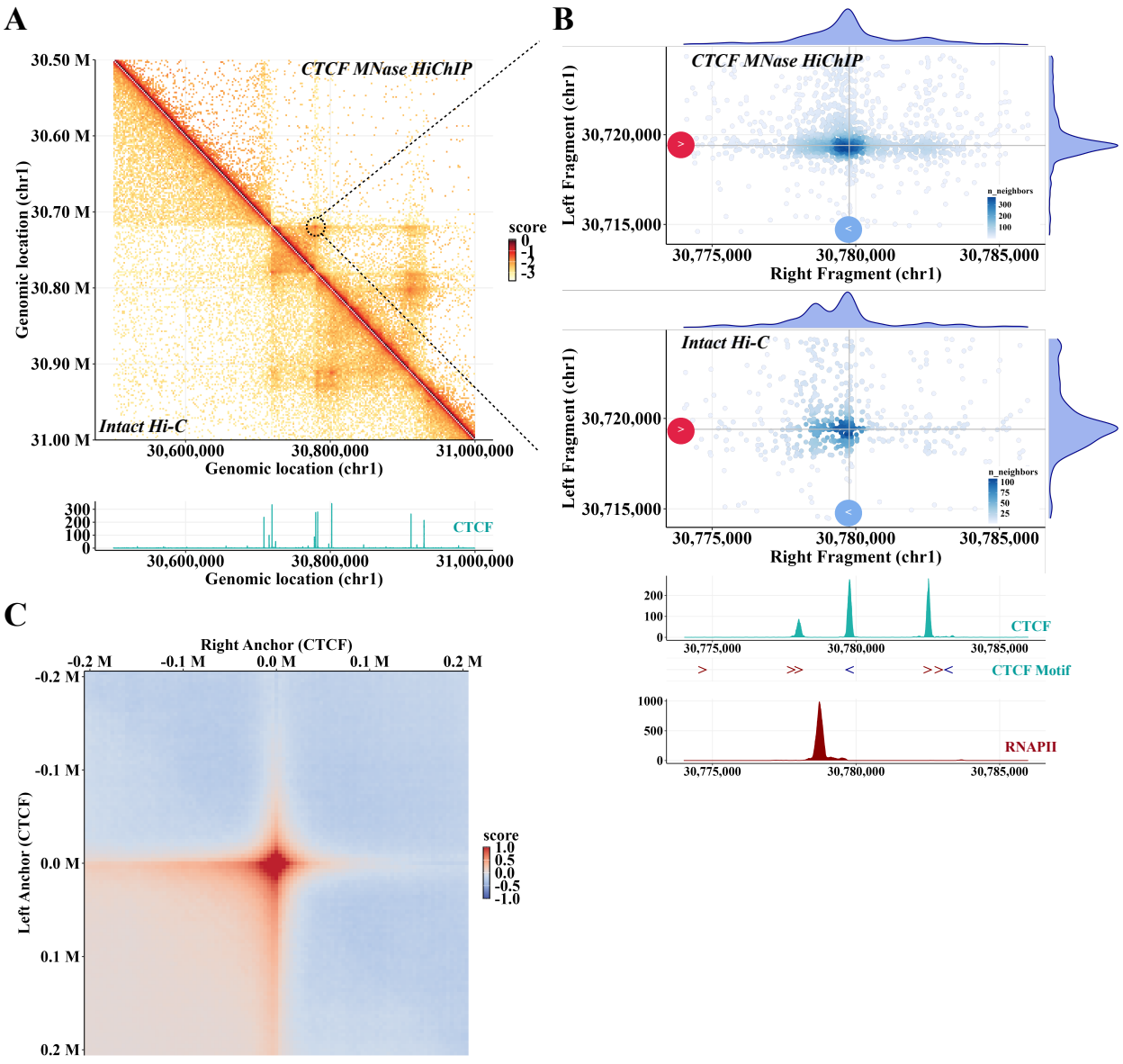


Figure A.1 Comparison of K562 CTCF MNase HiChIP to K562 Intact Hi-C. A K562 CTCF MNase HiChIP (above diagonal) and K562 Intact Hi-C (GSE237898, below diagonal) balanced contact maps (log10 scale, 2kb resolution) for a region on chromosome one with CTCF ChIP-

Seq annotation below (ENCFF168IFW). **B** Marginal densities and scatterplots colored by number of neighbors of left fragments vs right fragments (5' end of the read) for the loop circled in panel **(A)** between two convergent CTCF motifs: chr1:30,719,390-30,719,408 (+), chr1: 30,779,763-30,779,781 (-). The red (>) and blue (<) symbols respectively denote CTCF (+) and CTCF (-) motifs. CTCF ChIP-Seq (ENCFF168IFW), CTCF motifs, and RNAPII ChIP-Seq (ENCFF914WIS) annotations for the right anchor are shown below the scatterplots. **C** Metaplot of Intact Hi-C contacts aggregated over chromosome one CTCF HiChIP loops (n = 3645 loops, resolution = 4kb, linear scale, detrended contacts. Red represents more contacts than expected while blue represents less contacts than expected). Aggregated contacts are obtained using the HiContacts::aggregate R function.

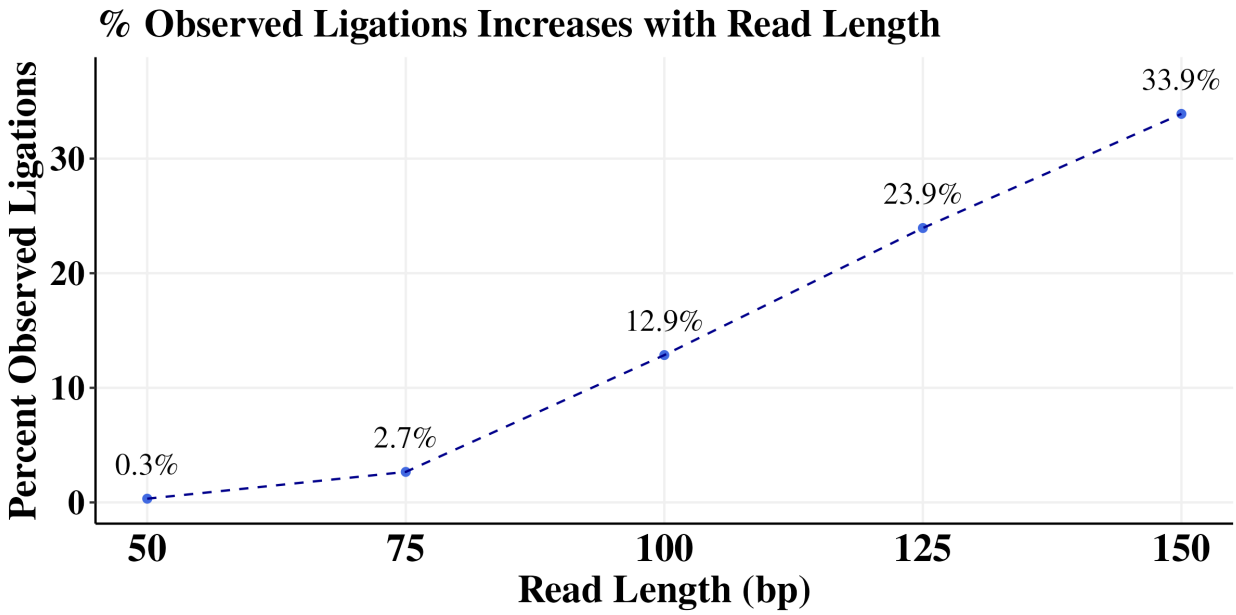


Figure A.2 Percent observed ligations increases with read length.

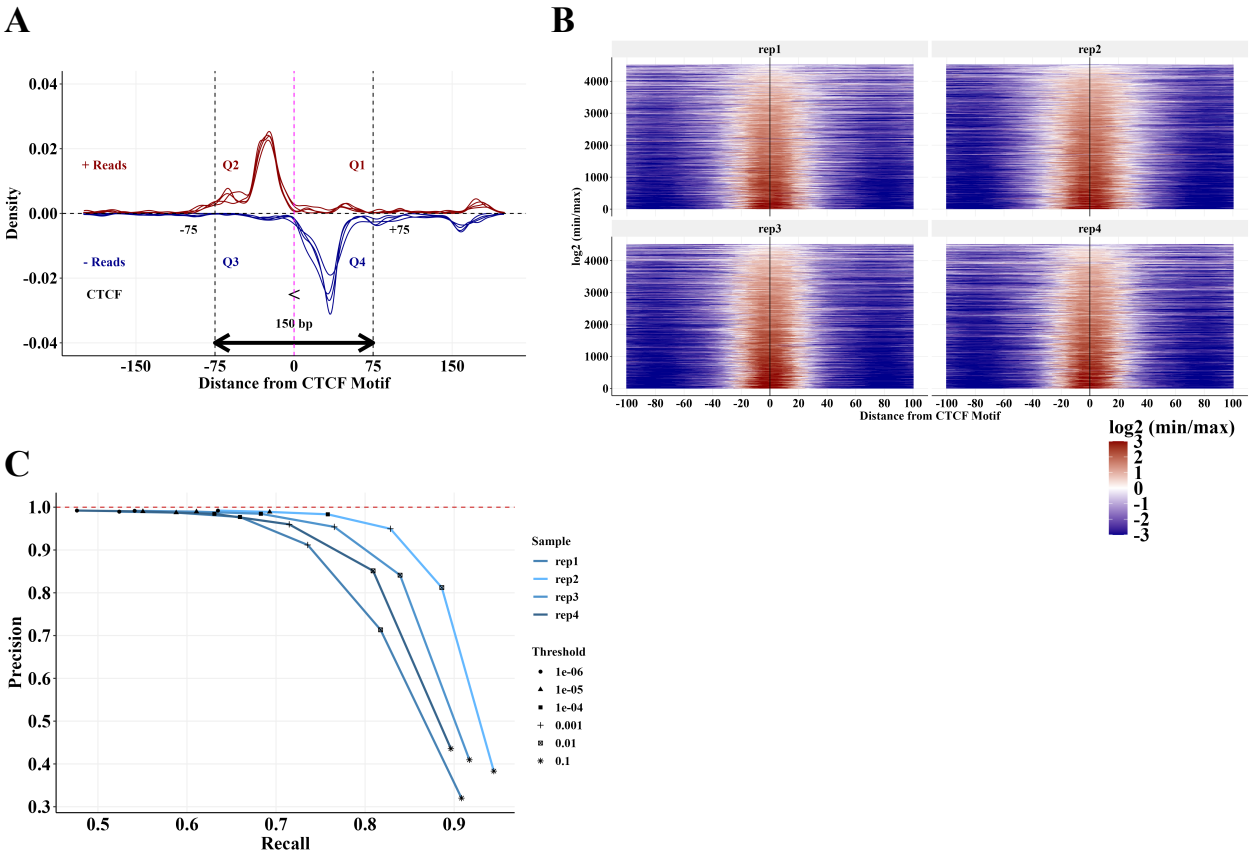


Figure A.3 Comparison of K562 CTCF MNase HiChIP replicates. A Figure 1.2D read distribution pattern at a CBS is observed across all four replicates. CBS (-) is located at chr1:30,779,763-30,779,781. **B** Figure 1.3A CAMEL statistic (log2 (min/max)) pattern observed across all four replicates. **C** Figure 1.3B (precision recall curves) by replicate.

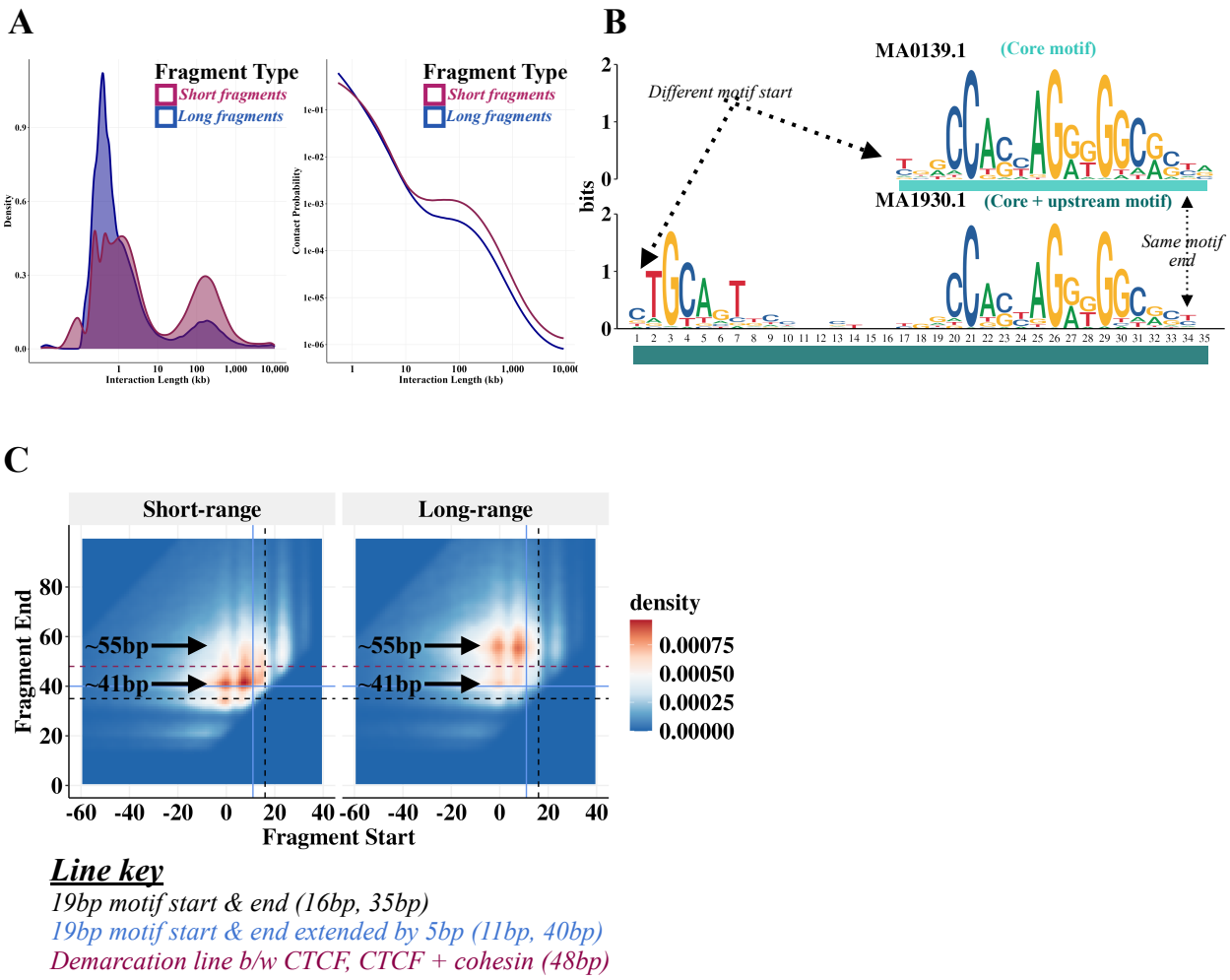


Figure A.4 Companion figure to Figure 1.4: CTCF and cohesin footprints. A 1D density (left) and P(S) curve (right) of the interaction lengths (kb) of fragments overlapping CBS motifs (+), starting at least 5bp before the motif start, and ending at least 5bp after the motif start. Long and short fragments are defined as greater than or equal to 120bp (long) and less than 120bp (short) respectively. **B** Comparison of JASPAR 19bp (MA0139.1) and 35bp (MA1930.1) CTCF motifs. Meme files (MA0139.1.meme, MA1930.1.meme) were downloaded from JASPAR, read into R (universalmotif::read_meme), and aligned into a ggplot compatible plot (universalmotif::view_motifs). **C** Figure 1.4C with annotations. 2D density of all fragments overlapping CBS (+) with fragment length less than 120bp. Fragment end and fragment start are aligned to the start of the long (35bp) CTCF motif (MA1930.1) such that 0 represents the start of the 35bp CTCF motif, 16bp represents the start of the 19bp CTCF motif (MA0139.1), and 35bp represents the end of both CTCF motifs. Annotation lines are added in with explanation in the "Line key", and arrows are drawn to show the locations of the fragment end for the CTCF and CTCF with cohesin footprints.

