UNIVERSITY OF CALIFORNIA SAN DIEGO

New insights into 3D chromatin organization and function

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Lina Zheng

Committee in charge:

Professor Wei Wang, Chair
Professor Gene Yeo, Co-Chair
Professor Graham McVicker
Professor Bing Ren
Professor Jing Yang

2022

The dissertation of Lina Zheng is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To my grandparents, my parents and my husband,

words can not express my gratitude and appreciation

for your endless love and unconditional support.

I am lucky to have you in my life.

My success is yours.

EPIGRAPH

Act enthusiastic and you will be enthusiastic.

–Dale Carnegie

TABLE OF CONTENTS

DISSERTATION APPROVAL PAGE iii

DEDICATION iv

EPIGRAPH v

TABLE OF CONTENTS vi

LIST OF FIGURES xi

LIST OF TABLES xiv

ACKNOWLEDGEMENTS xv

VITA xviii

ABSTRACT OF THE DISSERTATION xx

INTRODUCTION 1

References 6

Chapter 1. Noncoding loci without epigenomic signals can be essential for maintaining global chromatin organization and cell viability 13

1.1 Abstract 13

1.2 Introduction 14

1.3 Materials and Methods 16

1.3.1 Network construction and hub identification 16

1.3.2 Hub screening and validation 18

1.3.3 Hi-C library preparation and data analysis 24

1.3.4 Bulk RNA-seq and data analysis 26

1.3.5 scRNA-seq and data analysis 27

1.4 Results 31

1.4.1 FCN networks are resistant to random attacks but vulnerable to targeted attacks 32

1.4.2 Cancer-related mutations alter 3D hub contacts 34

1.4.3 Targeted deletion of hubs can significantly affect cell viability.........36

1.4.4 Cell death caused by hub deletion does not result from disruption of functional elements or off-target effects..39

1.4.5 Deletion of essential hubs can alter the global chromatin structure.41

1.4.6 Deletion of essential hubs can up-regulate apoptotic genes 42

1.4.7 Deletion of essential hubs can alter gene expression in distal regions 43

1.4.8 The global impact of hub deletion suggests that hubs might be potential noncoding therapeutic targets...44

1.5 Discussion 47

1.6 Acknowledgements 51

1.7 Figures..53

1.8 Supplementary Figures 61

1.9 Supplementary Tables...74

1.10 References 80

Chapter 2. Regulatory elements can be essential for maintaining broad chromatin organization and cell viability...85

2.1 Abstract...85

2.2 Introduction...86

2.3 Materials and Methods 88

2.3.1 Predicting high-resolution regulatory element contacts by EpiTensor 88

2.3.2 Identification of sample-specific degree/sample-specific GV...........88

2.3.3 Cell culture 89

2.3.4 Design and construction of the CRISPR-Cas9 pgRNA library 89

2.3.5 CRISPR-Cas9 pgRNA library screening 91

2.3.6 Identification of functional hotspots involved in cell growth and proliferation 91

2.3.7 Individual validation of functional hotspots by cell proliferation assay 93

2.3.8 Hi-C library preparation and data analysis 94

2.3.9 Evaluation of the potential off-target effects by the CRISPR-Cas9 system through whole genome sequencing (WGS) 96

2.3.10 Bulk RNA-seq and data analysis 97

2.3.11 Single-cell RNA-seq and data analysis 98

2.3.12 Validation of synthetic lethal pairs by cell proliferation assay 100

2.4 Results 102

2.4.1 Small-world network formed by 3D contacts between promoters and enhancers 102

2.4.2 Mutations in hotspot enhancers and promoters could alter 3D contacts 103

2.4.3 CRISPR/Cas9 library screening identified hotspots essential for cell growth and survival 106

2.4.4 The essentiality of hotspots does not result from any association with essential genes 109

2.4.5 Deleting essential hotspots can affect broad chromatin organization 110

2.4.6 Essential hotspots tend to reside in dense chromatin structures ... 113

2.4.7 Hotspot deletion synergistically affects gene expression 113

2.5 Discussion 116

2.6 Acknowledgements 119

2.7 Figures..121

2.8 Supplementary Figures 128

2.9 Supplementary Tables..135

2.10 References 139

Chapter 3. Regulation associated modules reflect 3D genome modularity associated with chromatin activity..145

3.1 Abstract..145

3.2 Introduction 147

3.3 Materials and Methods 150

3.3.1 Regulatory associated modules (RAM) identification 150

3.3.2 Hi-C data analysis..151

3.3.3 Lamina-associated domains (LADs) data processing 152

3.3.4 Cohesin degradation analysis...152

3.3.5 Enrichment of upregulated genes in enhancer-promoter pairs occurring in the same RAM of K562 but split in the background cell type by Hypergeometric Test..153

3.3.6 Enrichment of the somatic variants in cancer cRAM boundaries compared to the TAD boundaries assessed by Two-sample Proportion Tests. 153

3.3.7 Somatic mutations and structural variation analysis for cancer patients from PCAWG 154

3.3.8 Motif analysis 154

3.4 Results..155

3.4.1 Regulation associated modules (RAMs) detected by the histone modification peaks 155

3.4.2 Characterization of the consensus regulation associated modules (cRAMs) 157

3.4.3 RAMs are functional units 158

3.4.4 RAMs are insensitive to cohesin degradation 160

3.4.5 Extrachromosomal DNA (ecDNA) from cancer patients majorly originated from intra-RAM 161

3.4.6 Deletion of the cRAM boundaries are predicted to alter the 3D chromatin structures 162

3.4.7 Somatic genetic variations enriched in regulation associated modules boundaries 163

3.5 Discussion 166

3.6 Acknowledgements 168

3.7 Figures 169

3.8 Supplementary Figures 176

3.9 Supplementary Tables 183

3.10 References 191

Chapter 4. Concluding Remarks 197

LIST OF FIGURES

Figure 1.1 Characterization of the FCNs and hub nodes 53

Figure 1.2 Identification of essential hubs for cell growth and proliferation in the K562 cell line through pgRNA-mediated fragment deletion 55

Figure 1.3 Characterization of the deleted hub and the impact of its deletion on the global chromatin structure..57

Figure 1.4 Hub deletion induces global changes in gene expression 58

Figure 1.5 The concurrent alterations of 3D chromatin structure and gene expressions after hub deletion..59

Figure S1.1 Hub characterization..61

Figure S1.2 Formation or disruption of cell type-specific hubs......................................63

Figure S1.3 Histone modifications and open chromatin signals around the hub regions in different cell types 64

Figure S1.4 The correlations between replicates in the functional screening for essential hubs in the K562 cell line and hub distribution in each chromosome.............................65

Figure S1.5 Validation of essential hubs in K562 cells through fragment deletion.........67

Figure S1.6 Validation of hub_22_7 in multiple cancer cell lines through fragment deletion 68

Figure S1.7 Comprehensive genomic and epigenomic data of the essential hub_22_7 in K562 cells 69

Figure S1.8 Interchromosomal Hi-C contacts in wild-type K562 cells............................70

Figure S1.9 TADs in the wild-type and hub-deleted K562 cells 71

Figure S1.10 Hi-C contacts 72

Figure S1.11 The expression levels of TOP3B in different cell lines/tissues and H3K27ac signals of the enhancer regions interacting with the TOP3B promoter. 73

Figure 2.1 Small-world network analysis and mutation effects on 3D contact for hotspot enhancers and promoters 121

Figure 2.2 Identification of essential hotspots for cell growth and proliferation in the K562 cell line through pgRNA deletion-based CRISPR screening.......................................123

Figure 2.3 Validation of candidate essential hotspot loci in K562 cells and multiple cell lines 124

Figure 2.4 Deletion of an essential hotspot impacts broad chromatin structure...........126

Figure 2.5 Synergistic change in gene expression after hotspot deletion 127

Figure S2.1 The correlations between replicates in the functional screening for essential hotspots in the K562 cell line...128

Figure S2.2 Validation of essential hotspots in K562 cells and multiple cancer cell lines as assessed by fragment deletion 129

Figure S2.3 WGS analysis of the off-target effects of the pgRNA targeting hotspot_10_25 in the genome browser view...130

Figure S2.4 Hi-C contact maps, A/B compartments and TADs before and after hotspot_10_25 deletion 131

Figure S2.5 Contact map (chr10) of wild-type K562 at a 1-Mb resolution....................132

Figure S2.6 Distributions of the effective diameter, modularity and chromatin loops...133

Figure S2.7 The growth effects of gene pairs that showed downregulated expression according to single-cell RNA-seq after hotspot_10_25 deletion in K562 cells 134

Figure 3.1 Regulation associated module (RAM) identification...................................169

Figure 3.2 Characterization of the cRAMs and boundaries...171

Figure 3.3 RAM is a functional unit 172

Figure 3.4 The association of the cRAM boundaries with the 3D chromatin structure and cancer somatic variants...174

Figure S3.1 The frequency of the consensus RAM boundaries in the 93 normal samples 176

Figure S3.2 The frequency of the consensus RAM boundaries in the 19 cancer samples 177

Figure S3.3 Consensus RAMs 178

Figure S3.4 The K562 enhancer-promoter interactions over K562 RAMs and cancer cRAMs 179

Figure S3.5 RAMs are resistant to cohesion degradation 180

Figure S3.6 Predicted chromatin structure change upon deletion of cRAM boundaries and TAD boundaries in H1-hESC cell line 181

Figure S3.7 Gene Ontology Molecular Functions of the genes within 2.5kb from the enriched motifs overlapping with somatic indels 182

LIST OF TABLES

Table S1.1 Number of hubs selected in each cell line. 74

Table S1.2 The percentage of cell line-specific nodes in all nodes, degree-GV-correlated nodes and degree-GV-correlated hubs. 75

Table S1.3 Additional Hi-C data with 20-kb resolution 76

Table S1.4 The chromosome distributions for differentially expressed genes 77

Table S1.5 chromHMM states distribution in K562 essential hubs 78

Table S1.6 Summary of WGS analysis for hub_22_7 deletion in K562 cells.................79

Table S2.1 Summary of WGS analysis for hotspot_10_25 deletion in K562 cells 135

Table S2.2 Hi-C comparison between the wild-type and hotspot_10_25-deleted K562 cells 136

Table S2.3 Comparison of chromatin loops before/after hotspot deletion 137

Table S2.4 Comparison of chromatin loops before/after hotspot deletion in chr10: 11-17 Mb and chr10: 68-76 Mb 138

Table S3.1 Collected ChIP-seq samples from ROADMAP and ENCODE 183

Table S3.2 Upregulated gene expressions involved in enhancer-promoter interactions occurred in the same RAM in K562 but in different GM12878 RAMs 188

Table S3.3 Upregulated gene expressions involved in enhancer-promoter interactions occurred in the same RAM in K562 but in different HEPG2 RAMs..............................189

Table S3.4 DEGs annotated with Gene Ontology Molecular Function of “Calcium Ion Binding” 190

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor Dr. Wei Wang for his guidance throughout my Ph.D. education and scientific insight into many problem-solving processes in this dissertation. His enthusiasm towards scientific research deeply motivated me to become an energetic scientist in bioinformatics.

I would like to thank Dr. Gene Yeo, Dr. Bing Ren, Dr. Graham McVicker and Dr. Jing Yang for taking the time to serve on my dissertation committee. I indeed appreciate their constructive feedback and suggestions for my dissertation. I am also grateful to Dr. Wensheng Wei's lab for their collaborative efforts and experimental work in Chapter1 and Chapter2.

Furthermore, I would like to thank my fellow lab members Dr. Bo Ding, Dr. Zhao Chen, Dr. Ying Zhao and Mr. Peiyao Wu for their guidance, assistance and contributions to the work included in this dissertation. I would also like to thank my families and friends for their love and support.

The introduction is, in part, based on material as it appears in "Noncoding loci without epigenomic signals can be essential for maintaining global chromatin organization and cell viability." Bo Ding; Ying Liu; Zhiheng Liu; Lina Zheng; Ping Xu; Zhao Chen; Peiyao Wu; Ying Zhao; Qian Pan; Yu Guo; Wensheng Wei; Wei Wang. Science Advances, 7(45), eabi6020, 2021. DOI: 10.1126/sciadv.abi6020. The introduction is also, in part, based on the material as it appears in “Regulatory elements can be essential for maintaining broad chromatin organization and cell viability.” Ying Liu; Bo Ding; Lina Zheng; Ping Xu; Zhiheng Liu; Zhao Chen; Peiyao Wu; Ying Zhao; Qian Pan; Yu Guo; Wei Wang; Wensheng Wei. Nucleic Acids Research, Oxford University Press, 2022. gkac197, https://doi.org/10.1093/nar/gkac197. The introduction is also, in part, based on material as it appears in “Regulation associated modules reflect 3D genome modularity associated with chromatin activity.” Lina Zheng; Wei Wang. bioRxiv, 2022. https://doi.org/10.1101/2022.03.02.482718. This paper is in submission. The dissertation author was the primary investigator and author of all these papers.

Chapter 1, is a reformatted reprint of the material as it appears in “Noncoding loci without epigenomic signals can be essential for maintaining global chromatin organization and cell viability.” Bo Ding; Ying Liu; Zhiheng Liu; Lina Zheng; Ping Xu; Zhao Chen; Peiyao Wu; Ying Zhao; Qian Pan; Yu Guo; Wensheng Wei; Wei Wang. Science Advances, 7(45), eabi6020, 2021. DOI: 10.1126/sciadv.abi6020. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is a reformatted reprint of the material as it appears in “Regulatory elements can be essential for maintaining broad chromatin organization and cell viability.” Ying Liu; Bo Ding; Lina Zheng; Ping Xu; Zhiheng Liu; Zhao Chen; Peiyao Wu; Ying Zhao; Qian Pan; Yu Guo; Wei Wang; Wensheng Wei. Nucleic Acids Research, Oxford University Press, 2022. gkac197, https://doi.org/10.1093/nar/gkac197. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reformatted reprint of the material as it appears in “Regulation associated modules reflect 3D genome modularity associated with chromatin activity.” Lina Zheng; Wei Wang. bioRxiv, 2022. https://doi.org/10.1101/2022.03.02.482718. In submission. The dissertation author was the primary investigator and author of this paper.

VITA

2008-2012	B.S. in Pharmaceutical Science, China Pharmaceutical University, Nanjing, China
2013-2016	M.S. in Statistics concentration on biostatistics, San Diego State University
2018-2022	Graduate Researcher, University of California San Diego
2017-2022	Ph.D. in Bioinformatics and Systems Biology, University of California San Diego

PUBLICATIONS

L. Zheng, W. Wang. Regulation associated modules reflect 3D genome modularity associated with chromatin activity. *bioRxiv.* (2022) 2022.03.02.482718. https://doi.org/10.1101/2022.03.02.482718. (*Nature Communications,* in review.)

Y. Liu*, B. Ding*, **L. Zheng***, P. Xu*, Z. Liu, Z. Chen, P. Wu, Y. Zhao, Q. Pan, Y. Guo, W. Wang, W. Wei. Regulatory elements can be essential for maintaining broad chromatin organization and cell viability. *Nucleic Acids Research,* 2022. gkac197, https://doi.org/10.1093/nar/gkac197*.* (* equal contribution)

B. Ding*, Y. Liu*, Z. Liu*, **L. Zheng***, P. Xu, Z. Chen, P. Wu, Y. Zhao, Q. Pan, Y. Guo, W. Wei, W. Wang. Noncoding loci without epigenomic signals can be essential for maintaining global chromatin organization and cell viability. *Science Advances.* 7 (2021) eabi6020. (* equal contribution)

B. Ding*, **L. Zheng***, W. Wang. Assessment of Single Cell RNA-Seq Normalization Methods. *G3: Genes, Genomes, Genetics*. 7 (2017) 2039–2045. (* equal contribution)

B. Ding*, **L. Zheng***, Y. Zhu, N. Li, H. Jia, R. Ai, A. Wildberg, W. Wang, Normalization and noise reduction for single cell RNA-seq experiments, *Bioinformatics*. 31 (2015) 2225–2227. (* equally contribution)

R. Ainsworth, D. Hammaker, G. Nygaard, C. Ansalone, C. Machado, K. Zhang, **L. Zheng**, L. Carrillo, A. Wildberg, A. Kuhs, M. Svensson, D. Boyle, G. Firestein, W. Wang. Systems-biology analysis of rheumatoid arthritis fibroblast-like synoviocytes reveals cell line-specific transcription factor function. *Nature Communications,* (2022). In revision.

X. Ren, M. Wang, B. Li, K. Jamieson, **L. Zheng**, I.R. Jones, B. Li, M.A. Takagi, J. Lee, L. Maliskova, T.W. Tam, M. Yu, R. Hu, L. Lee, A. Abnousi, G. Li, Y. Li, M. Hu, B. Ren, W.

Wang, Y. Shen. Parallel characterization of cis-regulatory elements for multiple genes using CRISPRpath. *Science Advances.* 7 (2021) eabi4360.

M. Wang, K. Zhang, V. Ngo, C. Liu, S. Fan, J.W. Whitaker, Y. Chen, R. Ai, Z. Chen, J. Wang, **L. Zheng**, W. Wang. Identification of DNA motifs that regulate DNA methylation. *Nucleic Acids Research.* 47 (2019) 6753–6768.

Y. Zhu, Z. Chen, K. Zhang, M. Wang, D. Medovoy, J.W. Whitaker, B. Ding, N. Li, **L. Zheng**, W. Wang. Constructing 3D interaction maps from 1D epigenomes. *Nature Communications.* 7 (2016) 10812.

H.R. Dueck, R. Ai, A. Camarena, B. Ding, R. Dominguez, O.V. Evgrafov, J.-B. Fan, S.A. Fisher, J.S. Herstein, T.K. Kim, J.M.H. Kim, M.-Y. Lin, R. Liu, W.J. Mack, S. McGroty, J.D. Nguyen, N. Salathia, J. Shallcross, T. Souaiaia, J.M. Spaethling, C.P. Walker, J. Wang, K. Wang, W. Wang, A. Wildberg, **L. Zheng**, R.H. Chow, J. Eberwine, J.A. Knowles, K. Zhang, J. Kim. Assessing characteristics of RNA amplification methods for single cell RNA sequencing. *BMC Genomics*. 17 (2016) 966.

ABSTRACT OF THE DISSERTATION

New insights into 3D chromatin organization and function

by

Lina Zheng

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2022

Professor Wei Wang, Chair

Professor Gene Yeo, Co-Chair

Precise and delicate 3D genome organization is fundamental to cellular homeostasis. Big data generated by high-throughput technologies has granted great opportunities to deepen our understanding of chromatin organization and function, particularly for non-coding DNA sequences which have long been considered “junk DNAs”. This dissertation aims at illuminating the structural importance of the non-coding DNA sequences and elucidating the modular organization of the 3D genome from histone modifications. First, I performed network analysis on Hi-C 3D contact data to identify non-coding DNA regions forming many spatial contacts with other regions (“hubs”) without epigenetic signals that could maintain the global 3D chromatin structure. Furthermore, I employed a small-world network on epigenetic histone modification data to identify a group of active enhancers and promoters harboring many 3D contacts (“hotspots”) which can maintain broad 3D chromatin organization beyond enhancer-promoter interactions. Deletion of hubs and hotspots can produce a profound impact on 3D chromatin organization and cell viability. In addition, through investigation of the histone modification across cell types, I identified the regulation associated modules (“RAMs”) that can not only reflect the modular organization of the 3D genome but also be better aligned with the chromatin function. These studies provide new insights into 3D genome organization and function, navigating future efforts in the mechanistic investigation of the 3D genome.

INTRODUCTION

In eukaryotes, chromatin is efficiently packaged and organized in the nucleus. Generally, 3D chromatin organization is formed in a hierarchical structure starting from a DNA double helix: the DNA double helix firstly winds around histones to form nucleosomes; the nucleosomes are further packaged with the help of cellular components to form higher-order structures, such as chromatin loops(*1*–*4*), topologically associated domains (TADs)(*5*, *6*), compartments(*2*, *7*), and chromosome territories(*8*). 3D genome organization is fundamental to cellular homeostasis(*5*, *6*, *9*–*12*). Aberrant chromatin organization has been reported to be associated with human diseases(*13*–*15*).

Non-coding DNA sequences, barely encoding any proteins but accounts for most of the human genome, have been known to play important roles in many biological processes recently(*16*–*18*). Growing evidence has shown that non-coding regions are associated with human diseases, such as enhancers, promoters, boundaries of topologically associated domains, transposable elements, and non-coding RNAs(*5*, *6*, *9*, *19*–*26*). Despite spectacular progress in the functional roles of the non-coding regions in gene regulation, the contributions from non-coding regions in maintaining proper 3D chromatin organization are still not well investigated, particularly for the regions without any epigenetic signals.

In Chapter1, colleagues and I showed that the non-coding regions without epigenetic signals can maintain global 3D chromatin organization. Using the scale-free network analysis on the Hi-C contact data and CRISPR-CAS9 high throughput screening experiment, we identified dozens of non-coding regions without epigenetic signals essential to cell fitness. Deletion of such locus can produce a profound impact on 3D chromatin organizations such as the flips between A/B compartments, TAD alterations, and enhancer-promoter interactions changes revealed by Hi-C analysis. Single-cell RNA-seq analysis also showed that many apoptosis genes are upregulated and distal gene expressions are changed upon deletion of such locus.

Promoters and enhancers are cis-regulatory elements (CREs) that are critical for the spatiotemporal regulation of gene expressions(*27–30*). The activated enhancers are involved in direct physical contact with both the nearby active promoters(*31–33*) and the distal active promoters in the linear genome(*28*, *34–39*). Therefore, a precise and delicate 3D genome chromatin organization is essential to forming and maintaining promoter-enhancer interactions. Many studies have reported that an altered chromatin organization can result in inappropriate enhancer-promoter interactions and lead to dysregulation of gene expressions(*9*, *40–44*).

Moreover, advanced microscopes have observed that the transcription factors and polymerases are grouped to form high transcriptional compact clusters instead of evenly scattered along the genome(*45*). The transcriptional condensates can further form liquid-liquid phase separation in vivo and in vitro(*46–48*). Disruption of the phase separation has been observed to affect the chromatin organization(*49*). These observations suggested that the transcription machinery is important to 3D genome organization.

Although the functional roles of regulatory elements in gene regulation have been intensively studied in the past few years, the contributions from enhancers and promoters in maintaining broad 3D chromatin organization beyond enhancer-promoter interactions have not been well studied. In Chapter2, colleagues and I presented the structural importance of regulatory elements in maintaining broad chromatin structures. Through small world networks analysis on enhancer-promoter interactions, we identified a group of regulatory elements harboring many 3D contacts ("hotspots"). Using the CRISPR-CAS9 screening followed by Hi-C and single-cell RNA-seq analysis, we identified dozens of hotspots that can be essential to maintaining a broad chromatin structure and cell viability. Our study illuminated a previously unidentified role of regulatory elements, particularly enhancers, in maintaining a broad chromatin organization.

Post-translational modifications on histones, including methylation(*50*–*52*), acetylation(*53*–*55*), phosphorylation(*56*, *57*) and ubiquitination(*58*, *59*), play critical roles in chromatin organization and function. Histone modifications are involved in gene regulation(*60*, *61*). For instance, acetylation of the lysine residue at N-terminal position 27 of the histone H3 protein (H3K27ac) serves as an indicator of the strong transcription activities; trimethylation of the lysine residue at N-terminal position 9 of the histone H3 protein (H3K9me3) acts as an indicator of the gene silence and low transcription activities. In addition, histone modifications can also regulate the chromatin structures by controlling the chromatin conformation loose or tight. Previous studies have shown that histone acetylation and phosphorylation can change the conformation by effectively altering the charges on the residues. For example, acetylation on the N-terminal of lysine can neutralize the positive charge on the residues and thus leads to a loose conformation or euchromatin(*62*, *63*). Other modifications, such as histone methylation can alter the conformation by being recognized by other proteins(*62*). One example is H3K9me3. H3K9me3 can recruit heterochromatin protein 1 (*HP1*), which can compact chromatin into heterochromatin(*62*).

Histone modifications have been observed to be associated with the 3D genome spatial domains(*2*, *7*, *64–68*). At a relatively small scale of chromatin organization, transcriptionally active TADs and polycomb-repressed TADs are often correlated with separate histone marks(*68*). At a large scale of chromatin organization beyond TADs, the “A compartment” is often enriched in active histone marks, and the “B compartment” is often in repressive histone marks (*2*, *7*). Through the genome-wide clustering algorithms, the A/B compartment could be further clustered into six commonly known subtypes: A1, A2, B1, B2, B3, and B4 subcompartments, which have distinct functions. These subcompartments are often enriched in different histone mark patterns(*2*, *69*). Based on the correlation between the histone marks and the 3D spatial organization, a few computational models have been built to predict the 3D genome organization from histone marks(*70–76*).

Previous studies have intensively investigated the association between histone modifications and the existing 3D genome organization modules. However, an unexplored field is to illustrate the 3D genome organization modules directly from epigenetic histone modifications to better understand the relationship between chromatin organization and function. In Chapter3, I presented a newly identified regulation associated module ("RAM") from histone modifications. Through intensive characterizations and comparative analysis of RAMs and previously known spatial domains such as TADs and A/B compartments, RAM can reflect the spatial modular organization of the 3D genome and be better aligned with chromatin function. Perturbation of RAM boundaries is predicted to produce more severe alterations in chromatin structure than TADs from deep learning models. The enrichment of somatic indels in RAM boundaries highlights the critical role of RAM in pathogenesis.

The introduction is, in part, based on material as it appears in "Noncoding loci without epigenomic signals can be essential for maintaining global chromatin organization and cell viability." Bo Ding; Ying Liu; Zhiheng Liu; Lina Zheng; Ping Xu; Zhao Chen; Peiyao Wu; Ying Zhao; Qian Pan; Yu Guo; Wensheng Wei; Wei Wang. Science Advances, 7(45), eabi6020, 2021. DOI: 10.1126/sciadv.abi6020. The introduction is also, in part, based on the material as it appears in "Regulatory elements can be essential for maintaining broad chromatin organization and cell viability." Ying Liu; Bo Ding; Lina Zheng; Ping Xu; Zhiheng Liu; Zhao Chen; Peiyao Wu; Ying Zhao; Qian Pan; Yu Guo; Wei Wang; Wensheng Wei. Nucleic Acids Research, Oxford University Press, 2022. gkac197, https://doi.org/10.1093/nar/gkac197. The introduction is also, in part, based on material as it appears in "Regulation associated modules reflect 3D genome modularity associated with chromatin activity." Lina Zheng; Wei Wang. bioRxiv, 2022. https://doi.org/10.1101/2022.03.02.482718. This paper is in submission. The dissertation author was the primary investigator and author of all these papers.

References

1. F. Jin, Y. Li, J. R. Dixon, S. Selvaraj, Z. Ye, A. Y. Lee, C.-A. Yen, A. D. Schmitt, C. A. Espinoza, B. Ren, A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*. **503**, 290–294 (2013).

2. S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, E. L. Aiden, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. **159**, 1665–1680 (2014).

3. N. Mandal, W. Su, R. Haber, S. Adhya, H. Echols, DNA looping in cellular repression of transcription of the galactose operon. *Genes Dev.* **4**, 410–418 (1990).

4. J. Fraser, I. Williamson, W. A. Bickmore, J. Dostie, An Overview of Genome Organization and How We Got There: from FISH to Hi-C. *Microbiology and Molecular Biology Reviews*. **79** (2015), pp. 347–372.

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

6. E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Blüthgen, J. Dekker, E. Heard, Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*. **485**, 381–385 (2012).

7. E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, J. Dekker, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. **326**, 289–293 (2009).

8. T. Cremer, M. Cremer, Chromosome territories. *Cold Spring Harb. Perspect. Biol.* **2**, a003889 (2010).

9. D. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, S. Mundlos, Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*. **161**, 1012–1025 (2015).

10. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Journal Club for Condensed Matter Physics* (2010), , doi:10.36471/jccm_february_2010_02.

11. T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, P. Fraser, Single-cell Hi-C reveals cell-to-cell variability in chromosome

structure. *Nature*. **502**, 59–64 (2013).

12. B. Bonev, G. Cavalli, Organization and function of the 3D genome. *Nature Reviews Genetics*. **17** (2016), pp. 661–678.

13. A.-L. Valton, J. Dekker, TAD disruption as oncogenic driver. *Curr. Opin. Genet. Dev.* **36**, 34–40 (2016).

14. E. Giorgio, D. Robyr, M. Spielmann, E. Ferrero, E. Di Gregorio, D. Imperiale, G. Vaula, G. Stamoulis, F. Santoni, C. Atzori, L. Gasparini, D. Ferrera, C. Canale, M. Guipponi, L. A. Pennacchio, S. E. Antonarakis, A. Brussino, A. Brusco, A large genomic deletion leads to enhancer adoption by the lamin B1 gene: a second path to autosomal dominant adult-onset demyelinating leukodystrophy (ADLD). *Hum. Mol. Genet.* **24**, 3143–3154 (2015).

15. D. Hnisz, A. S. Weintraub, D. S. Day, A.-L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker, R. A. Young, Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science*. **351**, 1454–1458 (2016).

16. B. S. Gloss, M. E. Dinger, Realizing the significance of noncoding functionality in clinical genomics. *Exp. Mol. Med.* **50**, 97 (2018).

17. B. E. Engelhardt, C. D. Brown, Diving deeper to predict noncoding sequence function. *Nat. Methods*. **12** (2015), pp. 925–926.

18. N. Carey, *Junk DNA: A Journey Through the Dark Matter of the Genome* (Icon Books, 2015).

19. H. Giral, U. Landmesser, A. Kratzer, Into the Wild: GWAS Exploration of Non-coding RNAs. *Front Cardiovasc Med*. **5**, 181 (2018).