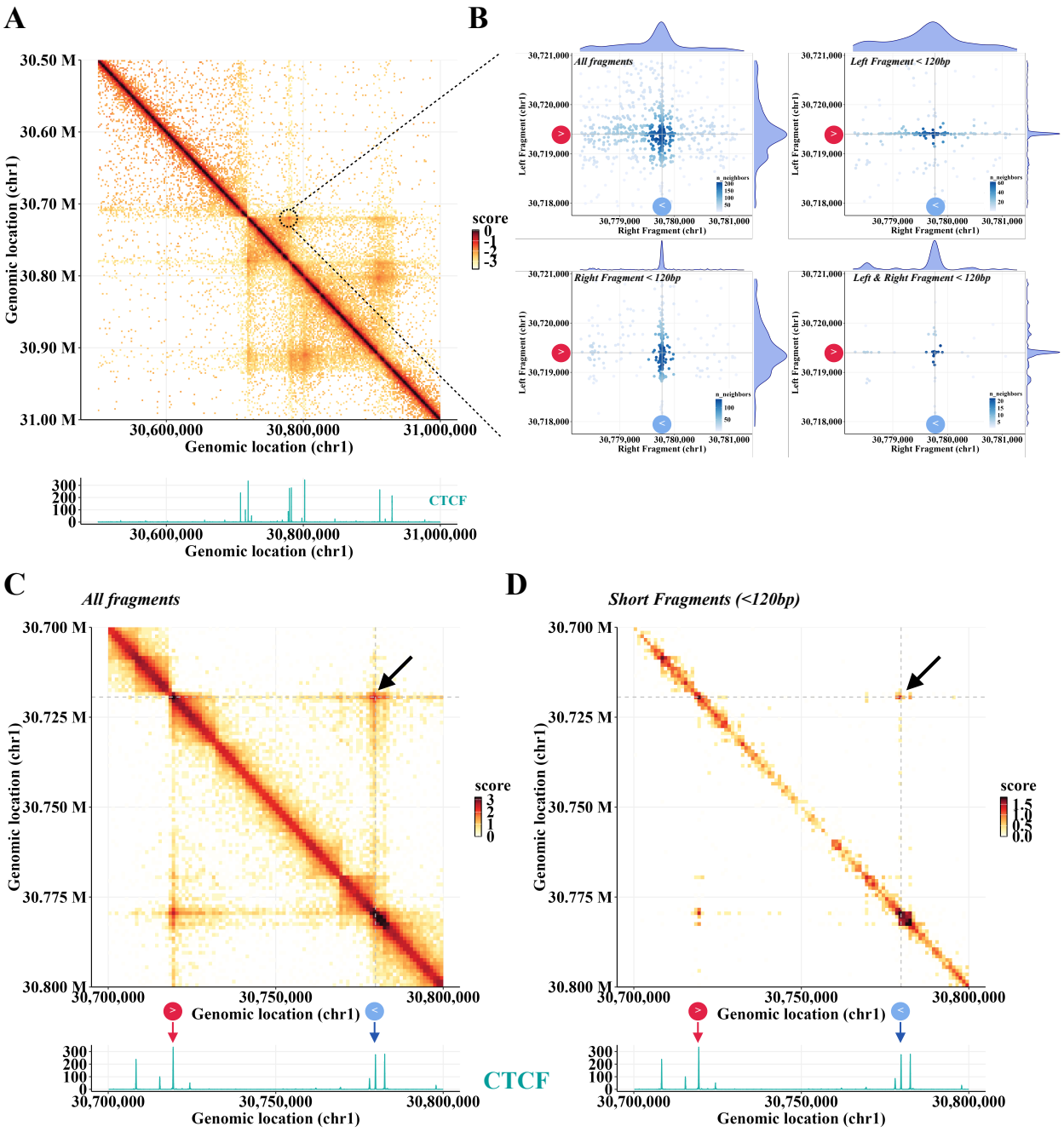


Figure A.5 Filtering to short fragments enables identification of CTCF-mediated contacts. **A** K562 CTCF MNase HiChIP balanced contact map (log10 scale, 2kb resolution) for chr1:30,500,000-31,000,000 with CTCF ChIP-Seq annotation below (ENCFF168IFW). **B** Marginal densities and scatterplots colored by number of neighbors of left fragments vs right fragments (midpoint of fragment) for the loop circled in panel **(A)** between two convergent CTCF motifs: chr1:30,719,390-30,719,408 (+), chr1: 30,779,763-30,779,781 (-). Each panel

shows the result of filtering based on fragment length; from left to right top to bottom (also labeled above): all fragments, left fragment < 120bp, right fragment < 120bp, and both left and right fragment < 120bp. The red (>) and blue (<) symbols respectively denote CTCF (+) and CTCF (-) motifs for panels **(B, C, D)**. **C** Raw contact map (log10 scale, 1kb resolution) of all fragments for a zoomed in view (chr1:30,700,000-30,800,000) of the loop circled in panel **(A)**, with CTCF ChIP-Seq and motif annotations below. **D** Raw contact map (log10 scale, 1kb resolution) for only short fragments (left fragment and right fragment < 120bp) for the same region as panel **(C)**, with CTCF ChIP-Seq annotation below.

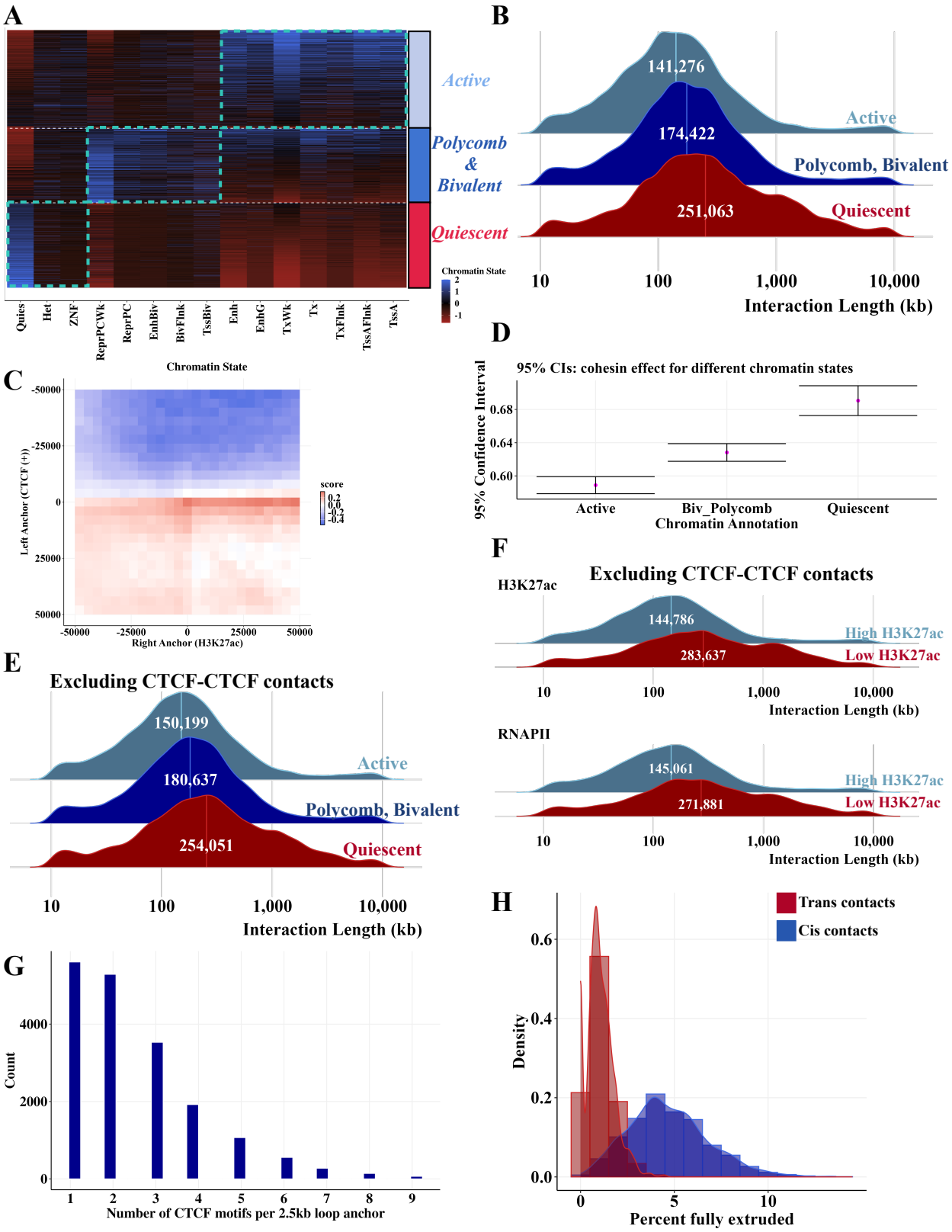

A
Active
Polycomb & Bivalent
Quiescent
Chromatin State
Quies
Het
ZNF
ReprPCWk
ReprPC
EnhBiv
BivFlnk
TssBiv
Enh
EnhG
TxWk
Tx
TxFlnk
TssAFlnk
TssA
B
141,276
174,422
251,063
Active
Polycomb, Bivalent
Quiescent
10
100
1,000
10,000
Interaction Length (kb)
C
Chromatin State
Left Anchor (CTCF (+))
Right Anchor (H3K27ac)
score
D
95% CIs: cohesin effect for different chromatin states
95% Confidence Interval
Active
Biv_Polycomb
Quiescent
Chromatin Annotation
E
Excluding CTCF-CTCF contacts
150,199
180,637
254,051
Active
Polycomb, Bivalent
Quiescent
Interaction Length (kb)
F
Excluding CTCF-CTCF contacts
H3K27ac
144,786
283,637
High H3K27ac
Low H3K27ac
RNAPII
145,061
271,881
G
Count
Number of CTCF motifs per 2.5kb loop anchor
H
Trans contacts
Cis contacts
Density
Percent fully extruded

Figure A.6 Companion figure to Figure 1.5: Cohesin extrudes further through quiescent regions than active regions. A ChromHMM annotated 1 MB regions downstream of CBS cluster into three main groups: active, polycomb & bivalent, and quiescent chromatin. Each 1 MB region is characterized by x% TssA (active TSS), y% TssAFlnk (flanking active TSS), z% TxFlnk (flanking transcription), etc. as defined using ChromHMM. Each row corresponds to a CTCF binding site (N = 10,906), and each column shows the normalized fraction of downstream DNA with a specific chromHMM annotation**. B** Ridge plots for the same data graphed in Figure 1.5C with averages marked for each chromatin type. This demonstrates how we obtain the estimates reported in Figure 1.5F. **C** Metaplot of Intact Hi-C contacts (GEO: GSE237898) aggregated over chromosome one CTCF (+):H3K27ac contacts. All chr1 H3K27ac peaks (ENCODE: ENCFF544LXB) between 10kb and 100kb downstream of CBS (+) are included in this aggregated plot (N = 2732, resolution = 4kb, linear scale, detrended contacts). Contacts are aggregated over all such CTCF:H3K27ac pairwise contacts regardless of whether there is a loop between the CTCF and H3K27ac (to keep this aggregation unbiased). Greater than 10kb was chosen to filter to long-range contacts, and less than 100kb was chosen to keep CTCF:H3K27ac contacts in the same TAD. Red represents more contacts than expected while blue represents less contacts than expected. Aggregated contacts are obtained using the HiContacts::aggregate R function. **D** Controlling for locus-specific variation with linear mixed models does not attenuate the relationship between chromatin state and extruded loop size. **E** Ridge plots for the same data graphed in **(B)** but excluding contacts within 1kb of CTCF motifs on the negative strand (convergent CTCF-CTCF contacts). **F** Ridge plots for the same data graphed in Figure 1.5E but excluding contacts within 1kb of CTCF motifs on the negative strand (convergent CTCF-CTCF contacts). **G** Distribution of the number of CTCF motifs in a 2.5kb loop anchor. **H** Estimates of the fully extruded state obtained using trans (red) and cis (blue) contacts. The cis contact density histogram is equivalent to Figure 1.5A. Estimate is obtained using left fragments that overlap *CAMEL* identified CBS (+) and have an interaction length greater than 10kb with start and end at least 5 bp from motif start and end, length < 120, and extended fragment end. For each CBS with at least 50 long-range TF-protected fragments overlapping the motif, % convergent is calculated as the number of interaction partners overlapping CTCF (-) motifs / total number of fragments at motif. Because this estimate is conditional on CTCF binding at the anchor, we divide estimates by two to account for the ~50% occupancy of CTCF[33].

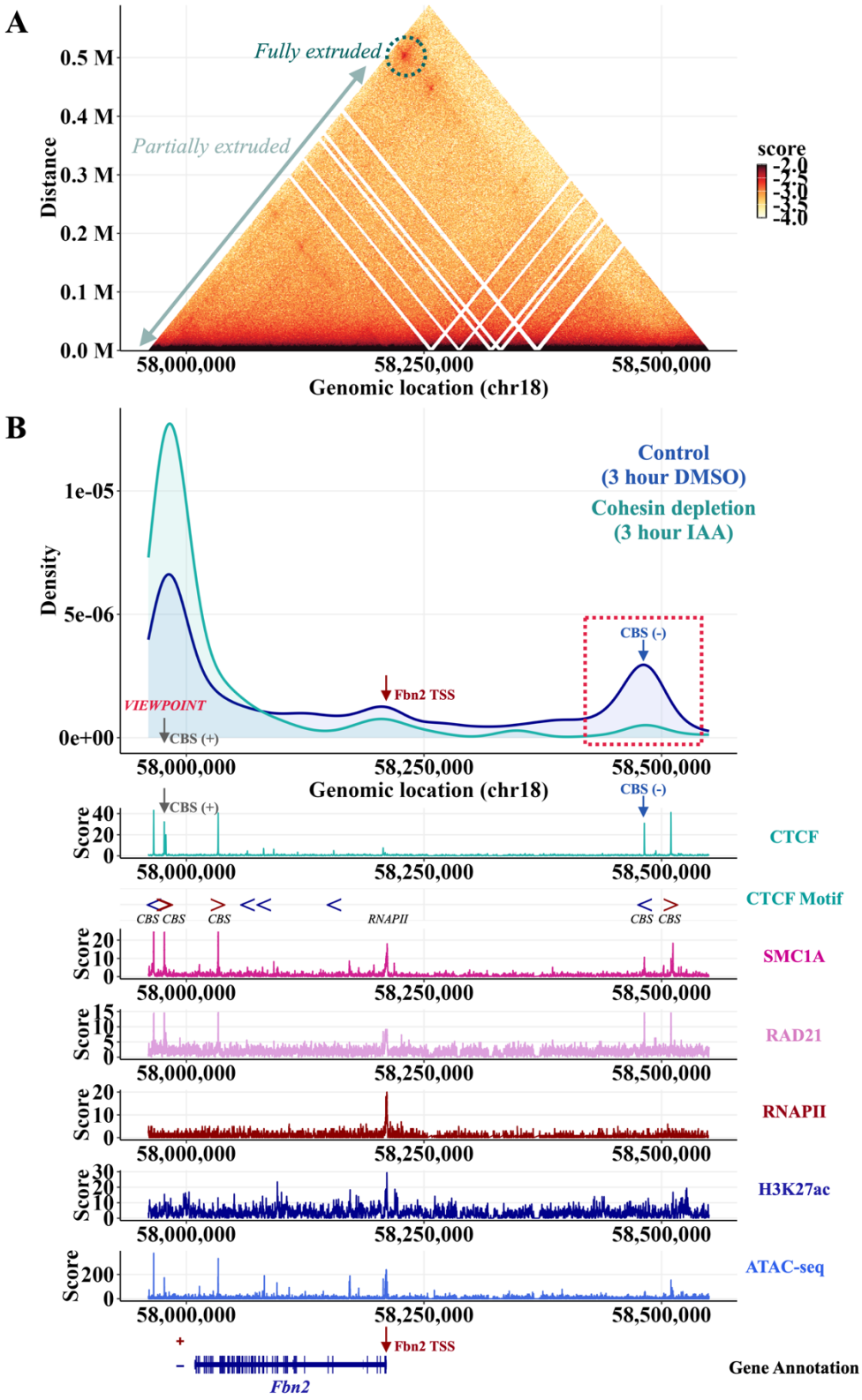

A
Fully extruded
Partially extruded
Distance
0.5 M
0.4 M
0.3 M
0.2 M
0.1 M
0.0 M
score
58,000,000
58,250,000
58,500,000
Genomic location (chr18)
B
Control
(3 hour DMSO)
Cohesin depletion
(3 hour IAA)
Density
1e-05
5e-06
0e+00
CBS (-)
VIEWPOINT
CBS (+)
Fbn2 TSS
Genomic location (chr18)
Score
CTCF
CTCF Motif
CBS CBS
CBS
RNAPII
SMC1A
RAD21
RNAPII
H3K27ac
ATAC-seq
Fbn2 TSS
Fbn2
Gene Annotation

Figure A.7 mESC *Fbn2* CBS:RNAPII contact lost upon depleting cohesin. A mESC RCMC *Fbn2* TAD balanced contact map (log10 scale, 1kb resolution), visualized with HiContacts::plotMatrix. **B** 1D density for left fragments (<120bp) overlapping the CBS (+) anchoring the left side of the *Fbn2* TAD for control (3 hour DMSO) and cohesin depletion (3 hour IAA). The CBS (+) coordinates are chr18:57,976,797-57,976,815 and the CBS (-) coordinates are chr18:58,481,866-58,481,884 (mm10). The viewpoint (CBS (+)), *Fbn2* TSS, and CBS (-) are marked on the 1D density. ChIP-seq annotations downloaded from GEO and gene annotation (plotted with plotgardener::plotGenes) are included below the 1D density. The GEO accession IDs for the ChIP-seq files are as follows: CTCF (GSM3508478), ATAC-seq (GSE98390), H3K27ac (GSM2417096), RNAPII (GSM6809981), RAD21 (GSE137272), and SMC1A (GSE137272). We use mm10 coordinates for all mESC figures in this paper; files downloaded in other mouse genomes were converted to mm10 using CrossMap.

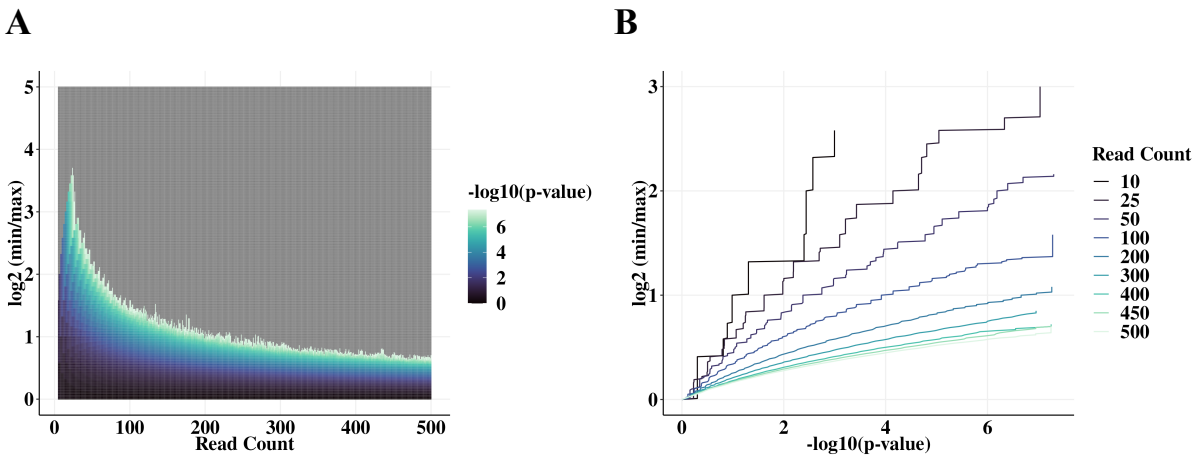


Figure A.8 The probability of observing a high CAMEL statistic under the null hypothesis is higher at low read counts. A Simulated log2 (min/max) vs read count colored by -log10(p-value) heat map. **B** Simulated log2 (min/max) vs -log10(p-value) for a few example read counts.