20. L. Gao, Y. Uzun, P. Gao, B. He, X. Ma, J. Wang, S. Han, K. Tan, Identifying noncoding risk variants using disease-relevant gene regulatory networks. *Nat. Commun.* **9**, 702 (2018).

21. X. Xiao, H. Chang, M. Li, Molecular mechanisms underlying noncoding risk variations in psychiatric genetic studies. *Mol. Psychiatry*. **22**, 497–511 (2017).

22. W. F. Ooi, A. M. Nargund, K. J. Lim, S. Zhang, M. Xing, A. Mandoli, J. Q. Lim, S. W. T. Ho, Y. Guo, X. Yao, S. J. Lin, T. Nandi, C. Xu, X. Ong, M. Lee, A. L.-K. Tan, Y. N. Lam, J. X. Teo, A. Kaneda, K. P. White, W. K. Lim, S. G. Rozen, B. T. Teh, S. Li, A. J. Skanderup, P. Tan, Integrated paired-end enhancer profiling and whole-genome sequencing reveals recurrent CCNE1 and IGF2 enhancer hijacking in primary gastric adenocarcinoma. *Gut*. **69** (2020), pp. 1039–1052.

23. A. Claringbould, J. B. Zaugg, Enhancers in disease: molecular basis and emerging treatment strategies. *Trends Mol. Med.* **27**, 1060–1073 (2021).

24. L. Pasquali, K. J. Gaulton, S. A. Rodríguez-Seguí, L. Mularoni, I. Miguel-Escalada, İ. Akerman, J. J. Tena, I. Morán, C. Gómez-Marín, M. van de Bunt, J. Ponsa-Cobas, N. Castro, T. Nammo, I. Cebola, J. García-Hurtado, M. A. Maestro, F. Pattou, L. Piemonti, T. Berney, A. L. Gloyn, P. Ravassard, J. L. G. Skarmeta, F. Müller, M. I. McCarthy, J. Ferrer, Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat. Genet.* **46**, 136–143 (2014).

25. L. M. Payer, K. H. Burns, Transposable elements in human genetic disease. *Nature Reviews Genetics*. **20** (2019), pp. 760–772.

26. M. Esteller, Non-coding RNAs in human disease. *Nature Reviews Genetics*. **12** (2011), pp. 861–874.

27. R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jørgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Müller, A. R. R. Forrest, P. Carninci, M. Rehli, A. Sandelin, An atlas of active enhancers across human cell types and tissues. *Nature*. **507**, 455–461 (2014).

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

29. N. D. Tippens, A. Vihervaara, J. T. Lis, Enhancer transcription: what, where, when, and why? *Genes Dev.* **32**, 1–3 (2018).

30. J. L. Plank, A. Dean, Enhancer function: mechanistic and genome-wide insights come together. *Mol. Cell*. **55**, 5–14 (2014).

31. J. van Arensbergen, B. van Steensel, H. J. Bussemaker, In search of the determinants of enhancer-promoter interaction specificity. *Trends Cell Biol.* **24**, 695–702 (2014).

32. J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, B. E. Bernstein, Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. **473**, 43–49 (2011).

33. I. Chepelev, G. Wei, D. Wangsa, Q. Tang, K. Zhao, Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res.* **22**, 490–503 (2012).

34. J. Fitz, T. Neumann, M. Steininger, E.-M. Wiedemann, A. C. Garcia, A. Athanasiadis, U. E. Schoeberl, R. Pavri, Spt5-mediated enhancer transcription directly couples enhancer activation with physical promoter interaction. *Nat. Genet.* **52**, 505–515

(2020).

35. A. Pombo, N. Dillon, Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* **16**, 245–257 (2015).

36. H. K. Long, S. L. Prescott, J. Wysocka, Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell*. **167**, 1170–1187 (2016).

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

38. Y. Ghavi-Helm, F. A. Klein, T. Pakozdi, L. Ciglar, D. Noordermeer, W. Huber, E. E. M. Furlong, Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*. **512**, 96–100 (2014).

39. A. Sanyal, B. R. Lajoie, G. Jain, J. Dekker, The long-range interaction landscape of gene promoters. *Nature*. **489**, 109–113 (2012).

40. S. S. P. Rao, S.-C. Huang, B. Glenn St Hilaire, J. M. Engreitz, E. M. Perez, K.-R. Kieffer-Kwon, A. L. Sanborn, S. E. Johnstone, G. D. Bascom, I. D. Bochkov, X. Huang, M. S. Shamim, J. Shin, D. Turner, Z. Ye, A. D. Omer, J. T. Robinson, T. Schlick, B. E. Bernstein, R. Casellas, E. S. Lander, E. L. Aiden, Cohesin Loss Eliminates All Loop Domains. *Cell*. **171**, 305–320.e24 (2017).

41. J. Zuin, J. R. Dixon, M. I. J. A. van der Reijden, Z. Ye, P. Kolovos, R. W. W. Brouwer, M. P. C. van de Corput, H. J. G. van de Werken, T. A. Knoch, W. F. J. van IJcken, F. G. Grosveld, B. Ren, K. S. Wendt, Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 996–1001 (2014).

42. A. S. Hansen, I. Pustova, C. Cattoglio, R. Tjian, X. Darzacq, CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *Elife*. **6** (2017), doi:10.7554/eLife.25776.

43. J. Y. Xiao, A. Hafner, A. N. Boettiger, How subtle changes in 3D structure can create large changes in transcription. *Elife*. **10** (2021), doi:10.7554/eLife.64320.

44. W. A. Flavahan, Y. Drier, B. B. Liau, S. M. Gillespie, A. S. Venteicher, A. O. Stemmer-Rachamimov, M. L. Suvà, B. E. Bernstein, Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*. **529**, 110–114 (2016).

45. Z. Liu, W. R. Legant, B.-C. Chen, L. Li, J. B. Grimm, L. D. Lavis, E. Betzig, R. Tjian, 3D imaging of Sox2 enhancer clusters in embryonic stem cells. *Elife*. **3**, e04236 (2014).

46. A. Boija, I. A. Klein, B. R. Sabari, A. Dall'Agnese, E. L. Coffey, A. V. Zamudio, C. H. Li, K. Shrinivas, J. C. Manteiga, N. M. Hannett, B. J. Abraham, L. K. Afeyan, Y. E. Guo, J. K. Rimel, C. B. Fant, J. Schuijers, T. I. Lee, D. J. Taatjes, R. A. Young, Transcription Factors Activate Genes through the Phase-Separation Capacity of

Their Activation Domains. *Cell*. **175**, 1842–1855.e16 (2018).

47. B. R. Sabari, A. Dall'Agnese, A. Boija, I. A. Klein, E. L. Coffey, K. Shrinivas, B. J. Abraham, N. M. Hannett, A. V. Zamudio, J. C. Manteiga, C. H. Li, Y. E. Guo, D. S. Day, J. Schuijers, E. Vasile, S. Malik, D. Hnisz, T. I. Lee, I. I. Cisse, R. G. Roeder, P. A. Sharp, A. K. Chakraborty, R. A. Young, Coactivator condensation at super-enhancers links phase separation and gene control. *Science*. **361** (2018), doi:10.1126/science.aar3958.

48. S. Chong, C. Dugast-Darzacq, Z. Liu, P. Dong, G. M. Dailey, C. Cattoglio, A. Heckert, S. Banala, L. Lavis, X. Darzacq, R. Tjian, Imaging dynamic and selective low-complexity domain interactions that control gene transcription. *Science*. **361** (2018), doi:10.1126/science.aar2555.

49. S. V. Ulianov, A. K. Velichko, M. D. Magnitov, A. V. Luzhin, A. K. Golov, N. Ovsyannikova, I. I. Kireev, A. S. Gavrikov, A. S. Mishin, A. K. Garaev, A. V. Tyakht, A. A. Gavrilov, O. L. Kantidze, S. V. Razin, Suppression of liquid–liquid phase separation by 1,6-hexanediol partially compromises the 3D genome organization in living cells. *Nucleic Acids Research*. **49** (2021), pp. 10524–10541.

50. A. J. Bannister, R. Schneider, T. Kouzarides, Histone Methylation. *Cell*. **109** (2002), pp. 801–806.

51. J. R. Whetstine, Histone Methylation. *Handbook of Cell Signaling* (2010), pp. 2389–2397.

52. E. L. Greer, Y. Shi, Histone methylation: a dynamic mark in health, disease and inheritance. *Nat. Rev. Genet.* **13**, 343–357 (2012).

53. J. Gräff, L.-H. Tsai, Histone acetylation: molecular mnemonics on the chromatin. *Nat. Rev. Neurosci.* **14**, 97–111 (2013).

54. K. E. Neely, J. L. Workman, Histone acetylation and chromatin remodeling: which comes first? *Molecular Genetics and Metabolism*. **76** (2002), pp. 1–5.

55. D. E. Sterner, S. L. Berger, Acetylation of histones and transcription-related factors. *Microbiol. Mol. Biol. Rev.* **64**, 435–459 (2000).

56. S. J. Nowak, V. G. Corces, Phosphorylation of histone H3: a balancing act between chromosome condensation and transcriptional activation. *Trends in Genetics*. **20** (2004), pp. 214–220.

57. D. Rossetto, N. Avvakumov, J. Côté, Histone phosphorylation: a chromatin modification involved in diverse nuclear events. *Epigenetics*. **7**, 1098–1108 (2012).

58. S. G. Swygert, C. L. Peterson, Chromatin dynamics: interplay between remodeling enzymes and histone modifications. *Biochim. Biophys. Acta*. **1839**, 728–736 (2014).

59. F. Mattiroli, L. Penengo, Histone Ubiquitination: An Integrative Signaling Platform in

Genome Stability. *Trends Genet.* **37**, 566–581 (2021).

60. T. Kouzarides, Chromatin Modifications and Their Function. *Cell*. **128** (2007), pp. 693–705.

61. B. Li, M. Carey, J. L. Workman, The Role of Chromatin during Transcription. *Cell*. **128** (2007), pp. 707–719.

62. A. J. Bannister, T. Kouzarides, Regulation of chromatin by histone modifications. *Cell Research*. **21** (2011), pp. 381–395.

63. C. Demetriadou, C. Koufaris, A. Kirmizis, Histone N-alpha terminal modifications: genome regulation at the tip of the tail. *Epigenetics Chromatin*. **13**, 29 (2020).

64. G. J. Filion, J. G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L. D. Ward, W. Brugman, I. J. de Castro, R. M. Kerkhoven, H. J. Bussemaker, B. van Steensel, Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell*. **143**, 212–224 (2010).

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

66. M. H. Nichols, V. G. Corces, Principles of 3D compartmentalization of the human genome. *Cell Rep.* **35**, 109330 (2021).

67. N. Sikorska, T. Sexton, Defining Functionally Relevant Spatial Chromatin Domains: It is a TAD Complicated. *J. Mol. Biol.* **432**, 653–664 (2020).

68. Q. Szabo, F. Bantignies, G. Cavalli, Principles of genome folding into topologically associating domains. *Sci Adv*. **5**, eaaw1668 (2019).

69. K. Xiong, J. Ma, Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nat. Commun.* **10**, 5069 (2019).

70. Y. Zhu, Z. Chen, K. Zhang, M. Wang, D. Medovoy, J. W. Whitaker, B. Ding, N. Li, L. Zheng, W. Wang, Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.* **7**, 10812 (2016).

71. J. Huang, E. Marco, L. Pinello, G.-C. Yuan, Predicting chromatin organization using histone marks. *Genome Biol.* **16**, 162 (2015).

72. H. Ashoor, X. Chen, W. Rosikiewicz, J. Wang, A. Cheng, P. Wang, Y. Ruan, S. Li, Graph embedding and unsupervised learning predict genomic sub-compartments from HiC chromatin interaction data. *Nat. Commun.* **11**, 1173 (2020).

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

74. E. Sefer, C. Kingsford, Semi-nonparametric modeling of topological domain formation from epigenetic data. *Algorithms Mol. Biol.* **14**, 4 (2019).

75. Y. Chen, Y. Wang, Z. Xuan, M. Chen, M. Q. Zhang, De novo deciphering three-dimensional chromatin interaction and topological domains by wavelet transformation of epigenetic profiles. *Nucleic Acids Res.* **44**, e106 (2016).

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

Chapter 1. Noncoding loci without epigenomic signals can be essential for maintaining global chromatin organization and cell viability

1.1 Abstract

Most noncoding regions of the human genome do not harbor any annotated element and are even not marked with any epigenomic or protein binding signal. However, an overlooked aspect of their possible role in stabilizing 3D chromatin organization has not been extensively studied. To illuminate their structural importance, we started with the noncoding regions forming many 3D contacts (referred to as hubs) and performed a CRISPR library screening to identify dozens of hubs essential for cell viability. Hi-C and single-cell transcriptomic analyses showed that their deletion could significantly alter chromatin organization and affect the expressions of distal genes. This study revealed the 3D structural importance of noncoding loci that are not associated with any functional element, providing a previously unknown mechanistic understanding of disease-associated genetic variations (GVs). Furthermore, our analyses also suggest a possible approach to develop therapeutics targeting disease-specific noncoding regions that are critical for disease cell survival.

1.2 Introduction

Noncoding sequences of the human genome, such as noncoding RNAs (ncRNAs), enhancers, and transposons, are known to be critical for many biological processes and are thus functionally important. Despite the great progress in uncovering new roles of these noncoding elements, most of the human genome remains unannotated. As the three-dimensional (3D) organization of the genome is essential for regulating transcription and other cellular functions(*1–6*), an overlooked aspect of noncoding sequences is their "structural importance" in forming and maintaining the proper 3D chromatin structure, particularly for those that are not marked by any epigenetic signal or annotated with any functional unit.

In protein function analysis, some residues could be important, if they are essential for maintaining the proper conformation(*7*), even though they may not be directly involved in the protein's enzymatic activity or interaction with ligands. Similarly, noncoding genomic sequences could play critical roles in stabilizing the proper chromatin structure, although they do not harbor any enhancer or transcription factor (TF) binding site. Previous studies have shown that changing the noncoding sequences could alter chromatin organization; for instance, deletion of some boundary sequences of topologically associating domains (TADs)(*1*, *2*) causes aberrant gene transcription, leading to disease(*3*). TAD boundaries can be considered a special case, but the structural importance of noncoding sequences, particularly those not associated with TADs or any functional elements, has not been fully investigated.

Deleting a noncoding sequence and examining a phenotypic readout such as cell viability can directly assess its importance. High-throughput genetic screening by the CRISPR-Cas9 system has been effectively applied to analyzing long ncRNAs (lncRNAs)(*8*, *9*), enhancers, and promoters(*10–12*). However, it is still prohibitive to delete each 5-kb segment in the genome for thorough screening, and random selection of deletion loci is inefficient. For example, less than 3% of lncRNAs were reported to be essential for cell growth and survival(*8*, *9*), and this percentage is expected to be even lower for unannotated noncoding loci. A reasonable strategy is to start with the genomic loci involved in many chromatin contacts, hereinafter referred to as hubs, because disrupting these hubs would potentially lead to a relatively profound perturbation to the chromatin organization.

Here, we performed network analysis on Hi-C 3D contact data and identified a group of loci as hubs. Through a high-throughput CRISPR-Cas9 library screening by targeted deletion, we found that some hubs without any epigenetic marks were essential for cell growth and survival. We examined the impacts of hub deletion on the global chromatin structure and gene expression using Hi-C and single-cell RNA sequencing (scRNA-seq) technologies.

1.3 Materials and Methods

1.3.1 Network construction and hub identification

Evaluating the significance of Hi-C interaction pairs. We collected the raw reads, scale factors for vanilla coverage (VC) normalization, and the expected normalized reads for interaction pairs from the Hi-C experiments provided by Rao et al.(*13*)(GSE63525). The raw read, R_{ij} , between fragments F_i and F_j was first divided by both sequence distances between F_i and F_j and obtained the expected normalized reads for the scale factors S_{F_i} and S_{F_j} for VC normalization, $R_{ij}^{norm} = \frac{R_{ij}}{S_{F_i}S_{F_j}}$. Then, we calculated the distance R_{ij}^{exp} . Last, the significance of the interaction between F_i and F_j was evaluated using the P value of the normalized read R_{ij}^{norm} calculated on the basis of a Poisson distribution (*4*) with an expectation equal to R_{ij}^{exp}(*14*).

Hub identification. We identified hubs in each Fragment Contact Network (FCN) using a Z-score of its degree, $Zscore = \frac{d_i - \mu}{\sigma}$, where d_i is the degree of the ith fragment and μ and σ are the average and standard deviation of the degrees of all nodes in a chromosome of a cell line. We used a Z-score cut-off of 2.0 to select hubs that accounted for less than 10% of the total nodes(*14*) (**Table S1.1**).

Epigenetic signal/gene enrichment in hubs and nonhubs. The peaks of six histone modifications (H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K27me3, and H3K9me3) and ATAC-seq peaks were counted in the hub/nonhub regions in the six cell lines (GM12878, HMEC, HUVEC, IMR90, NHEK, and K562). They were downloaded from www.encodeproject.org/. The KBM7 cell line was not included in the analysis because of the lack of enough histone modification ChIP-seq data. Distributions of the overlapping histone modification and ATAC-seq peaks were compared between hubs and nonhubs, and *P* values were calculated using matched-pairs *t* test.

To check the gene enrichment in the hub region and the entire genome, the annotated genes in hg19 genome downloaded from the UCSC genome browser overlapped with the whole genome (all 563,566 5-kb fragments covered in the Hi-C data in the entire genome), union of the hubs (union of all 87,324 hubs in the seven cell lines), and common hubs (8025 common hubs were found in the seven cell lines).

Cell line specificity of the node degree distribution. For 5-kb resolution Hi-C data in the five cell lines, we used a correlation-based method to evaluate cell type specificity. (i) The degree of each node was represented as a vector containing the degree *z* score values calculated for the five cell lines that had both GV and Hi-C data (GM12878, HMEC, HUVEC, IMR90, and K562). (ii) For cell type specificities, there are $2^5 = 32$ possible vectors, including 2 with no cell line specificity (0,0,0,0,0), (1,1,1,1,1); 5 specific to one cell line (1,0,0,0,0), (0,1,0,0,0)...(0,0,0,0,1); 10 specific to two cell lines (1,1,0,0,0), (1,0,1,0,0)...(0,0,0,1,1); 10 specific to three cell lines (1,1,1,0,0), (1,0,1,1,0)...(0,0,1,1,1); and 5 specific to four cell lines (1,1,1,1,0), (1,0,1,1,1)...(0,1,1,1,1). (iii) For each node, we calculated the Pearson correlation between the degree vector and these cell line specificity vectors. If the best correlation coefficient was larger than a threshold of 0.9 (P

< 0.006), then we assigned the node with the corresponding cell line specificity(*14*). (**Table S1.2**)

For 20-kb resolution Hi-C data in 12 normal cell lines and 2 cancer cell lines (**Table S1.3**), we used a distribution-based method to evaluate the cell type specificities. (i) The degree of each node was represented as a vector containing the degree z score values calculated in all cell lines that had both GV and Hi-C data. (ii) For each node, we assumed that the normalized degrees obey a Gaussian distribution across normal cell lines, and we calculated the mean and SD. (iii) On the basis of the mean and SD for each node, we calculated the *z* score for each cell line, i.e., the cell line specificity z score. A node was considered cell line–specific if the absolute value of the cell line specificity z score was greater than 1(*14*). The “Network construction and hub identification” section was presented in an earlier and limited preprint version of this study deposited in BioRxiv(*14*).

1.3.2 Hub screening and validation

Cell culture. K562, H1975, and NAMALWA cells were cultured in RPMI 1640 medium (Gibco). 293T, HeLa, A549, and Huh7.5.1 cells were cultured in Dulbecco’s modified Eagle’s medium (Gibco). All cells were supplemented with 10% fetal bovine serum (Biological Industries) with 1% penicillin/streptomycin and cultured in 5% CO2 at 37°C.

Design and construction of the CRISPR-Cas9 pgRNA library. To validate the importance of the hub regions, we sorted the hub regions with PLT and selected the top

700 all-cell line hubs and top 300 K562-specific hubs. Among them, 960 hubs were suitable for designing pgRNAs for CRISPR-Cas9 screening. For each hub, up to 20 pgRNAs were designed to target 1-kb upstream and 1-kb down- stream regions flanking the two boundaries of the 5-kb segment. To ensure the cleavage accuracy and efficacy, we required sgRNAs in each pair to contain at least two mismatches to any other loci in the human genome, and their GC contents are between 0.2 and 0.8. For all the possible pgRNAs obtained from the selected sgRNAs, we removed those that may delete any promoter or exon of protein-coding genes, and we ensured that the cut site of each sgRNA is at least 30 base pairs (bp) away from the exon-intron boundary of the coding genes. We also designed 473 pgRNAs deleting the promoter region and first exon of 29 ribosomal genes as positive controls, and 100 pgRNAs targeting the AAVS1 locus as well as 100 nontargeting pgRNAs as negative controls, which were obtained from our previous library(*15*). As a result, the hub deletion library contained 17,476 pairs of gRNAs targeting 960 hub loci. The 128-nt oligonucleotides containing pgRNA coding sequences were designed, synthesized (Agilent Technologies Inc.), and cloned into the lentiviral expression vector following the two-step cloning method as previously described(*15*), with a minimum representation of 150 transformed colonies per pgRNA in each cloning step.

CRISPR-Cas9 pgRNA library screening. K562 cells stably expressing Cas9 were infected with pgRNA library lentiviruses at an MOI of <0.3 (1000× to 1500× coverage of the library), and two replicates were arranged. Seventy-two hours after infection, enhanced green fluorescent protein–positive (EGFP+) cells were selected by fluorescence-activated cell sorting (FACS; day 0). For each replicate, the harvested cells were divided into a day 0 control group and an experimental group, which was further maintained at a minimum coverage of 1500× for 30 days. Then, cells from each group with 1500× library coverage were, respectively subjected to genomic DNA extraction, PCR amplification of sgRNA-coding sequences, and high-throughput sequencing analysis (Illumina HiSeq 2500 and HiSeq X Ten platform) as previously described(*15*).

Identification of functional hubs. Sequencing reads were mapped to the pgRNA library and further normalized to reads per million for each barcoded gRNA. After calculating the quantile of pgRNA counts from two replicates, we removed noisy pgRNAs if a pgRNA's quantile difference of two replicates was in either 3% tail of the distribution. Then, $\log_2$FC between the experimental and control groups was calculated for each pgRNA, and 100 negative control genes were generated by randomly sampling 20 *AAVS1*-targeting pgRNAs with replacement. Two scores for each set of hubs were calculated: (i) the mean $\log_2$FC of all pgRNAs in the set, denoted by FC_{hub}; and (ii) $-log_{10}P_{value}$of the one-sided Mann-Whitney U test of all pgRNAs in the set compared with pgRNAs targeting the AAVS1 locus, denoted by P_{hub}. The background distribution of these two scores was represented by the mean (μ_{FC} and μ_P) and SD (σ_{FC} and σ_P) of all negative control genes. Then, the essentiality of hubs was evaluated by the following function and hubs with the lowest I_{score}(≤ −1) were identified as essential hubs.

$$I_{score} = sign(\frac{FC_{hub}-\mu_{FC}}{\sigma_{FC}}) \times \left|\,|\frac{FC_{hub}-\mu_{FC}}{\sigma_{FC}}| + \frac{P_{hub}-\mu_P}{\sigma_P}\right|$$

To further avoid the potential issue of cell toxicity generated from multiple cleavages by some pgRNAs, we retrieved the GuideScan specificity score to evaluate each sgRNA(*16*). By calculating the harmonic mean of the two sgRNAs for each pgRNA, a specificity score was generated for each pgRNA. We kept only the identified essential hubs if their targeting pgRNAs had specificity scores > 0.1 and $log_2FC < -1$. Furthermore, to avoid the copy number effect on dropout screening, the copy number of each hub locus in the K562 cell line was analyzed on the basis of ENCODE consortium copy number data (www.encodeproject.org/files/ENCFF486MJU/). After further filtering hub loci with copy number amplification, the remaining hits were regarded as essential hubs.

Distance between hubs and centromeres. We calculated the distances between hubs and centromeres using their nearest boundaries and compared the distance distributions for essential and nonessential hubs. Chi-squared goodness of fit test was used to calculate the P value.

Individual validation of essential hubs by cell proliferation assay. For each candidate hub locus, two pgRNAs were used for the individual validations, and they were either newly designed or selected from the library showing consistent depletion in replicates. To ensure high targeting specificity of all the selected pgRNAs, we required that their specificity scores are all greater than 0.15, and the score of at least one pgRNA for each hub is greater than 0.2. For the newly designed pgRNA, we further required that they do not include ≥4-bp homopolymer stretches and that their GC contents are between 0.4 and 0.7. We also changed the deletion regions, which included each sgRNA targeting −1 to +0.5 kb, flanking the two boundaries of the 5-kb hub loci (− and + refer to the outer and inner hub directions, respectively). Other rules were the same as those used for the pgRNA design in the library screening.

All the pgRNAs targeting each hub to be validated were individually cloned into a lentiviral expression vector containing an EGFP selection marker. After virus packaging, the pgRNA lentiviruses were respectively transduced into K562 cells at an MOI of <1. The percentages of EGFP-expressing cells indicating the fraction of pgRNA-containing cells were quantified every 3 days by FACS. Cell proliferation of each sample was measured by normalizing the per- centage of EGFP+ cells at each time point to that at 3 days after in- fection (labeled day 0), which was the same as previously described(*9*, *15*). The experiments lasted for 15 days after the first FACS analysis, and at least 100,000 cells were analyzed.

WGS to evaluate off-target effects. K562 cells were lentivirally transduced with the pgRNA hub_22_7-pg2. The EGFP+ cells were collected by FACS sorting at day 8 after pgRNA infection at an MOI of <1, and the sorted cells were subjected to genomic DNA extraction. The WGS library was prepared following the manufacturer's instructions and sequenced using the Illumina HiSeq 4000 platform. Using the WGS data, we evaluated the deletion efficiency at the targeted locus and off-target effects.

We downloaded the K562 (wild-type) WGS data from ENCODE with accession codes ENCFF313MGL, ENCFF004THU, ENCFF506TKC, and ENCFF066GQD and then evaluated the potential off-target effects following the published procedures(*17*). We first generated putative off-target sites for hub_22_7 in the hg19 genome using Cas-OFFinder (23). We called the base mismatch type with at most four mismatches without considering any bulge (mismatch ≤ 4, bulge = 0). We also called bulge mismatch type with at most two mismatches with at maximum two bulges (mismatch ≤ 2, bulge ≤ 2). In total, we examined 455 potential off-target loci. To detect the candidate mutations and indels in the hub-deleted cells, variant calling was performed as described in genome analysis toolkit (GATK) Best Practices (https://gatk.broadinstitute.org/hc/en-us). Briefly, reads were aligned to the human reference genome (hg19) using BWA-0.7.17. Duplicated reads were then removed using GATK4 MarkDuplicatesSpark (https://gatk.broadinstitute.org/hc/en-us/articles/360037224932-MarkDuplicatesSpark). The reads were then processed via base quality score recalibration using GATK4. Germline mutations (compared to the hg19 reference genome) were called in both wild-type and hub-deleted cells by GTAK HaplotypeCaller (version 4.1.4.1) with the default parameters. SNVs and indels called by GATK4 Mutect2 (version 4.1.4.1) with the default parameters were used to assess off-target deletions.

We further confirmed no off-target effects using a different analysis software, BCFTOOLS suite (version 1.9, www.htslib.org/doc/ bcftools.html), to reexamine the single-nucleotide polymorphisms (SNPs) and indel sites from the WGS data. The mapped BAM file of K562 cells was piped into bcftools mpileup and bcftools call with default parameters. The called raw variant call format (VCF) file was filtered by a bcftools filter with “%QUAL < 30 || DP < 30” marked as low-quality variants. Homozygous variants were also removed from the raw VCF file with the parameter “GT = 1/1.” Gold standard indels

VCF of Mills and 1000G were downloaded from GATK Resource Bundle (https://gatk.broadinstitute.org/hc/en-us/articles/360035890811- Resource-bundle). The gold standard indels were also removed from the VCF file using bcftools isec with parameter "-n -1 -c all." There were no putative off-target sites found in the 13,809 indels obtained using bedtools intersect (https://bedtools.readthedocs.io).

1.3.3 Hi-C library preparation and data analysis

Hi-C library preparation. The pgRNA Hub_22_7-pg2 was delivered into K562 cells through lentiviral infection at an MOI of <1. EGFP+ cells were collected by FACS sorting at day 9 after infection, and the sorted cells were allowed to recover under normal cell culture conditions for 2 hours before proceeding to conduct the Hi-C library. One million cells were used for each Hi-C library preparation using an Arima-HiC kit (Arima Genomics, San Diego) following the manufacturer's instructions. Hi-C libraries were sequenced using the Illumina NovaSeq platform.

Hi-C data processing. The Hi-C raw FASTQ data were processed by the Juicer pipeline(*18*) with the default parameters. Hi-C reads were aligned to hg19 (GRCh37), and the reads with mapping quality score (MAPQ) < 30 were further trimmed. The output bam files were trans- formed into 5-kb, 10-kb, 25-kb, 50-kb, 100-kb, and 1-Mb resolution contact matrix. The contact matrix was then normalized by the VC method(*13*). The significance level of a given interaction pair was calculated from Poisson distribution fitting between the measured interaction reads and the expected reads by VC normalization.

Juicebox (https://aidenlab.org/juicebox/) and HiCExplorer(*19*, *20*) were used to visualize the processed Hi-C data.

Loop calling. In both wild-type K562 and hub_22_7-deleted K562 cells, the VC normalized Hi-C contact reads were processed by HiCCUPS with default parameters at 25-kb resolution for calling loops. (https:// github.com/aidenlab/juicer/wiki/HiCCUPS).

TAD calling. We used insulation score(*21*) to identify the TADs for K562 wild-type and hub_22_7 deletion cells in 10-kb resolution data. The HiCExplorer software was used to plot the TADs(*19*).

A/B compartment analysis. The A/B compartment analysis was conducted using 50-kb bins. The eigenvectors for each chromosome in both K562 wild-type and hub-deleted cells were extracted from the VC normalized Hi-C counts processed by the Juicer pipeline with the default parameters(*18*). The polymerase II (Pol II) ChIP-seq data in K562 cells were downloaded from ENCODE(*22*). The correlation between the first eigenvector of each chromosome and the Pol II peaks density was calculated, on the basis of which we determined the A and B compartments(*23*). We repeated this analysis in GM12878, HUVEC, IMR90, and NHEK. For HMEC, there were no Pol II ChIP-seq data available, and thus, we used TSS density for hg19 genome to assign A/B compartments.

Effective diameter comparison. The effective diameter was computed by SNAP software (https://snap.stanford.edu/snap/). We calculated the effective diameter deviation for each chromosome both before and after hub deletion and found that the deviation followed a Gaussian distribution by the Shapiro-Wilk normality test (P = 0.27 so that the null hypothesis of being normal distribution was accepted). Then, we calculated the P value for the deviation of each chromo- some on the basis of a Gaussian distribution and identified the significantly changed chromosome with P < 0.05.

Modularity comparison. The modularity was computed by SNAP software (https://snap.stanford.edu/snap/). We collected the modularity scores of each chromosome in the seven wild-type cell lines (GM12878, K562, HUVECs, IMR90, NHEK, KBM7, and HMEC) and found that the modularity score for each chromosome followed a Gaussian distribution (all P values ≥ 0.01 to accept the null hypothesis of being a Gaussian distribution in the Shapiro-Wilk normality test). Then, for each chromosome in hub-deleted K562 cells, we calculated the P value of its modularity score on the basis of chromosome-specific modularity distribution and identified significantly changed chromosomes with P < 0.05.

1.3.4 Bulk RNA-seq and data analysis

Bulk RNA-seq library preparation. The pgRNA AAVS1-pg1 targeting the AAVS1 locus was delivered into K562 cells at an MOI of <1. Then, 2 × 10^6 EGFP+ K562 cells were sorted by FACS 8 days after transfection. Total RNA was extracted using the RNeasy Mini Kit (QIAGEN, 79254) with three replicates. The RNA-seq libraries were further prepared following the NEBNext PolyA mRNA Magnetic Isolation Module [New England Biolabs (NEB), E7490S], NEBNext RNA First Strand Synthesis Module (NEB, E7525S), NEBNext mRNA Second Strand Synthesis Module (NEB, E6111S), and NEBNext Ultra DNA Library Prep Kit for Illumina (NEB, E7370L). All samples were subjected to next-generation sequencing (NGS) analysis using the Illumina HiSeq 4000 platform.

Bulk RNA-seq data processing. In the bulk RNA-seq library, the sequencing reads with Phred scores of ≥30 were aligned to the human reference genome (GRCh37/hg19) using HISAT2 (2.0.4)(*24*, *25*) and assembled and quantified by StringTie(1.3.5)(*24*, *26*). The gene read counts for each sample were further normalized by CPM.

1.3.5 scRNA-seq and data analysis

Single-cell library preparation. K562 cells infected with Hub_22_7-pg2 were FACS-sorted 8 days after lentivirus transduction for single-cell library preparation. The single-cell library was prepared with the established protocol de- scribed previously(*27*). Briefly, polyadenylated RNA was reverse transcribed through tailed oligo(dT) priming directly in whole-cell lysate (single droplet) using Moloney murine leukemia virus reverse transcriptase (MMLV RT) and temperature switch oligos. The resulting full-length complementary DNA (cDNA) contained the complete 5′ end of the mRNA, as well as an anchor sequence that served as a universal priming site for second-strand synthesis. The cDNA was preamplified using 15 cycles with Kapa HiFi HotStart ReadyMix. We used the Nextera DNA Sample Preparation Kit to generate single-cell libraries. The amplified cDNA was tagmented at 55°C for 5 min in a 20-μl reaction with 0.25 μl of transposase and 5 μl of Nextera reaction buffer. Five microliters of neutralization buffer was added to the tagmentation reaction mix to strip the transposase off the DNA, and the tagmented DNA was amplified by 12 cycles of standard Nextera PCR. Then the DNA was purified with 20 μl of Ampure beads (sample to beads ratio of 1:0.6). The prepared libraries were sequenced on an Illumina HiSeq 4000 instrument.

scRNA-seq processing. The FASTQ files were first mapped to the human reference genome (GRCh37/hg19) using Picard (2.17.0) (https://broadinstitute.github.io/picard/) and STAR (2.5.3a)(*28*). We used the Drop-seq processing pipeline developed by the McCarroll laboratory(*27*) to remove low-quality reads (lower than Q10) and PCR duplicates (identified by cell barcodes and molecular barcodes). The cells were descendingly ordered by read count. Reads from all the cells were pooled together to form a cumulative distribution. Cells with the most reads before the inflection point "knee" of the cumulative distribution were kept for the following analysis.

We calculated a *P* value for each gene to assess whether the change was significant. Each cell was first normalized by CPM. We calculated E_i, which is the sum of CPMs for a given gene across all the cells, and E_{total}, which is the sum of E_i for all the genes. We then computed P_i =E_i /E_{total}. In a given cell *j*, the normalized gene expression of all genes was assumed to follow a binomial distribution G_{ij}~*B* ($N_{j,}$ P_i) independently and identically, where G_{ij} is the expected reads of gene i in cell j and N_j is the total reads for cell j. We calculated a *P* value to evaluate how significantly each gene expression in each cell deviated from the expected value on the basis of the binomial distribution, which indicates its differential expression across cells. We also calculated the *P* value for genes in the negative control (*Δ*AAVS1) and wild-type bulk RNA-seq data the same way.

Single-cell trajectory branching and pseudotime analysis. Because hub deletion affected cell proliferation, we focused on analyzing the apoptosis genes annotated in the KEGG database (www. genome.jp/kegg/). Considering the noise in the scRNA-seq data, we selected apoptosis genes that showed differential expression in at least 10 to 15% of cells ($P < 0.05$). As a result, 93 apoptosis genes were identified in K562 cells with the essential hub chr22: 17,325,000 to 17,330,000 deleted. All the single-cell and bulk data were clustered with trajectory branching and pseudotime analysis using the Monocle R package(*29*, *30*). Monocle(*29*, *30*) assigned each cell a pseudotime value and a "state" on the basis of the segment of the trajectory according to the PQ tree algorithm. Cells with the same state were clustered together(*30*), and then relative gene expression in each cluster was computed.

DEGs identified from pseudotime analysis. To identify differentially expressed genes (DEGs) between state 1 and state 2 defined in the pseudotime analysis, a Wilcoxon rank sum test was applied to identify DEGs in state 2 compared to those in state 1 using a P value cutoff of 0.05. The chromosome distributions for these DEGs are listed in **Table S1.4**.

Investigation of the essentialities of DEGs from scRNA-seq data. Among the DEGs in chr22 upon hub_22_7 (chr22: 17,325,000 to 17,330,000) deletion, which were significantly decreased from state 1 to state 2, a top-ranked DEG *THOC5* was selected to analyze its importance on cell growth and proliferation in K562 cells. Three sgRNAs were designed to knock down its expression through the CRISPRi strategy, which were selected from the hCRISPRi-v2 library(*31*). These sgRNAs were also individually cloned into the lentiviral expression vector with an EGFP marker and then respectively transduced into K562 cells stably expressing dCas9-KRAB (Krüppel-associated box) protein at an MOI of <1. The cell proliferation assay was performed as previously described(*9*, *15*). The first time point of FACS analysis was 6 days after lentiviral infection, and the experiment lasted for 12 days.

1.4 Results

We first downloaded the 5-kb resolution Hi-C data in seven human cell lines [GM12878, human mammary epithelial (HMEC), human umbilical vein endothelial (HUVEC), IMR90, normal human epidermal keratinocytes (NHEK), K562, and KBM7](*13*) and identified significant intrachromosomal contact pairs (*P* value cutoff of e^{-20}, see **Materials and Methods**). We next assembled all the contacts in a chromosome for a certain cell line into a network, which is hereinafter referred to as the fragment contact network (FCN). In the FCN, each node is a 5-kb fragment, and each edge represents a 3D contact. The degree of a node reflects how many contacts it forms. We calculated the *z* score of each node's degree as $z\,score = \frac{d_i - \mu}{\sigma}$, where d_i is the degree of the ith fragment and μ and σ are the mean and standard deviation of the degrees of all nodes in a chromosome of a cell line. The nodes with a $z\,score$ ≥ 2.0 were considered "hubs", whereas the rest of the nodes were considered "nonhubs". (see **Materials and Methods**, **Table S1.1** and ref(*14*)). The hubs count for less than 10% of the total nodes in a given FCN.

Note that these contacts indicate the spatial closeness of the contacting loci, and they are not necessarily mediated by proteins or ncRNAs to form specific chromatin loops. An analogy is the core residues of a protein, which are located in the interior and form many contacts with other residues but do not necessarily have specific residue-residue interactions mediated by such as hydrogen bonds and electrostatic interactions; however, deleting these residues can disrupt the packing of the interior residues and thus distort the proper conformation required for the protein's normal function. Similarly, perturbing a hub may have the same impacts on the 3D genome structure by disrupting chromatin organization.

To illustrate the importance of the hubs, we first investigated their contribution to stabilizing the FCN and their association with genetic variations (GVs; in this study, we focused on single nucleotide variations hereinafter) in cancer. Then, we identified hubs essential for cell viability using CRISPR screening. Lastly, we illustrated the impact of hub deletion on chromatin structure and gene expression using Hi-C and scRNA-seq.

1.4.1 FCN networks are resistant to random attacks but vulnerable to targeted attacks

In this study, we focused on intrachromosomal contacts and constructed FCNs for each chromosome in each cell line, resulting in a total of 161 (= 23 × 7) FCNs for all chromosomes in the seven cell lines. We found that the degree distribution of FCN follows a power law (**Figure 1.1A**), indicating that FCNs are scale-free networks. FCNs are resistant to random attacks (random removal of nodes in the network) but vulnerable to targeted attacks (targeted removal of specific nodes) against high-degree nodes, as scale-free networks(*32*). The 161 FCNs have similar network parameters, such as effective diameters, which is the path length such that 90% of node pairs are at a smaller or equal distance apart. The most significant outlier was the FCN of chr9 in the leukemia cancer cell line K562, which had a significantly larger effective diameter than the rest (**Figure 1.1B, Figure S1.1A**). We also calculated the diameter by considering the translocation between chr9 and chr22 (Philadelphia translocation), and it was still significantly different from other chromosomes. We found that computationally removing high-degree nodes from chr9 of GM12878 normal cells led to a similar degree distribution of chr9 in K562 cancer cells, which suggests that the targeted perturbation shifted the FCN of a normal cell toward that of a cancer cell (**Figure 1.1C**). This analysis suggests that GVs in K562 cells likely target the high-degree nodes of chr9 and thus alter the network properties. We also confirmed that the high degree nodes (hubs) are crucial for stabilizing the contacts between their connecting nodes in the network (hereinafter defined as “neighbors”) (**Figure S1.1C-F**).

We next investigated the genomic and epigenomic signals in the identified hub regions in six cell lines (no epigenomic data for KBM7). Compared to the nonhub loci, hub loci had fewer peaks for five histone marks (H3K27ac, H3K27me3, H3K4me1, H3K4me3, and H3K36me3) and a comparable number of H3K9me3 peaks (**Figure 1.1D-F, Figure S1.1G-I**). We also observed less open chromatin (**Figure 1.1G**) and fewer annotated regions (including coding genes, ncRNA, and other annotated regions downloaded from https://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/refFlat.txt.gz) in hubs than in nonhub regions (**Figure 1.1H**)(*14*) (see **Materials and Methods**). Furthermore, we compared the epigenetic marks (H3K27ac, H3K4me3, H3K4me1, H3K27me3, and H3K9me3) and assay for transposase accessible chromatin with high-throughput sequencing (ATAC-seq) peaks in the upstream and downstream of the hubs in multiple cell types. We considered different distances away from the hub regions ranging from 0 to 50 kb in linear distance (**Figure S1.3**). Comparing with the upstream and downstream regions, hubs also have lower H3K27ac, H3K27me3, H3K4me1, H3K4me3, H3K36me3, and open chromatin signals and comparable H3K9me3 peak numbers. We also identified A/B compartments for hub and nonhub loci and found that hubs are enriched in B compartments (**Figure S1.1J**), which is consistent with the histone modification analysis. These observations suggest that hubs are similar to the core residues in proteins, both densely packed in the interior of the 3D structure. Therefore, perturbation to hubs by GVs and deletions could disrupt chromatin packing to affect the surrounding 3D organization of chromatin and propagate through the genome, leading to observable phenotypes such as disease formation and cell death.

Next, we examined whether the hubs found in normal cells have significantly different 3D contacts in cancers and whether these changes are associated with GVs. Then, we investigated whether and how deleting hubs can cause cell death.

1.4.2 Cancer-related mutations alter 3D hub contacts

As K562 is a cancer cell line, we investigated whether K562-specific GVs are related to changes in spatial contacts. After calculating the z score for each node's degree so that it is comparable across cell lines, we checked its specificity, i.e., whether the contact degree was specifically high in any particular cell line.(see **Materials and Methods**) When considering all the nodes, we did not observe any specificity bias toward K562 cells: In the 563,566 total nodes of the whole genome, 38.3% showed no specificity, 12.2% showed specificity in K562 cells, and the largest of the other specificities was 13.4% (**Figure 1.1J**, **Table S1.2**).

Next, we calculated the Pearson correlation coefficient between a node's degree and GV occurrence in the node across cell lines. We found that K562-specific GVs were associated with the degree changes in K562 cells: Among all 54,117 nodes with degree-GV Pearson correlation coefficients > 0.9 (referred to as degree-GV–correlated nodes; **Figure 1.1I**), 24,229 (44.72%) were K562 specific, i.e., the GV is only observed in K562, and the degree of the node shows significantly higher or lower degree in K562 than the other cell lines; as a comparison, the largest percentage for another cell type (HMEC) specificity was only 10.6% (5743 nodes) (**Figure 1.1J**, **Table S1.2**). This bias toward the only diploid cancer cell line K562 among the seven was even more obvious for hubs: For all the hubs identified in at least one of the seven cell lines, there were 8765 degree-GV–correlated hubs, among which 5379 (61.37%) were K562-specific compared to the largest percentage of 824 (9.4%) specific to another cell type (HMEC) (**Figure 1.1J**, **Table S1.2**). Together, these analyses suggest that K562-specific GVs tend to significantly change the contact degrees, particularly on hubs, which is consistent with the observation that the FCN is vulnerable to targeted GVs in hubs.

GVs can either disrupt hubs in normal cells or form new disease-specific hubs in cancer cells. We thus analyzed hub formation and disruption separately and found a strong correlation between GV and contact degree change in K562 cells for both scenarios. In particular, the percentages of hub disruption in chr9 of K562 cells (i.e., hubs found in the other four cell types but not in K562 cells) were 47.56 and 47.50% without and with consideration of translocation between chr9 and chr22, respectively (only the untranslocated part of chr9 was used for calculation). This was significantly higher than all other chromosomes in each cell line, whose range was between 0 and 18.6% (**Figure 1.1K**). Our analyses clearly show that the GVs in K562 cells severely disrupted the hubs on chr9 shared by other cell lines.

To confirm the generality of this observation, we extended our analysis to four normal cell lines (GM12878, HMEC, HUVEC, and IMR90) and three cancer cell lines (HepG2, HeLa-S3, and K562) that had both 20-kb resolution Hi-C and GV data. We also found a strong correlation between the degree and GV in cancers (**Figure 1.1L**), suggesting that cancer-specific GVs tended to significantly alter the 3D contacts of hubs.

1.4.3 Targeted deletion of hubs can significantly affect cell viability

The above analyses indicated that hubs are not necessarily directly involved in functional activities, but they can be crucial for stabilizing the chromatin structure and are thus functionally important. To further test this hypothesis, we selected 960 hub regions (each 5 kb in length) to examine their impacts on cell growth and survival in a high-throughput deletion screen with the highest partner linking tendency (PLT). These hubs are those likely to stabilize the contacts between neighbors, including 683 hubs present in all cell lines and 277 hubs specific to K562 cells. They are evenly distributed along the chromosomes (**Figure S1.4E**).

For screening, we constructed a paired-guide RNA (pgRNA) library(*15*) targeting the selected hubs mediated by the CRISPR-Cas9 system. Using lentiviral transduction at a low multiplicity of infection (MOI) of <0.3, we transfected the pgRNA library containing a total of 17,476 pgRNAs into K562 cells stably expressing the Cas9 protein. This library also included 473 pgRNAs targeting essential ribosomal genes as positive controls, 100 pgRNAs targeting the AAVS1 locus, and 100 nontargeting pgRNAs as negative controls. The library cells were cultured for 30 continuous days after transduction. We sequenced cells at day 0 (controls) and day 30 to determine the abundance of barcode-gRNA regions, which represent the corresponding pgRNAs (**Figure 1.2A**).

Distributions of pgRNA reads from the control/experimental group between two biological replicates were highly correlated (**Figure S1.4A-B**), and the scatter plot of each hub's mean fold change between replicates also showed a high correlation (Pearson correlation coefficient = 0.75) (**Figure S1.4C**). In the day 30 cell population, compared with nontargeting pgRNAs or those targeting AAVS1, we identified hub regions with significant depletion in their targeting pgRNAs, consistent with positive controls that target essential ribosomal genes. The fold changes of all pgRNAs targeting each hub were calculated, and their P values were computed by comparison with the AAVS1-targeting pgRNAs using the Mann-Whitney U test(*8*, *9*), which is focused on analyzing screening data with the in-library controls and could more accurately reflect the fitness effect of each locus. AAVS1-targeting pgRNAs were randomly sampled to generate a distribution of negative controls, which was used to compute the hubs' *P* values. Combining the mean fold change and corrected P values, an I_{score} was computed for each hub. Eventually, the hubs whose Iscore was less than or equal to −1 were considered essential hits. Overall, 77 hubs were selected in K562 cells whose deletion led to cell death or growth inhibition (**Figure 1.2B**).

It has been reported that multiple cleavages in genomic loci generated by Cas9 activity could lead to cellular toxicity and thus affect growth screen measurements(*33*–*36*). To minimize the potential off-target effects, we calculated the GuideScan specificity score(*16*) for each single guide RNA (sgRNA) of every pgRNA, which focused on assessing the specificities of sgRNAs with two or three mismatches to off-target loci that are commonly used in library screens, and generated a specificity score for each pgRNA. We found that pgRNA targeting AAVS1 with a specificity score ≤ 0.1 could lead to a significant dropout effect in K562 cells (**Figure S1.4D**). To further assure the target specificity, we selected only targeting pgRNAs with specificity scores > 0.1 and log2 (fold change) (log_2FC) < −1 for subsequent analysis (**Figure 1.2C**). Furthermore, hub loci with copy number amplification were also filtered out to minimize the effect due to multiple cleavages by certain pgRNAs(*37*). Using these stringent criteria, we identified 35 essential hubs in K562 cells (**Figure 1.2C**). We checked the location of essential hubs and found some of them located near the centromeres (**Figure S1.4E**), but they are not significantly closer to centromeres than the nonessential ones (P = 0.092, **Figure S1.4F**).

We then chose seven candidate hubs for individual validation in K562 cells. For each hub, two or three pgRNAs with high specificity scores were selected (see **Materials and Methods**). All but two identified hubs were validated to severely affect cell growth and proliferation in K562 cells (**Figure 1.2D-E, Figure S1.5**), indicating their functional roles in cell fitness. To further explore the cell type specificity of the essential hubs, we selected hub_22_7 (chr22: 17,325,000 to 17,330,000, hg19), which showed the most significant growth defect in K562 cells, and performed the same cell proliferation assay in five other cancer cell lines. Compared with negative controls targeting the AAVS1 locus, targeted deletion of hub_22_7 did not lead to significant cell death or cell growth inhibition in the following four tested cell lines: HeLa (cervical cancer cells), H1975 (non–small cell lung cancer cells), A549 (non–small cell lung cancer cells), and NAMALWA (Burkitt's lymphoma) (**Figure 1.2F, Figure S1.6A**). In the liver cancer cell line Huh7.5.1, deletion of hub_22_7 showed a weak effect on cell fitness compared with deletion of the essential gene RPL19 serving as the positive control (**Figure 1.2F, Figure S1.6A**). Overall, only the hub_22_7 locus exhibited a remarkable essential role in the K562 cell line. These results validate the essential hubs identified in the screen.

1.4.4 Cell death caused by hub deletion does not result from disruption of functional elements or off-target effects

To illuminate the mechanism of cell death induced by hub deletion, we first examined the functional annotation and epigenetic modifications in these regions. None of the essential hubs overlap with gene coding regions, ncRNA regions, or TAD boundaries. A total of 77.1% (27 of 35 including 3 of 5 individually validated hubs) of the essential hubs did not overlap with any histone modification or TF chromatin immunoprecipitation sequencing (ChIP-seq) peak (**Figure 1.3A**, an example of hub_22_7 in **Figure 1.3B** and full genomics and epigenomics signals for hub_22_7 in **Figure S1.7**). We also checked the ChromHMM states (the 18-state data downloaded from the Roadmap Epigenomics project (https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html#exp_18state) in the K562 essential hubs and found that 82.286% of them are in the quiescent/low states (**Table S1.5**). These observations indicated that the essentiality of these hubs did not result from the genes or regulatory elements they harbor.

We next evaluated the essential hub_22_7 to rule out the possibility that cell death was caused by off-target cleavage. Using the validated pgRNA hub_22_7-pg2 with high specificity, we first measured its deletion efficiency by real-time quantitative polymerase chain reaction (PCR) (**Figure S1.6B**) at each time point after pgRNA transduction and then performed whole-genome sequencing (WGS) to evaluate its potential off-target effect on the day showing the highest deletion efficiency (see **Materials and Methods**). We identified >3.7 million single-nucleotide variants (SNVs) and >890,000 indels compared to the hg19 reference genome (**Table S1.6**). The fact that we could successfully identify 87.4% germline mutations found in the published wild-type K562 cells (ENCODE database with the accession codes ENCFF313MGL, ENCFF004THU, ENCFF506TKC, and ENCFF066GQD) suggests reliable library quality. We manually checked the indels on 455 potential off-target loci and 2 on-target loci identified by Cas-OFFinder(*38*) using loose criteria (bulge = 0, mismatch ≤4; bulge ≤2, mismatch ≤2) to avoid missing any possible off-target site (see **Materials and Methods**). Significant indels were found in only 2 on-target loci and not found in any of the 455 putative off-target loci, indicating no off-target cleavage. These analyses confirmed that cell death caused by hub deletion did not result from off-target effects.

1.4.5 Deletion of essential hubs can alter the global chromatin structure

We next performed Hi-C analysis to examine the chromatin structure changes in hub_22_7-pg2–infected (hub_22_7-deleted) K562 cells. To characterize the global impact of hub deletion, we first constructed FCNs in the hub_22_7-deleted cells using the same criteria as in the wild-type cells and analyzed the changes in the network properties, including effective diameter and modularity, which is the difference between the fraction of edges observed within a group of nodes and the expected value in a random network.

By analyzing the effective diameters of the FCNs before and after the hub deletion, we found that chr9, chr10, and chr22 had significant changes ($P < 0.05$; see **Materials and Methods**, **Figure 1.3C**): chr22 and chr10 increased, while chr9 decreased upon hub deletion. The hub-deleted cells also showed significant changes in the modularity scores of chr9, ch10, chr16, and chr22 ($P < 0.05$; see **Materials and Methods**, **Figure 1.3D**). While the change in the hubs residing in chr22 and chr22-translocated chr9 in K562 cells was not unexpected, the unexpected impact on chr10 and chr16 illuminated the importance of the understudied interactions between chromosomes (**Figure S1.8**).

The increased diameter and modularity in chr22 suggest that hub deletion reduces long-range chromatin contacts and enhances modularization of the FCN, consistent with the overall Hi-C contact difference between the wild-type and hub-deleted cells (**Figure 1.3E**). We did find newly formed and disrupted chromatin loops (examples in **Figure 1.3E**) and merge or split of a small percent of TADs in the hub-deleted cells (examples of chr22:

24 to 26 Mb, 35 to 38 Mb, and 45 to 47 Mb in **Figure S1.9**). Together, deletion of a hub has a global impact on chromatin structure that can propagate to other chromosomes.

1.4.6 Deletion of essential hubs can up-regulate apoptotic genes

Next, we set out to identify genes whose expression was significantly affected by hub deletion. Cells transduced with pgRNAs have various rates toward cell death, and the cell population is thus heterogeneous. Therefore, we used single-cell analysis to define the different cell states in the population. We performed Drop-seq analysis(*27*) on hub_22_7-pg2–infected K562 cells and collected scRNA-seq data for 393 cells passing the quality control criteria. The bulk RNA-seq data of the wild-type and AAVS1-deletion K562 cells were included as controls. All the RNA-seq data were normalized using counts per million (CPM), and the scaled z score for each gene in each individual cell or bulk sample was calculated by fitting a binomial distribution (see **Materials and Methods**). The scaled z score matrix of single-cell and bulk RNA-seq data was used for the following analysis.

We performed trajectory branching and pseudotime analysis using Monocle(*29*, *30*). Given that cell viability was significantly affected upon hub_22_7 deletion, we analyzed 93 apoptosis genes documented in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (www.genome.jp/kegg/). The single cells together with bulk samples of the wild-type and AAVS1-deletion K562 cells were grouped into five cell states (**Figure 1.4A**). Both AAVS1-deletion and wild-type samples were assigned to state 1, indicating that AAVS1 deletion is a valid control. Single cells in state 1 resemble the wild-type cells at the low value of pseudotime, which is understandable because hub deletions were not synchronized in all cells. States 4 and 5 have the highest pseudotime values and thus are the most distinct from the wild-type state. Overall, the apoptosis genes showed increasing expression levels from state 2 to state 5 (see examples in **Figure 1.4B**). Because states 2, 4, and 5 are the leaf nodes in the trajectory tree that represent local minimum or maximum points, we clustered the apoptosis genes according to their expression profiles in these three states. Each of the gene clusters presented with unique patterns as they progressed toward cell apoptosis (**Figure 1.4C**). This scRNA-seq analysis depicted the transcriptomic progression toward cell death upon hub deletion in K562 cells.

1.4.7 Deletion of essential hubs can alter gene expression in distal regions

We noticed that multiple contacts between promoters and enhancers located at the opposite sides of the hub in the linear genome were disrupted upon hub deletion (**Figure 1.5A, Figure S1.10**), indicating that deleting a hub could affect transcriptional regulation. To investigate whether important genes in chr22(*31*, *39*) were affected, we compared the expression profiles of state 2 and state 1 and found significantly down-regulated genes upon hub deletion, including multiple essential genes whose gene knockdown would significantly affect the K562 cell viability (identified from previous genome-wide CRISPRi screening(*31*)), such as *ATXN10, THOC5, CHEK2,* and *HSCB* (**Figure 1.5B**). Notably, these genes are located distal (12 to 34 Mb away in the linear genome) from the deleted hub_22_7 loci. We confirmed the essentiality of *THOC5* in K562 cells through CRISPRi-based gene knockdown (**Figure 1.5B**). Furthermore, the high-resolution Hi-C data indicated that its promoter's interaction with enhancers was disrupted upon hub deletion (**Figure 1.5C**). They are located in compartment A (active) to compartment B (inactive) flip region (chr22: 29 to 32 Mb, **Figure 1.5C**), consistent with *THOC5* repression. These observations suggest that the chromatin structure alteration induced by hub deletion could affect the expression of distal genes, including those essential for cell viability.

1.4.8 The global impact of hub deletion suggests that hubs might be potential noncoding therapeutic targets

Given that deleting one essential hub can affect many genes, a new "one-drug–multiple-targets" therapeutic strategy may be developed to synergize different pathways. Namely, disease-specific noncoding regions, such as hubs that are essential in only cancer cells, could be potential therapeutic targets. In our screen, we identified a group of essential hubs specifically for K562 cells (**Figure 1.2E-F, Figure S1.6**). The deletion of hub_22_7 resulted in an approximately 80% decrease in the cell proliferation rate of K562 cells but nearly no significant effects on the other analyzed cell lines (**Figure 1.2F, Figure S1.6A**). As K562 cell is a leukemia cancer cell line, such K562-specific hubs could be potential therapeutic targets for chronic myelogenous leukemia. As shown above, deletion of this hub caused the down-regulation of many essential genes and the activation of apoptosis pathways. Therefore, this collective effect of killing cancer cells is more potent than targeting each individual pathway and would make it more difficult for cancer cells to develop drug resistance.

Furthermore, hub deletion also affected genes specifically expressed in K562 cells, although they are not essential for cell viability. For example, K562 cell-specific high expression of *TOP3B* (**Figure S1.11**), which plays important roles in the maintenance of gene stabilities and chromosome bridging(*39*, *40*), was down-regulated upon hub deletion due to the disruption of its promoter-enhancer interactions. By examining the ENCODE data in 23 cell lines/tissues, we found that the enhancers located at chr22: 17,125,000 to 17,130,000 were marked by H3K27ac in only K562 cells and another leukemia cell line, Dnd41 (**Figure S1.11**). The low expression of *TOP3B* in Dnd41 cells (**Figure S1.11**)(*37*) suggests that these enhancers may regulate only *TOP3B* in K562 cells. Therefore, deleting this hub can specifically down-regulate *TOP3B* in K562 cells.