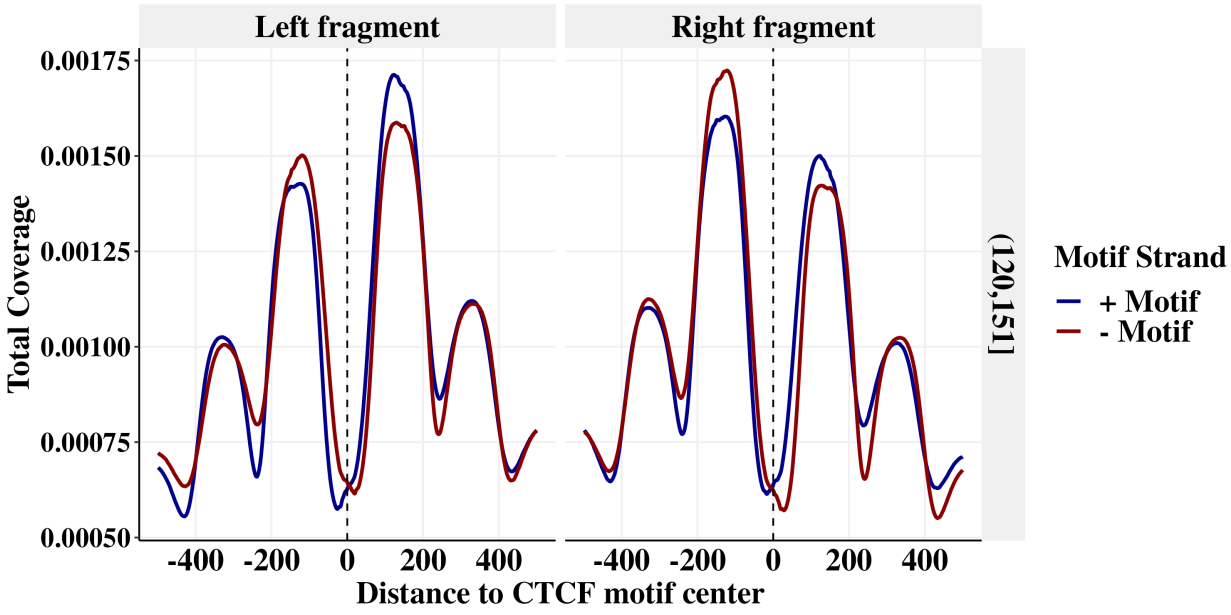


Figure A.9 Nucleosomes are preferentially positioned inside the loop. Coverage of long (between 120 and 151bp) fragments faceted by left or right fragment in an interaction pair and colored by the CTCF motif strand. Only fragments with interaction length > 10kb are included.

Appendix B

Supplemental material for Chapter 2

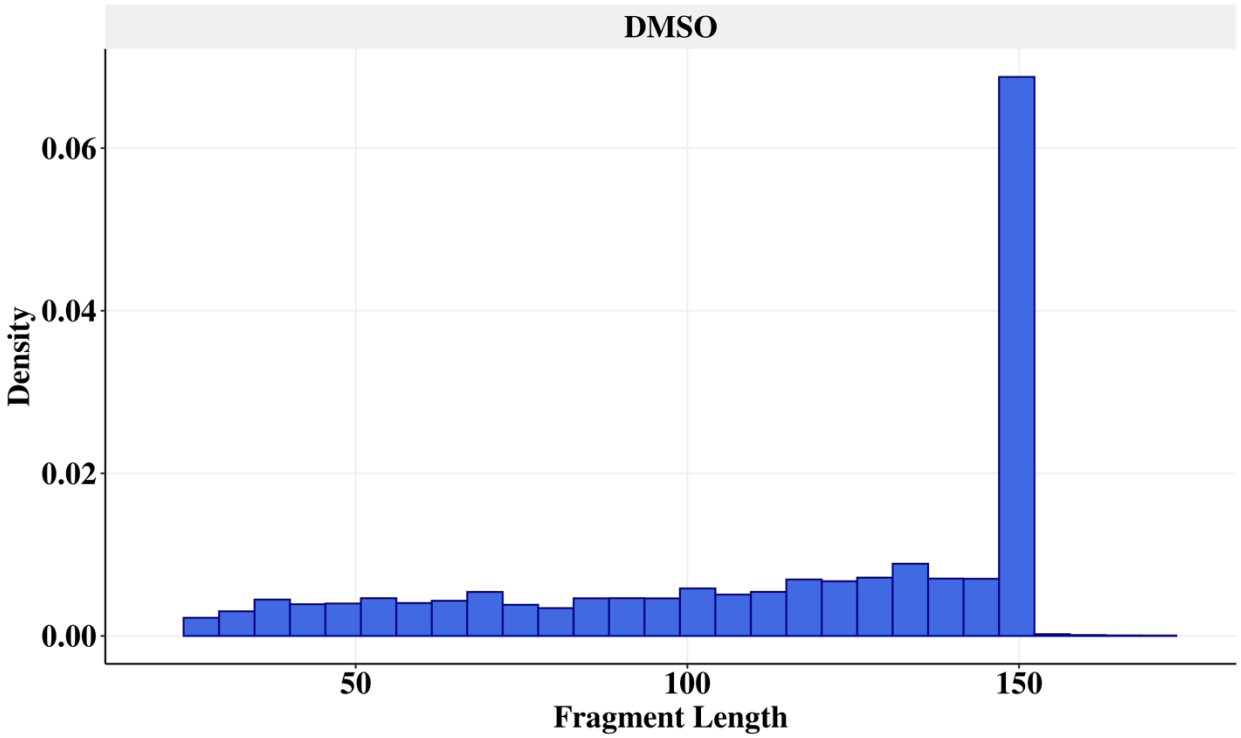


Figure B.1 Fragment length density histogram (chr1 fragments).

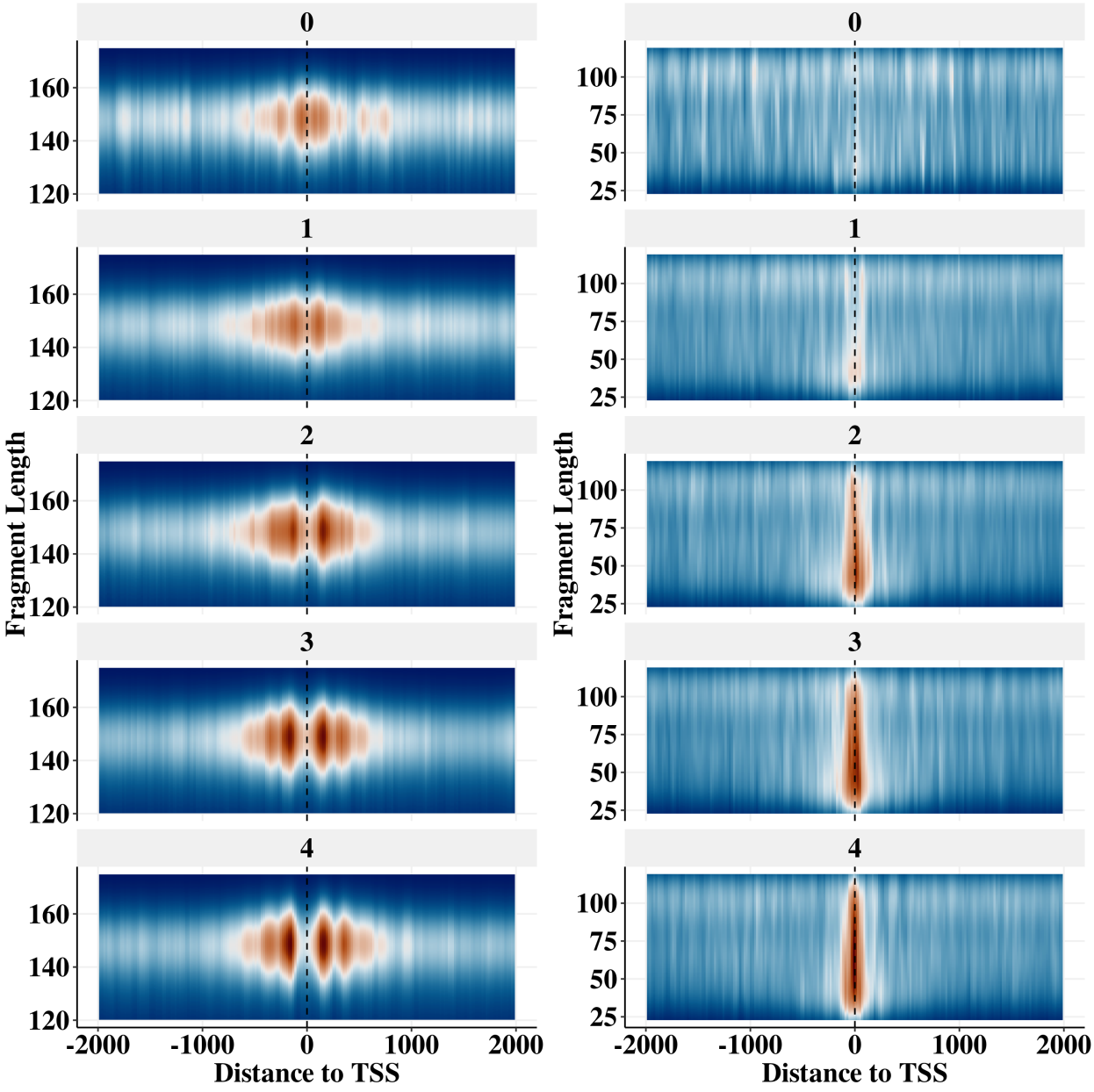


Figure B.2 Increased expression is associated with a larger NDR and TF-sized footprint enrichment at the TSS. Figure 2.1B stratified by fragment length and expression, where "0" is unexpressed genes and "1"-"4" are expression quartiles ("1" is the lowest expression quartile, "4" is the highest). X-coordinates are distance to TSS with respect to the positive strand (coordinates of gene (-) fragments are not flipped in this plot).

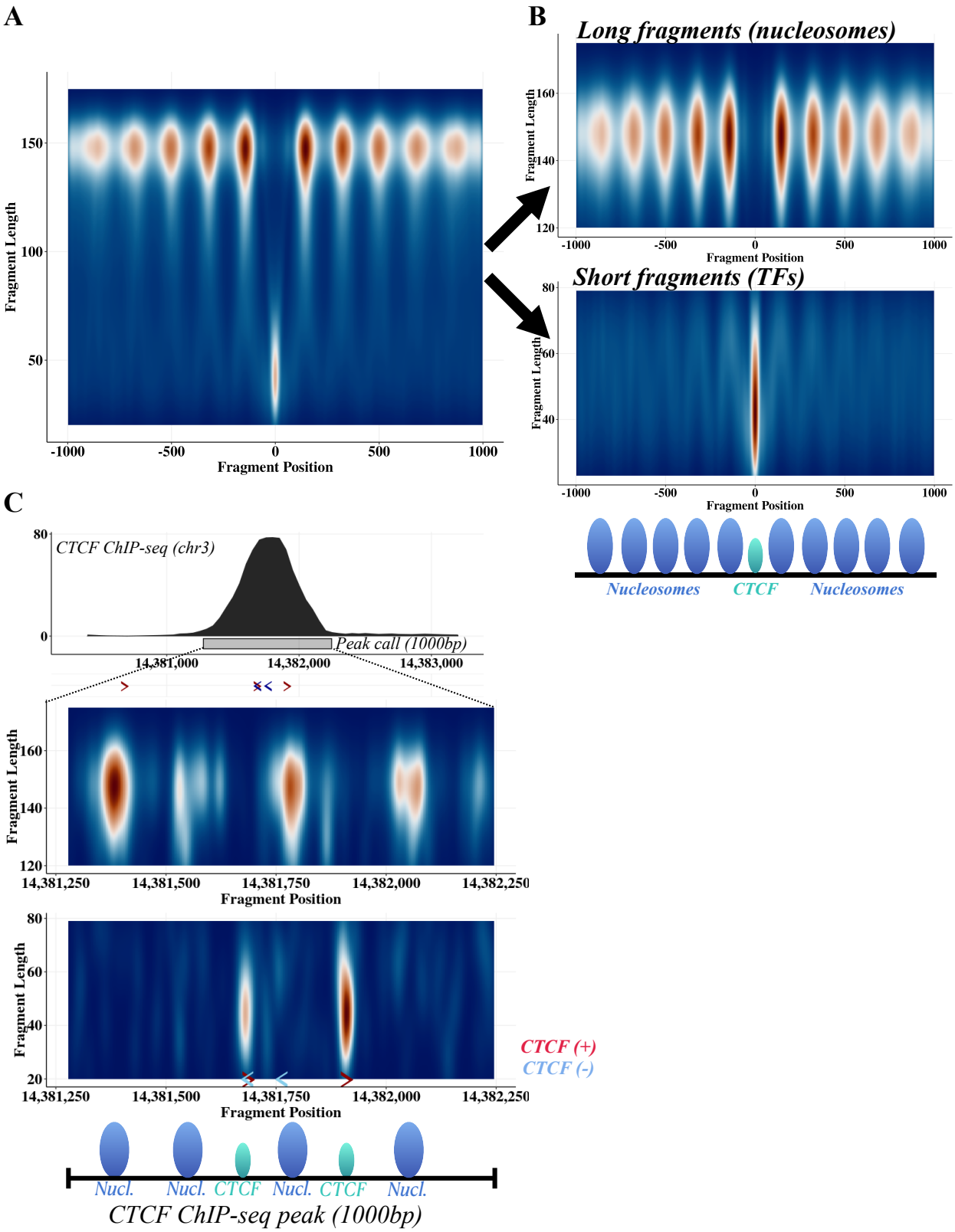


Figure B.3 High density of short fragments reveals transcription factor binding events. A Aggregate 2D density plot of all fragments within 1kb of CTCF binding events supported by

ChIP-seq. **B** Panel **(A)** split by fragment size distinguishes nucleosome occupancy from TF occupancy. **C** Locus-specific 2D density plot of fragments underlying a 1kb CTCF ChIP-seq peak identifies bound motifs.

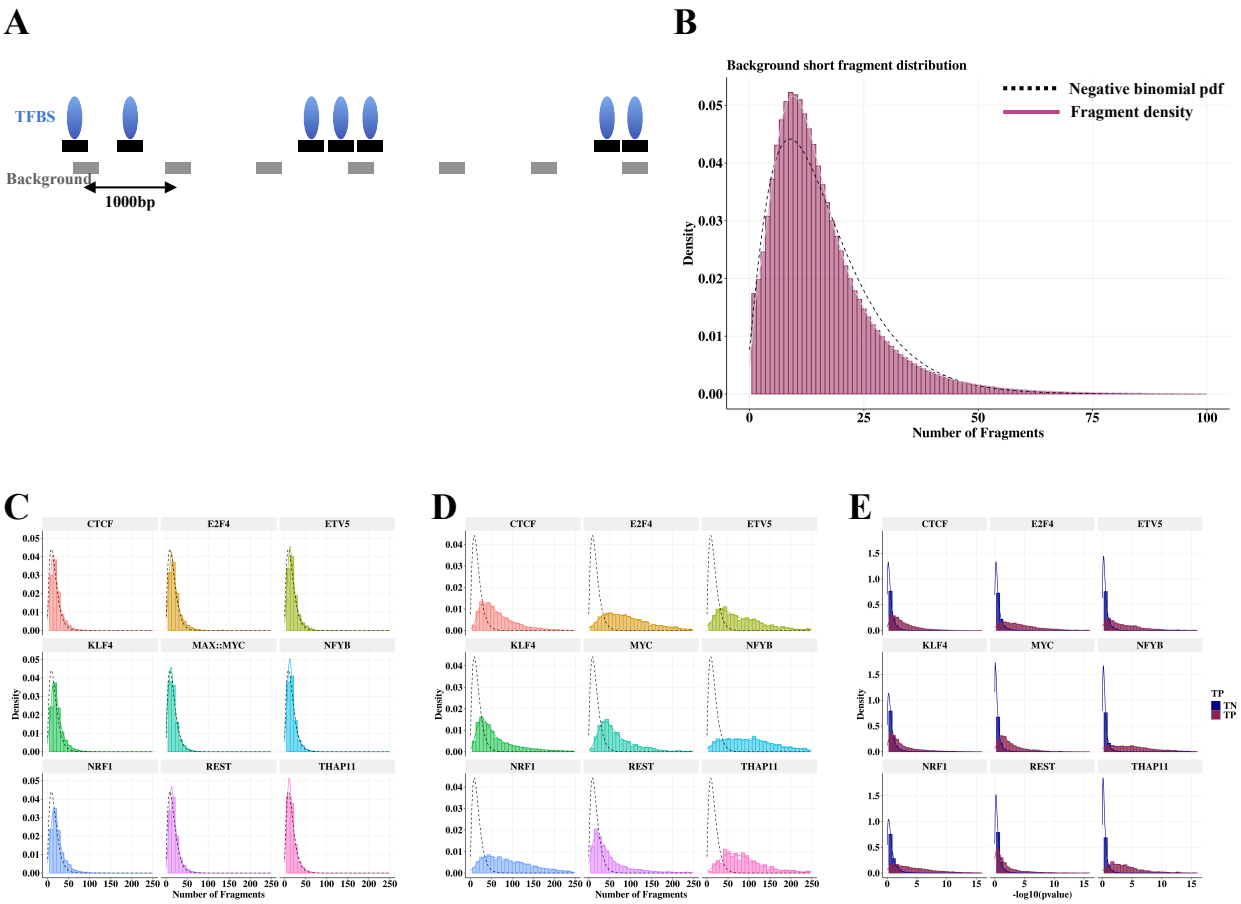


Figure B.4 Evaluating significance of short fragments at transcription factor motifs. A Schematic for identifying genome-wide background short fragment distribution. **B** Genome-wide short fragment density histogram with fitted negative binomial probability distribution function (black dotted line) overlaid. **C** Motif-specific comparison of short fragments at unoccupied motifs to fitted negative binomial background distribution from **(B)**. **D** Motif-specific comparison of short fragments at occupied motifs (using UniBind[72] as true positive set) to fitted negative binomial background distribution from **(B). E** Motif-specific comparison of p-values at TN (true negative, unoccupied motifs) and TP (true positive, occupied motifs).

Appendix C

Supplemental material for Chapter 3

Materials and Methods

Cell culture

HCT 116 cells (ATCC #CCL-247) were grown in DMEM (Gibco #11995-065) with 10% FBS (Sigma-Aldrich #F4135-500ML) and 1% penicillin/streptomycin (Thermo Fisher Scientific # 15140122) at 37 °C and passaged at 70-80% confluency using TryplE (Thermo Fisher Scientific #12604021). Cells were tested for mycoplasma contamination using MycoAlert Mycoplasma Detection kit (Lonza #LT07-318). For doxycycline-inducible cell lines, induction with doxycycline (Thermo Fisher Scientific #J60579.14) was performed for 5 hours at concentration of 100 ng/mL.