We also used Genomic Regions Enrichment of Annotations Tool (GREAT)(*41*) to search for pathways enriched (binomial false discovery rate Q ≤ 1 $\times e^{-5}$) in the loci whose Hi-C contacts (P $\leq 10^{-20}$) were significantly reduced upon hub_22_7 deletion in chr22 (**Figure 1.5D**). Notably, the *APOBEC3* family genes stood out, and in particular, *APOBEC3B* was significantly down-regulated from state 1 to state 2 (**Figure 1.5F**). This is likely due to the reduced interaction between the *APOBEC3B* promoter and its enhancers upon hub_22_7 deletion (**Figure 1.5E**). APOBEC3 enzymes were reported as therapeutic targets for cancer treatment(*42*, *43*), and their aberrant expression (e.g., higher expression of *APOBEC3B*) could cause cancerous mutagenesis leading to drug resistance or metastasis(*44–46*). Although *APOBEC3B* is not essential for K562 cell viability, its down-regulation could effectively reduce the mutation rate, which is crucial for developing a potent therapy. Together, deleting one hub may synergize with multiple pathways to kill cancer cells and simultaneously reduce the cancer’s mutation capability. This example suggests that the identification and deletion of cancer-specific hubs could open a new avenue for developing potent therapeutics.

1.5 Discussion

Noncoding genomic regions without any epigenetic mark, open chromatin, or TF binding have been overlooked in functional analysis. By analyzing the 3D contact networks derived from Hi-C data, we found that such noncoding regions without any mark can be in contact with many other loci and thus become hubs in the 3D contact network. Our simulated deletion of hubs in normal GM12878 cells shifted the 3D contact network toward the K562 cancer cell line. Our analysis also showed a strong correlation between 3D contact change and GV occurrence in the hubs of cancer cell lines, suggesting that cancer-specific GVs tend to significantly alter the 3D contacts of hubs. These results indicate that hubs likely play critical roles in normal cells, and noncoding disease-associated GVs can occur in hub regions to form or disrupt hubs in normal cells, which may cause aberrant cellular functions leading to diseases. Therefore, our analysis provides a new perspective to understand the mechanisms of noncoding GVs that do not overlap with any epigenetic mark, TF binding, or open chromatin but are tightly associated with diseases.

To further examine the importance of the hub regions, we deleted 960 hubs in K562 cells using a pgRNA CRISPR-Cas9 library. Through computational analysis combined with the in-library AAVS1 controls and stringent filtering to avoid the potential issues of off-target effects and copy number amplifications, we found that 35 hubs could affect cell growth or viability after targeted deletion. The percentage of hubs essential for cell fitness is comparable to those of essential lncRNAs (<3%)(*8*, *9*) and protein-coding genes (< 3%)(*47*), which further supports the importance of hubs. Five of seven loci were individually validated with multiple pgRNAs, and hub_22_7 was further validated to be specifically essential for cell fitness in the K562 cell line. Using WGS analysis, we also confirmed that the targeting pgRNA of hub_22_7 has no off-target effect across the genome.

To understand the impact of hub deletion, we focused on § validated hub hub_22_7 that has no epigenetic mark, TF binding, or open chromatin signal in K562 cells. This hub was randomly selected from the K562 essential hubs and could serve as a representative group of cell type–specific essential hubs. Hi-C analysis showed that deleting the 5-kb hub significantly altered the 3D contact networks, as quantified by the significant change in FCN properties, including diameter and modularity. The hub deletion effects were far beyond the contacting loci of the hub and indicate that the impact of hub deletion is global.

We speculate that this global impact may start from the disruption of chromatin packing around the deleted hub and propagate to affect distal chromatin looping and promoter-enhancer interactions. An analogy is mutation of a residue in the interior of a protein's structure that can significantly change the protein conformation, leading to protein dysfunction. Therefore, although hubs do not host or interact with any gene, the propagated effect can alter the transcription of distal genes, as shown by the scRNA-seq data, which are essential for cell viability by themselves or in combination with other affected nonessential genes. We recognize that it is difficult to prove the causal relationship between global chromatin organization change and cell proliferation or gene expression, which still remains technically challenging and worthy of future investigation. Nevertheless, this is the first study to observe that noncoding loci without any epigenetic signals are not junk DNA, which could contribute to maintaining the global chromatin structure.

Furthermore, we showed that hubs can be cancer specific, which indicates a possibility of developing treatments to target a specific cancer. We are aware that the present studies are in cell lines and that further analysis in tumor tissues is necessary to confirm the translational value. However, it is worth noting that, because the global impact of hub deletion can affect many genes located distal from each other in the genome, the identified cancer-specific hubs could be potential new therapeutic targets. Targeting these noncoding loci could leverage the synergistic effects of multiple mechanisms to develop potent therapeutics, and treatment resistance is harder to develop because it requires mutations to interfere with the large number of genes affected by hub inhibition. There is a long way to go to translate this discovery, and there are possible roadblocks such as targeting multiple genes/pathways that may lead to lack of specificity for developing new therapeutics. As there are much more noncoding loci than the genes, overcoming the potential pitfalls requires additional effort to better understand the mechanisms of these “dark matter” in the genome for treating disease. Our findings here suggest an exciting direction for further exploration given the fast advancement of genome editing and delivery technologies.

Together, we report here the first study to reveal that noncoding loci without any epigenetic mark, TF binding, or open chromatin signal can be essential for cell viability. The importance of these loci for global chromatin organization and their impact on distal gene expression upon deletion make them a potential new class of therapeutic targets that have not yet been found.

1.6 Acknowledgements

Chapter 1, is a reformatted reprint of the material as it appears in “Noncoding loci without epigenomic signals can be essential for maintaining global chromatin organization and cell viability”. Bo Ding[#], Ying Liu[#], Zhiheng Liu[#], Lina Zheng[#], Ping Xu, Zhao Chen, Peiyao Wu, Ying Zhao, Qian Pan, Yu Guo, Wensheng Wei and Wei Wang. Science Advances, 7(45), eabi6020, 2021. DOI: 10.1126/sciadv.abi6020. ([#] These authors contributed equally to this work) The dissertation author was the primary investigator and author of this paper.

We acknowledge the staff of the BIOPIC High-throughput Sequencing Center (Peking University) for assistance in NGS analysis; the National Center for Protein Sciences (Beijing) at Peking University for assistance with FACS and analysis; and H. Lv and L. Du for technical help. We acknowledge Y. Yu (Peking University) for assistance in preparing the NGS library. We acknowledge the staff of the UC San Diego IGM Genomics Center for sequencing services and the UC San Diego Human Embryonic Stem Cell Core Facility for cell sorting services. We acknowledge J. Xu (UC San Diego) for assistance in preparing a single-cell RNA-seq library.

W.Wa. and W.We. conceived and supervised the project. W.Wa., W.We., B.D., and Y.L. designed the experiments. B.D. and L.Z. constructed the network analysis and identified and characterized hub regions. Y.G. designed the pgRNA library for hub screening. Y.L. and P.X. performed the pgRNA library construction and screening. Y.L. performed the experiments, including individual validation of candidate hubs, DEGs,

WGS, and bulk RNA-seq, with the help of P.X. and Q.P. Z.L. performed the bioinformatics analysis of the screening data and designed the pgRNAs used for individual validation. P.W. and Z.C. performed the Hi-C experiments on hub-deleted cells. P.W. and Y.Z. performed scRNA-seq on hub-deleted cells. L.Z. and B.D. performed the bioinformatics analysis of the WGS, Hi-C, and single-cell RNA-seq data. B.D., Y.L., L.Z., W.We., and W.Wa. wrote the manuscript with contributions from all other authors.

1.7 Figures

Figure 1.1. Characterization of the FCNs and hub nodes. (A) Degree distribution of FCN. (**B**) The effective diameter of FCN remains largely unchanged with increasing network size. “Translocated” (untranslocated): network constructed by considering (not considering) the translocation between chr9 and chr22. (**C**) The degree distribution of chr9 in normal cell lines (GM12878 as an example) after removal of high-degree nodes is similar to that of K562 chr9. (**D-G**) Epigenomic signals in hubs and nonhubs: H3K27ac (**D**), H3K4me1 (**E**), H3K9me3 (**F**), and ATAC-seq (**G**) in diverse cell lines. (**H**) The percentages of the annotated regions (including coding genes, ncRNA, and other annotated regions at https://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/refFlat.txt.gz) in the whole genome, union of hubs (hubs appeared in at least one cell line), and common hubs (hubs appeared in all cell lines). Numbers above the bar plot are the number of nodes overlapping with gene regions (top) and the number of nodes in that category (bottom in parentheses). (**I**) Definition of degree-GV–correlated nodes. The example node has a high degree in K562 and low degrees in others, which is correlated with the GV profile with a SNP in K562 but none in others. (**J**) The distribution of cell line specificities in all nodes, degree-GV–correlated nodes, and degree-GV–correlated hubs. (**K**) The distribution of one cell type–specific hub and four cell type–specific hubs in chromosomes and cell lines. (**L**) The percentage of degree-GV–correlated nodes in normal cell lines and cancer cell lines.

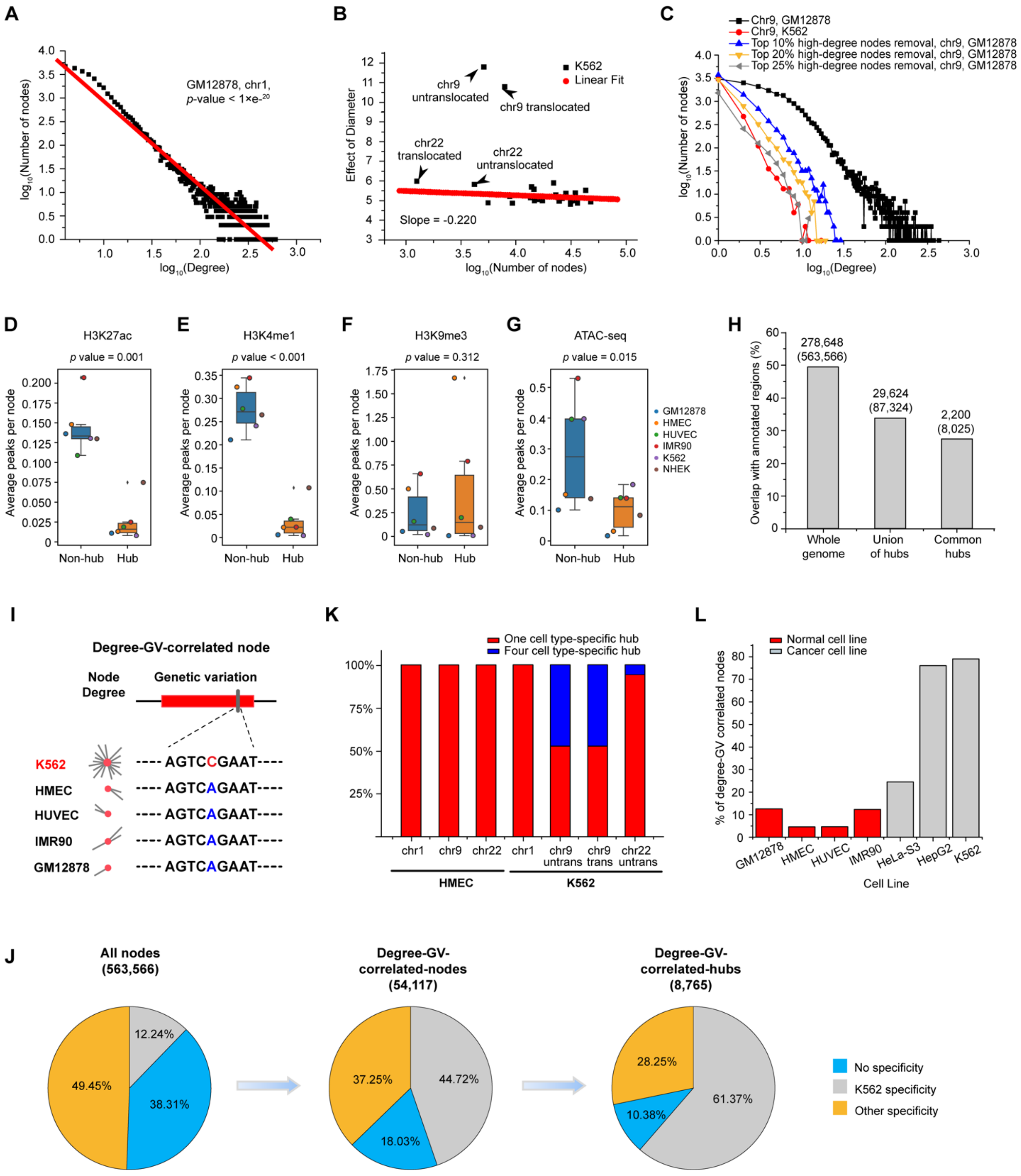
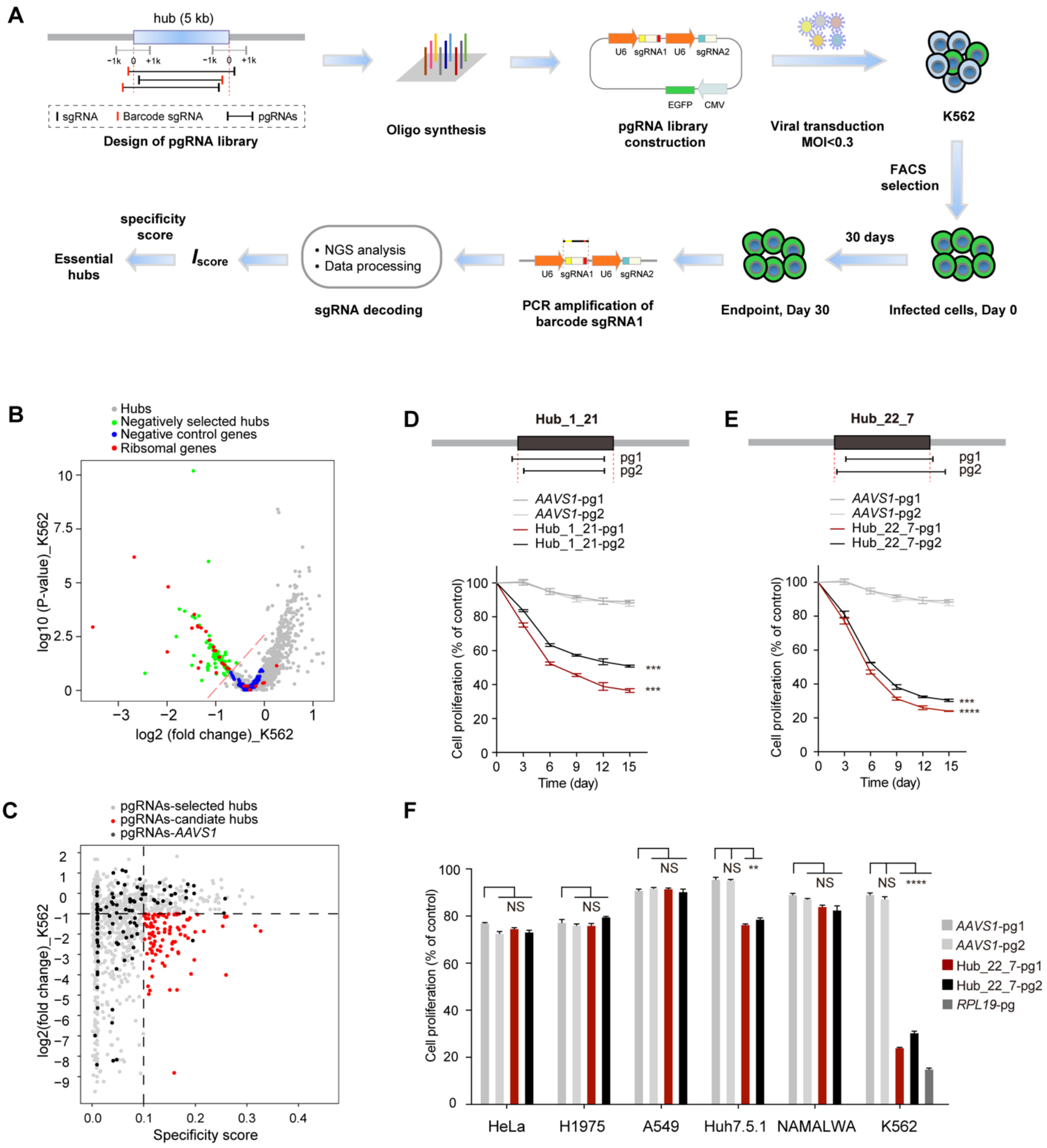

A
GM12878, chr1,
p-value < 1×e-20
log10(Number of nodes)
log10(Degree)
B
K562
Linear Fit
chr9 untranslocated
chr9 translocated
chr22 translocated
chr22 untranslocated
Slope = -0.220
Effect of Diameter
log10(Number of nodes)
C
Chr9, GM12878
Chr9, K562
Top 10% high-degree nodes removal, chr9, GM12878
Top 20% high-degree nodes removal, chr9, GM12878
Top 25% high-degree nodes removal, chr9, GM12878
log10(Number of nodes)
log10(Degree)
D
H3K27ac
p value = 0.001
E
H3K4me1
p value < 0.001
F
H3K9me3
p value = 0.312
G
ATAC-seq
p value = 0.015
Average peaks per node
Non-hub
Hub
GM12878
HMEC
HUVEC
IMR90
K562
NHEK
H
Overlap with annotated regions (%)
278,648 (563,566)
29,624 (87,324)
2,200 (8,025)
Whole genome
Union of hubs
Common hubs
I
Degree-GV-correlated node
Node Degree
Genetic variation
K562 ---- AGTCCGAAT----
HMEC ---- AGTCAGAAT----
HUVEC ---- AGTCAGAAT----
IMR90 ---- AGTCAGAAT----
GM12878 ---- AGTCAGAAT----
K
One cell type-specific hub
Four cell type-specific hub
100%
75%
50%
25%
chr1
chr9
chr22
chr1
chr9 untrans
chr9 trans
chr22 untrans
HMEC
K562
L
Normal cell line
Cancer cell line
% of degree-GV correlated nodes
GM12878
HMEC
HUVEC
IMR90
HeLa-S3
HepG2
K562
Cell Line
J
All nodes (563,566)
12.24%
38.31%
49.45%
Degree-GV-correlated-nodes (54,117)
44.72%
18.03%
37.25%
Degree-GV-correlated-hubs (8,765)
61.37%
10.38%
28.25%
No specificity
K562 specificity
Other specificity

Figure 1.2. Identification of essential hubs for cell growth and proliferation in the K562 cell line through pgRNA-mediated fragment deletion. (**A**) Schematic of the pgRNA library design, cloning, and functional screening of selected hub loci. CMV, cytomegalovirus. (**B**) Volcano plot of the fold change and P value of hubs in the K562 cell line. The dotted red line represents Iscore = −1. (**C**) Selection of candidate essential hubs by pgRNA fold change and specificity score. Essential hits were selected with specificity score > 0.1, log2 (fold change) (log2FC) < −1. (**D** and **E**) Validation of top-ranked essential hubs in K562 by cell proliferation assay. AAVS1-pg1 and AAVS1-pg2 are pgRNAs targeting AAVS1 as negative controls. Asterisk (*) represents P values compared with AAVS1-pg1 at day 15, calculated by two-tailed Student's t test, and adjusted by Benjamini-Hochberg procedure. (**F**) Validation of hub_22_7 in multiple cancer cell lines, including A549, H1975, HeLa, Huh7.5.1, and NAMALWA. Asterisk (*) represents P values compared with AAVS1-pg1 at day 15, calculated by two-tailed Student's t test, and adjusted by Bonferroni correction accounting for multiple testings. Data are presented as the means ± SD. (n = 3). **P < 0.01, ***P < 0.001, and ****P < 0.0001. NS, not significant.



A
hub (5 kb)
−1k 0 +1k −1k 0 +1k
I sgRNA I Barcode sgRNA ⊢⊣ pgRNAs
Design of pgRNA library
Oligo synthesis
U6 sgRNA1 U6 sgRNA2
EGFP CMV
pgRNA library construction
Viral transduction MOI<0.3
K562
FACS selection
30 days
Infected cells, Day 0
Endpoint, Day 30
U6 sgRNA1 U6 sgRNA2
PCR amplification of barcode sgRNA1
• NGS analysis
• Data processing
sgRNA decoding
I_{score}
specificity score
Essential hubs
B
● Hubs
● Negatively selected hubs
● Negative control genes
● Ribsomal genes
log10 (P-value)_K562
10
7.5
5
2.5
0
−3 −2 −1 0 1
log2 (fold change)_K562
C
● pgRNAs-selected hubs
● pgRNAs-candiate hubs
● pgRNAs-AAVS1
log2(fold change)_K562
2 1 0 −1 −2 −3 −4 −5 −6 −7 −8 −9
0.0 0.1 0.2 0.3 0.4
Specificity score
D
Hub_1_21
pg1
pg2
AAVS1-pg1
AAVS1-pg2
Hub_1_21-pg1
Hub_1_21-pg2
Cell proliferation (% of control)
100 80 60 40 20 0
0 3 6 9 12 15
Time (day)
***
***
E
Hub_22_7
pg1
pg2
AAVS1-pg1
AAVS1-pg2
Hub_22_7-pg1
Hub_22_7-pg2
Cell proliferation (% of control)
100 80 60 40 20 0
0 3 6 9 12 15
Time (day)
***
****
F
Cell proliferation (% of control)
100 80 60 40 20 0
NS
NS
NS
NS **
NS
NS ****
HeLa H1975 A549 Huh7.5.1 NAMALWA K562
AAVS1-pg1
AAVS1-pg2
Hub_22_7-pg1
Hub_22_7-pg2
RPL19-pg

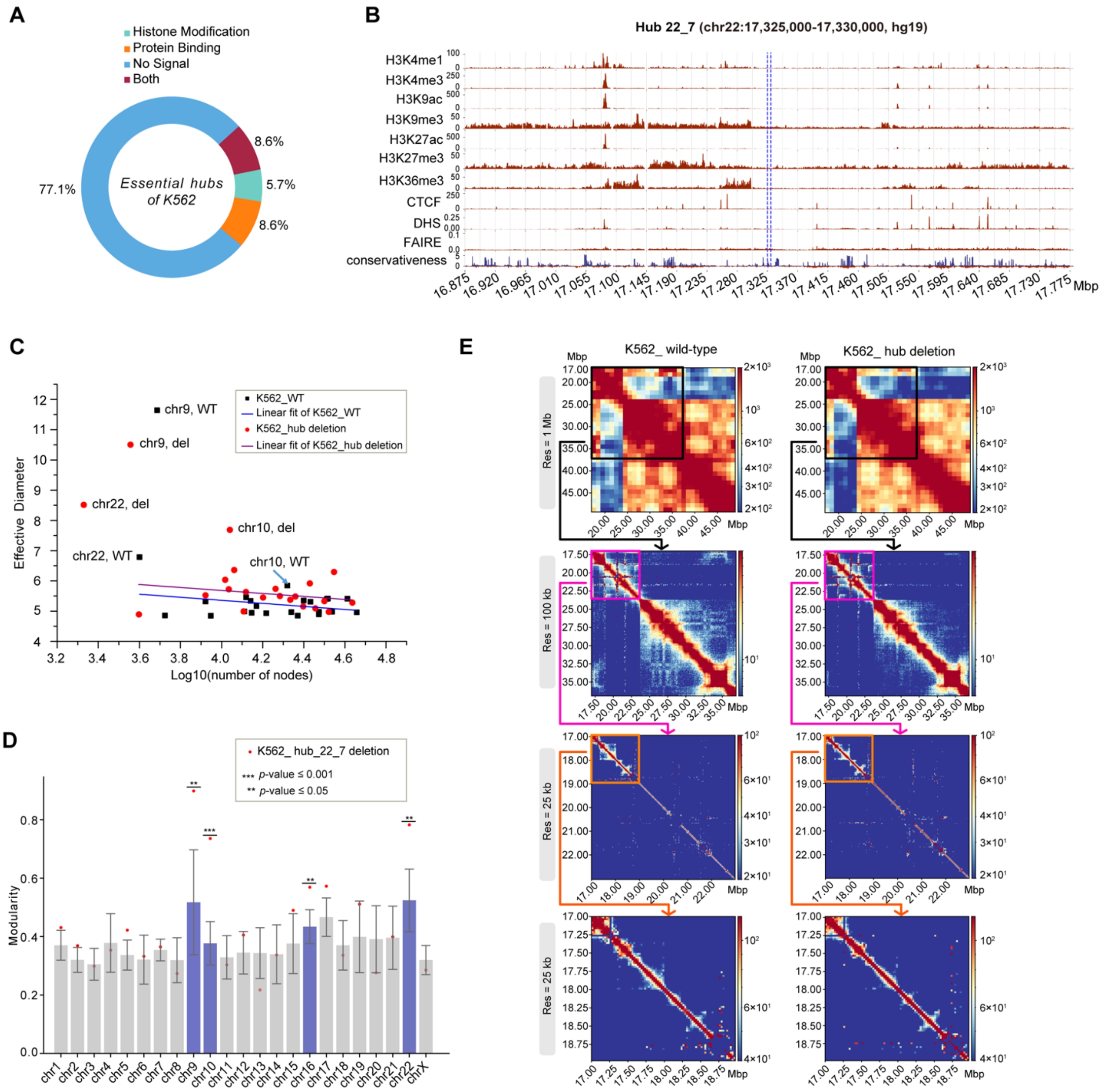


Figure 1.3. Characterization of the deleted hub and the impact of its deletion on the global chromatin structure. (**A**) Overlap of essential hubs in K562 cells with the peaks of 10 histone marks (H3K27ac, H3K4Me1, H3K4me3, H3K27me3, H3K9me3, H3K36me3, H3K4me2, H3K79me2, H3K9ac, and H3K20me1) and 151 TFs. (**B**) Histone marks, CTCF (CCCTC-binding factor) binding, open chromatin, DHS, DNase hypersensitivity; FAIRE, formaldehyde-assisted isolation of regulatory elements, and conservation score (100 vertebrates basewise conservation by PhyloP) on chr22: 17,325,000 to 17,330,000 (see Figure S1.6 for all signals). (**C**) Effective diameters versus log10(number of nodes) the wild-type (WT) and hub_22_7-deleted K562 cells. (**D**) Modularity scores in the seven wild-type cell lines for 23 chromosomes. Red dots, hub_22_7 deletion. (**E**) Hi-C contacts for wild-type and hub_22_7-deleted K562 cells at 1-Mb, 100-kb, and 25-kb resolutions.

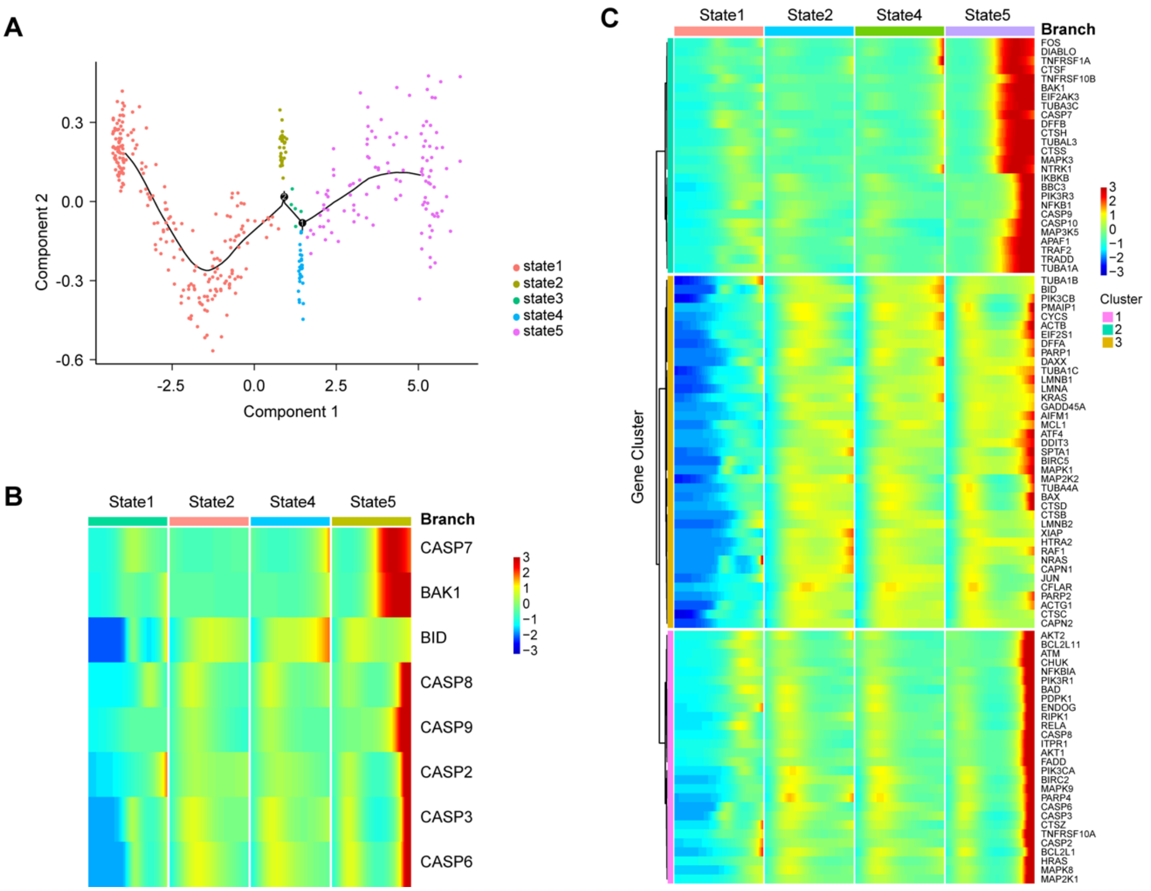
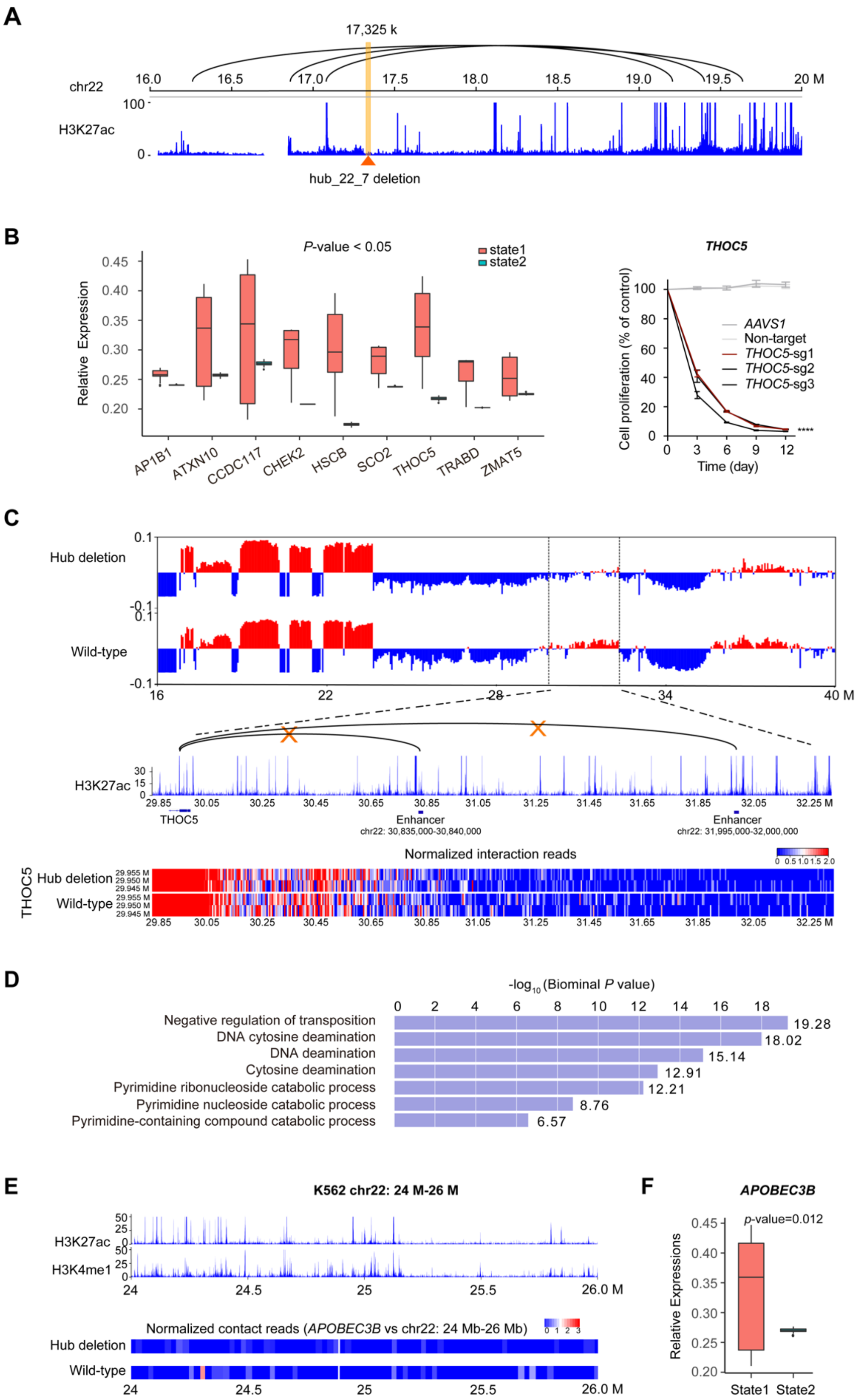


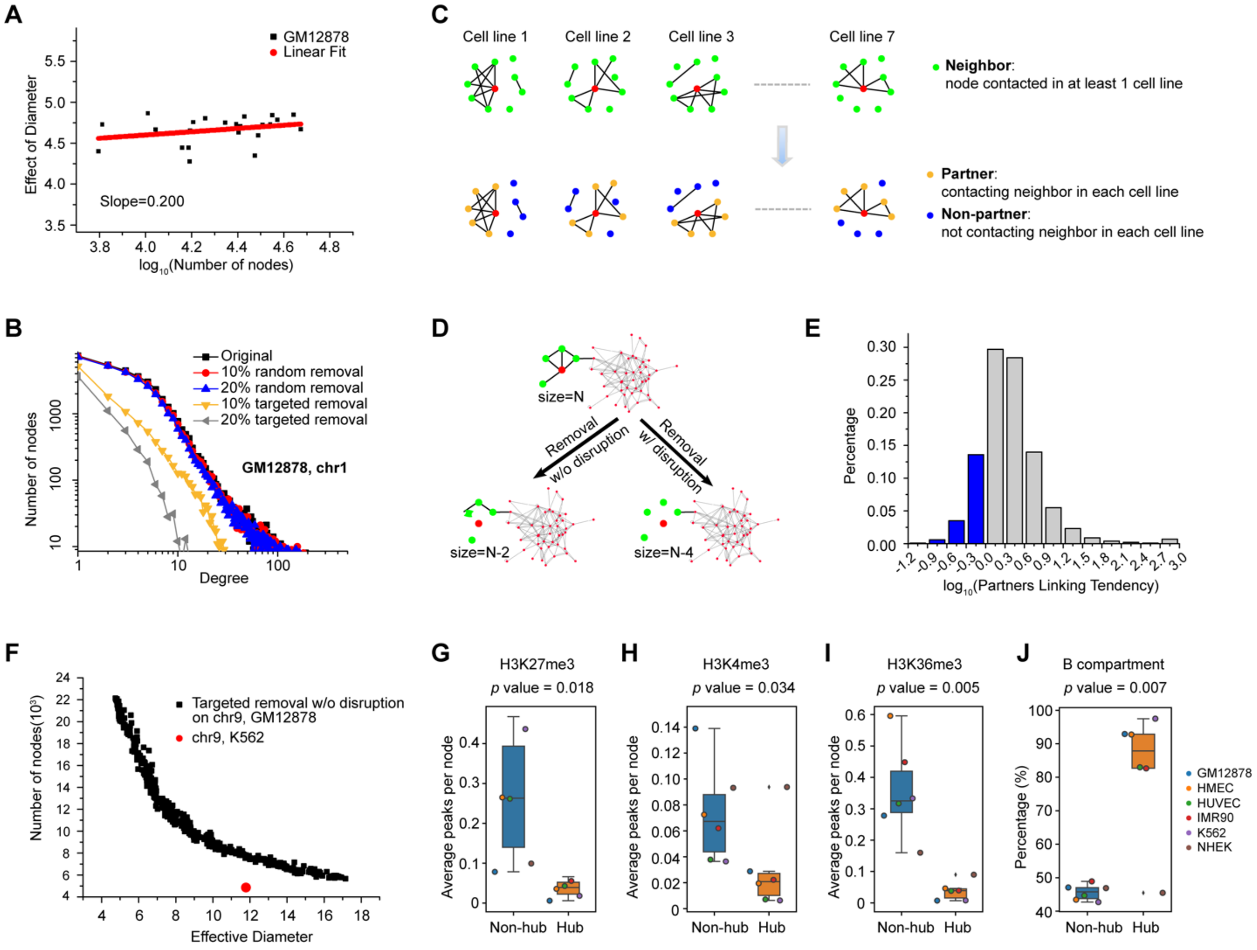
Figure 1.4. Hub deletion induces global changes in gene expression. (**A**) Pseudotime clusters of hub_22_7-deleted and wild-type K562 cells based on apoptosis gene expression. (**B**) Examples of typical apoptosis gene expressions (CASP2, CASP3, CASP6, CASP7, CASP8, CASP9, BAK1, and BID) under different cell states defined by pseudotime analysis. (**C**) Global analysis of 93 KEGG apoptosis gene expression levels in states 1, 2, 4, and 5. Genes were clustered into three groups.

Figure 1.5. The concurrent alterations of 3D chromatin structure and gene expressions after hub deletion. (**A**) Disruption of enhancer-promoter interactions upon hub_22_7 deletion. (**B**) Essential genes of K562 cells located on chr22 with significantly down-regulated expression ($P < 0.05$) in state 2 compared to state 1. (**C**) A/B compartment change (50-kb resolution) upon hub deletion. Multiple enhancer-promoter contacts with *THOC5* were disrupted in the compartment changing region (chr22: 29,850,000 to 32,350,000). (**D**) Gene Ontology biological process pathways associated with loci whose 3D contacts were disrupted by hub deletion. (**E**) The 3D contacts between the *APOBEC3B* promoter and enhancers located on chr22: 24,000,000 to 26,000,000 were significantly decreased in hub-deleted cells. The enhancers were identified using the overlapping peaks of H3K27ac and H3K4me1 in the wild-type cells. (**F**) The relative expression level of *APOBEC3B* from state 1 to state 2.


A
17,325 k
chr22
16.0
16.5
17.0
17.5
18.0
18.5
19.0
19.5
20 M
100
H3K27ac
0
hub_22_7 deletion
B
P-value < 0.05
state1
state2
Relative Expression
0.45
0.40
0.35
0.30
0.25
0.20
AP1B1
ATXN10
CCDC117
CHEK2
HSCB
SCO2
THOC5
TRABD
ZMAT5
THOC5
Cell proliferation (% of control)
100
80
60
40
20
0
AAVS1
Non-target
THOC5-sg1
THOC5-sg2
THOC5-sg3
****
0
3
6
9
12
Time (day)
C
Hub deletion
Wild-type
0.1
-0.1
16
22
28
34
40 M
H3K27ac
30
15
0
29.85
30.05
30.25
30.45
30.65
30.85
31.05
31.25
31.45
31.65
31.85
32.05
32.25 M
THOC5
Enhancer
chr22: 30,835,000-30,840,000
Enhancer
chr22: 31,995,000-32,000,000
Normalized interaction reads
0
0.5
1.0
1.5
2.0
THOC5
Hub deletion
Wild-type
29.955 M
29.950 M
29.945 M
D
-log10 (Biominal P value)
0
2
4
6
8
10
12
14
16
18
Negative regulation of transposition
DNA cytosine deamination
DNA deamination
Cytosine deamination
Pyrimidine ribonucleoside catabolic process
Pyrimidine nucleoside catabolic process
Pyrimidine-containing compound catabolic process
19.28
18.02
15.14
12.91
12.21
8.76
6.57
E
K562 chr22: 24 M-26 M
H3K27ac
H3K4me1
50
25
0
24
24.5
25
25.5
26.0 M
Normalized contact reads (APOBEC3B vs chr22: 24 Mb-26 Mb)
0
1
2
3
Hub deletion
Wild-type
F
APOBEC3B
p-value=0.012
Relative Expressions
0.45
0.40
0.35
0.30
0.25
0.20
State1
State2

1.8 Supplementary Figures

Figure S1.1. Hub characterization. (**A**) The effective diameter of FCN remains almost unchanged with increasing network size (taking GM12878 cells as an example). (**B**) Random removal of nodes did not change, while targeted removal significantly altered the degree distribution (chr1 in GM12878 cells shown as an example). (**C**) Neighbours of a node are those directly linked to it in at least one cell line. In a particular cell line, the interacting neighbours of a node are called its partners, and neighbours not interacting with the node in this cell line are called non-partners. (**D**) Removal of a node with or without disrupting the contacts between its neighbours. Removal with disruption and removal without disruption refer to removing a node disrupting and not disrupting contacts between its neighbours respectively. Removal with disruption decreases the network size more significantly than removal without disruption. (**E**) The distribution of Partner Linking Tendency (PLT) for all the nodes in the 7 cell lines. PLT shows the difference in the contact ratio between partners and non-partners for each node in all cell lines. Blue bars, PLT < 1.0; grey bars, PLT > 1.0. This suggests that the removal of most of the nodes will disrupt the contacts between their neighbours. (**F**) The relationship between the number of nodes and effective diameter with targeted removal without disruption. The red dot represents chr9 of K562 cells, whose size is much smaller than the simulated removal without disruption. (**G-J**) Comparison of epigenomic signals between hubs and non-hubs, including H3K27me3 (**G**), H3K4me3 (**H**), and H3K36me3 (**I**). The data were obtained from the NIH Roadmap Epigenetics Project. (**J**) Comparison of B-compartment percentage between hubs and non-hubs. The data were calculated from Hi-C.


A
GM12878
Linear Fit
Effect of Diameter
5.5
5.0
4.5
4.0
3.5
Slope=0.200
3.8
4.0
4.2
4.4
4.6
4.8
log$_{10}$(Number of nodes)
C
Cell line 1
Cell line 2
Cell line 3
Cell line 7
Neighbor:
node contacted in at least 1 cell line
Partner:
contacting neighbor in each cell line
Non-partner:
not contacting neighbor in each cell line
B
Original
10% random removal
20% random removal
10% targeted removal
20% targeted removal
GM12878, chr1
Number of nodes
1000
100
10
1
10
100
Degree
D
size=N
Removal
w/o disruption
Removal
w/ disruption
size=N-2
size=N-4
E
Percentage
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-1.2
-0.9
-0.6
-0.3
0.0
0.3
0.6
0.9
1.2
1.5
1.8
2.1
2.4
2.7
3.0
log$_{10}$(Partners Linking Tendency)
F
Targeted removal w/o disruption on chr9, GM12878
chr9, K562
Number of nodes(10^3)
24
22
20
18
16
14
12
10
8
6
4
4
6
8
10
12
14
16
18
Effective Diameter
G
H3K27me3
p value = 0.018
Average peaks per node
0.4
0.3
0.2
0.1
0
Non-hub
Hub
H
H3K4me3
p value = 0.034
Average peaks per node
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0
Non-hub
Hub
I
H3K36me3
p value = 0.005
Average peaks per node
0.6
0.5
0.4
0.3
0.2
0.1
0
Non-hub
Hub
J
B compartment
p value = 0.007
Percentage (%)
100
90
80
70
60
50
40
Non-hub
Hub
GM12878
HMEC
HUVEC
IMR90
K562
NHEK

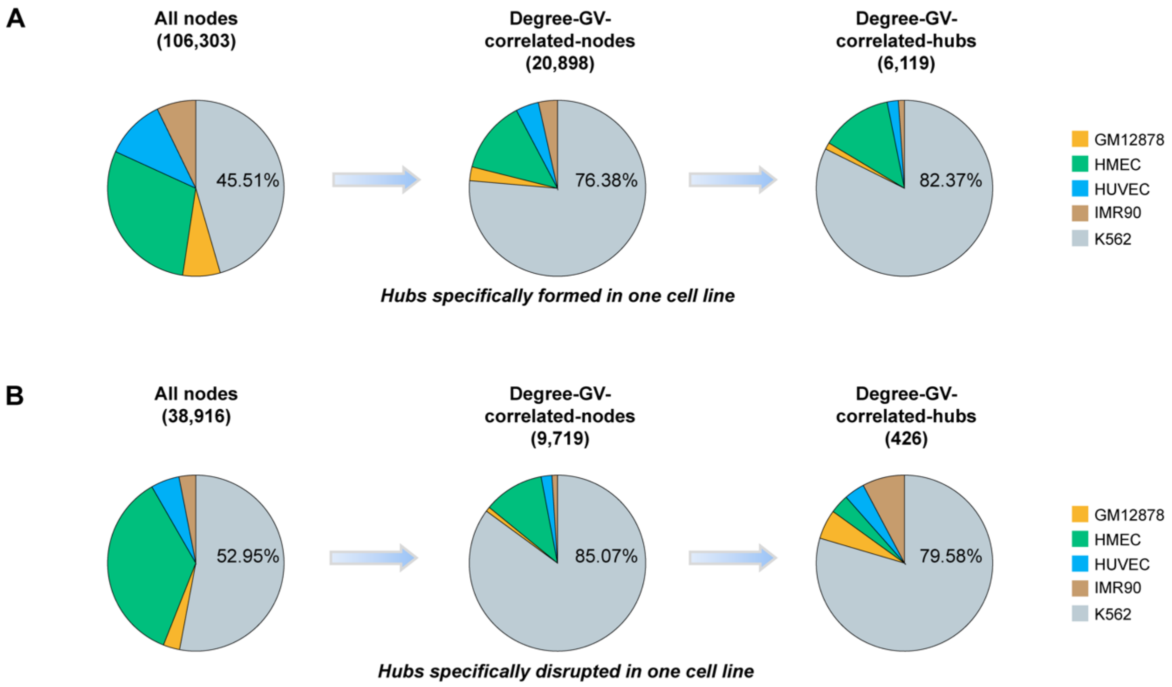


Figure S1.2. Formation or disruption of cell type-specific hubs. (**A**) The distribution of one cell line specificities, i.e., cell type-specific interaction formation, in all nodes, degree-GV-correlated nodes and degree-GV-correlated hubs (**Table S1.2**). (**B**) The distribution of four cell line specificities, i.e., cell type-specific interaction disruption, in all nodes, degree-GV-correlated-nodes and degree-GV-correlated-hubs.

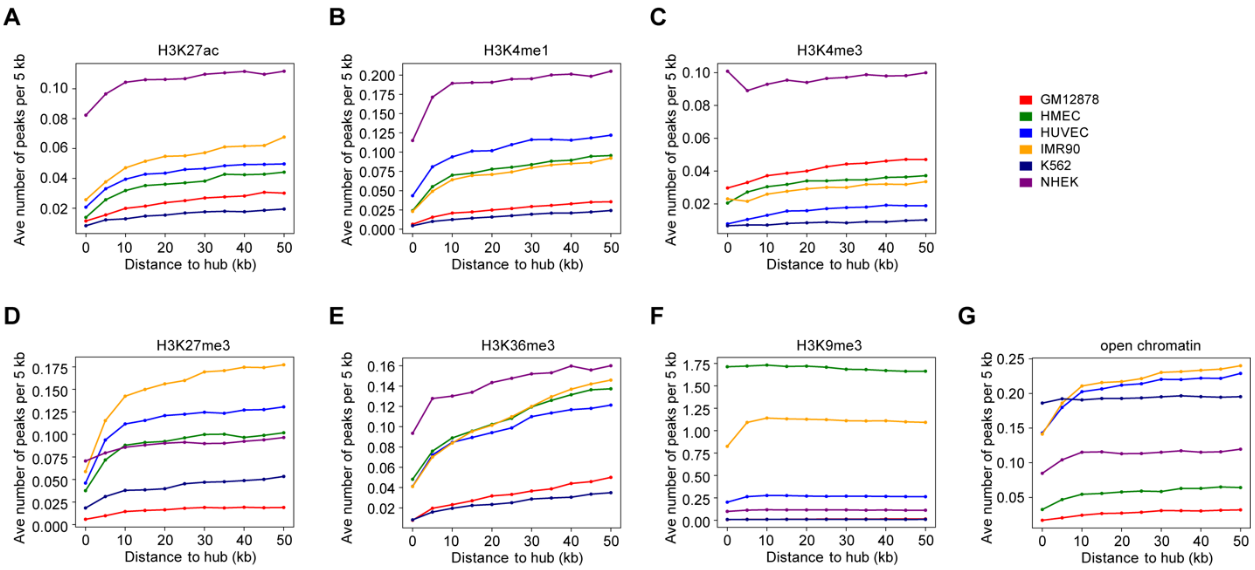
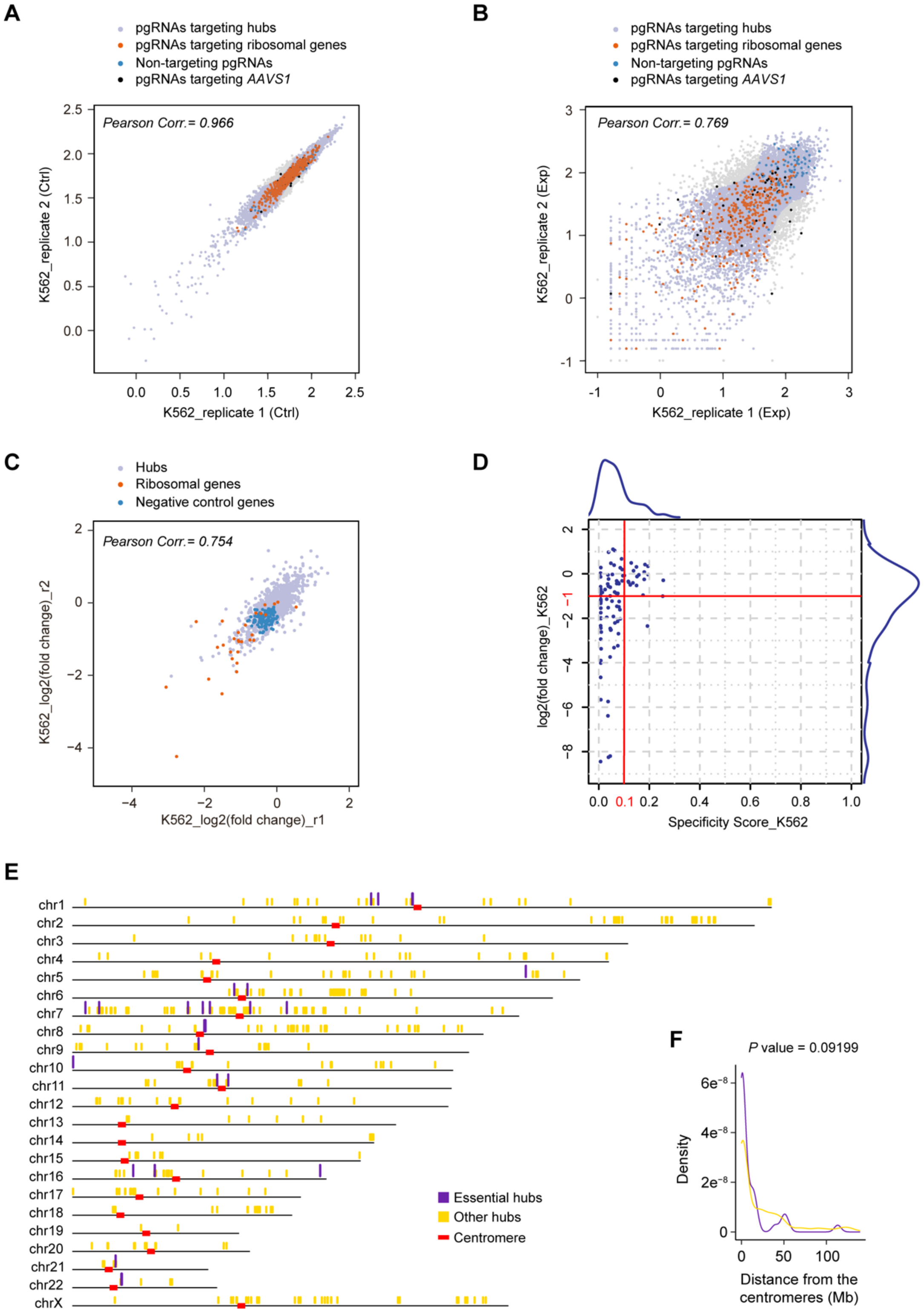


Figure S1.3. Histone modifications and open chromatin signals around the hub regions (0-50 kbp upstream and downstream) in different cell types. Average peak numbers in the 5-kb bins of H3K27ac (**A**), H3K4me1 (**B**), H3K4me3 (**C**), H3K27me3 (**D**), H3K36me3 (**E**), H3K9me3 peaks (**F**), ATAC-seq (**G**).

Figure S1.4. The correlations between replicates in the functional screening for essential hubs in the K562 cell line and hub distribution in each chromosome. (**A** and **B**) Scatter plots of normalized pgRNA read counts of the hub libraries including day 0 control samples (Ctrl) (**A**) and day 30 experimental samples (Exp) (**B**) in the K562 cell line. The Pearson correlation coefficients (Pearson corr.) between the two biologically independent replicates of each sample are also presented. (**C**) Scatter plots of pgRNA fold changes between the two biological replicates in K562 cells. (**D**) The distribution of pgRNAs targeting the AAVS1 locus with different log2(fold change) and specificity scores. (**E**) Hub distribution in each chromosome. The lines above the chromosomes represent the locations of hubs, among which, the purple vertical lines indicate the essential hubs and the yellow lines indicate other hubs. The red horizontal blocks in the chromosomes represent the locations of the centromeres. (**F**) The density of essential hubs and other hubs with different distances (in Mb) from the centromeres. The purple line indicates the density of essential hubs and the yellow line indicates the density of other hubs.


A
pgRNAs targeting hubs
pgRNAs targeting ribosomal genes
Non-targeting pgRNAs
pgRNAs targeting *AAVS1*
Pearson Corr.= 0.966
K562_replicate 2 (Ctrl)
2.5
2.0
1.5
1.0
0.5
0.0
0.0 0.5 1.0 1.5 2 2.5
K562_replicate 1 (Ctrl)
B
pgRNAs targeting hubs
pgRNAs targeting ribosomal genes
Non-targeting pgRNAs
pgRNAs targeting *AAVS1*
Pearson Corr.= 0.769
K562_replicate 2 (Exp)
3
2
1
0
-1
-1 0 1 2 3
K562_replicate 1 (Exp)
C
Hubs
Ribosomal genes
Negative control genes
Pearson Corr.= 0.754
K562_log2(fold change)_r2
2
0
−2
−4
−4 −2 0 2
K562_log2(fold change)_r1
D
log2(fold change)_K562
2
0
−1
−2
−4
−6
−8
0.0 0.1 0.2 0.4 0.6 0.8 1.0
Specificity Score_K562
E
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
Essential hubs
Other hubs
Centromere
F
P value = 0.09199
Density
6e−8
4e−8
2e−8
0
0 50 100
Distance from the centromeres (Mb)

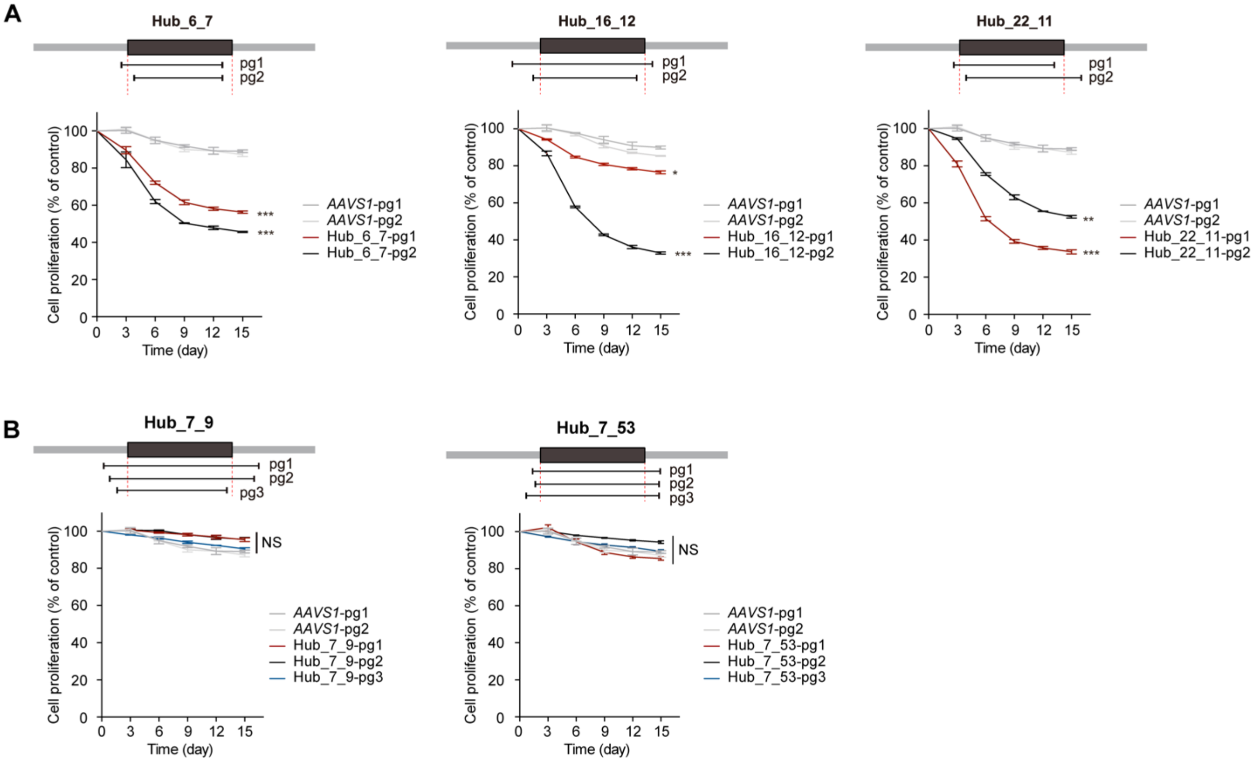


Figure S1.5. Validation of essential hubs in K562 cells through fragment deletion. (**A**) Validation of three additional hubs that were confirmed to be essential for cell viability in K562 cells. (**B**) Validation of two selected hubs that showed no significant effect on cell viability in K562 cells. The rule of designing pgRNAs for each hub and the method for determining their effects on cell growth or proliferation were the same as described in Figure 1.2D-E. Asterisks represent p-value compared with AAVS1_pg1 at day 15, which were calculated by two-tailed Student's t-test and adjusted for multiple comparisons by the Benjamini-Hochberg method. Data are presented as the mean ± s.d. (n = 3 biologically independent samples). *p-value < 0.05; **p-value < 0.01; ***p-value < 0.001; ****p-value < 0.0001; NS, not significant.