Region Capture Micro-C (RCMC) and Micro-C

RCMC was conducted following the procedure of Goel et al.[20]. For each RCMC experiment, 50 million cells were trypsinized using TryplE, quenched with cell culture media, washed with PBS, and crosslinked with disuccinimidyl glutarate (ThermoScientific #20593) for 30 minutes followed by 16% formaldehyde (Thermo Fisher Scientific #28906) for 10 minutes. MNase (Worthington Biochem #LS004798) was titrated to produce an 80% monomer to 20% dimer ratio, and MNase amounts were chosen independently for each sample based on the titration results. Dinucleosome-sized fragments were selected by cutting 250-400 bp bands from agarose gels. Following library preparation, Micro-C libraries were sequenced using paired-end 2x150 bp sequencing prior to the region capture step on an Illumina Novaseq 6000 sequencer at the St. Jude Computational Biology and Genomics Laboratory. For RCMC, the *MYC* locus was captured using a custom biotinylated 80-mer probe panel (Twist Biosciences) tiling the region from 126,800,000 to 129,750,000 bp in the hg38 human genome assembly with no gaps between probes. Libraries were

sequenced using paired-end 2x50 base pair or 2x150 base pair sequencing on an Illumina Novaseq 6000 sequencer at the St. Jude Computational Biology and Genomics Laboratory.

RCMC 3D contact data analysis

RCMC data was proceeded as described in Goel et al. with minor modifications. Read quality was verified using FastQC. The RCMC library collected from wild-type HCT 116 cells was sequenced using 2x50 bp for RCMC data analysis. The data was processed through the following pipeline. Briefly, .fastq files (50 bp reads) were aligned to the reference genome (hg38) using bowtie2, filtered for pairs within the region to create .pair files (pairtools select), and merging .pair files to generate .cool files (pairtools merge, pairix) and using cooler zoomify to create final .mcool files for visualization purposes.

RCMC contact maps were visualized in cooltools (v.0.5.1) and aligned with genomic tracks generated in coolbox (v.0.3.8). Contact maps were normalized using the standard iterative correction and eigenvector decomposition (ICE) balancing method implemented in cooltools and applied to the entire capture locus[20].

Identification of RCMC loop regions

Detection of 3D chromatin loops was performed using the Mustache (Multi-scale Detection of Chromatin Loops from Hi-C and Micro-C Maps using Scale-Space Representations) tool (v.1.3.1)[155]. Loops were detected at a resolution of 1 kb and a threshold of $p<0.1$ was used to define statistical significance. From the set of significant loops, “RCMC loop regions” were defined as continuous stretches including two or more overlapping or adjoining loop anchors. Single loop anchors without an additional neighboring or overlapping anchor were considered low-confidence and excluded from the analysis. The final RCMC loop regions were collapsed into a single .BED file.

Annotation of regulatory elements

ENCODE Candidate Cis-Regulatory Elements (cCREs) from HCT 116 were downloaded from ENCODE (https://www.encodeproject.org). Encode enhancers comprise proximal enhancer-like signatures (pELS)

and distal enhancer-like signatures (dELS). For CTCF annotation, CTCF-only labels were addressed for analysis. Low H3K4me3 and H3K27ac signals with high DNase was labeled as low enhancer signal as annotated by ENCODE. All cCRE classifications can be viewed at: https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=mm10&g=encodeCcreCombined.

Hi-C data analysis

Publicly available Hi-C data[156] was obtained from the GEO under accession number GSE199607. Contact maps were normalized via ICE balancing in cooltools. The data were visualized at the minimum bin size for this dataset of 5 kb.

Generation of RCMC footprint maps

The RCMC library collected from wild-type HCT 116 cells was sequenced using 2x150 bp to allow for footprinting. The data was processed through the following pipeline in Terra. Briefly, fastq files (150 bp reads) were aligned to the reference genome (hg38) using BWA-MEM, ligations were obtained (pairtools parse), the resulting pairsam file was sorted (pairtools sort), PCR duplicates were removed (pairtools dedup), and finally the file was split into pairs and bam files (pairtools split). We then filtered to both fragments of an interaction pair in the captured region (chr8:126,800,000-129,750,000) for use in all downstream analyses, as is recommended in the original RCMC pipeline by Goel *et al.*[20].

To obtain the start and end of each fragment, we took the minimum of each fragment's 5' and 3' position (fragment start) and the maximum of each fragment's 5' and 3' position (fragment end). The default behavior of pairtools parse only provides the 5' end of the read; to output both the 5' and 3' ends, we included the flag *--add-columns pos5,pos3* to the pairtools parse command. Fragment lengths are the difference between the fragment start and fragment end.

The resulting pairs file was collapsed into a GRanges object for further analysis in R. The ligated pairs were split to be analyzed separately, and the coordinates of each fragment middle were used for subsequent analysis. Fragments with sizes greater than 120 bp were assumed to correspond to nucleosomes, while

fragments with sizes between 20 bp and 80 bp were used to assemble footprint maps of TFs. Footprints density maps of TFs were created by graphing the fragment length (fragment end - fragment start) vs the fragment mid-points (1/2 * (fragment start + fragment end)) using a custom 2D kernel density function in R. This function smooths the fragment lengths and the fragment mid-points to aggregate similar fragments together into a more easily distinguished footprint. This approach automatically normalizes read densities in different genomic windows by dividing by the number of fragments, allowing for comparisons between different genomic regions.

Footprint detection

Footprints were identified automatically using a custom algorithm written in R (v.4.4.1). Footprinting was conducted within RCMC loop regions, as defined in "Identification of RCMC loop regions". For each loop region, the footprinting procedure was carried out as follows. A 2D fragment density map was computed for the entire loop region as described in "Generation of RCMC footprint maps", using a variable kernel size n that scaled according to the length L of the region by the equation $n=\alpha/L$, where α is an empirically determined scaling parameter. A density threshold t_{intens} was used to perform binary thresholding of the 2D density map of fragments between 20 and 80 bp in length. The binarized 2D map was then collapsed to a 1D binary track using an OR operation applied to each genomic bin, so as to register a given position in the genome as a "1" if any fragment length between 20 and 80 bp registered above the threshold, and "0" otherwise. This track was dilated with a 3 bp window to remove any small gaps. Contiguous regions within the dilated binary track were identified as separate footprints, and the results were stored in a hierarchical S4 object in R containing RCMC loop regions within the two microcompartments of interest, with each loop region containing detected footprints. The values for the parameters used in this analysis were $\alpha=0.83$ and $t_{intens}=0.208$. TFs were assigned using the pipeline described below.

TF footprint density was calculated using by averaging .bed file density over 2 kb windows.

Assigning TFs to footprints using publicly available ChIP-seq and DNA binding motifs

Publicly available ChIP-seq datasets analyzed by ENCODE (https://www.encodeproject.org) were used to assign 18 TFs in HCT 116 to footprints. For each TF: ATF3 (ENCFF388DEA), CEBPB (ENCFF705PPP), CTCF (ENCFF388PVO), ELF1 (ENCFF805TRK), EGR1 (ENCFF329FHN), FOSL1 (ENCFF045ROE), JUND (ENCFF647MFN), MAX (ENCFF842QPT), REST (ENCFF825IKF), SP1 (ENCFF205TXT), SRF (ENCFF449SAT), TCF7L2 (ENCFF957AFR), TEAD4 (ENCFF317YMX), USF1 (ENCFF427NMO), YY1 (ENCFF658WIT), ZBTB33 (ENCFF884WBU), ZNF274 (ENCFF457LUQ), ZFX (ENCFF102UZB) bigWig files for fold change over control from two isogenic replicates were downloaded. bigWigSummary was used to calculate average TF signal and standard deviation for ChIP-signal in our region of interest (chr8:127195000-128195000) at 100,000 bin resolution. The signal threshold for each TF bigwig file was set to average + 0.5 * standard deviation. bigWigAverageOverbed was used to overlap ChIP-signal with footprints and average ChIP signal for each footprint was reported. The threshold generated for each TF was applied to footprint average signal to determine if the footprint signal exceeded the threshold. Footprint ChIP signals exceeding TF-specific thresholds were used to assign footprints to TFs.

DNA binding motifs for TFs exceeding ChIP-seq thresholds in RCMC footprints were used to assign TFs to basepair-specific locations within footprints. The locations of vertebrate JASPAR database motifs were identified using the Find Individual Motif Occurrences (FIMO) tool with a p-value cutoff of $p<10^{-2}$. Motifs were considered for assignment if at least half of the motif fell within the footprint coordinates. The complete set of motifs in the vicinity of E36 are displayed in Figure C.5B. If multiple motifs for a TF were detected within a footprint, only the most significant motif was selected for visualization and structural analysis in Figure 3.2. The resulting set of motifs are a model of a TF's favored binding site within a footprint where TF binding is confirmed by ChIP-seq, and were used to dock TF molecular structures on enhancer DNA as described below.

Molecular docking of TF structures on enhancers

As described above, molecular docking of TF structures on enhancers was performed using a set of motifs modeling favored TF binding sites within footprints where TF binding is confirmed by ChIP-seq. Molecular docking was performed using PyMOL (v. 2.4.1). Co-crystal structures were obtained from the Protein Data Bank (PDB) for EGR1 (4f2a), YY1 (1ubd), MYC (1nkp), MAX (1nkp), TEAD4 (5gzb), SRF (1srs), ZBTB33 (5vmw), and USF1 (1an4). A published docked crystal structure was used for SP1[157]. Water molecules, metal atoms, and ligands were manually removed from these structures. Structures for FOSL1, JUND, CTCF, were TCF7L2 were generated using the AlphaFold3 webserver (https://alphafoldserver.com) with the canonical protein isoform as the input. Intrinsically disordered portions of the protein or structured elements not bound to DNA were manually removed from AlphaFold-predicted structures. Each protein-DNA structure was manually docked on its cognate motif in the DNA sequence by aligning the phosphate backbone of the DNA double helix. In cases where two TFs clashed sterically, only one of the TFs was kept for illustrative purposes.

Supplemental method for assigning TFs to footprints using DNA binding motifs alone

DNA binding motifs for TFs in the JASPAR database were used to assign TFs to footprints. The locations of motifs within the 2.8 Mb *MYC* TAD were identified using FIMO with a p-value cutoff of $p<10^{-4}$. The motifs that correspond to TFs whose RNA-seq expression exceeded a threshold t_{exp} (transcripts per million) became candidates for assignments inside of binarized footprints. Because the boundaries of binarized footprints correspond to a decay in the density of the centers of protected fragments, motifs were considered to reside inside of a footprint if the center of the motif occurred within the boundaries of the binarized footprint. Motifs were stored within their corresponding footprint in the S4 object and ranked according to RNA-seq expression of the corresponding TF from highest to lowest expression within the footprint. The rank of each motif is given by an index *m.* The resulting data structure can be used to assign the top *m* motifs to a footprint. An expression threshold of $t_{exp}=1$ was used in this analysis. The final data structure was flattened to produce Table C.2.

DNase-seq and DNase I footprints

DNase-seq peaks data in HCT 116 cells was obtained from ENCODE datasets under the experiment ID ENCSR000ENM (accession number ENCFF240LRP). DNase I footprints in HCT 116 cells were obtained from ENCODE datasets under the experiment IDs ENCSR000ENM (accession numbers ENCFF625YUY and ENCFF909DCR) and ENCSR811COL (accession numbers ENCFF434OGU and ENCFF024YRO).

ATAC-seq and estimation of ATAC-seq footprints

ATAC-seq data in HCT 116 cells was obtained from ENCFF296ZZB. The number of estimated ATAC-seq footprints in the locus was calculated as follows. The number of bases covered by ATAC-seq peaks in the MYC TAD was calculated (117,415 bp). Based on spacing observed in RCMC footprints, DNase I footprints, and in Hu et al.[109], we assumed that this data should yield one footprint per 50 to 150 bp at a sufficiently high read depth. This resulted in an estimate of 782-2,348 footprints in the MYC TAD.

PRO-cap data

PRO-cap data in HCT 116 cells was obtained from ENCODE experiment ENCSR360MUR (accession numbers ENCFF568VYK and ENCFF652SYZ).

Aggregate TF ChIP-seq analysis

Aggregate TF ChIP-seq signal (Figure 3.1C) was calculated by summing the normalized ChIP-seq signals for the 18 TFs listed in "Assigning TFs to footprints using publicly available ChIP-seq and DNA binding motifs" using bedtools 'makewindows' and 'coverage'.

Analysis of SNP proximity to enhancers

Positions of stitched enhancers in 86 human cell and tissue types were downloaded[107]. Their hg19 coordinates were converted into hg38 coordinates using liftOver[158] with parameter -ends=1000. The collapsed union of these converted enhancer sets was used for downstream proximity analyses.

Version 1.0 (All Associations) of the EMBL-EBI genome-wide association study database containing single-nucleotide polymorphisms associated with traits was downloaded (date 2024-11-03). This file was

processed as previously[107]. Briefly, rows with reference SNP identifiers, single positions in the reference genome, and that were reproduced in at least two independent studies were extracted, resulting in 67,288 trait-SNP associations. To quantify the relative proportion of trait categories represented in the trait-associated SNP database, traits with >100 associated SNPs were manually grouped.

Coding SNPs were filtered using coding exons of protein-coding genes. The hg38 positions of RefSeq genes were acquired from the UCSC Genome Browser, and a BED file of all exons was made using bedToExons -cdsOnly[158]. SNPs with positions in the hg38 reference genome were filtered against these exons using bedtools intersect -v[159]. The remaining SNPs were considered non-coding.

Non-coding SNPs whose coordinates overlapped the collapsed union of hg38 stitched enhancer coordinates were determined using bedtools intersect. For non-overlapping SNPs, the distance to the nearest enhancer was determined using bedtools closest.

Nuclear isolation and RNase treatment prior to RCMC

65 million wild type HCT 116 cells were trypsinized and washed twice with ice-cold PBS. We chose to isolate nuclei, in contrast using permeabilized intact cells as in prior attempts to conduct Hi-C after RNase treatment[160]. Cells were subsequently resuspended with 5 mL of cytoskeletal buffer (CSK, 10 mM PIPES pH 6.8, 100 mM KCl, 100 mM sucrose, 3 mM $MgCl_2$, with cOmplete EDTA-free protease inhibitor cocktail) supplemented with 0.5% Triton X-100 and incubated on ice for 3 minutes. Next, permeabilized nuclei were pelleted by centrifugation at 600 xg for 3 minutes at 4 °C and washed once with 5 mL of CSK buffer without Triton X-100. After the wash, nuclei were resuspended in 5 mL of CSK buffer supplemented with 250 U of RNase A/T1 cocktail (Thermo Fisher Scientific #AM2286) and incubated at 37 °C for 20 minutes with vigorous shaking. A parallel control experiment used equal volume of a “Mock” carrying buffer for the RNase enzyme (10 mM HEPES pH 7.2, 20 mM NaCl, 0.1% Triton X-100, 1 mM EDTA, 50% glycerol v/v) instead of the RNase A/T1 cocktail. Treated nuclei were pelleted by centrifugation at 1,800 x g for 3 minutes at 4 °C. Obtained nuclear pellets were then used as input to the DSG/formaldehyde

crosslinking step described in “Region Capture Micro-C”, and subsequent steps of RCMC, with subsequent calculations assuming 50 million nuclei were present at the start of crosslinking.

Confirmation of nuclear RNA degradation via reverse transcriptase quantitative polymerase chain reaction (RT-qPCR)

Total RNA was extracted following RNase or Mock treatment using the Qiagen RNeasy Mini kit (Qiagen #74104) with DNase I digestion to remove genomic DNA. Primers were designed targeting *PVT1*, *MYC* and *MYC* pre-mRNA. The *MYC* pre-mRNA primers were designed to detect the junction between the second exon and the second intron of the *MYC* gene. One-step RT-qPCR was performed using the NEB Luna Universal One-Step RT-qPCR kit (New England Biolabs #ES3005S). The sequences of the RT-qPCR primers are given below:

PVT1 F: GCCCCTTCTATGGGAATCACTA, *PVT1* R: GGGGCAGAGATGAAATCGTAAT, *MYC* F: AGCTGCTTAGACGCTGGATTTT, *MYC* R: TCGAGGTCATAGTTCCTGTTGG, *MYC* pre mRNA F: CACCACCAGCAGCGACTCT, *MYC* pre-mRNA R: CAATGAAAATGGGAAAGGTATCCA

Analysis of RCMC contact maps after RNase treatment

Sensitivity of chromatin contacts to RNase treatment was quantified by analyzing RCMC 3D contact maps using cooltools. Briefly, .mcool files were imported for RNase and Mock treated samples at a resolution of 10 kb without performing ICE balancing. Raw read counts in each matrix were rescaled by dividing the number of read counts in each bin by the total number of reads in the region of interest. Matrices containing rescaled read counts were divided to calculate a ratio of RNase to Mock read density. The resulting matrix was visualized on a base-10 log scale, where a lower value indicates that a particular genomic region was more sensitive to loss of 3D contact after RNase treatment.

Nuclear isolation and RNase treatment prior to mass spectrometry

20 million HCT 116 cells integrated with HA-Flag-GFP-MYC constructs were induced with doxycycline, collected, and washed twice with ice-cold PBS. Cells were subsequently resuspended with 2 mL of cytoskeletal buffer (CSK, 10 mM PIPES pH 6.8, 100 mM KCl, 100 mM sucrose, 3 mM $MgCl_2$, with cOmplete EDTA-free protease inhibitor cocktail) supplemented with 0.5% Triton X-100 and incubated on ice for 3 minutes. Next, permeabilized cells were pelleted by centrifugation at 600g for 3 minutes at 4 °C and washed once with 2 mL of CSK buffer without Triton X-100. After the wash, cells were resuspended in 2 mL of "CSK" buffer supplemented with 100 U of RNase A/T1 cocktail (Thermo Fisher Scientific #AM2286) and incubated at 37 °C for 20 minutes with vigorous shaking. A parallel control experiment used equal volume of "Mock" buffer (10 mM HEPES pH 7.2, 20 mM NaCl, 0.1% Triton X-100, 1 mM EDTA, 50% glycerol v/v) instead of the RNase A/T1 cocktail. Treated cells were pelleted by centrifugation at 1,800 xg for 3 minutes at 4 °C. Obtained cell pellets were used as input for subsequent experiments.

Chromatin extraction

Chromatin extraction was carried out using "Subcellular Protein Fractionation Kit for Cultured Cells" (Thermo Fisher Scientific #78840) essentially in accordance with manufacturer instructions. In brief, RNase and "Mock" treated pellets were resuspended in 250 uL nuclear extraction buffer (NEB) with protease inhibitor and incubated at 4 °C for 30 minutes. Next, material was pelleted by centrifugation at 5,000 x g for 5 minutes, and the resulting pellet was resuspended in NEB supplemented with MNase, $CaCl_2$, and protease inhibitors. Subsequent MNase digestion was done by incubation at 37 °C for 5 minutes. Resulting insoluble material was pelleted by centrifugation at 16,000 xg for 5 minutes, and supernatant was kept as the chromatin-associated fraction.