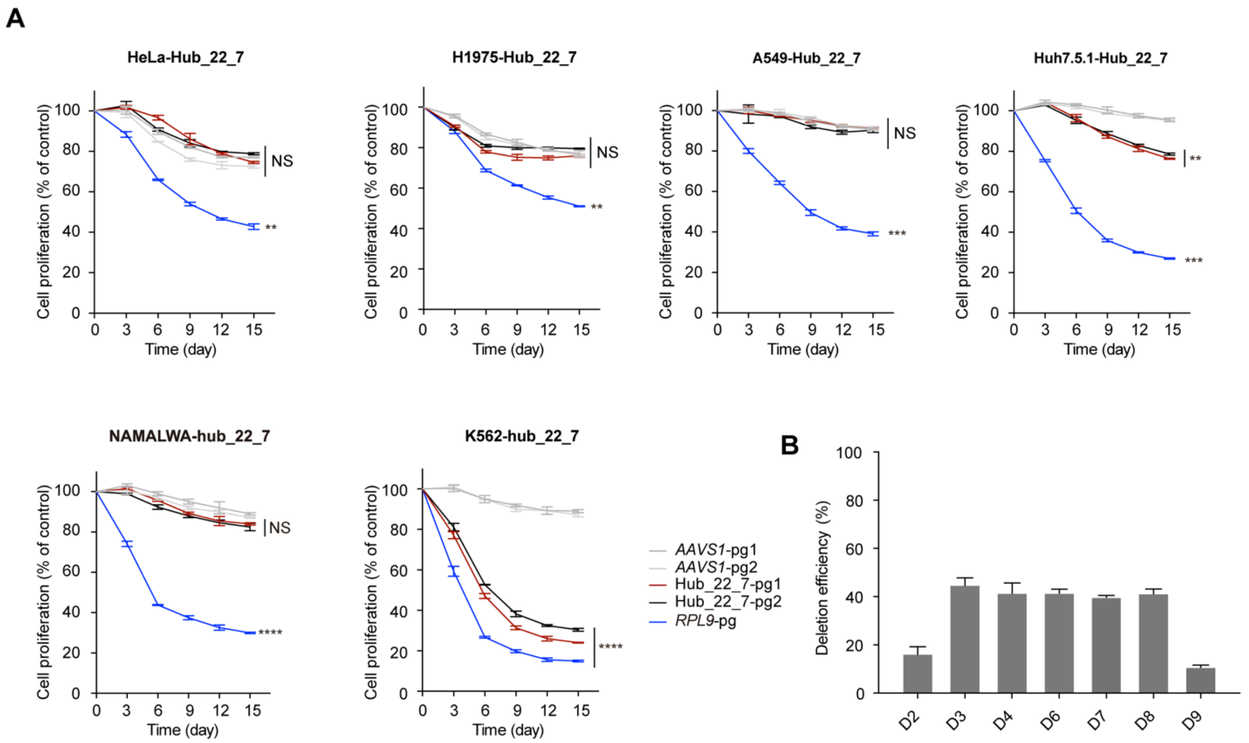


Figure S1.6. Validation of hub_22_7 in multiple cancer cell lines through fragment deletion. (**A**) The performance of pgRNAs targeting hub_22_7, the AAVS1 locus and the essential protein-coding gene RPL19 in various cell lines (HeLa, H1975, A549, Huh7.5.1, NAMALWA and K562). The method for determining their effects on cell growth or proliferation was the same as that described in Figure 1.2D-E. Data are presented as the mean ± s.d. (n = 3 biologically independent samples). *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001; NS, not significant. (**B**) Deletion efficiency of the pgRNA targeting hub_22_7 in K562 cells on different days post lentiviral transfection.

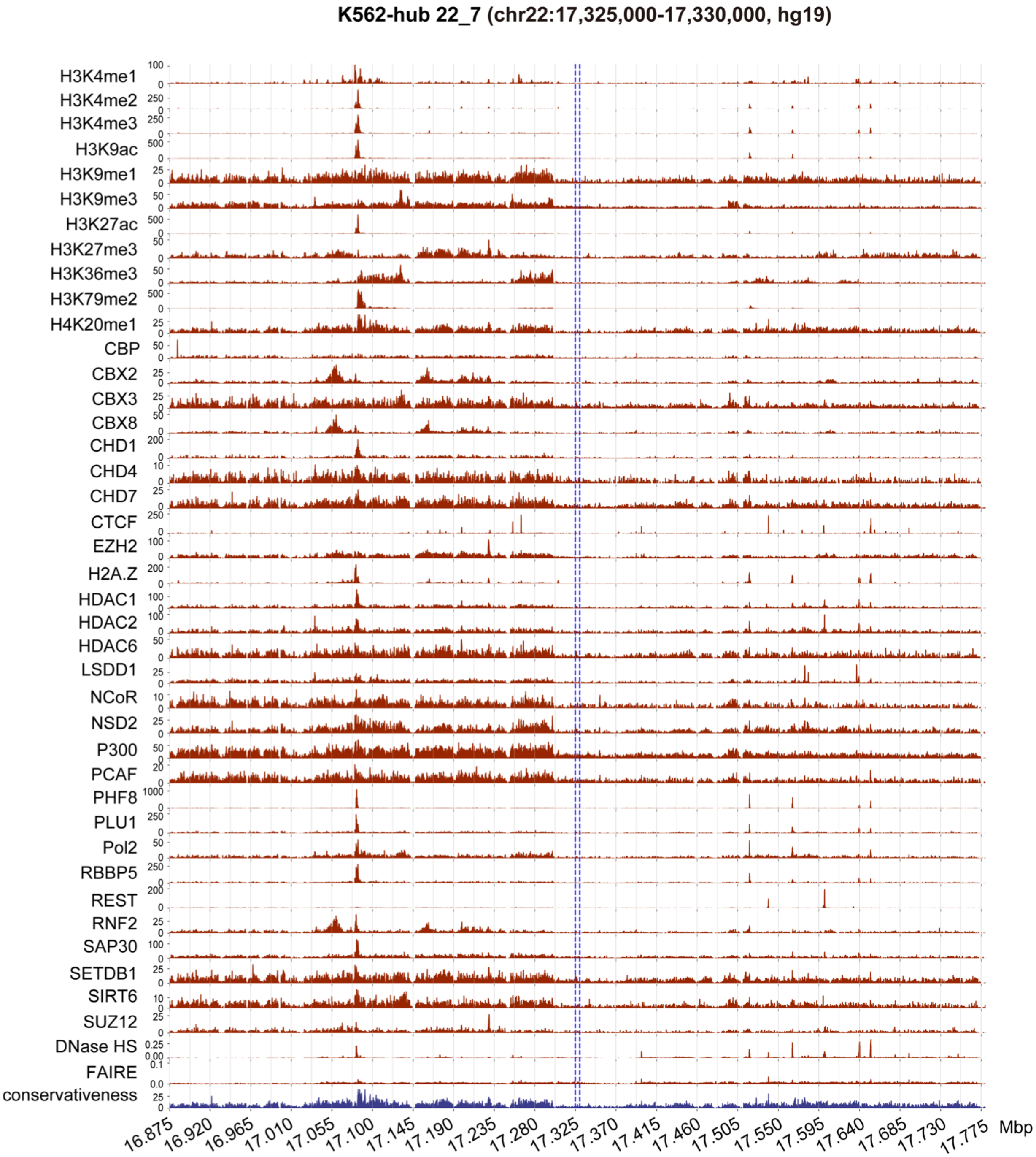


Figure S1.7. Comprehensive genomic and epigenomic data of the essential hub_22_7 (chr22:17,325,000-17,330,000) in K562 cells.

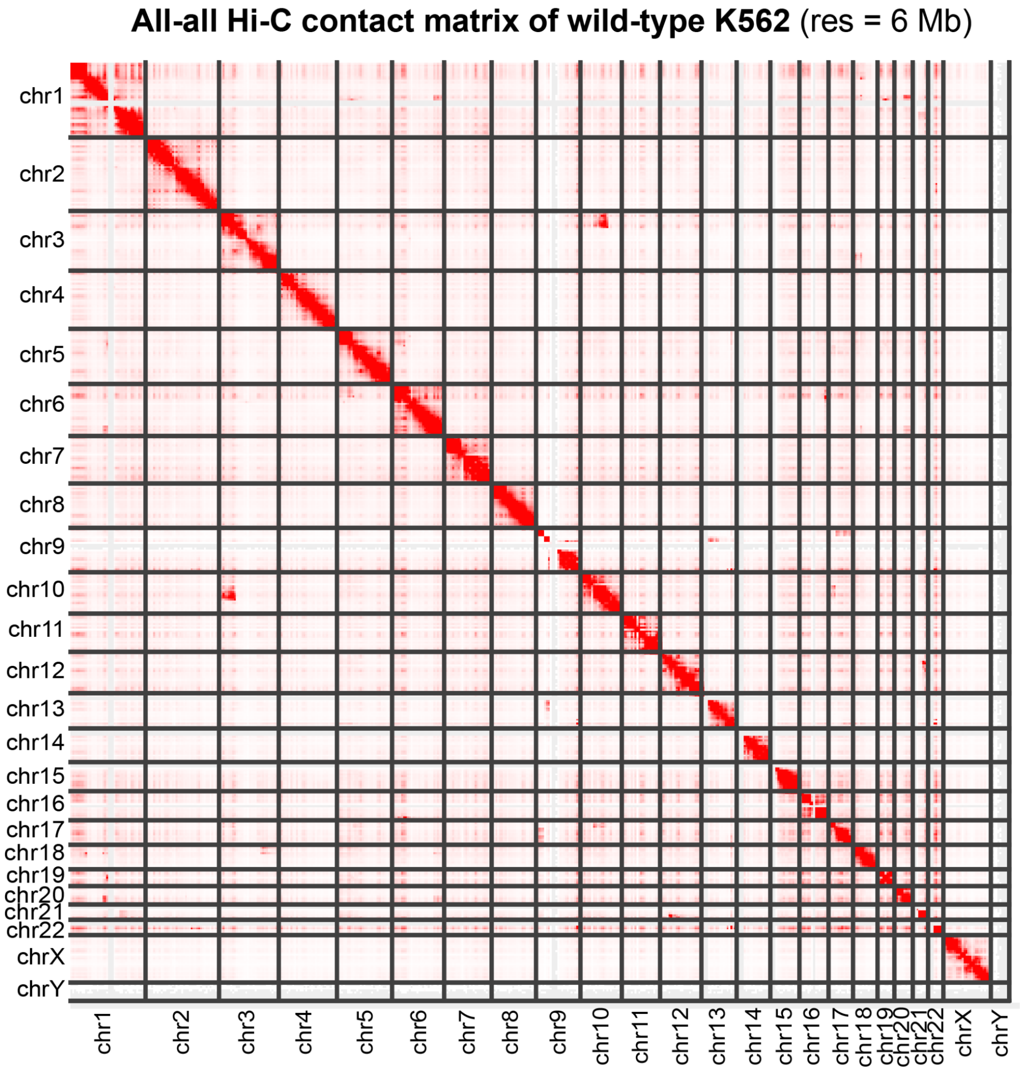
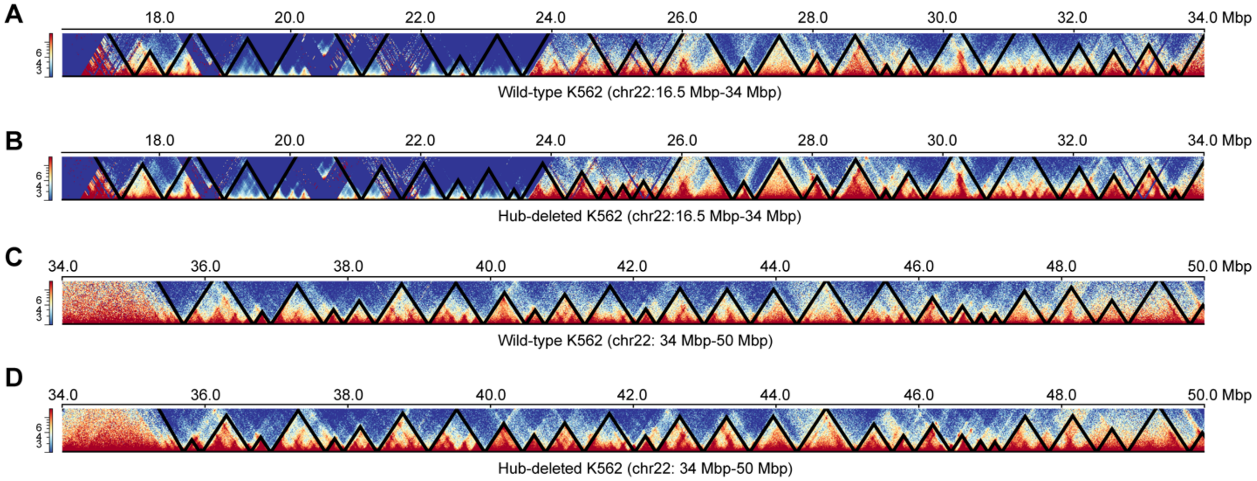


Figure S1.8. Interchromosomal Hi-C contacts in wild-type K562 cells.



Wild-type K562 (chr22:16.5 Mbp-34 Mbp)

Hub-deleted K562 (chr22:16.5 Mbp-34 Mbp)

Wild-type K562 (chr22: 34 Mbp-50 Mbp)

Hub-deleted K562 (chr22: 34 Mbp-50 Mbp)

Figure S9. TADs in the wild-type and hub-deleted K562 cells. (**A** and **B**) TADs in chr22: 16.5 Mbp-34 Mbp. (**C** and **D**) TADs in chr22: 34 Mbp-50 Mbp.

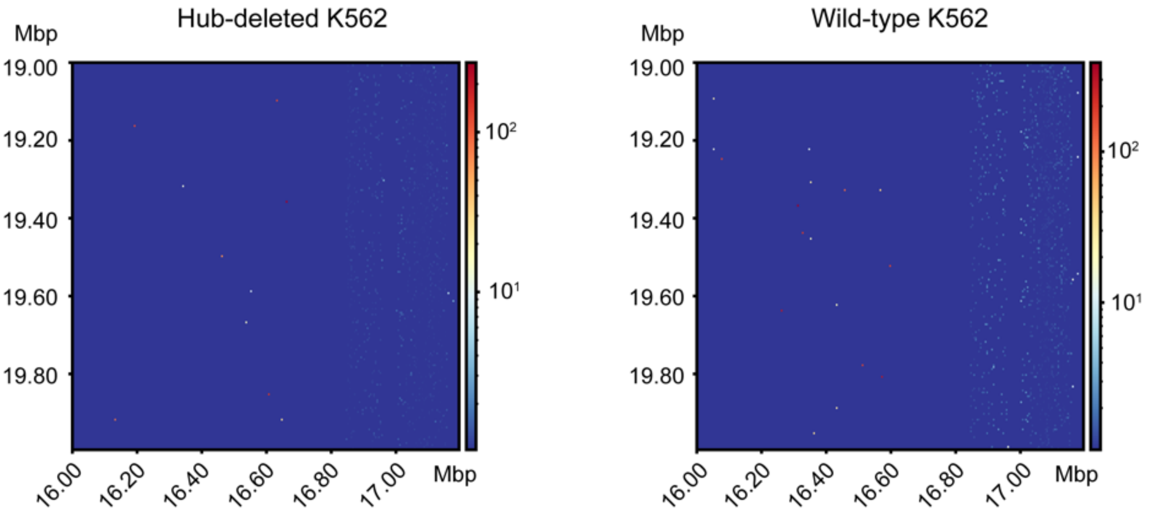


Figure S1.10. Hi-C contacts between chr22: 16 M-17 M and chr22: 19 M-20 M in the hub-deleted (left) and wild-type (right) K562 cells.

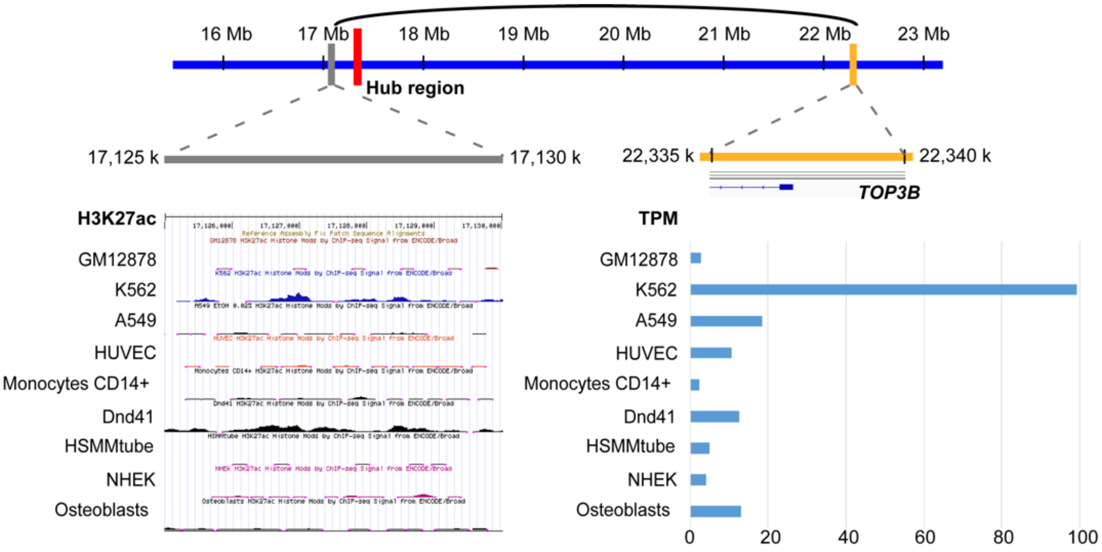


Figure S1.11. The expression levels of *TOP3B* in different cell lines/tissues and H3K27ac signals of the enhancer regions interacting with the *TOP3B* promoter.

1.9 Supplementary Tables

Table S1.1 Number of hubs selected in each cell line.

Cell Line	Number of hubs	Number of total nodes	Ratio
GM12878	36,371	545,098	0.067
HMEC	37,452	465,032	0.081
HUVEC	33,817	483,049	0.07
IMR90	32,183	495,532	0.065
NHEK	26,437	446,942	0.059
K562	44,295	480,934	0.092
KBM7	37,453	488,283	0.077

Table S1.2 The percentage of cell line-specific nodes in all nodes, degree-GV-correlated nodes and degree-GV-correlated hubs.

Specificity	All nodes	degree-GV-correlated-nodes	degree-GV-correlated-hubs
Total	563,566	54,177	8,765
No specificity	215,893	9,766	910
GM12878	8,482	615	99
HMEC	45,196	3,871	824
HUVEC	13,712	1,074	142
IMR90	8,841	828	101
K562	68,988	24,229	5,379
GM12878, HMEC	7,572	370	58
GM12878, HUVEC	8,232	398	45
GM12878, IMR90	31,295	1,979	106
GM12878, K562	16,954	1,269	256
HMEC, HUVEC	16,867	1,070	116
HMEC, IMR90	10,858	600	102
HMEC, K562	75,344	5,743	372
HUVEC, IMR90	13,915	678	85
HUVEC, K562	14,296	1,060	90
IMR90, K562	7,121	627	80

Table S1.3 Additional Hi-C data with 20-kb resolution.

Samples	Experiment ID	Status
transverse colon, adult male (54 years)	ENCSR079IDJ	Normal
gastrocnemius medialis, adult female (51 years)	ENCSR974ADY	Normal
transverse colon, adult female (53 years)	ENCSR504OTV	Normal
gastrocnemius medialis, adult female (53 years)	ENCSR089CCK	Normal
gastrocnemius medialis, adult male (37 years)	ENCSR125MJP	Normal
transverse colon, adult male (37 years)	ENCSR295BDK	Normal
gastrocnemius medialis, adult male (54 years)	ENCSR479USL	Normal
transverse colon, adult female (51 years)	ENCSR424WMG	Normal
HeLa-S3	ENCSR693GXU	Cancer
HepG2	ENCSR194SRI	Cancer

Table S1.4 The chromosome distributions for differentially expressed genes.

chromosome	DEG.number
chr1	631
chr10	246
chr11	351
chr12	345
chr13	111
chr14	207
chr15	179
chr16	291
chr17	365
chr18	78
chr19	471
chr2	417
chr20	168
chr21	59
chr22	142
chr3	336
chr4	210
chr5	266
chr6	325
chr7	270
chr8	188
chr9	230
chrX	245

Table S1.5 chromHMM states distribution in K562 essential hubs.

chromHMM states	description	Percentage(%)
State1	Active TSS	0
State2	Flanking TSS	0.229
State3	Flanking TSS Upstream	0
State4	Flanking TSS Downstream	0
State5	Strong transcription	0
State6	Weak transcription	6.057
State7	Genic enhancer1	0
State8	Genic enhancer2	0
State9	Active Enhancer 1	0.229
State10	Active Enhancer 2	0.229
State11	Weak Enhancer	2.4
State12	ZNF genes & repeats	0
State13	Heterochromatin	0
State14	Bivalent/Poised TSS	0
State15	Bivalent Enhancer	0
State16	Repressed PolyComb	0
State17	Weak Repressed PolyComb	8.571
State18	Quiescent/Low	82.286

Table S1.6 Summary of WGS analysis for hub_22_7 deletion in K562 cells.

CellLine	K562-wt	K562-del
Genome	hg19	hg19
Genome Coverage	42x	27x
Mapping Rate	94.39%	80.19%
SNV+INDEL(Compared.to.hg19)	4.7M	4.1M
SNV+INDEL(Compared.to.hg19) confirmation ratio	87.40%	
Variations in Putative Off-Target loci (compared to wt)	N/A	0

1.10 References

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

2. E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Blüthgen, J. Dekker, E. Heard, Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*. **485**, 381–385 (2012).

3. D. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, S. Mundlos, Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*. **161**, 1012–1025 (2015).

4. E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, J. Dekker, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. **326**, 289–293 (2009).

5. T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, P. Fraser, Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*. **502**, 59–64 (2013).

6. J. Dekker, M. A. Marti-Renom, L. A. Mirny, Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.* **14**, 390–403 (2013).

7. G. D. Rose, P. J. Fleming, J. R. Banavar, A. Maritan, A backbone-based theory of protein folding. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 16623–16633 (2006).

8. S. J. Liu, M. A. Horlbeck, S. W. Cho, H. S. Birk, M. Malatesta, D. He, F. J. Attenello, J. E. Villalta, M. Y. Cho, Y. Chen, M. A. Mandegar, M. P. Olvera, L. A. Gilbert, B. R. Conklin, H. Y. Chang, J. S. Weissman, D. A. Lim, CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science*. **355** (2017), doi:10.1126/science.aah7111.

9. Y. Liu, Z. Cao, Y. Wang, Y. Guo, P. Xu, P. Yuan, Z. Liu, Y. He, W. Wei, Genome-wide screening for functional long noncoding RNAs in human cells by Cas9 targeting of splice sites. *Nat. Biotechnol.* (2018), doi:10.1038/nbt.4283.

10. D. R. Simeonov, B. G. Gowen, M. Boontanrart, T. L. Roth, J. D. Gagnon, M. R. Mumbach, A. T. Satpathy, Y. Lee, N. L. Bray, A. Y. Chan, D. S. Lituiev, M. L. Nguyen,

R. E. Gate, M. Subramaniam, Z. Li, J. M. Woo, T. Mitros, G. J. Ray, G. L. Curie, N. Naddaf, J. S. Chu, H. Ma, E. Boyer, F. Van Gool, H. Huang, R. Liu, V. R. Tobin, K. Schumann, M. J. Daly, K. K. Farh, K. M. Ansel, C. J. Ye, W. J. Greenleaf, M. S. Anderson, J. A. Bluestone, H. Y. Chang, J. E. Corn, A. Marson, Discovery of stimulation-responsive immune enhancers with CRISPR activation. *Nature*. **549**, 111–115 (2017).

11. C. P. Fulco, M. Munschauer, R. Anyoha, G. Munson, S. R. Grossman, E. M. Perez, M. Kane, B. Cleary, E. S. Lander, J. M. Engreitz, Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science*. **354**, 769–773 (2016).

12. Y. Diao, R. Fang, B. Li, Z. Meng, J. Yu, Y. Qiu, K. C. Lin, H. Huang, T. Liu, R. J. Marina, I. Jung, Y. Shen, K.-L. Guan, B. Ren, A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods*. **14**, 629–635 (2017).

13. S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, E. L. Aiden, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. **159**, 1665–1680 (2014).

14. B. Ding, L. Zheng, D. Medovoy, W. Wang, Targeted mutations on 3D hub loci alter spatial interaction environment. *bioRxiv* (2015), p. 030999.

15. S. Zhu, W. Li, J. Liu, C.-H. Chen, Q. Liao, P. Xu, H. Xu, T. Xiao, Z. Cao, J. Peng, P. Yuan, M. Brown, X. S. Liu, W. Wei, Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. *Nat. Biotechnol.* **34**, 1279–1286 (2016).

16. A. R. Perez, Y. Pritykin, J. A. Vidigal, S. Chhangawala, L. Zamparo, C. S. Leslie, A. Ventura, GuideScan software for improved single and paired CRISPR guide RNA design. *Nat. Biotechnol.* **35**, 347–349 (2017).

17. C. Smith, A. Gore, W. Yan, L. Abalde-Atristain, Z. Li, C. He, Y. Wang, R. A. Brodsky, K. Zhang, L. Cheng, Z. Ye, Whole-genome sequencing analysis reveals high specificity of CRISPR/Cas9 and TALEN-based genome editing in human iPSCs. *Cell Stem Cell*. **15**, 12–13 (2014).

18. N. C. Durand, M. S. Shamim, I. Machol, S. S. P. Rao, M. H. Huntley, E. S. Lander, E. L. Aiden, Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst*. **3**, 95–98 (2016).

19. F. Ramírez, V. Bhardwaj, L. Arrigoni, K. C. Lam, B. A. Grüning, J. Villaveces, B. Habermann, A. Akhtar, T. Manke, High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, 189 (2018).

20. J. Wolff, L. Rabbani, R. Gilsbach, G. Richard, T. Manke, R. Backofen, B. A. Grüning,

Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **48**, W177–W184 (2020).

21. E. Crane, Q. Bian, R. P. McCord, B. R. Lajoie, B. S. Wheeler, E. J. Ralston, S. Uzawa, J. Dekker, B. J. Meyer, Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*. **523**, 240–244 (2015).

22. K. Y. Yip, C. Cheng, N. Bhardwaj, J. B. Brown, J. Leng, A. Kundaje, J. Rozowsky, E. Birney, P. Bickel, M. Snyder, M. Gerstein, Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13**, R48 (2012).

23. R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, L. Chen, Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* **30**, 90–98 (2011).

24. M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, S. L. Salzberg, Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).

25. D. Kim, B. Langmead, S. L. Salzberg, HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*. **12**, 357–360 (2015).

26. M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, S. L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

27. E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, S. A. McCarroll, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. **161**, 1202–1214 (2015).

28. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T. R. Gingeras, STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. **29**, 15–21 (2013).

29. C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, J. L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

30. X. Qiu, A. Hill, J. Packer, D. Lin, Y.-A. Ma, C. Trapnell, Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods*. **14**, 309–315 (2017).

31. M. A. Horlbeck, L. A. Gilbert, J. E. Villalta, B. Adamson, R. A. Pak, Y. Chen, A. P. Fields, C. Y. Park, J. E. Corn, M. Kampmann, J. S. Weissman, Compact and highly

active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife*. **5** (2016), doi:10.7554/eLife.19760.

32. R. Albert, H. Jeong, A. L. Barabasi, Error and attack tolerance of complex networks. *Nature*. **406**, 378–382 (2000).

33. A. J. Aguirre, R. M. Meyers, B. A. Weir, F. Vazquez, C.-Z. Zhang, U. Ben-David, A. Cook, G. Ha, W. F. Harrington, M. B. Doshi, M. Kost-Alimova, S. Gill, H. Xu, L. D. Ali, G. Jiang, S. Pantel, Y. Lee, A. Goodale, A. D. Cherniack, C. Oh, G. Kryukov, G. S. Cowley, L. A. Garraway, K. Stegmaier, C. W. Roberts, T. R. Golub, M. Meyerson, D. E. Root, A. Tsherniak, W. C. Hahn, Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discov.* **6**, 914–929 (2016).

34. D. M. Munoz, P. J. Cassiani, L. Li, E. Billy, J. M. Korn, M. D. Jones, J. Golji, D. A. Ruddy, K. Yu, G. McAllister, A. DeWeck, D. Abramowski, J. Wan, M. D. Shirley, S. Y. Neshat, D. Rakiec, R. de Beaumont, O. Weber, A. Kauffmann, E. R. McDonald 3rd, N. Keen, F. Hofmann, W. R. Sellers, T. Schmelzle, F. Stegmeier, M. R. Schlabach, CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities but Generate False-Positive Hits for Highly Amplified Genomic Regions. *Cancer Discov.* **6**, 900–913 (2016).

35. D. W. Morgens, M. Wainberg, E. A. Boyle, O. Ursu, C. L. Araya, C. K. Tsui, M. S. Haney, G. T. Hess, K. Han, E. E. Jeng, A. Li, M. P. Snyder, W. J. Greenleaf, A. Kundaje, M. C. Bassik, Genome-scale measurement of off-target activity using Cas9 toxicity in high-throughput screens. *Nat. Commun.* **8**, 15178 (2017).

36. J. Tycko, M. Wainberg, G. K. Marinov, O. Ursu, G. T. Hess, B. K. Ego, Aradhana, A. Li, A. Truong, A. E. Trevino, K. Spees, D. Yao, I. M. Kaplow, P. G. Greenside, D. W. Morgens, D. H. Phanstiel, M. P. Snyder, L. Bintu, W. J. Greenleaf, A. Kundaje, M. C. Bassik, Mitigation of off-target toxicity in CRISPR-Cas9 screens for essential non-coding elements. *Nat. Commun.* **10**, 4063 (2019).

37. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature*. **489**, 57–74 (2012).

38. S. Bae, J. Park, J.-S. Kim, Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics*. **30**, 1473–1475 (2014).

39. T. Wang, K. Birsoy, N. W. Hughes, K. M. Krupczak, Y. Post, J. J. Wei, E. S. Lander, D. M. Sabatini, Identification and characterization of essential genes in the human genome. *Science*. **350** (2015), pp. 1096–1101.

40. T. Zhang, M. Wallis, V. Petrovic, J. Challis, P. Kalitsis, D. F. Hudson, Loss of TOP3B leads to increased R-loop formation and genome instability. *Open Biol.* **9**, 190222 (2019).

41. C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, G. Bejerano, GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).

42. S. Venkatesan, R. Rosenthal, N. Kanu, N. McGranahan, J. Bartek, S. A. Quezada, J. Hare, R. S. Harris, C. Swanton, Perspective: APOBEC mutagenesis in drug resistance and immune escape in HIV and cancer evolution. *Ann. Oncol.* **29**, 563–572 (2018).

43. M. E. Olson, R. S. Harris, D. A. Harki, APOBEC Enzymes as Targets for Virus and Cancer Therapy. *Cell Chem Biol*. **25**, 36–49 (2018).

44. C. Swanton, N. McGranahan, G. J. Starrett, R. S. Harris, APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity. *Cancer Discov.* **5**, 704–712 (2015).

45. N. Roper, S. Gao, T. K. Maity, A. R. Banday, X. Zhang, A. Venugopalan, C. M. Cultraro, R. Patidar, S. Sindiri, A.-L. Brown, A. Goncearenco, A. R. Panchenko, R. Biswas, A. Thomas, A. Rajan, C. A. Carter, D. E. Kleiner, S. M. Hewitt, J. Khan, L. Prokunina-Olsson, U. Guha, APOBEC Mutagenesis and Copy-Number Alterations Are Drivers of Proteogenomic Tumor Evolution and Heterogeneity in Metastatic Thoracic Tumors. *Cell Rep.* **26**, 2651–2666.e6 (2019).

46. Y. Zhang, R. Delahanty, X. Guo, W. Zheng, J. Long, Integrative genomic analysis reveals functional diversification of APOBEC gene family in breast cancer. *Hum. Genomics*. **9**, 34 (2015).

47. F. M. Behan, F. Iorio, G. Picco, E. Gonçalves, C. M. Beaver, G. Migliardi, R. Santos, Y. Rao, F. Sassi, M. Pinnelli, R. Ansari, S. Harper, D. A. Jackson, R. McRae, R. Pooley, P. Wilkinson, D. van der Meer, D. Dow, C. Buser-Doepner, A. Bertotti, L. Trusolino, E. A. Stronach, J. Saez-Rodriguez, K. Yusa, M. J. Garnett, Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature*. **568**, 511–516 (2019).

Chapter 2. Regulatory elements can be essential for maintaining broad chromatin organization and cell viability

2.1 Abstract

Increasing evidence shows that promoters and enhancers could be related to 3D chromatin structure, thus affecting cellular functions. Except for their roles in forming canonical chromatin loops, promoters and enhancers have not been well studied regarding the maintenance of broad chromatin organization. Here, we focused on the active promoters/enhancers predicted to form many 3D contacts with other active promoters/enhancers (referred to as hotspots) and identified dozens of loci essential for cell growth and survival through CRISPR screening. We found that the deletion of an essential hotspot could lead to changes in broad chromatin organization and the expression of distal genes. We showed that the essentiality of hotspots does not result from their association with individual genes that are essential for cell viability but rather from their association with multiple dysregulated non-essential genes to synergistically impact cell fitness.

2.2 Introduction

Promoters and enhancers are regulatory elements that control gene expression in response to intra- and extracellular signals(*1–4*). In many cases, activated enhancers appear to engage in direct physical contact with their nearby promoters (*5–7*). However, there are also enhancers whose interacting promoters are distally located in the linear genome(*2*, *8–13*), and they are brought to spatial proximity by such as chromatin looping (*14–17*), protein oligomerization(*2*, *18*, *19*) or Pol II tracking along chromatin(*2*, *20*). These observations on long-range enhancer-promoter interactions highlight the important impact of 3D chromatin structure on the activities of these regulatory elements.

Recently, an increasing number of investigations on chromosome spatial structures have indicated that enhancer-promoter interactions play pivotal roles in forming specific 3D structures. Imaging analyses showed that transcription factors (TFs) and polymerases are not evenly distributed in the nucleus but rather concentrated in certain regions to form spatial clusters; these regions are associated with high transcriptional activities and a more compact chromatin structure(*21–23*). Transcription could also affect the 3D topology, and a recent study reported that transcription elongation can be critical for chromatin organization(*24*). These studies suggested a mutual relationship between promoter/enhancer activity and 3D chromatin structure. The discovery that these regulatory elements positioning in such spatial clusters with active transcription could contribute to maintaining broad chromatin structures therefore has become an emerging question.

Given these observations and evidence, we hypothesize the following: if a promoter or enhancer is positioned in 3D space in a manner pivotal for the maintenance and stabilization of the surrounding 3D chromatin structure, perturbing such elements may impact chromatin organization beyond their specific direct enhancer-promoter interaction; namely, perturbation of such promoter or enhancer would significantly alter broad chromatin organization and disrupted regulation of multiple direct and indirect target genes simultaneously.

To investigate our hypothesis, we started with active promoters/enhancers that likely form many 3D spatial contacts with other active promoters or enhancers, referred to as hotspots hereinafter. Using our previously published algorithm EpiTensor(*25*), we identified hotspots at a high resolution of 200 bp based on their covariation of epigenetic marks across cell types. Interestingly, cancer-specific genetic variations (we focused on single nucleotide variations) were discovered to have a significantly higher chance of residing in hotspot regions. Through high-throughput CRISPR-Cas9 library screening of hotspots by targeted deletion, dozens of noncoding loci were identified as essential for cell growth and survival, referred to as essential hotspots. We then evaluated the impact of the 3D chromatin structure by Hi-C technology and transcriptome patterns using single cell RNA-seq upon knocking out hotspots. Importantly, we found that deleting a hotspot enhancer could alter broad chromatin organization beyond chromatin looping, which has not been reported before. Deletion of the hotspot would further impact the expression levels of multiple genes concurrently, which exhibited synergistic effects to affect cell fitness.

2.3 Materials and Methods

2.3.1 Predicting high-resolution regulatory element contacts by EpiTensor

Active promoters marked by H3K27ac and H3K4me3 and active enhancers marked by H3K27ac and H3K4me1 were identified in 73 normal and 5 cancer cells/tissues that were available in the Roadmap Epigenomics project(*26*). The 3D contacts between these active promoters/enhancers in each cell/tissue were predicted by EpiTensor(*25*) with an EpiTensor score cut-off ≥ sqrt (25000). In a given cell/tissue sample, these contacts were assembled into a regulatory element interaction network (REIN) in which each node represents a promoter/enhancer and an edge represents a predicted contact.

2.3.2 Identification of sample-specific degree/sample-specific GV

A distribution-based method was used to evaluate the cell type specificities for degree: 1) For each node, we collected the normalized degree in all samples that had epig23enomic data (73 normal and 5 cancer in total); then we calculated the mean and standard deviation for each node across the normal samples, under the assumption that the normalized degrees of normal samples obey a Gaussian distribution; finally, the Z-score for each node in each sample, i.e. the sample-specific degree Z-score, was calculated using the mean and standard deviation. A node was considered to have sample-specific degree if the absolute value of the sample-specific degree Z-score was greater than 1.

We identified sample-specific GVs with a similar method: for each GV, we first calculated mean and standard deviation of B-allele frequency in all normal samples (45 in total); then, Z-score for each GV in each sample, i.e. the sample-specific allele frequency Z-score, was calculated using the mean and standard deviation. A GV was considered as sample-specific if the absolute value of the sample-specific allele frequency Z-score was greater than 1.

2.3.3 Cell culture

K562, H1975 and NAMALWA cells were cultured in RPMI 1640 medium (Gibco), and 293T, HeLa, A549 and Huh7.5.1 cells were cultured in Dulbecco's modified Eagle's medium (DMEM, Gibco). All media were supplemented with 10% fetal bovine serum (FBS, Biological Industries) and 1% penicillin/streptomycin, and cells were cultured with 5% CO_2 at 37°C.

2.3.4 Design and construction of the CRISPR-Cas9 pgRNA library

To explore the cellular function of hotspots, we selected 751 hotspots identified in the K562 cell line. For each hotspot, the designed sgRNAs targeted 100-bp inside regions and 1-kb outside regions flanking the two boundaries of hotspot loci. If there were not enough sgRNAs satisfying the following design rules, sgRNAs were searched among the 5-kb outside regions flanking each boundary. All the PAM motifs in the targeting regions were scanned to identify available sgRNA targeting sites. All the selected sgRNAs are located in noncoding regions and satisfy all the following conditions: (1) the targeting sequence is unique for the intended locus; (2) the sgRNA contains at least 2 mismatches to any other locus in the human genome; and (3) the GC content of the sgRNA ranges from 20% to 80%. We enumerated all possible pgRNAs from the selected sgRNAs and then retained those satisfying these conditions: (1) the two sgRNAs respectively targeted 100-bp inside regions and 1-kb (or 5-kb) outside regions flanking each hotspot boundary; (2) the deletion regions should not overlap with any promoter or exonic region of protein-coding genes; and (3) the sgRNA targeting sites are at least 30 bp away from the exon-intron boundary of protein-coding genes. The gRNA pairs were designed with one unique gRNA serving as a decoding barcode, and up to 20 pgRNAs were designed for each locus.

Finally, 14,399 pairs of gRNAs targeting 751 hotspots were generated for the hotspot deletion library together with 473 pgRNAs targeting the promoter regions (5 kb upstream of the transcription start site) and the first exon of 29 ribosomal genes (serving as positive controls) and 100 pgRNAs targeting the *AAVS1* locus and 100 non-targeting pgRNAs from a previous library(*27*) (serving as negative controls). According to the two-step cloning method(*27*), 128-nt oligonucleotides containing pgRNA coding sequences were synthesized (Agilent Technologies, Inc.), cloned into a lentiviral expression vector harbouring an EGFP selection marker (with a minimum representation of 150 transformed colonies per pgRNA in each cloning step) and further packaged as previously described(*27*).

2.3.5 CRISPR-Cas9 pgRNA library screening

To ensure the infection at 1,000-1,500 cells per pgRNA with an MOI of < 0.3, K562 cells stably expressing Cas9 were seeded in duplicate in T-175 flasks (Corning). Twenty-four hours later, each replicate was infected by the pgRNA library lentiviruses supplemented with 8 μg/ml polybrene. Seventy-two hours post infection, EGFP$^+$ cells were sorted by FACS (Day 0 control group). For each replicate, the initial EGFP$^+$ pool (1500-fold coverage) was isolated for DNA extraction, and the same number of cells as the experimental group was maintained at a minimum coverage of 1,500 cells per pgRNA at each passage for 30 days. Then, cells from each condition with 1500x library coverage were respectively subjected to genomic DNA extraction, PCR amplification of sgRNA-coding sequences and high-throughput sequencing analysis (Illumina HiSeq2500 platform) as previously described(*27*).

2.3.6 Identification of functional hotspots involved in cell growth and proliferation

The raw pgRNA counts were extracted from paired-end sequencing FASTQ files by bash script based on AWK. Since the low reads in the control groups affect the analysis confidence, pgRNAs with raw reads of less than 5 were eliminated from the following analysis. The total counts were further normalized to adjust the sequence depth of each replicate in the control and experimental groups. To further filter noisy pgRNAs, we removed pgRNAs whose quantile difference of two replicates was in either 3% tail of the distribution, and 100 negative control genes were generated by randomly sampling 20 *AAVS1*-targeting pgRNAs with replacement. In each replicate, we calculated the fold change between the experimental and control group for each pgRNA, and the mean fold change of all targeting pgRNAs for each hotspot. Then, the fold changes in the two replicates were averaged for each specific hotspot. In summary, two features for each set of hotspots were calculated: 1) the mean log_2FC ($\log_2$(fold change)) of all pgRNAs in the set, denoted by $FC_{hotspot}$; and 2) the $-log_{10}P_{value}$ of two-sided Mann-Whitney U test of all pgRNAs in the set compared with pgRNAs targeting the *AAVS1* locus, denoted by $P_{hotspot}$. To consider both the fold change and *P* value, we defined a screen score for the hotspots as follows:

$$Screen\ score\ =\ sign(LFC)\ \times\ \left|\left|\frac{LFC\ -\ \mu_{LFC}}{\sigma_{LFC}}\right|\ +LP\right|$$

where LFC is the log_2FC, $\mu_{LFC}$is the mean of the LFC, $\sigma_{LFC}$is the standard deviation of the LFC, and LP is the $-log_{10}P_{value}$. Hotspots with screen scores of less than -2.5 were identified as essential hotspots.

To further avoid the potential issue of cellular toxicity generated from multiple cleavages by some pgRNAs, we retrieved the GuideScan specificity score (a score reflecting the sgRNA cutting specificity) to evaluate each sgRNA(*28*). A specificity score was further assigned for each pgRNA, which was calculated as half of the harmonic mean of the specificity scores of the two sgRNAs. The formula is as follows:

$$pgRNA_{specific\ score}\ =\ \frac{1}{\frac{1}{sgRNA1_{specific\ score}}+\frac{1}{sgRNA2_{specific\ score}}}.$$

From the identified essential hotspots through the above analysis, those targeting pgRNAs were further selected, whose specificity score is > 0.1 and $\log_2$(fold change) is < -1. To further avoid the copy number effects on drop-out screening, the copy number of each hotspot locus in the K562 cell line was analysed based on ENCODE consortium copy number data (https://www.encodeproject.org/files/ENCFF486MJU/). After filtering hotspot loci with copy number amplification, the remaining hits were regarded as essential hotspots.

2.3.7 Individual validation of functional hotspots by cell proliferation assay

For each candidate hotspot without immediate overlap with the promoter or gene body of protein-coding genes, two or three pgRNAs were used for the individual validation, which were selected from the library that were consistently depleted or newly designed. To ensure the targeting specificity of all the selected pgRNAs, we required that the specificity scores are all greater than 0.15 and that the score of at least one pgRNA for each hotspot is more than 0.2. For the newly designed pgRNA, to ensure the cleavage efficiency, we further required that they don't include ≥ 4-bp homopolymer stretches, and their GC contents are between 0.4 and 0.7. We further ensured that each sgRNA targeting site is 400 bp inside and 1 kb outside the two boundaries of the hotspot loci. All the pgRNAs targeting each hotspot locus to be validated were individually cloned into a lentiviral expression vector containing an EGFP selection marker. The cell proliferation assay was performed as previously described(*27*). The experiments lasted for 15 days after the first FACS analysis, and at least 10,000 cells were analysed.

For the hotspots overlapping with the promoter or within the intron of possible essential protein-coding genes, three pgRNAs were selected for subsequent validation. The cDNA of each neighbouring coding gene was cloned into a lentiviral vector containing a puromycin selection marker and individually transduced into K562 cells. Three days after virus infection, the cells with candidate gene overexpression were enriched by puromycin treatment, and then the corresponding pgRNAs targeting the neighbouring hotspot were respectively transduced into these cells as well as into wild-type K562 control cells. The cell proliferation assay was performed as described above.

2.3.8 Hi-C library preparation and data analysis

Hi-C library preparation. The pgRNA Hotspot_10_25-pg2 was delivered into K562 cells via lentiviral infection at an MOI of < 1. EGFP-positive cells were then collected by FACS sorting at day 9 post infection. Before the Hi-C library preparation, the sorted cells were allowed to recover under normal cell culture conditions for 2 h. Finally one million cells were used for Hi-C library preparation by the Arima-HiC kit (Arima Genomics, San Diego) following the manufacturer's instructions. The K562 hotspot_10_25 Hi-C library was sequenced using the Illumina NovaSeq platform.

Hi-C data processing. An in-house pipeline Juicer(*29*) was implemented to process the Hi-C data. Hi-C contact reads were first aligned to hg19 (GRCh37), and the reads were reserved if MAPQ greater than 30. Then, the vanilla coverage (VC) method(*14*) was applied to the Hi-C raw reads. Between the expected VC-normalized reads and the observed VC-normalized reads, we conducted a Poisson distribution fitting. The normalized contacts were considered significant if the p-value is ≤ 0.05. HiCExplorer(*30*, *31*) and HiCPlotter(*32*) were utilized to visualize the processed Hi-C data.

Chromatin loops identification. The HiCCUPS software (https://github.com/aidenlab/juicer/wiki/HiCCUPS) was utilized to call the loops at 10-kb resolution in both the wild-type and hotspot-deleted K562 cells. All the other parameters in HiCCUPS were set to default.

Topological associated domain (TAD) identification. The Insulation Score method was used to call the TAD for the wild-type and hotspot_10_25-deleted cells at 10 kb resolution. TADs were visualized using HiCExplorer.

A/B compartment analysis. We performed the A/B compartment analysis on the wildtype and hotspot_10_25 deleted cells at 50-kb resolution. The eigenvectors for each individual chromosome were extracted from VC-normalized Hi-C reads using the Juicer pipeline(*29*). All the parameters were set to default. To determine the direction of A or B compartments in each chromosome, the K562 Pol II peak file was obtained from ENCODE (https://www.encodeproject.org/). A correlation score between the first eigenvector of each chromosome and the K562 Pol II peak density in 50 kb-sized bins was calculated.

Hi-C comparison. We used HiCRep(*33*, *34*) to calculate the Stratum-adjusted correlation coefficient (SCC) to measure the Hi-C reproducibility. We performed HiCcompare R bioconductor package(*35*) to detect the Hi-C contact differences across all the chromosomes. All the analyses were done at 25-kb resolution.

2.3.9 Evaluation of the potential off-target effects by the CRISPR-Cas9 system through whole genome sequencing (WGS)

K562 cells were infected with the validated pgRNA hotspot_10_25-pg2 at an MOI of < 1. Eight days after lentiviral infection, the pgRNA-infected cells were sorted by FACS, and were further subjected to genomic DNA extraction. The whole genome sequencing (WGS) library was prepared following the manufacturer's instructions and sequenced using the Illumina HiSeq 4000 platform. Using the WGS data, we evaluated the potential off-target effects after targeted deletion of hotspot_10_25.

The K562 (wild-type) WGS data were downloaded as controls from ENCODE with accession code ENCFF313MGL, ENCFF004THU, ENCFF506TKC and ENCFF066GQD. A strict off-target evaluation was conducted according to the whole-genome sequencing approach(*36*, *37*). The putative off-target sites for hotspot_10_25 were output by Cas-OFFinder in the hg19 genome(*36*). To avoid missing any potential off-target locus, we considered two scenarios to detect the potential off-target loci: 1) no more than 4 base mismatches without any bulge mismatch (mismatch ≤ 4, bulge = 0) and 2) no more than 2 base mismatches with no more than 2 bulge mismatches 2 (mismatch ≤ 2, bulge ≤ 2). In total, we examined 746 potential off-target loci. In order to detect the candidate mutations and indels in the K562 wildtype and hotspot_10_25 deleted K562 cells, we performed variant call according to the approaches described in GATK Best Practices (https://gatk.broadinstitute.org/hc/en-us). The sequencing reads were firstly aligned to the human reference genome (hg19) using BWA-0.7.17. Then we used the GATK4 tools MarkDuplicatesSpark (https://gatk.broadinstitute.org/hc/en-us/articles/360037224932-

MarkDuplicatesSpark) to remove the duplicated reads. Finally, the reads were processed via base quality score recalibration with the GATK4 tools. Germline mutations (compared to the hg19 reference genome) were called in both wild-type and hotspot_10_25 deleted K562 cells by GTAK HaplotypeCaller (version 4.1.4.1) with default parameters. SNVs and indels in pgRNA-infected K562 cells compared to wild-type K562 cells were identified via the tools GATK Mutect2 (version 4.1.4.1) with default parameters. These SNVs and indels were further compared with generated putative off-target loci.

For further confirmation, we applied the BCFTOOLS suite (version 1.9, http://www.htslib.org/doc/bcftools.html) to call variants. BCFTOOLS mpileup and call commands with default settings were used to generate raw variants. Then, variants with "%QUAL < 30 || DP < 30" were marked as low-quality variants by the BCFTOOLS filter command and filtered out in addition to the homozygous variants with the feature "GT = 1/1". We also used the BCFTOOLS isec command with parameter "-n -1 -c all" to filter the Mills and 1000G gold standard indels obtained from the GATK resource bundle (https://gatk.broadinstitute.org/hc/en-us/articles/360035890811-Resource-bundle). The putative off-target sites generated by Cas-OFFinder were checked with the variants called by the above BCFTOOLS pipelines, and no overlaps were found.

2.3.10 Bulk RNA-seq and data analysis

We downloaded the K562 bulk RNA-seq data with pgRNA targeting *AAVS1*, which was generated by our previously published research(*38*) (GEO accession code GSE176503). In the bulk RNA-seq library, the sequencing reads were aligned to the human reference genome (GRCh37/hg19) using HISAT2 (2.0.4)(*39–41*) and assembled and quantified by StringTie (1.3.5)(*39*, *42*).

2.3.11 Single-cell RNA-seq and data analysis

Single-cell library preparation. K562 cells were infected with the validated pgRNA hotspot_10_25-pg2. Eight days after lentiviral infection, the pgRNA-infected cells were subjected to FACS for single-cell library preparation. The single-cell library was prepared according to a previously established Drop-seq protocol(*43*). PolyA+ RNA was reverse transcribed through tailed oligo-dT priming directly in whole-cell lysates (single droplet) using Moloney Murine Leukaemia Virus Reverse Transcriptase (MMLV RT) and temperature switch oligos. The resulting full-length cDNA contains the complete 5′ end of the mRNA as well as an anchor sequence that serves as a universal priming site for second strand synthesis. The cDNA was pre-amplified using 15 cycles with Kapa HiFi Hotstart Readymix and then tagmented at 55°C for 5 min in a 20 µl reaction following the Illumina Nextera DNA preparation kit. 5 microliters of neutralization buffer was added to the tagmentation reaction mix to quench the reaction. The tagmented DNA was amplified by 12 cycles of standard Nextera PCR. The DNA was then purified with Ampure beads (sample to beads ratio of 1:0.6). The prepared hotspot_10_25-deleted single cell library of K562 was sequenced on an Illumina HiSeq 4000 instrument.

Single cell RNA-seq processing. The single cell RNA-seq data were processed using the Drop-seq pipeline developed by the McCarroll lab(*43*). Low-quality reads (lower than Q10) and PCR duplicates were removed. Cells were ranked in descending order by the total number of read counts. Cells ranked before the inflection point of the cumulative distribution were selected for the following analysis. Each cell was first normalized by counts per million (CPM). The value E_i was computed as the sum of CPMs for a given gene across all the cells. $E_{total}$was calculated as the sum of E_i for all the genes. Then, a P_i was computed as P_i =E_i /E_{total}. In a given cell *j,* the normalized gene expression of all genes was assumed to independently and identically follow the binomial distribution G_{ij} ~*B* ($N_{j,}$ P_i), where G_{ij} is the expected read of gene *i* in cell *j* and N_j is the total read for cell *j*. A p-value was computed to evaluate how each gene expression in each cell significantly deviated from the expectation based on the binomial distribution. We also calculated *p*-values for genes in the negative control (Δ*AAVS1*) and wild-type bulk RNA-seq data in the same way.

Single-cell trajectory branching and pseudotime analysis. Because hotspot deletion severely hampered cell proliferation, we focused on analysing the apoptosis-related genes annotated in the KEGG database(*44*). The 99 apoptosis-related genes that showed differential expression upon deleting hotspot_10_25 (chr10: 74,123,469-74,124,868) in at least 10% ~ 15% of cells (*p*-value < 0.05) were selected. All the normalized single cells and bulk data were clustered with trajectory branching and pseudotime analysis using Monocle(*45*, *46*). Monocle assigned a specific pseudotime value and a “state” to each cell. Cells with the same “state” and similar pseudotime were clustered together(*45*), and then the relative gene expression in each cluster was computed.

Differentially expressed genes identified from pseudotime analysis. To identify differentially expressed genes (DEGs) pairwise between different cell states, a *Wilcoxon* Rank-Sum Test(*47*) was used to identify genes that showed significantly up- and downregulated in the cell state pair.

2.3.12 Validation of synthetic lethal pairs by cell proliferation assay

Selection of the targeting sgRNA for each gene. To explore the synthetic lethal pairs among the four significantly downregulated genes located within the same TAD of hotspot_10_25 after hotspot deletion, we first determined the targeting sgRNA to ensure efficient knockdown of each gene. Three sgRNAs were selected to target the promoter region of each gene from the hCRISPRi-v2 library(*48*), and a non-targeting sgRNA was set as a control. These sgRNAs were further cloned into the lentiviral expression vector with an EGFP selection marker and then transduced into K562 cells stably expressing KRAB-dCas9 protein through lentiviral infection. Three days after infection, the EGFP-positive cells were sorted by FACS, and the total RNA of each sample was extracted using an RNeasy Mini Kit (QIAGEN 79254). cDNA was synthesized from 2 µg of total RNA using the Quantscript RT Kit (TIANGEN KR103-04), and real-time qPCR was performed with TB Green™ Premix Ex Taq™ II (Tli RNaseH Plus, TAKARA) to detect the expression of each indicated gene as well as of the reference gene *GAPDH*. The sgRNAs showing the most significant knockdown effect were selected for subsequent experiments to evaluate the synergistic effect.

Evaluation of the growth effect of each individual gene and gene pair in K562

cells. The above four selected sgRNAs were grouped into six gRNA pairs targeting six gene pairs. The four sgRNAs and six pgRNAs were respectively cloned into the lentiviral expression vector with an EGFP selection marker and then transduced into K562 cells stably expressing KRAB-dCas9 protein at an MOI of < 1. The cell proliferation assay was performed as described above. The first time point of FACS analysis was at 6 days after lentiviral infection, and the experiment lasted for another 12 days.

2.4 Results

2.4.1 Small-world network formed by 3D contacts between promoters and enhancers

To identify regulatory element promoters or enhancers that are likely to be important for chromatin organization, we set out to classify such elements involved in many interactions with other loci in the genome. We first identified active promoters (marked by H3K27ac and H3K4me3) and enhancers (marked by H3K27ac and H3K4me1) in 73 normal and 5 cancer cell lines/tissues with all 3 marks using data from the Roadmap Epigenomics project(*26*). The 3D contacts between these active promoters/enhancers in each cell line/tissue predicted by EpiTensor(*25*) were assembled into a regulatory element interaction network (referred to as REIN hereinafter), in which nodes are promoters/enhancers and edges represent 3D contacts (see Materials and Methods). We resorted to computational prediction by EpiTensor(*25*) because Hi-C data with sufficient resolution to define the interactions between promoters and enhancers were rare. We have previously shown that chromatin contacts could be successfully predicted by EpiTensor(*25*), which detects epigenetic covariation patterns between promoter-enhancer, promoter-promoter and enhancer-enhancer pairs at 200-bp resolution via tensor analysis. Such a covariation indicates that possible 3D contacts can be formed between active regulatory elements in a cell type-specific manner. Therefore, when considering spatial contacts in a particular cell type or tissue, we only considered those formed between active promoters and/or enhancers, as marked by open chromatin or H3K27ac, because these contacts are likely to establish functional regulation. EpiTensor predictions were shown to be highly concordant with the Hi-C, ChIA-PET and eQTL results in different cell types(*25*).

In REIN, each node represents a promoter/enhancer in the given cell line/tissue, and each edge represents a contact predicted by EpiTensor. The degree of a given node reflects its total contacts. We examined the topological properties of REIN using SNAP software(*49*). Through computational simulations, the cluster coefficients of REIN (the percentage of node pairs that connected when they were connected to another node) were found to be similar to an equivalent (same number of nodes and edges) regular lattice network(*50*), and their path length (the largest required number of steps between node pairs) was similar to that of the equivalent random network(*51*) (**Figure 2.1A**). These properties showed that the REINs are small-world networks. Small-world networks are characterized by robustness in that they are resistant to random attacks (random removal of nodes) but vulnerable to targeted attacks (removal of specific nodes) on high-degree nodes that have significantly more contacts than the other nodes(*52*). We selected the top 10% high-degree nodes in REIN as “hotspots” for further analysis.

2.4.2 Mutations in hotspot enhancers and promoters could alter 3D contacts

We collected all the genomic loci identified as hotspots in at least one cell line/tissue. In total, we found 48,110 regions, the majority of which are enhancers, and 12,754 of them overlap with promoter regions (1 kb around the transcription start sites). Consistent with our previous analysis(*25*), these loci tend to be active (overlapping with H3K27ac signals) in more cell types than the non-hotspot loci (**Figure 2.1B**). We noticed that the number of interactions a hotspot forms varies significantly across cell types/tissues, and on average, a locus was identified as a hotspot only in 7 out of 78 cell types/tissues. Particularly, promoter hotspots are shared by more cell types/tissues (on average 17 out of 78) than enhancer hotspots (on average 4 out of 78), which is not unexpected as enhancers are known to be cell type/tissue specific.

Given the importance of high-degree nodes in a small-world network, mutations in hotspot loci may have a severe impact on the network structure. To investigate this possibility, we analysed the loci that are active in all examined cell lines but show a significant change in degree. We first identified nodes with sample-specific degrees: using the degree numbers of each node in all 73 normal samples as the background distribution, we identified nodes that are active in a specific sample and whose degree also significantly deviates from the mean. We then determined the sample-specific genetic variations (GVs). We collected 1,197,917 GVs in 45 normal and 17 cancer samples (DCC accession number ENCFF105JRY). For each GV in each sample, if its B-allele frequency significantly deviates from the mean in the 45 normal samples, we considered this GV specific to the sample (see Materials and Methods).