Proteolytic digestion and peptide extraction for mass spectrometry

Concentrations of proteins in Nuclear Extraction Reagent from the NE-PER™ Nuclear and Cytoplasmic Extraction Reagents kit (Thermo Fisher) were determined using the Pierce 660 nm Protein Assay (Thermo Fisher). Prior to the purification of proteins via the SP3 as in Hughes *et al.*[161], 25 µg protein was diluted in

90 µL 1X SDS lysis buffer (5 % (w/v) SDS, 100mM TEAB pH 8.5, 40mM CAA, 10mM TCEP) in Low Protein Binding Microcentrifuge Tubes (Thermo Scientific). Proteins were denatured at 70 °C for 15 min prior to the addition of 10 µL SP3 beads. Proteins were bound to the SP3 beads by adding 500 µL 100 % (v/v) ethanol followed by three washing steps with 500 µL 80 % (v/v) ethanol. After that, 1 µg trypsin/LysC mix was added in 1 µL 100 mM TEAB and proteins were digested overnight in a shaking incubator at 37 °C at 115 RPM. The following day, an additional 1 µg trypsin/LysC mix was added in 1 µL 100mM TEAB and the digestion continued for 4 hours at 37 °C. Peptide digests were purified using stage tips as in Rappsilber *et al.*[162]. Purified peptides were then dried using a speed-vac concentrator and resuspended in 12 µL 0.2% (v/v) formic acid in MS-grade water. Peptide concentrations were determined using the Pierce Quantitative Fluorometric Peptide Assay (Thermo Scientific) following the manufacturer's directions.

Liquid chromatography mass spectrometry (LC/MS) data collection

LC-MS/MS data collection was executed on the Vanquish Neo nanoLC system paired with an Orbitrap Eclipse mass spectrometer, equipped with a FAIMS Pro Interface and an Easy Spray ESI source, all by Thermo Fisher Scientific (Waltham, MA, USA).

NanoLC separation utilized an Acclaim PepMap trap column (75µm x 2 cm) combined with an EasySpray ES902 column (75 µm x 25 cm, 100 Å) from Thermo Fisher Scientific. The peptide extracts were injected with a volume of 5 µL. Peptide separation was conducted with a mobile phase consisting of 0.1% (v/v) formic acid in water (solution A) and 0.1% (v/v) formic acid in 80% (v/v) acetonitrile (solution B), flowing at 300 nL/min, while the column temperature was held constant at 40 °C. The column underwent initial conditioning with 3% solution B, followed by a gradual linear gradient up to 40% solution B over a 60-minute period. Any remaining peptides adhering to the C18 resin were subsequently washed off with 95% solvent B for 6.0 minutes.

Using the MS in positive mode, the ion source temperature was set to 305 °C and 2000 V, and ionized peptides were passed through the FAIMS Pro unit at -50 V. Mass spectra were collected in MS1 mode with a resolution of 120,000, spanning the mass range of 350-2000 m/z. This was carried out using standard automatic gain control settings and automatic injection time. For MS2 data acquisition, the MS was operated in DIA mode with a resolution of 30,000. MS2 spectra were collected across a precursor mass range of m/z 375-1200 using isolation windows of m/z 25, overlaps of m/z 0.5, custom AGC target (1000 units), 30% normalized collision energy, and 30,000 resolution.

Analysis of LC/MS data

To analyze the proteome data, the DIA-NN 1.9 software platform[163] was utilized. The LC-MS data files were processed with following settings: the Homo sapiens FASTA database (UP000005640) was used for library generation and library-free searches. Trypsin/P was the chosen protease, allowing for one missed cleavage and N-terminal methionine excision, with carbamidomethylation of cysteines. Peptides considered ranged from 7 to 30 amino acids in length. The precursor charge ranged from 1 to 4, with m/z values between 300 and 1800, and fragment ions had an m/z range of 200-1800. The precursor false discovery rate (FDR) was set to 1.0%. Protein interference was analyzed at the gene level using the heuristic protein inference for filtering. A single-pass mode neural network classifier was used. Quantitation was performed using the QuantUMS (high precision) algorithm, with RT-dependent cross-run normalization. The library was generated through IDs, RT, & IM profiling, and both speed and RAM usage were optimized for optimal results.

Statistical analysis

Genomic feature overlap analysis was conducted using a permutation-based statistical test. For overlap between enhancers and RCMC loop regions, the genomic coordinates of RCMC loop regions were randomly permuted 100 times while keeping feature sizes constant. The overlap between these permuted regions and the locations of ENCODE enhancers was calculated using bedtools ‘intersect’. The percentage

of overlapping features was averaged over these 100 trials and compared to the observed overlap using a binomial test conducted in PRISM (v.9). A p-value threshold of 0.05 was used to define statistical significance.

Mass spectrometry results were analyzed using an unpaired two-tailed Student's t-test (n=4 per condition) in a custom R script. A p-value threshold of 0.05 was used to define statistical significance without multiple hypothesis correction.

The significance of RCMC loop anchors was assessed using Mustache as described in "Identification of RCMC loop regions." A p-value threshold of 0.1 was used to define statistical significance.

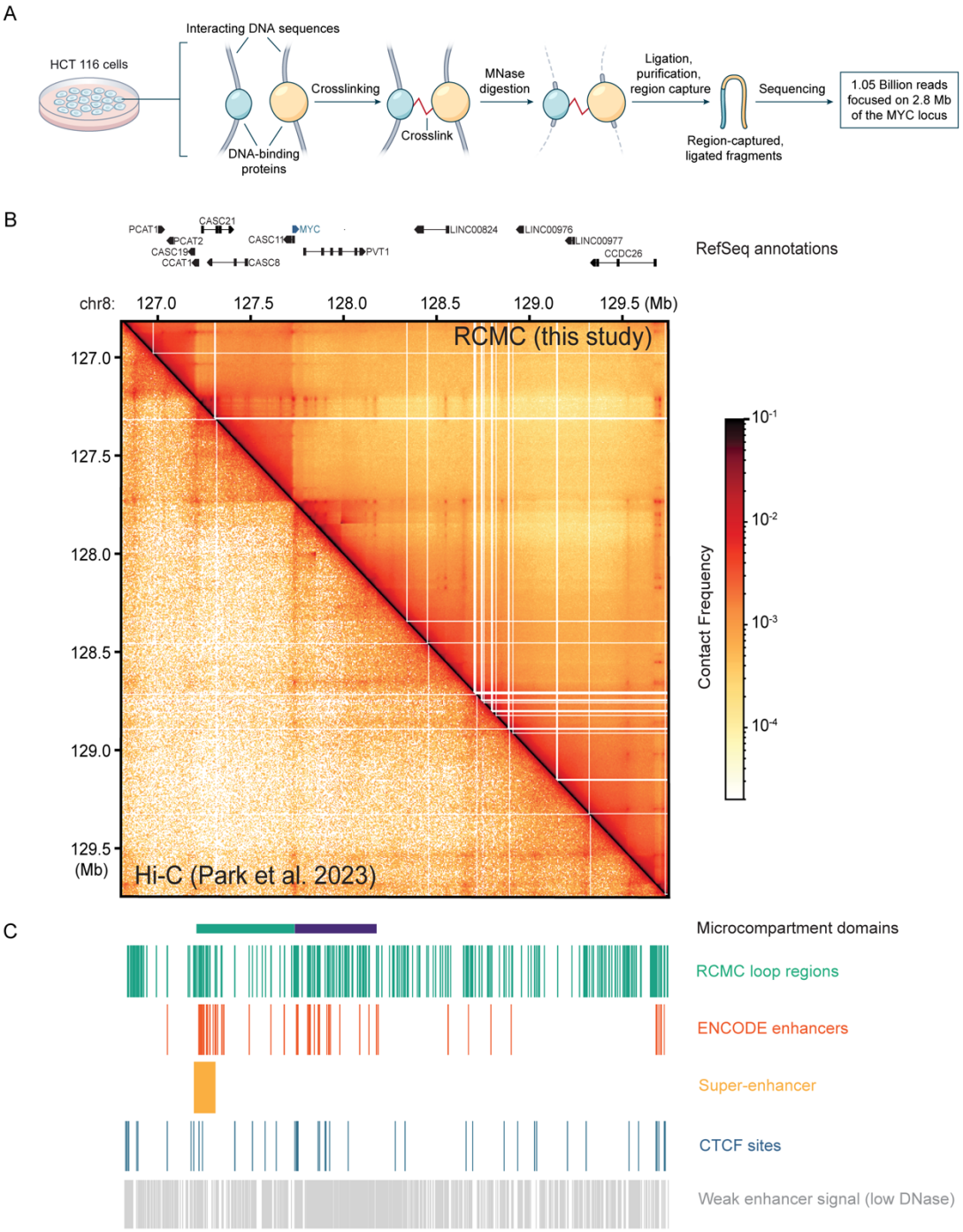


Figure C.1 Region Capture Micro-C (RCMC) reveals the 3D chromatin structure of the 2.8 Mb *MYC* TAD. A Schematic of Region Capture Micro-C conducted on HCT 116 colorectal cancer cells. **B** RCMC chromatin contact map of the MYC TAD. RefSeq annotations of protein-coding genes and long noncoding RNAs are shown above the contact heatmap. The RCMC contact map is symmetric about the diagonal, allowing multiple datasets to be compared. (Above matrix diagonal) RCMC data collected in this study (bin size = 1 kb). (Below matrix diagonal) Hi-C data collected in Park et al. (bin size = 5 kb). The locations of two microcompartment domains are shown below the contact map (green and purple bars). **C** Locations

of RCMC contact sites (green tick marks) and ENCODE annotations of enhancers (orange tick marks), CTCF sites (blue tick marks), and weak enhancer signals (grey tick marks).

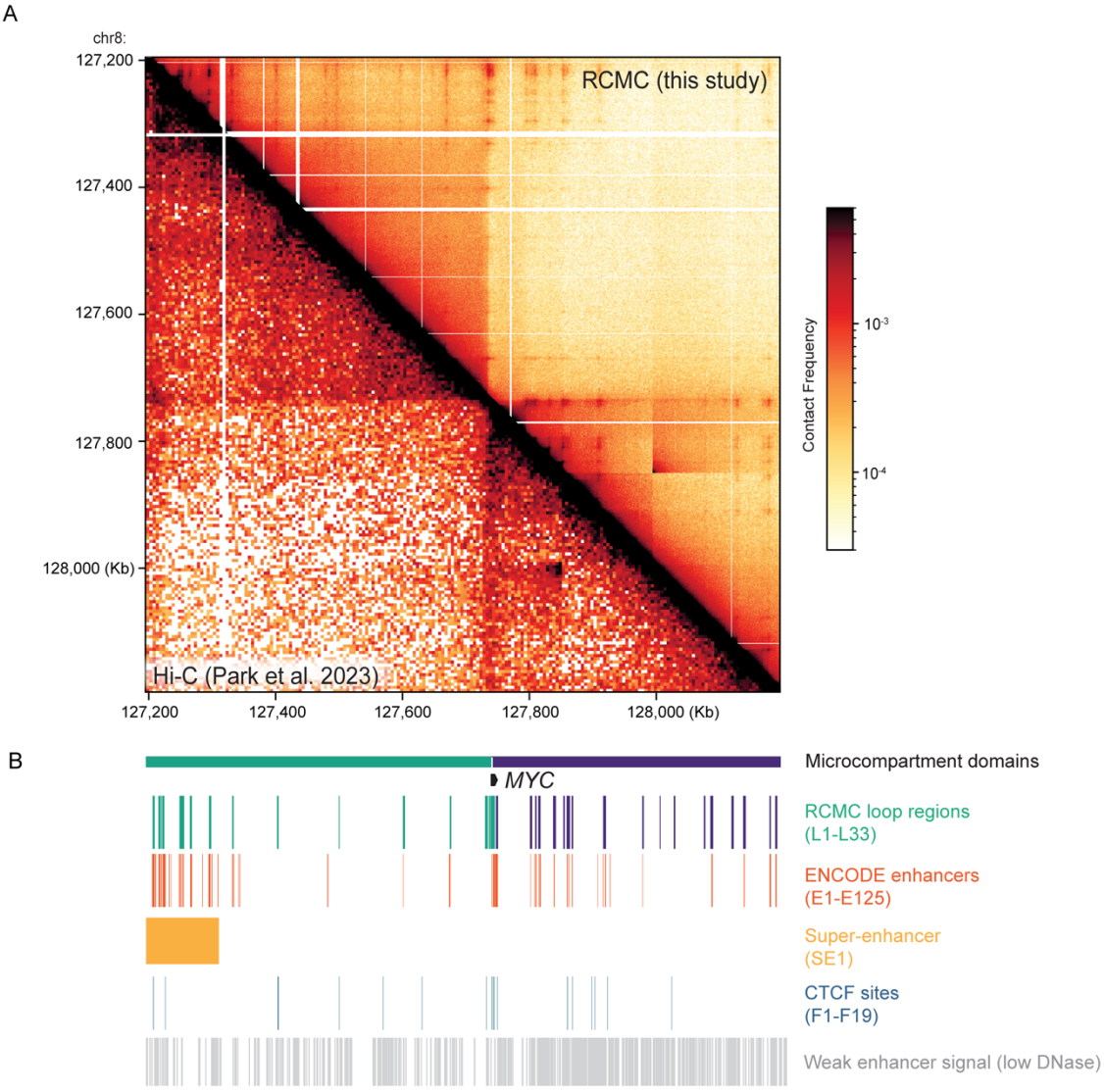


Figure C.2 Microcompartment domains in the 2.8 Mb *MYC* TAD. A Contact map for the *MYC*-proximal microcompartment domains. (Above matrix diagonal) RCMC data collected in this study (bin size = 1 kb). (Below matrix diagonal) Hi-C data collected in Park et al., displayed at the minimum bin size for this dataset (bin size = 5 kb). **B** Locations of RCMC contact sites (green and purple tick marks) and ENCODE annotations of enhancers (orange tick marks), super-enhancer annotation (yellow tick marks), and CTCF sites (blue tick marks) in the *MYC*-proximal microcompartment domains (green and purple bars) spanning 555 and 445 kb, respectively.

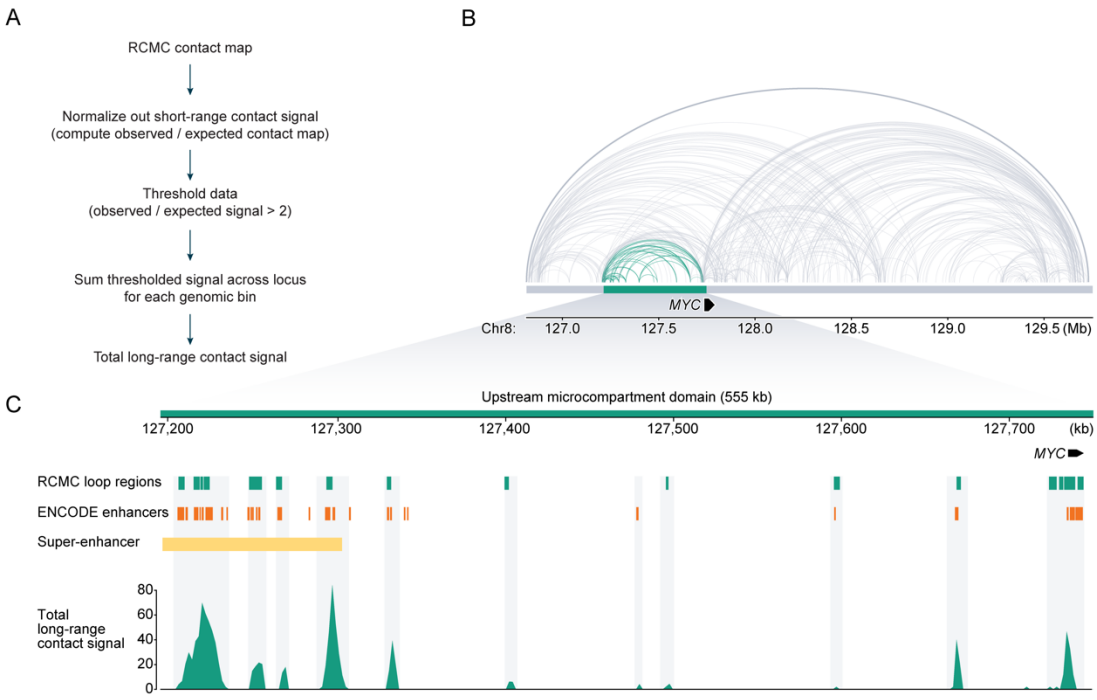


Figure C.3 Regulatory elements preferentially engage in long-range interactions. A Schematic of method for calculating total long-range contact signal. **B** Map of RCMC 3D loops across *MYC* locus for navigational purposes (same as Figure 3.1B). **C** Total long-range contact signal in the microcompartment domain upstream of *MYC*, in relation to RCMC loop regions, ENCODE enhancers, and super-enhancers in HCT 116 cells.

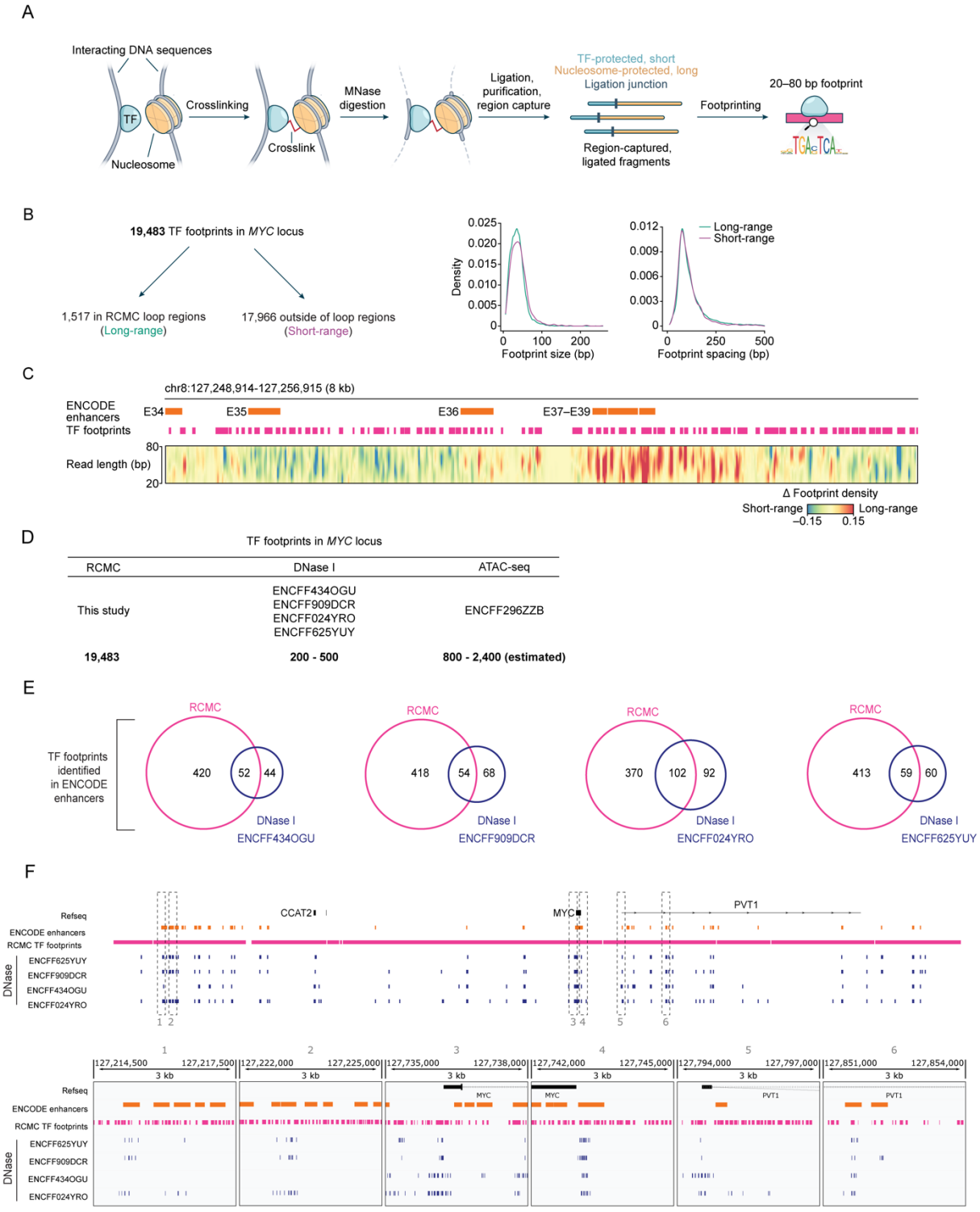


Figure C.4 Transcription factor footprinting and benchmarking against DNase I and ATAC-seq footprinting. A Schematic of TF footprinting using RCMC. Interacting DNA sequences containing TFs and nucleosomes are crosslinked to fix spatially proximal DNA-DNA interactions. Digestion of DNA with MNase results in DNA fragments protected from digestion where proteins are bound. Proximity ligation

joins spatially proximal DNA sequences. DNA is purified and a region capture step is used to enrich fragments within the specified locus (here, the 2.8 Mb *MYC* locus). Fragments protected by TFs (20-80 bp) are preserved during size selection when they are ligated to the longer nucleosome-sized fragments (120-180 bp). TF-sized 20-80 bp fragments are used to identify TF footprints as described in Figure 3.2A. **B** (Left) TF footprints can be the result of long-range 3D contacts or short-range contacts. Long-range contacts can be identified by RCMC loop regions, while short-range contacts occur outside of loop regions. (Right) Distribution of TF footprint size and spacing between footprint centers at found at long-range contacts and short-range contacts. **C** Long-range TF footprints tend to occur near ENCODE enhancers. Top track: Multiple enhancers (orange) within an 8 kb window. Middle track: TF footprints (pink). Bottom track: Heatmap displaying the difference between footprint maps of long-range contacts and short-range contacts (see Methods). Long-range footprints are indicated by enrichment of red signal, while short-range footprints are indicated by enrichment of blue signal. **D** Benchmarking total TF footprint detection in the *MYC* locus by RCMC against DNase I and ATAC-seq footprinting. DNase I footprints are based on footprint counts from the datasets indicated. ATAC-seq footprints are estimated based on the dataset indicated (see Methods). **E** Benchmarking TF footprint detection in ENCODE enhancers within the MYC locus by RCMC (pink) against DNase footprinting (dark blue) for the datasets indicated. **F** ENCODE enhancers frequently lack DNase footprints but contain RCMC footprints. (Top) Locations of ENCODE enhancers (orange), RCMC footprints (pink), and DNase footprints (dark blue) and locations of six zoom-in boxes. (Bottom) Six zoom-ink boxes of 3 kb showing enhancer footprinting by RCMC and DNase.

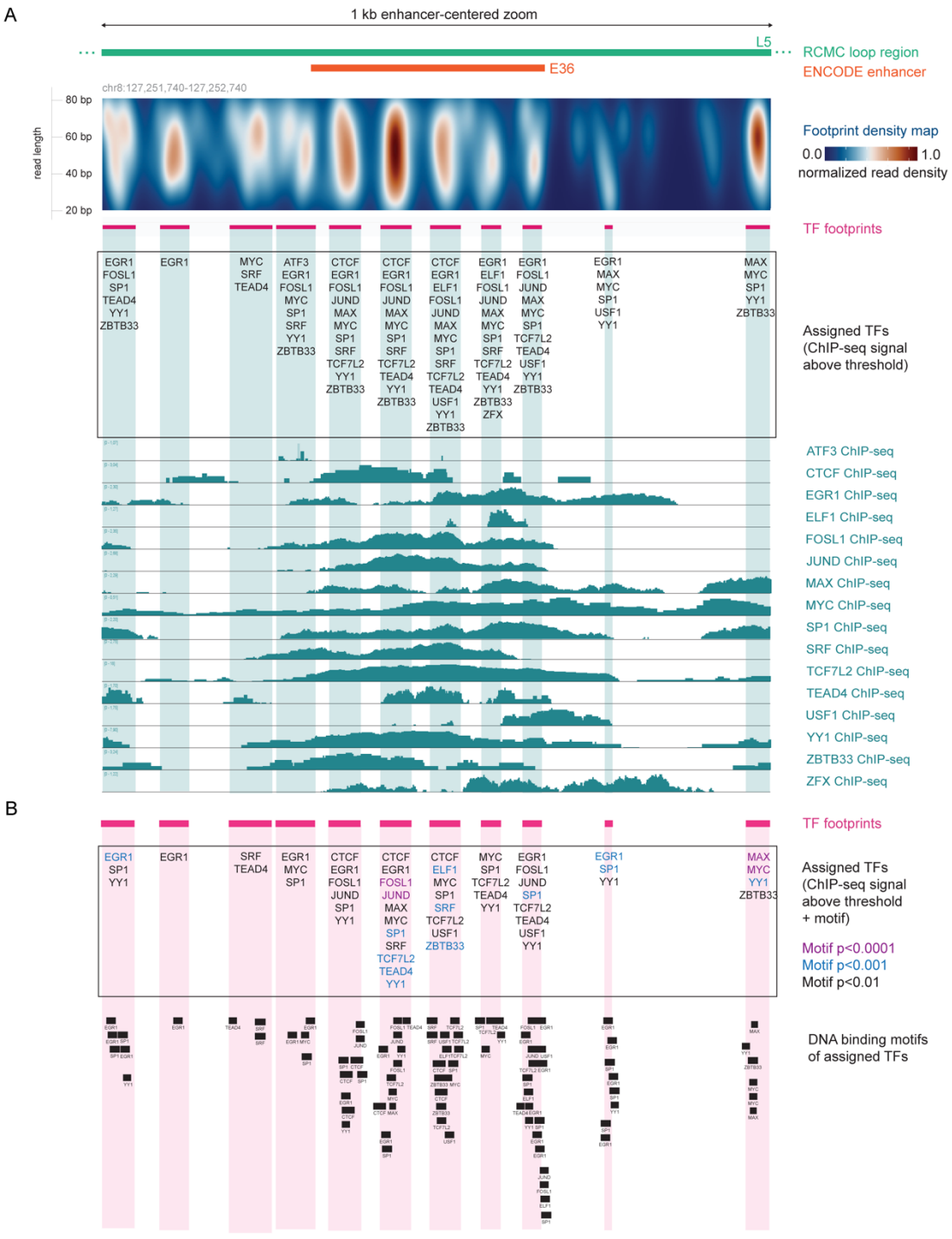


Figure C.5 ChIP-seq data and DNA binding motifs support TF assignments. A ChIP-seq signals used for TF assignment in the vicinity of ENCODE enhancer E36, as shown in Figure 3.2C. TFs whose ChIP-seq signal in TF footprints within this region exceeded their respective threshold are listed in the box, with ChIP-seq tracks shown below. **B** Motifs used for TF assignment in the vicinity of ENCODE enhancer E36, as shown in Figure 3.2C. Motifs were identified for the factors listed at each footprint above exceeding ChIP-seq thresholds (see C.5A). Factor names are colored according to the p-value of their motif as assigned

by FIMO (purple: p<0.0001, blue, p<0001, black, p<0.01). Only motifs falling within footprints are shown below (at least half of the motif occurs within the boundaries of the footprint). If multiple motifs for the same TF were found within a footprint, the most significant motif was used for illustrative purposes in Figure 3.2D. A motif for ATF3 was not found within the region.

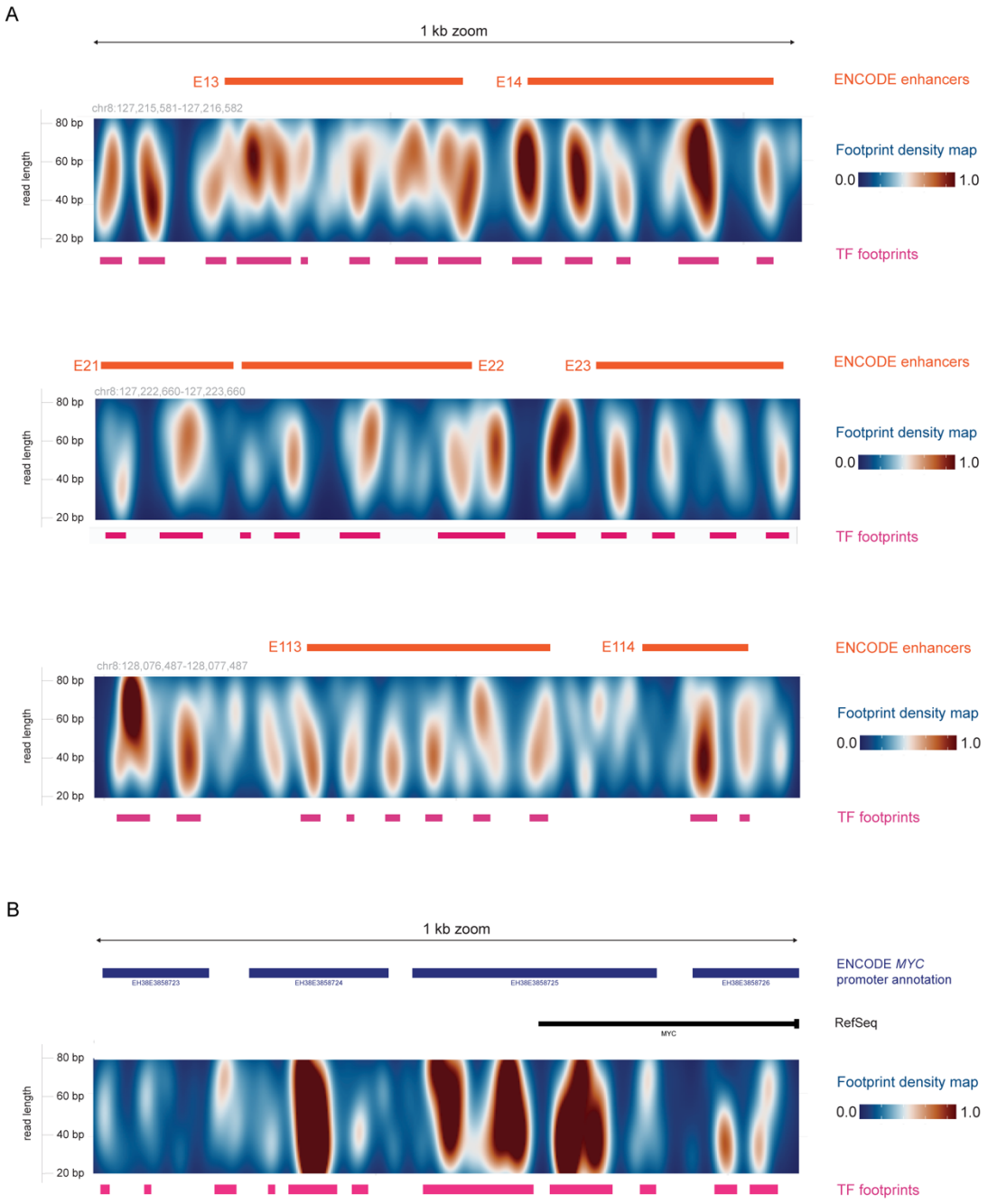


Figure C.6 TF footprints at *MYC* enhancers and *MYC* promoter. A TF footprints maps in three 1 kb regions containing ENCODE enhancers. Top: Enhancers E13-14. Middle: Enhancers E22 and E23, with partial view of E21. Bottom: Enhancers E113, E114. Footprint density maps are separately normalized by region. **B** TF footprint map in the 1 kb region containing the *MYC* promoter. ENCODE promoter

annotations are shown in dark blue. RefSeq annotation of the *MYC* 5'UTR and a portion of exon 1 is shown in black.

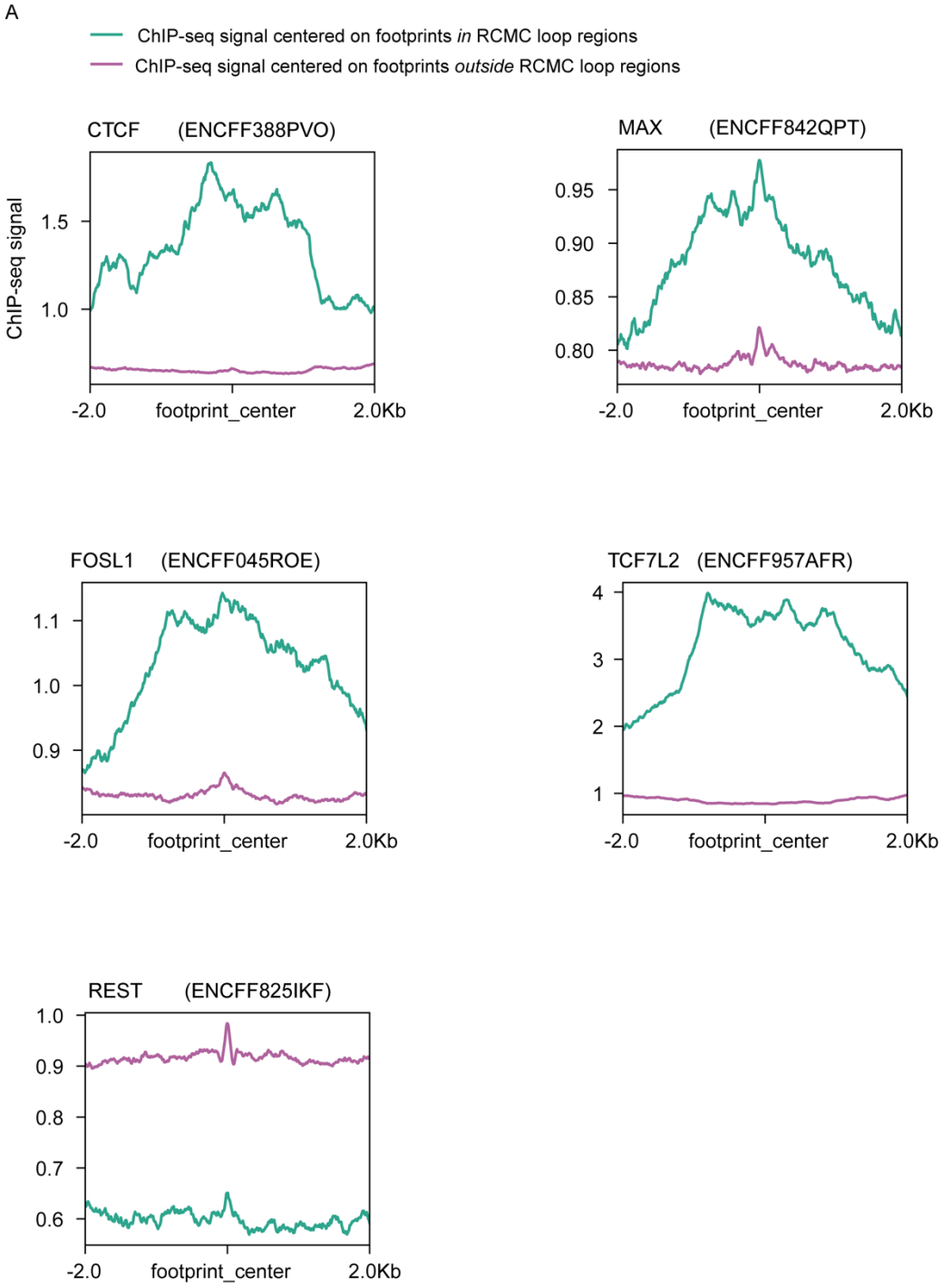


Figure C.7 TF ChIP-seq signal density differs inside and outside of enhancer-associated RCMC loop regions. A Metaplots of ChIP-seq signal density centering on TF footprints found in long-range contacts

(RCMC loop regions, green line) and short-range contacts (non-loop regions, purple line). Factors shown (left to right and top to bottom): CTCF, MAX, FOSL1, TCF7L2, REST. Y-axis represents aggregated ChIP-seq signal, x-axis shows 2 kb to each side of aggregated footprint centers.

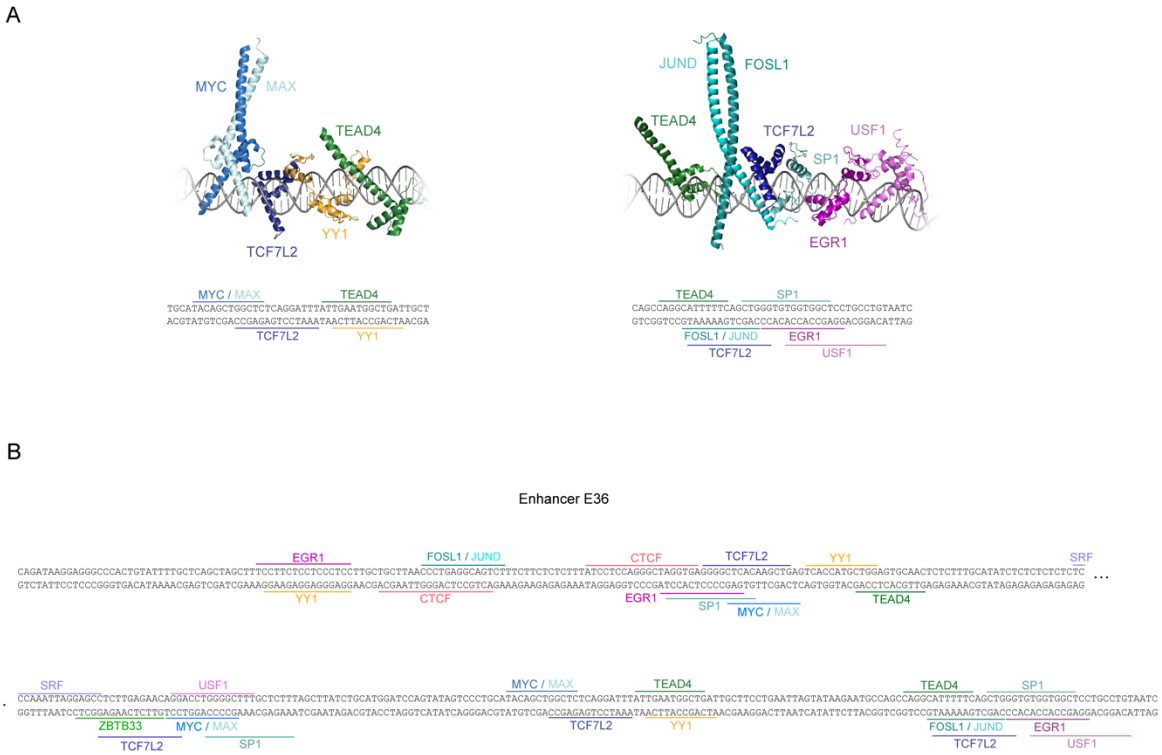


Figure C.8 Nano-assembly model of an enhancer. A Structural models for TFs in the remaining two of the five E36 enhancer footprints (related to Figure 3.2D). The models consist of a subset of the TFs listed in (C), docked on their motifs using available co-crystal structures (for EGR1, YY1, MYC, MAX, TEAD4, SRF, ZBTB33, USF1), published docked crystal structures (for SP1), or TF-DNA structures generated by AlphaFold3 (for FOSL1, JUND, CTCF, TCF7L2). Some TFs were not displayed in the model due to steric constraints. DNA sequences and locations of TF DNA binding motifs are shown beneath each assembly. **B** Complete DNA sequence of ENCODE enhancer E36 including basepair-resolution assignments of TFs in nano-assembly footprints. Partial sequences are shown in Figure 3.2D and Figure C.8A, along with structural models matching each sequence.

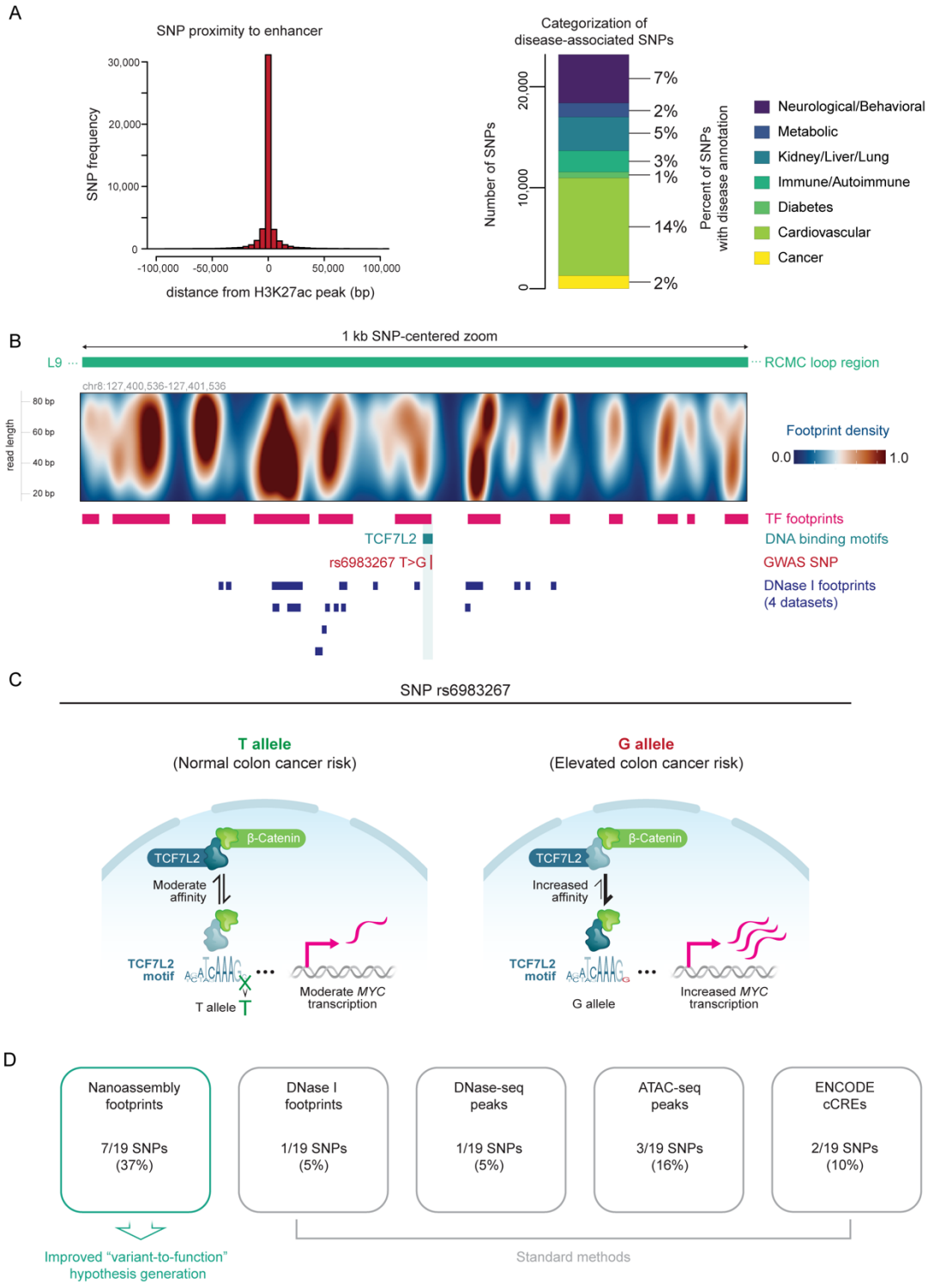


Figure C.9 Nano-assemblies facilitate interpretation of disease-associated genetic variation. A Analysis of 67,288 single nucleotide polymorphisms (SNPs). (Left) Proximity of SNPs to enhancers defined by H3K27ac signal. SNPs located within enhancers are assigned to the 0 bin. 61,845 (93%) of the

SNPs are non-coding, and 4,847 (7%) are coding. (Right) Bar chart representing the composition of the SNP catalogue according to phenotypic traits or diseases. SNPs mapping to disease-associated traits shown represent 34% of the total number of SNPs analyzed here. **B** Nano-assembly footprints identify sites of genomic variation and are consistent with proposed mechanisms. Location of a prominent risk SNP for colorectal cancer within the context of a nano-assembly model. The rs6983267 SNP (red) occurs within a TCF7L2 DNA binding motif (teal) at the edge of a footprint assigned to TCF7L2. The footprint occurs within RCMC loop region L9 (green). Binding of the TCF7L2 factor to its motif is not detected in four DNase I footprinting datasets (blue) (see Methods for accession numbers). **C** Model of the putative mechanism for a prominent SNP in the *MYC* locus associated with increased colon cancer risk. A thymine (T allele) at the rs6983267 SNP position creates a suboptimal TCF7L2 DNA binding motif and is associated with moderate TCF7L2 binding and basal *MYC* expression in response to Wnt/β-catenin signaling in the colon. Substitution with a guanine (G allele) creates a high-affinity TCF7L2 binding motif, leading to increased binding of this transcriptional activator and elevated *MYC* expression[138]. The HCT 116 colon cancer cell line studied here contains the cancer-associated G allele. **D** Nano-assembly footprints assist with variant-to-function studies of genomic variants. GWAS analysis identified 19 SNPs in the *MYC* locus associated with colorectal cancer or co-inherited with risk-associated SNPs. Nano-assembly footprinting reveals footprints at 7 of these 19 SNPs, enabling hypothesis generation related to TF binding. These footprints reveal notably more candidates for variant-to-function studies than standard methods such as identification of DNase footprints, DNase-seq hypersensitive site peaks, ATAC-seq peaks, and ENCODE candidate cis-regulatory elements (cCREs).

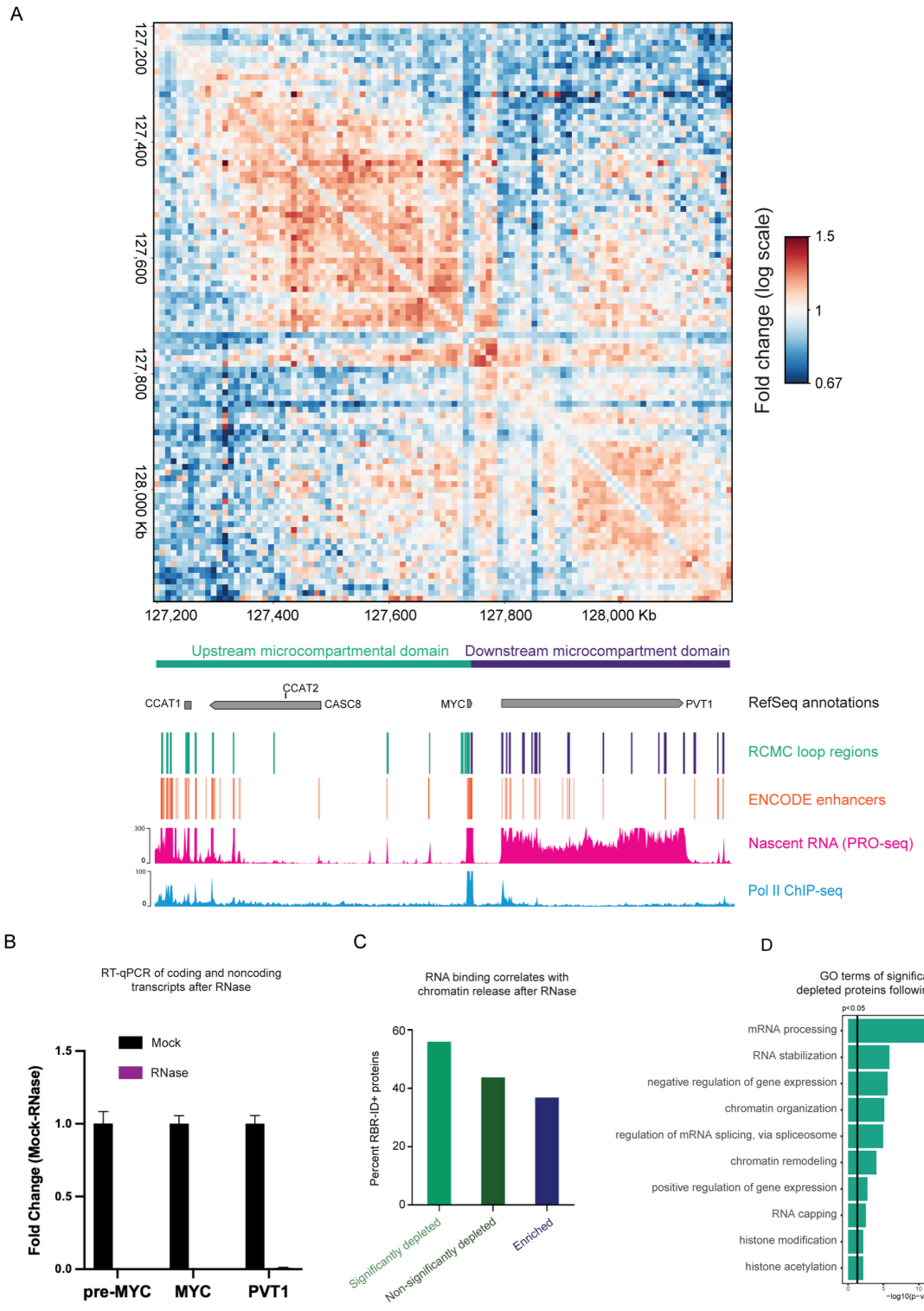


Figure C.10 RNA-dependent features of chromatin architecture and TF assemblies. A (Top) Difference between RNase and Mock-treated RCMC contact maps for the upstream and downstream *MYC* microcompartment domains shown in Figure 3.1A. Color bar represents the fold change between RNase and Mock read density per bin (without ICE balancing), red corresponds to increased signal after RNase

treatment while blue corresponds to diminished signal after RNase treatment (bin size = 10 kb). (Middle) Annotation of microcompartments, RefSeq RNA annotations, RCMC loop regions, ENCODE enhancers. (Bottom) PRO-seq and Pol II ChIP-seq signals indicate sites of active transcription. **B** RT-qPCR quantification of the levels of *PVT1* lncRNA, *MYC* mRNA, and *MYC* pre-mRNA in Mock- and RNase-treated isolated nuclei. RNA levels following RNase (purple) are too low to be visible on the plot. **C** Correlation between protein chromatin association after RNase treatment as detected by MS and RNA binding as detected via RNA binding region identification (RBR-ID). Bar plot y-axis shows the percentage of proteins in each category of the MS experiment that also crosslink to RNA in the RBR-ID assay. RBR-ID was conducted in K562 cells[113]. **D** Gene ontology (GO) terms for proteins significantly depleted from chromatin following RNase treatment. The top 10 GO terms are displayed, ranked by p-value.

See supplemental files (for large tables).

Table C.1 RCMC loop regions assigned to enhancer and promoter elements defined by ENCODE. Table displaying RCMC loop region IDs and coordinates, assigned to ENCODE enhancer and promoter element IDs along with enhancer coordinates (left column). Right column lists loop region IDs and coordinates, assigned to ENCODE CTCF element IDs and corresponding CTCF element coordinates in hg38.

Table C.2 Footprint assignment using TF DNA-binding motifs alone. Table listing RCMC loop region IDs and coordinates, with embedded footprint coordinates. Footprints are assigned to TFs based on DNA-binding motifs for factors meeting an RNA expression threshold in HCT 116 cells of >1 transcript per million. For each footprint, the TF name, motif coordinates, and expression level in HCT 116 cells are reported.

Table C.3 Nascent RNA expression at studied enhancers. Table displaying enhancer IDs, enhancer coordinates, and ENCODE IDs, followed by RPKM-normalized expression levels for HCT 116 PRO-seq replicates 1 and 2. Each replicate is labeled 'TRUE' if expression exceeds a cutoff of 10 RPKM, and 'FALSE' otherwise.

Table C.4 TFs with RNA-binding protein characteristics. Table displaying protein group, protein names, gene name, and protein description, followed by M (Mock treated) and R (RNase treated) mass spectrometry replicate values, mean values for each group, $\log_2$ fold change, t-test p-value and if protein is characterized as RBR-ID in Oksuz et al. 2023.

References

1. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**, 2038–2049 (2016).

2. Lupiáñez, D. G. *et al.* Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell* **161**, 1012–1025 (2015).

3. Flavahan, W. A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature* **529**, 110–114 (2016).

4. Hirayama, T., Tarusawa, E., Yoshimura, Y., Galjart, N. & Yagi, T. CTCF Is Required for Neural Development and Stochastic Expression of Clustered Pcdh Genes in Neurons. *Cell Rep.* **2**, 345–357 (2012).

5. Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* **47**, 818–821 (2015).

6. Li, Y. *et al.* The structural basis for cohesin–CTCF-anchored loops. *Nature* **578**, 472–476 (2020).

7. Narendra, V. *et al.* CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* **347**, 1017–1021 (2015).

8. Huang, H. *et al.* CTCF mediates dosage- and sequence-context-dependent transcriptional insulation by forming local chromatin domains. *Nat. Genet.* **53**, 1064–1074 (2021).

9. Nora, E. P. *et al.* Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* **169**, 930-944.e22 (2017).

10. Kubo, N. *et al.* Promoter-proximal CTCF-binding promotes long-range-enhancer dependent gene activation. *Nat. Struct. Mol. Biol.* **28**, 152–161 (2021).

11. Schuijers, J. *et al.* Transcriptional Dysregulation of MYC Reveals Common Enhancer-Docking Mechanism. *Cell Rep.* **23**, 349–360 (2018).

12. Cerda-Smith, C. G. *et al.* Integrative PTEN Enhancer Discovery Reveals a New Model of Enhancer Organization. 2023.09.20.558459 Preprint at https://doi.org/10.1101/2023.09.20.558459 (2023).

13. Davidson, I. F. *et al.* DNA loop extrusion by human cohesin. *Science* **366**, 1338–1345 (2019).

14. Gabriele, M. *et al.* Dynamics of CTCF- and cohesin-mediated chromatin looping revealed by live-cell imaging. *Science* **376**, 496–501 (2022).

15. Mach, P. *et al.* Cohesin and CTCF control the dynamics of chromosome folding. *Nat. Genet.* **54**, 1907–1918 (2022).

16. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).

17. Yang, J.-H. *et al.* Loss of epigenetic information as a cause of mammalian aging. *Cell* **186**, 305-326.e27 (2023).

18. Kim, H.-M. *et al.* Forkhead box protein D2 suppresses colorectal cancer by reprogramming enhancer interactions. *Nucleic Acids Res.* **51**, 6143–6155 (2023).

19. Xiao, L. *et al.* Targeting SWI/SNF ATPases in enhancer-addicted prostate cancer. *Nature* **601**, 434–439 (2022).

20. Goel, V. Y., Huseyin, M. K. & Hansen, A. S. Region Capture Micro-C reveals coalescence of enhancers and promoters into nested microcompartments. *Nat. Genet.* 1–9 (2023) doi:10.1038/s41588-023-01391-1.

21. Kim, K. L. *et al.* Dissection of a CTCF topological boundary uncovers principles of enhancer-oncogene regulation. *Mol. Cell* **84**, 1365-1376.e7 (2024).

22. Goel, V. Y. *et al.* Dynamics of microcompartment formation at the mitosis-to-G1 transition. 2024.09.16.611917 Preprint at https://doi.org/10.1101/2024.09.16.611917 (2024).

23. Kasinathan, S., Orsi, G. A., Zentner, G. E., Ahmad, K. & Henikoff, S. High-resolution mapping of transcription factor binding sites on native chromatin. *Nat. Methods* **11**, 203–209 (2014).

24. Henikoff, J. G., Belsky, J. A., Krassovsky, K., MacAlpine, D. M. & Henikoff, S. Epigenome characterization at single base-pair resolution. *Proc. Natl. Acad. Sci.* **108**, 18318–18323 (2011).

25. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *eLife* **6**, e21856 (2017).

26. Gutin, J. *et al.* Fine-Resolution Mapping of TF Binding and Chromatin Interactions. *Cell Rep.* **22**, 2797–2807 (2018).

27. Hua, P. *et al.* Defining genome architecture at base-pair resolution. *Nature* **595**, 125–129 (2021).

28. Bhattacharyya, S., Chandra, V., Vijayanand, P. & Ay, F. Identification of significant chromatin contacts from HiChIP data by FitHiChIP. *Nat. Commun.* **10**, 4221 (2019).

29. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.* **48**, D882–D889 (2020).

30. Chereji, R. V., Bryson, T. D. & Henikoff, S. Quantitative MNase-seq accurately maps nucleosome occupancy levels. *Genome Biol.* **20**, 198 (2019).

31. Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The Insulator Binding Protein CTCF Positions 20 Nucleosomes around Its Binding Sites across the Human Genome. *PLOS Genet.* **4**, e1000138 (2008).

32. Joseph, S. R. *et al.* Competition between histone and transcription factor binding regulates the onset of transcription in zebrafish embryos. *eLife* **6**, e23326 (2017).

33. Cattoglio, C. *et al.* Determining cellular CTCF and cohesin abundances to constrain 3D genome models. *eLife* **8**, e40164 (2019).

34. Dozmorov, M. G. *et al.* CTCF: an R/bioconductor data package of human and mouse CTCF binding sites. *Bioinforma. Adv.* **2**, vbac097 (2022).

35. Bailey, T. L. STREME: accurate and versatile sequence motif discovery. *Bioinformatics* **37**, 2834–2840 (2021).

36. Nakahashi, H. *et al.* A Genome-wide Map of CTCF Multivalency Redefines the CTCF Code. *Cell Rep.* **3**, 1678–1689 (2013).

37. Boyle, A. P. *et al.* High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.* **21**, 456–464 (2011).

38. Nora, E. P. *et al.* Molecular basis of CTCF binding polarity in genome folding. *Nat. Commun.* **11**, 5612 (2020).

39. Gassler, J. *et al.* A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture. *EMBO J.* **36**, 3600–3618 (2017).

40. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–2492 (2017).

41. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **41**, 827–841 (2013).

42. Banigan, E. J. *et al.* Transcription shapes 3D chromatin organization by interacting with loop extrusion. *Proc. Natl. Acad. Sci.* **120**, e2210480120 (2023).

43. Barshad, G. *et al.* RNA polymerase II dynamics shape enhancer–promoter interactions. *Nat. Genet.* 1–11 (2023) doi:10.1038/s41588-023-01442-7.

44. Faure, A. J. *et al.* Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. *Genome Res.* **22**, 2163–2175 (2012).

45. Hsieh, T.-H. S. *et al.* Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. *Mol. Cell* **78**, 539-553.e8 (2020).

46. Lengronne, A. *et al.* Cohesin relocation from sites of chromosomal loading to places of convergent transcription. *Nature* **430**, 573–578 (2004).

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

48. Hou, C., Li, L., Qin, Z. S. & Corces, V. G. Gene Density, Transcription, and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains. *Mol. Cell* **48**, 471–484 (2012).

49. Bonev, B. *et al.* Multiscale 3D Genome Rewiring during Mouse Neural Development. *Cell* **171**, 557-572.e24 (2017).

50. Oh, S. *et al.* Enhancer release and retargeting activates disease-susceptibility genes. *Nature* **595**, 735–740 (2021).

51. Guckelberger, P. *et al.* Cohesin-mediated 3D contacts tune enhancer-promoter regulation. 2024.07.12.603288 Preprint at https://doi.org/10.1101/2024.07.12.603288 (2024).

52. Pownall, M. E. *et al.* Chromatin expansion microscopy reveals nanoscale organization of transcription and chromatin. *Science* **381**, 92–100 (2023).

53. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

54. Abdennur, N. *et al.* Pairtools: from sequencing data to chromosome contacts. *bioRxiv* 2023.02.13.528389 (2023) doi:10.1101/2023.02.13.528389.

55. Zhao, H. *et al.* CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).

56. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

57. Murtagh, F. & Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? *J. Classif.* **31**, 274–295 (2014).

58. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The selection and function of cell type-specific enhancers. *Nat. Rev. Mol. Cell Biol.* **16**, 144–154 (2015).

59. Zhou, H. Y. *et al.* A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes Dev.* **28**, 2699–2711 (2014).

60. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735 (2003).

61. Bulger, M. & Groudine, M. Functional and Mechanistic Diversity of Distal Transcription Enhancers. *Cell* **144**, 327–339 (2011).

62. Li, Y. *et al.* The structural basis for cohesin–CTCF-anchored loops. *Nature* **578**, 472–476 (2020).

63. Kim, S., Yu, N.-K. & Kaang, B.-K. CTCF as a multifunctional protein in genome regulation and gene expression. *Exp. Mol. Med.* **47**, e166–e166 (2015).

64. Phillips, J. E. & Corces, V. G. CTCF: Master Weaver of the Genome. *Cell* **137**, 1194–1211 (2009).

65. Hsieh, T.-H. S. *et al.* Enhancer–promoter interactions and transcription are largely maintained upon acute loss of CTCF, cohesin, WAPL or YY1. *Nat. Genet.* **54**, 1919–1932 (2022).

66. Weintraub, A. S. *et al.* YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell* **171**, 1573-1588.e28 (2017).

67. Cubeñas-Potts, C. & Corces, V. G. Architectural Proteins, Transcription, and the Three-dimensional Organization of the Genome. *FEBS Lett.* **589**, 2923–2930 (2015).

68. Krietenstein, N. *et al.* Ultrastructural Details of Mammalian Chromosome Architecture. *Mol. Cell* **78**, 554-565.e7 (2020).

69. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat. Commun.* **10**, 1930 (2019).

70. Sept, C. E. *et al.* High-resolution CTCF footprinting reveals impact of chromatin state on cohesin extrusion dynamics. 2023.10.20.563340 Preprint at https://doi.org/10.1101/2023.10.20.563340 (2023).

71. Cutter, A. R. & Hayes, J. J. A brief review of nucleosome structure. *FEBS Lett.* **589**, 2914–2922 (2015).

72. Puig, R. R., Boddie, P., Khan, A., Castro-Mondragon, J. A. & Mathelier, A. UniBind: maps of high-confidence direct TF-DNA interactions across nine species. *BMC Genomics* **22**, 482 (2021).

73. Adachi, K. *et al.* Esrrb Unlocks Silenced Enhancers for Reprogramming to Naive Pluripotency. *Cell Stem Cell* **23**, 266-275.e6 (2018).

74. Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516–520 (2011).

75. Nardini, M. *et al.* Sequence-Specific Transcription Factor NF-Y Displays Histone-like DNA Binding and H2B-like Ubiquitination. *Cell* **152**, 132–143 (2013).

76. Oldfield, A. J. *et al.* NF-Y controls fidelity of transcription initiation at gene promoters through maintenance of the nucleosome-depleted region. *Nat. Commun.* **10**, 3072 (2019).

77. Di Giammartino, D. C. *et al.* KLF4 is involved in the organization and regulation of pluripotency-associated three-dimensional enhancer networks. *Nat. Cell Biol.* **21**, 1179–1190 (2019).

78. Cao, S. *et al.* A Distal CCAAT/NUCLEAR FACTOR Y Complex Promotes Chromatin Looping at the FLOWERING LOCUS T Promoter and Regulates the Timing of Flowering in Arabidopsis[W][OPEN]. *Plant Cell* **26**, 1009–1017 (2014).

79. Huang, X. *et al.* Molecular condensation of the CO/NF-YB/NF-YC/FT complex gates floral transition in Arabidopsis. *EMBO J.* **44**, 225–250 (2025).

80. Tang, S. C., Vijayakumar, U., Zhang, Y. & Fullwood, M. J. Super-Enhancers, Phase-Separated Condensates, and 3D Genome Organization in Cancer. *Cancers* **14**, 2866 (2022).

81. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13–23 (2017).

82. Bhat, P., Honson, D. & Guttman, M. Nuclear compartmentalization as a mechanism of quantitative control of gene expression. *Nat. Rev. Mol. Cell Biol.* **22**, 653–670 (2021).

83. Misteli, T. The Self-Organizing Genome: Principles of Genome Architecture and Function. *Cell* **183**, 28–45 (2020).

84. Oudelaar, A. M. & Higgs, D. R. The relationship between genome structure and function. *Nat. Rev. Genet.* **22**, 154–168 (2021).

85. Yang, J. H. & Hansen, A. S. Enhancer selectivity in space and time: from enhancer–promoter interactions to promoter activation. *Nat. Rev. Mol. Cell Biol.* **25**, 574–591 (2024).

86. Markaki, Y. *et al.* Xist nucleates local protein gradients to propagate silencing across the X chromosome. *Cell* **184**, 6174-6192.e32 (2021).

87. Quinodoz, S. A. *et al.* RNA promotes the formation of spatial compartments in the nucleus. *Cell* **184**, 5775-5790.e30 (2021).

88. Cramer, P. Organization and regulation of gene transcription. *Nature* **573**, 45–54 (2019).

89. Roeder, R. G. 50+ years of eukaryotic transcription: an expanding universe of factors and mechanisms. *Nat. Struct. Mol. Biol.* **26**, 783–791 (2019).

90. Vos, S. M. Understanding transcription across scales: From base pairs to chromosomes. *Mol. Cell* **81**, 1601–1616 (2021).

91. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).

92. Field, A. & Adelman, K. Evaluating Enhancer Function and Transcription. *Annu. Rev. Biochem.* **89**, 213–234 (2020).

93. Furlong, E. E. M. & Levine, M. Developmental enhancers and chromosome topology. *Science* **361**, 1341–1345 (2018).

94. Kim, S. & Shendure, J. Mechanisms of Interplay between Transcription Factors and the 3D Genome. *Mol. Cell* **76**, 306–319 (2019).

95. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

96. Richter, W. F., Nayak, S., Iwasa, J. & Taatjes, D. J. The Mediator complex as a master regulator of transcription by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **23**, 732–749 (2022).

97. Bulger, M. & Groudine, M. Looping versus linking: toward a model for long-distance gene activation. *Genes Dev.* **13**, 2465–2477 (1999).

98. Chen, Z. *et al.* Increased enhancer–promoter interactions during developmental enhancer activation in mammals. *Nat. Genet.* **56**, 675–685 (2024).

99. Du, M. *et al.* Direct observation of a condensate effect on super-enhancer controlled gene bursting. *Cell* **187**, 331-344.e17 (2024).

100. Li, G. *et al.* Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell* **148**, 84–98 (2012).

101. Panigrahi, A. K., Lonard, D. M. & O'Malley, B. W. Enhancer-promoter entanglement explains their transcriptional interdependence. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2216436120 (2023).

102. Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F. & Laat, W. de. Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. *Mol. Cell* **10**, 1453–1465 (2002).

103. Zuin, J. *et al.* Nonlinear control of transcription through enhancer–promoter interactions. *Nature* **604**, 571–577 (2022).

104. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

105. Corradin, O. & Scacheri, P. C. Enhancer variants: evaluating functions in common disease. *Genome Med.* **6**, 85 (2014).

106. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).

107. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).

108. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190–1195 (2012).

109. Hu, Y. *et al.* Multiscale footprints reveal the organization of cis-regulatory elements. *Nature* **638**, 779–786 (2025).

110. Panne, D., Maniatis, T. & Harrison, S. C. An atomic model of the interferon-beta enhanceosome. *Cell* **129**, 1111–1123 (2007).

111. Thanos, D. & Maniatis, T. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* **83**, 1091–1100 (1995).

112. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).

113. Oksuz, O. *et al.* Transcription factors interact with RNA to regulate genes. *Mol. Cell* **83**, 2449-2463.e13 (2023).

114. Karnuta, J. M. & Scacheri, P. C. Enhancers: bridging the gap between gene control and human disease. *Hum. Mol. Genet.* **27**, R219–R227 (2018).

115. Parker, S. C. J. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 17921–17926 (2013).

116. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).

117. Aboreden, N. G. *et al.* LDB1 establishes multi-enhancer networks to regulate gene expression. *BioRxiv Prepr. Serv. Biol.* 2024.08.23.609430 (2024) doi:10.1101/2024.08.23.609430.

118. Boehning, M. *et al.* RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nat. Struct. Mol. Biol.* **25**, 833–840 (2018).

119. Boija, A. *et al.* Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* **175**, 1842-1855.e16 (2018).

120. Cho, W.-K. *et al.* Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* **361**, 412–415 (2018).

121. Henninger, J. E. *et al.* RNA-Mediated Feedback Control of Transcriptional Condensates. *Cell* **184**, 207-225.e24 (2021).

122. Nair, S. J. *et al.* Phase separation of ligand-activated enhancers licenses cooperative chromosomal enhancer assembly. *Nat. Struct. Mol. Biol.* **26**, 193–203 (2019).

123. Sabari, B. R. *et al.* Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **361**, eaar3958 (2018).

124. Lovén, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334 (2013).

125. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).

126. Lorenzi, L. *et al.* The RNA Atlas expands the catalog of human non-coding RNAs. *Nat. Biotechnol.* **39**, 1453–1465 (2021).

127. Sartorelli, V. & Lauberth, S. M. Enhancer RNAs are an important regulatory layer of the epigenome. *Nat. Struct. Mol. Biol.* **27**, 521–528 (2020).

128. Kent, N. A., Adams, S., Moorhouse, A. & Paszkiewicz, K. Chromatin particle spectrum analysis: a method for comparative chromatin structure analysis using paired-end mode next-generation DNA sequencing. *Nucleic Acids Res.* **39**, e26 (2011).

129. Rao, S., Ahmad, K. & Ramachandran, S. Cooperative binding between distant transcription factors is a hallmark of active enhancers. *Mol. Cell* **81**, 1651-1665.e4 (2021).

130. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

131. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).

132. Choudhuri, A. *et al.* Common variants in signaling transcription-factor-binding sites drive phenotypic variability in red blood cell traits. *Nat. Genet.* **52**, 1333–1345 (2020).

133. Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA Binding Variation. *Cell* **166**, 538–554 (2016).

134. Mansour, M. R. *et al.* An Oncogenic Super-Enhancer Formed Through Somatic Mutation of a Noncoding Intergenic Element. *Science* **346**, 1373–1377 (2014).

135. Oldridge, D. A. *et al.* Genetic predisposition to neuroblastoma mediated by a LMO1 super-enhancer polymorphism. *Nature* **528**, 418–421 (2015).

136. Soldner, F. *et al.* Parkinson-associated risk variant in enhancer element produces subtle effect on target gene expression. *Nature* **533**, 95–99 (2016).

137. Stankey, C. T. *et al.* A disease-associated gene desert directs macrophage inflammation through ETS2. *Nature* **630**, 447–456 (2024).

138. Sur, I. K. *et al.* Mice Lacking a Myc Enhancer That Includes Human SNP rs6983267 Are Resistant to Intestinal Tumors. *Science* **338**, 1360–1363 (2012).

139. Tehranchi, A. K. *et al.* Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. *Cell* **165**, 730–741 (2016).

140. Fernandez-Rozadilla, C. *et al.* Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. *Nat. Genet.* **55**, 89–99 (2023).

141. Arner, E. *et al.* Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science* **347**, 1010–1014 (2015).

142. Hah, N., Murakami, S., Nagari, A., Danko, C. G. & Kraus, W. L. Enhancer transcripts mark active estrogen receptor binding sites. *Genome Res.* **23**, 1210–1223 (2013).

143. Kim, T.-K. *et al.* Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**, 182–187 (2010).

144. Li, W. *et al.* Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. *Nature* **498**, 516–520 (2013).

145. Sigova, A. A. *et al.* Transcription factor trapping by RNA in gene regulatory elements. *Science* **350**, 978–981 (2015).

146. Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B. & Sharp, P. A. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**, 360–363 (2013).

147. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322**, 1845–1848 (2008).

148. Jindal, G. A. & Farley, E. K. Enhancer grammar in development, evolution, and disease: dependencies and interplay. *Dev. Cell* **56**, 575–587 (2021).

149. Junion, G. *et al.* A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell* **148**, 473–486 (2012).

150. Kulkarni, M. M. & Arnosti, D. N. Information display by transcriptional enhancers. *Dev. Camb. Engl.* **130**, 6569–6575 (2003).

151. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* **13**, 613–626 (2012).

152. Morin, J. A. *et al.* Sequence-dependent surface condensation of a pioneer transcription factor on DNA. *Nat. Phys.* **18**, 271–276 (2022).

153. Meier, D. J. Theory of block copolymers. I. Domain formation in A-B block copolymers. *J. Polym. Sci. Part C Polym. Symp.* **26**, 81–98 (1969).

154. Feng, H., Lu, X., Wang, W., Kang, N.-G. & Mays, J. W. Block Copolymers: Synthesis, Self-Assembly, and Applications. *Polymers* **9**, 494 (2017).

155. Roayaei Ardakany, A., Gezer, H. T., Lonardi, S. & Ay, F. Mustache: multi-scale detection of chromatin loops from Hi-C and Micro-C maps using scale-space representation. *Genome Biol.* **21**, 256 (2020).

156. Park, D. S. *et al.* High-throughput Oligopaint screen identifies druggable 3D genome regulators. *Nature* **620**, 209–217 (2023).

157. Chou, C.-C., Rajasekaran, M. & Chen, C. An effective approach for generating a three-Cys2His2 zinc-finger-DNA complex model by docking. *BMC Bioinformatics* **11**, 334 (2010).

158. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590-598 (2006).

159. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma. Oxf. Engl.* **26**, 841–842 (2010).

160. Barutcu, A. R., Blencowe, B. J. & Rinn, J. L. Differential contribution of steady-state RNA and active transcription in chromatin organization. *EMBO Rep.* **20**, e48068 (2019).

161. Hughes, C. S. *et al.* Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* **14**, 68–85 (2019).

162. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).

163. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 (2020).