The nodes with a sample-specific degree containing at least one sample-specific GV, which are called degree-GV correlated nodes (**Figure 2.1C**), are good candidates to investigate the relationship between GV and degree. We first analysed 4 normal (GM12878, H1, HEK293 and IMR90) and 4 cancer cell lines (HeLaS3, HepG2, K562 and MCF-7) and found that degree-GV correlated nodes are more frequently observed in cancer cell lines than in normal cell lines (**Figure 2.1D**). First, we collected 21,064 nodes, which contain at least one sample-specific GV and show specific high/low degree in at least one of the 78 samples (73 normal and 5 cancer samples). Note that GV and high/low degree do not necessarily occur in the same cell line. For example, the degree of a node can be significantly high in GM12878, while the allele frequency of GV covered by this node is significantly high in K562. We found that the majority (62.59%) showed specificities in both cancer and normal cells, among which 18.36% were specific to cancer and 19.05% specific to normal cells (**Figure 2.1F**). Similarly, among the 629,547 cell-specific GVs, 58.86% showed specificities in both cancer and normal cells, 32.08% to only cancer cells and 9.06% to only normal cells (**Figure 2.1E**). However, the degree-GV-correlated nodes were dominated by cancer-specific nodes (87.18%), compared to 8.53% in both cancer and normal cells and 4.29% only in normal cells (**Figure 2.1F**). We observed the same trend for degree-GV-correlated hotspots, including 86.52% cancer-specific, 4.52% normal-specific, and 8.96% in both cancer and normal cells (**Figure 2.1F**). In summary, the majority of degree-GV-correlated nodes appear in cancer cells.

We further examined two groups of nodes in 8 distinct cell lines: one group had a significantly higher degree in one cell line than in other cell lines, which indicates cell-type-specific contact formation (one-cell-type-specific nodes), and the other had a significantly lower degree in one cell line than in the others, which indicates cell-type specific contact disruption (seven-cell-type-specific nodes). The percentages of HepG2-specific nodes and K562-specific nodes in the one-cell-type-specific group (cell-type-specific contact formation) are 28.5% and 7.4% in all nodes, 64.4% and 17.4% in degree-GV-correlated nodes and 64.3% and 18.1% in degree-GV-correlated hotspots, respectively (**Figure 2.1G**). Similarly, the percentages of HeLa-S3-specific nodes and K562-specific nodes in the seven-cell-type-specific group (cell-type-specific interaction disruption) are 41.8% and 13.2% in all nodes, 49.2% and 37.8% in degree-GV-correlated nodes and 61.3% and 27.1% in degree-GV-correlated hotspots, respectively (**Figure 2.1H**). Taken together, our analyses suggested that cancer-specific GVs are highly correlated with the node-degree change that alters the REIN.

2.4.3 CRISPR/Cas9 library screening identified hotspots essential for cell growth and survival

To further investigate the function of hotspots, 751 hotspots identified as enhancers were randomly selected for targeted deletion to analyse their impact on cell growth and survival. These hotspots do not overlap with coding regions of any protein-coding gene or noncoding RNA. In total, 14,399 paired gRNAs (pgRNAs) were designed to delete these loci (see Materials and Methods), including 473 positive control pgRNAs targeting 29 ribosomal genes, 100 negative control pgRNAs targeting the *AAVS1* locus and 100 non-targeting pgRNAs. Through lentivirus infection at a low MOI (Multiplicity of Infection), the pgRNA library was transduced into K562 cells stably expressing Cas9 protein. The pgRNA-infected samples were FACS-sorted 3 days post infection, serving as the control group, and then continuously cultured for 30 days to obtain the experimental group. The library cells from the control and experimental groups were sequenced to determine the abundance of each pgRNA (**Figure 2.2A**). The read distribution of pgRNAs showed a high correlation between the two biologically independent replicates for all groups (**Figure S2.1A-C**), indicating high reproducibility.

Compared with the control group, pgRNAs targeting ribosomal genes and hotspots in Day 30 experimental cells were both decreased more than those targeting the *AAVS1* locus and non-targeting pgRNAs. For all the pgRNAs of each hotspot, we calculated their fold changes and *P* values by comparing them with the pgRNAs targeting *AAVS1* using the Mann-Whitney U test(*53*, *54*), which is focused on analysing screening data with the in-library controls and could more accurately reflect the fitness effect of each locus. By randomly sampling the pgRNAs targeting *AAVS1*, we generated a distribution of negative controls and further computed the hotspots' *P* values. The screen score of each hotspot was calculated by combining its mean fold change and corrected *P* values (see **Materials and Methods**), and 49 hotspots with screen scores ≤ -2.5 were considered to significantly affect cell fitness upon deletion (**Figure 2.2B**). To avoid cellular toxicity caused by potential off-target effects(*55*–*58*), we assessed the specificities of sgRNAs with 2 or 3 mismatches to off-target loci using the GuideScan specificity score and calculated the specificity score for each pgRNA (see Materials and Methods)(*28*). Because *AAVS1*-targeting pgRNAs with specificity scores ≤ 0.1 could cause a dropout effect in K562 cells (**Figure 2.2C** and **Figure S2.1D**), we only kept pgRNAs with specificity scores > 0.1 and $\log_2$(fold change) < -1 for subsequent analysis. Furthermore, hotspots with copy number amplification were also removed to avoid cell death caused by multiple cleavages(*59*). Using such stringent criteria, we identified 43 hotspots essential for the cell fitness of K562 cells (**Figure 2.2C**).

Based on the ranking of the screen score, 7 top-ranked hotspots in K562 cells were chosen for individual validation by cell proliferation assay. None of them overlapped with any promoter, protein-coding gene or noncoding RNA. Three or two pgRNAs with high targeting specificity were separately constructed for each hotspot, and the cell proliferation assay was performed as previously reported(*54*). We found that deletion of these hotspots led to significant cell death or cell growth inhibition (**Figure 2.3A** and **Figure S2.2A**), which was consistent with the screening results, indicating that these hotspots played critical roles in cell fitness.

In our design, we assured that the deletion regions were not associated with any coding regions of protein-coding genes, but there were a few essential hotspots located near the promoter regions or in the introns of coding genes. To rule out the possibility of affecting the expressions of certain genes essential for cell growth and survival (essential genes) after hotspot deletion, we further investigated two identified hotspots located near the gene promoter or in the intronic regions, whose deletion may affect the expression of the corresponding genes. For hotspot_19_32 located in the intron of an essential gene *GATAD2A*, we chose 2 highly specific pgRNAs to respectively delete this locus in K562 cells and observed significant cell growth inhibition (**Figure 2.3B**). Importantly, we found that overexpression of *GATAD2A* did not rescue the cell death caused by hotspot deletion (**Figure 2.3B**), indicating that the hotspot deletion itself has a profound impact on cell growth. By detecting the expression level of the *GATAD2A* gene under each condition by real-time qPCR, we confirmed that the gene was successfully overexpressed in K562 cells and that the cell death caused by the hotspot deletion was not rescued by *GATAD2A* overexpression (**Figure 2.3C**). A similar result was obtained for hotspot_1_36, which is located approximately 3 kb upstream of the transcriptional start site of an essential gene *SLC2A1*. We performed the cell proliferation assay using 2 pgRNAs in wild-type K562 cells and K562 cells stably overexpressing *SLC2A1*. A similar level of influence on cell fitness was observed in both conditions for each pgRNA deletion, and real-time qPCR further confirmed that the growth phenotype was not due to affecting the expression level of *SLC2A1* (**Figure S2.2B-C**).

To further assess the essentialities of the identified K562-essential hotspots in other cancer cell lines, we chose hotspot_10_25 (chr10: 74,123,469-74,1248,68), which showed a significant growth defect in K562 if deleted, for parallel validations in HeLa (cervical cancer cells), H1975 (non-small cell lung cancer cells), A549 (non-small cell lung cancer cells) and NAMALWA (Burkitt's lymphoma) cells. Surprisingly, compared with the negative control *AAVS1*-targeting pgRNAs, hotspot_10_25 showed no significant effect in any of the five tested cell lines, indicating that its role in K562 cells is cell-type specific (**Figure 2.3D, Figure S2.2D**).

2.4.4 The essentiality of hotspots does not result from any association with essential genes

To understand how these identified essential hotspots exert their functional roles, we first examined whether direct interaction with essential genes determines the essentialities of these hotspots. We retrieved the essential genes whose knockdown would lead to cell death according to the CRISPRi-based screen(*48*) and identified all possible spatial contacts they formed that were detected by Hi-C (p-value ≤ 0.05) in wild-type K562 cells(*60*). There is no distinction between essential and non-essential hotspots regarding their association with essential genes (the *Wilcoxon* Rank-Sum test *p*-value = 0.98, indicating no significant difference) (**Figure 2.3E**). We also performed the same comparison using the spatial contacts predicted by EpiTensor and reached the same conclusion (the *Wilcoxon* Rank-Sum test *p*-value = 0.61) (**Figure 2.3E**). According to the above analysis, the essentiality of hotspots is not determined by their direct contact with essential genes.

2.4.5 Deleting essential hotspots can affect broad chromatin organization

We next investigated whether deleting hotspots affects chromatin organization. We selected hotspot_10_25 (chr10: 74,123,469-74,124,868) for further analysis, which showed unique essentiality in K562 cells (**Figure 2.3A, Figure 2.3D, Figure S2.2D**) yet does not interact with any essential protein-coding gene identified in the previous CRISPRi screening(*48*) in the Hi-C analysis. We first performed whole genome sequencing (WGS) to confirm that there was no off-target effect. The validated pgRNA hotspot_10_25-pg2 was chosen (**Figure 2.3A**), and the WGS library was generated 8 days after pgRNA infection in K562 cells. Compared to the hg19 human genome, we identified 4.1 million germline mutations in hotspot_10_25-deleted K562 cells, which showed 86.2% consistency with the published wild-type K562 WGS data. The high percentage of the germline mutation discovery rate indicated good quality of the library. We used Cas-OFFinder to identify 746 potential off-target loci with loose cut-off values (base mismatch ≤ 4, bulge ≤ 2) to avoid missing any possible off-target loci. We manually examined the putative off-target loci with the indels detected from the edited cells that differed from the wild-type cells (**Table S2.1**). Except for the significant indels found in the two on-target loci with clear cleavages in the pileup reads in the genome browser view (**Figure S2.3**), there was no cleavage in the pileup reads on any of the putative off-target loci (two examples of possible off-target loci are shown in **Figure S2.3**). These analyses confirmed that the cell growth defects did not result from off-target effects.

We subsequently performed Hi-C analysis on the edited K562 cells and compared it with the wild-type cells(*60*) (see Materials and Methods). The 100 kb-resolution Hi-C contact maps of the wild-type and hotspot-deleted cells are overall similar (**Figure S2.4A**), and no distinct flips between A and B compartments were observed on the entire chr10 at 50-kb resolution (**Figure S2.4B**). We compared the chromosome-wide similarity and detected differential Hi-C contacts using HiCRep(*33*, *34*) and HiCcompare(*35*). The high Stratum-adjusted correlation coefficient (SCC > 0.7) and the small percentage of differential Hi-C contacts suggested an overall similarity between the wild-type and hotspot_10_25-deleted K562 cells (**Table S2.2**). Topologically associated domains (TADs) largely remained similar, with a few TADs merge and split in the regions of chr10: 12230000-14540000, chr10: 15210000-15910000, chr10: 71220000-72220000*,* chr10: 89850000-91010000, chr10: 95290000-96350000 and chr10: 99450000-100140000 (**Figure S2.4C**). Interestingly, using HiCCUPS (https://github.com/aidenlab/juicer/wiki/HiCCUPS), we found that hotspot deletion did affect chromatin loops (**Table S2.3**). We next investigated whether deleting a hotspot could affect relatively broad genomic regions in spatial proximity. Using a sliding window with a bin step of 1 Mb and a flanking region of 2 Mb in the linear genome, we assembled all the Hi-C contacts (5-kb resolution with *p*-value ≤ 0.05) in each 5-Mb window into a sub-network. The modularity score and effective diameter were computed for each of these sub-networks in the wild-type and hotspot-deleted K562 cells. These two metrics of all the sliding windows of chr10 showed a high correlation between before and after hotspot deletion, with *Pearson* correlation coefficients of 0.84 and 0.91 for the effective diameter and modularity, respectively (**Figure 2.4A-B**). Notably, significant changes were observed on the 6-8 Mb regions surrounding the deleted hotspot for both effective diameter (chr10: 69-75 Mb for bin 71 and 72) and modularity (chr10: 68-76 Mb for bin 70, 71 and 73) (**Figure 2.4A-B**). Some other genomic regions interacting with the hotspot neighboring regions were also affected, such as chr10: 11-17 Mb (bin 13 and 14), showing a significant change in modularity (**Figure 2.4B, Figure S2.5**). We further examined the Hi-C contact alteration within chr10: 11-17 Mb and chr10: 68-76 Mb (**Figure 2.4C and 2.4E**). In the region of chr10: 12-14 Mb with observed Hi-C contact changes (**Figure 2.4C**), we also found consistent TAD splits (**Figure S2.4C**), disruption and formation of chromatin loops upon hotspot deletion (**Table S2.4**). These chromatin changes led to alteration of promoter-enhancer interactions, such as the enhanced and weakened contacts between the *CELF2*, *RSU1*, *FAM149B1* and *CCAR1* promoters and their interacting enhancers upon hotspot deletion (**Figure 2.4D and 2.4F**). Notably, these affected promoters and enhancers are not only located close to but also can be as far as 62 Mb away from the deleted hotspot_10_25 located at chr10: 74,123,469-74,124,868. These observations showed that hotspot deletion resulted in broad alterations in chromatin structure beyond its linear neighbor genome.

2.4.6 Essential hotspots tend to reside in dense chromatin structures

If essential hotspots are critical for maintaining the chromatin structure in the spatial neighbourhood, it is likely that the 3D contacts around them are dense. Therefore, we compared the sub-network effective diameters, modularity and chromatin loops in the 5-Mb regions centered at the essential and non-essential hotspots in wild-type K562. We found that essential hotspots were surrounded by higher effective diameters (*Wilcoxon* Rank-Sum test, *p*-value = 7.3E-6), higher modularities (*p*-value = 0.1) and higher loop densities (*p*-value = 4.7E-4) than non-essential hotspots (**Figure S2.6A-C**). In fact, using these three metrics in wild-type K562 cells, a random forest classification model could distinguish essential and non-essential hotspots with an AUC of 0.73 in 10-fold cross validations. This result resonates with the above observations and suggests that hotspots are pivotal for stabilizing dense chromatin contacts in the spatial neighborhood.

2.4.7 Hotspot deletion synergistically affects gene expression

We next performed single-cell RNA-seq using Drop-seq(*43*) to analyze the changes in gene expression upon hotspot_10_25 deletion. We transduced the individually validated pgRNA hotspot_10_25-pg2 (**Figure 2.3A**) targeting this essential hotspot into K562 cells, among which 482 single cells passed the quality control. We also included the bulk RNA-seq data of wild-type and *AAVS1*-deleted cells as controls. All the data were normalized together to make them comparable (see Materials and Methods). As deletion of this hotspot has an impact on cell viability or cell growth, we focused on genes related to apoptosis pathways to confirm their activation. We selected 99 apoptosis-related genes documented in the KEGG database and clustered the cells into five states by trajectory branching and pseudotime analysis using Monocle (**Figure 2.5A**)(*46*). The wild-type and *AAVS1*-deleted K562 cells (negative control) were located in state 1, suggesting that cells in this state resemble the wild-type cells. The apoptosis genes fell into three groups, with distinct expression patterns along the pseudotime but overall all increasing from cell state 1 to cell state 5, for example, *CASP*2, *CASP8*, *CASP9*, and *CASP10* in cluster 1, *CASP6* in cluster 2, and *CASP7* in cluster 3 (**Figure 2.5B**). Taken together, the single cell transcriptomic analysis showed that apoptosis pathways are activated upon hotspot deletion.

To investigate the impact of hotspot deletion on the spatial neighborhood, we analyzed the genes whose promoters were predicted to interact with the essential hotspot_10_25 by EpiTensor. Among the 14 genes located within the same TAD of hotspot_10_25, 4 showed significantly downregulated (*p*-value < 0.05) in the progression from state 1 to 5, including *P4HA1* (downregulated from state 1 to 2, **Figure 2.5C**) and *DNAJB12*, *ASCC1* and *ECD* (downregulated from state 2 to 4, **Figure 2.5D-E**, **Figure S2.7A**). By individually knocking down each gene by CRISPR interference (**Figure 2.5F, Figure S2.7B**), only *ECD* knockdown showed a weak impact on cell growth, and all the other genes showed no detectable effects (**Figure 2.5G, Figure S2.7C**). As the hotspot interacted with multiple genes, we investigated whether knocking down a pair of genes would have a synergistic effect on cell growth. Applying the CRISPRi strategy, we knocked down 6 pairs of genes in K562 cells using paired gRNAs, respectively. We found that simultaneous knockdown of *P4HA1-ECD* and *ASCC1-ECD* showed a much more significant impact on cell growth (**Figure 2.5G**). These results indicated that disrupting hotspot_10_25 could affect the expression levels of multiple interacting genes, and their synergistic effect could lead to cell death. Note that we were limited to examining pairs of genes, but hotspot deletion can affect multiple genes together with more significant synergic effects.

2.5 Discussion

In this study, we analyzed the hotspot promoters/enhancers that were predicted by EpiTensor(*25*) to form a large number of 3D contacts with other promoters/enhancers. The unsupervised learning method EpiTensor predicts all possible 3D contacts of promoter-promoter, promoter-enhancer and enhancer-enhancer pairs. In a particular cell line, we focused on the predicted contacts between active promoters/enhancers denoted by histone marks. The hotspot promoters/enhancers are defined by their 3D contacts with many active promoters/enhancers, which makes them a class of high-degree nodes in the REIN. We showed that REIN is a small-world network that is vulnerable to targeted perturbation to high-degree nodes. Therefore, it is reasonable to infer that hotspots can be important for stabilizing REIN and the 3D contacts formed between active promoters/enhancers.

We found that the occurrence of genetic variations (GVs) is much more strongly correlated with the alteration of 3D contact degrees (degree-GV-correlated nodes) in the hotspots in cancer cells than in normal cells. Furthermore, we showed that cancer-specific hotspots (only formed or disrupted in one particular cancer cell) are enriched with degree-GV-correlated nodes. Taken together, these observations suggest that GVs occurring in hotspots can lead to chromatin structure changes and dysregulated cellular functions.

To confirm the functional importance of the hotspots, we performed CRISPR/Cas9 library screening on hotspot enhancers by paired-gRNA deletion in the K562 cancer cell line. By calculating the screen score for each hotspot and further filtering loci with potential off-target effects or copy number amplifications, we identified 43 hotspots essential for cell growth and survival. Nine randomly selected loci were individually validated by cell proliferation assay, including 7 top-ranked hotspots in K562 cells without any overlap with coding genes and 2 loci overlapped with promoter or intronic regions of certain genes. We further identified that hotspot_10_25 was essential for cell fitness specifically in K562 cells through multiple validations in four other cancer cell lines.

We thus selected a hotspot enhancer (hotspot_10_25) as a representative of cancer-specific hits for in-depth analysis, which ensured that no off-target cleavages occurred through WGS analysis. Note that this selected hotspot is not unique compared to the other essential hotspots, and the insights obtained here are expected to be generalizable. Hi-C and scRNA-seq analyses showed that deleting this 1.4-kb long hotspot could impact a broad chromatin structure of 8-Mb regions surrounding the hotspot and affect the expression of numerous distal genes not even directly associated with the hotspot. These observations indicate that the hotspot enhancer has a pivotal role in chromatin organization beyond forming chromatin loops.

Importantly, this hotspot does not directly interact with any essential gene, and thus, the cell death resulting from its deletion is not due to directly disrupting the expression of essential genes. Single cell RNA-seq revealed that hotspot_10_25 deletion could affect the expression levels of multiple interacting genes located within the same TAD of the hotspot. By knocking down individuals and pairs of these genes, we found that although none of these dysregulated genes has a significant impact on cell fitness individually, altered expression of gene pairs showed significant synergistic effects leading to cell death.

We have revealed the understudied “structural importance” of noncoding regulatory elements, especially enhancers. We are aware that establishing the causal relationship between broad chromatin organization changes and cell proliferation or gene expression remains technically challenging. However, to our knowledge, this is the first report about the observations that enhancers could maintain a broad chromatin organization, which goes far beyond the direct interaction between promoters and enhancers. A logical inference towards the causal relationship based on our observations is that the impact of hotspot deletion is propagated through the 3D contact network and could impact distal genes that are important for cell fitness.

2.6 Acknowledgements

Chapter 2, in full, is a reformatted reprint of the material as it appears in “Regulatory elements can be essential for maintaining broad chromatin organization and cell viability.” Ying Liu[#]; Bo Ding[#]; Lina Zheng[#]; Ping Xu[#]; Zhiheng Liu; Zhao Chen; Peiyao Wu; Ying Zhao; Qian Pan; Yu Guo; Wei Wang; Wensheng Wei. Nucleic Acids Research, Oxford University Press, 2022. gkac197, https://doi.org/10.1093/nar/gkac197. ([#] These authors contributed equally to this work.) The dissertation author was the primary investigator and author of this paper.

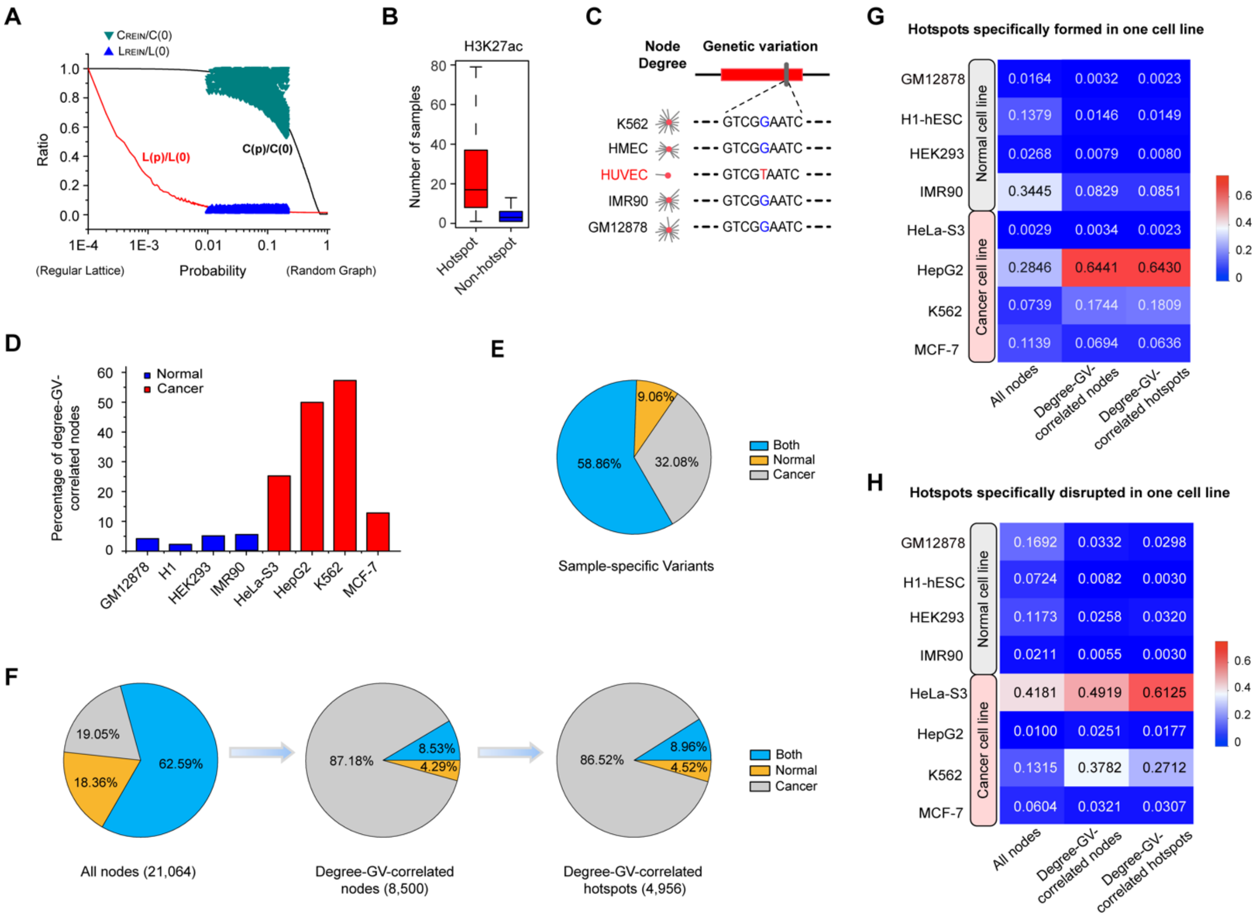
We acknowledge the staff of the BIOPIC High-throughput Sequencing Center (Peking University) for their assistance in next-generation sequencing analysis, the National Center for Protein Sciences (Beijing) at Peking University for their assistance with fluorescence-activated cell sorting and analysis, and Dr. Hongxia Lv and Ms. Liying Du for their technical help. We acknowledge Dr. Ying Yu (Peking University) for her assistance in preparing the NGS library. We acknowledge the staff of the UC San Diego IGM Genomics Center for sequencing services and the UC San Diego Human Embryonic Stem Cell Core Facility for cell sorting services. We acknowledge Ms. Jia Xu (UC San Diego) for her assistance in preparing a single cell RNA-seq library.

W. Wei and W. Wang conceived and supervised the project. W. Wei, W. Wang, Y.L. and B.D. designed the experiments. B.D. and L.Z. constructed network analysis and identified and characterized hotspot regions. Y.G. designed the pgRNA library for hotspot screening. Y.L. and P.X. performed the pgRNA library construction and screening. Y.L.

performed the experiments, including individual validation of candidate hotspots in multiple cell lines, whole-genome sequencing (WGS), bulk RNA-seq and examination of the synergistic effects with the help of P.X. and Q.P. Z.L. performed the bioinformatics analysis of the screening data and designed the pgRNAs used for individual validation. P.W. and Z.C. performed the Hi-C experiments on hotspot-deleted K562 cells. P.W. and Y.Z. performed single cell RNA-seq on hotspot-deleted K562 cells. L.Z. and B. D performed the bioinformatics analyses of the WGS, Hi-C and single-cell RNA-seq data. Y.L., B.D., L.Z., W. Wang and W. Wei wrote the manuscript with contributions from all other authors.

2.7 Figures

Figure 2.1. Small-world network analysis and mutation effects on 3D contact for hotspot enhancers and promoters. (**A**) The path length and cluster coefficient of REINs compared with equivalent regular lattice networks and equivalent random graph networks. (**B**) Comparison of H3K27ac peaks between hotspots and non-hotspots in 121 cell lines, primary cells and tissues characterized by the NIH Roadmap Epigenetics Project. (**C**) Definition of degree-GV-correlated nodes. In this example, the node has a low degree in the HUVEC cell line and a high degree in other cell lines, which is correlated with the GV profile with a G > T SNP in HUVEC that is not present in other cell lines. (**D**) The percentage of degree-GV-correlated nodes in normal cell lines and cancer cell lines. (**E**) The distribution of GV specificities in samples. Normal, cancer and both indicate GVs with specificities only in normal cells, only in cancer cells and in both cell types, respectively. (**F**) The distribution of normal or cancer cell line specificities in the nodes which contain at least one sample-specific GV and show specific high/low degree in at least one of the 78 samples (those nodes denoted as "All nodes"), degree-GV-correlated nodes and degree-GV-correlated hotspots. Note that for the nodes in the first group (All nodes), GV and high/low degree do not necessarily occur in the same cell line. For example, the degree of a node can be significantly high in GM12878, while the allele frequency of GV covered by this node is significantly high in K562. (**G**) The distribution of one-cell-line hotspot formation in all nodes, degree-GV-correlated nodes and degree-GV-correlated hotspots. (**H**) The distribution of one-cell-line hotspot disruption in all nodes, degree-GV-correlated nodes and degree-GV-correlated hotspots.


A
CREIN/C(0)
LREIN/L(0)
1.0
0.8
0.6
0.4
0.2
0.0
Ratio
L(p)/L(0)
C(p)/C(0)
1E-4
1E-3
0.01
0.1
1
(Regular Lattice)
Probability
(Random Graph)
B
H3K27ac
80
60
40
20
0
Number of samples
Hotspot
Non-hotspot
C
Node Degree
Genetic variation
K562
HMEC
HUVEC
IMR90
GM12878
GTCGGAATC
GTCGGAATC
GTCGTAATC
GTCGGAATC
GTCGGAATC
D
60
50
40
30
20
10
0
Percentage of degree-GV-correlated nodes
Normal
Cancer
GM12878
H1
HEK293
IMR90
HeLa-S3
HepG2
K562
MCF-7
E
9.06%
58.86%
32.08%
Both
Normal
Cancer
Sample-specific Variants
F
19.05%
18.36%
62.59%
87.18%
8.53%
4.29%
86.52%
8.96%
4.52%
Both
Normal
Cancer
All nodes (21,064)
Degree-GV-correlated nodes (8,500)
Degree-GV-correlated hotspots (4,956)
G
Hotspots specifically formed in one cell line
Normal cell line
Cancer cell line
GM12878 0.0164 0.0032 0.0023
H1-hESC 0.1379 0.0146 0.0149
HEK293 0.0268 0.0079 0.0080
IMR90 0.3445 0.0829 0.0851
HeLa-S3 0.0029 0.0034 0.0023
HepG2 0.2846 0.6441 0.6430
K562 0.0739 0.1744 0.1809
MCF-7 0.1139 0.0694 0.0636
All nodes
Degree-GV-correlated nodes
Degree-GV-correlated hotspots
0.6
0.4
0.2
0
H
Hotspots specifically disrupted in one cell line
Normal cell line
Cancer cell line
GM12878 0.1692 0.0332 0.0298
H1-hESC 0.0724 0.0082 0.0030
HEK293 0.1173 0.0258 0.0320
IMR90 0.0211 0.0055 0.0030
HeLa-S3 0.4181 0.4919 0.6125
HepG2 0.0100 0.0251 0.0177
K562 0.1315 0.3782 0.2712
MCF-7 0.0604 0.0321 0.0307
All nodes
Degree-GV-correlated nodes
Degree-GV-correlated hotspots
0.6
0.4
0.2
0

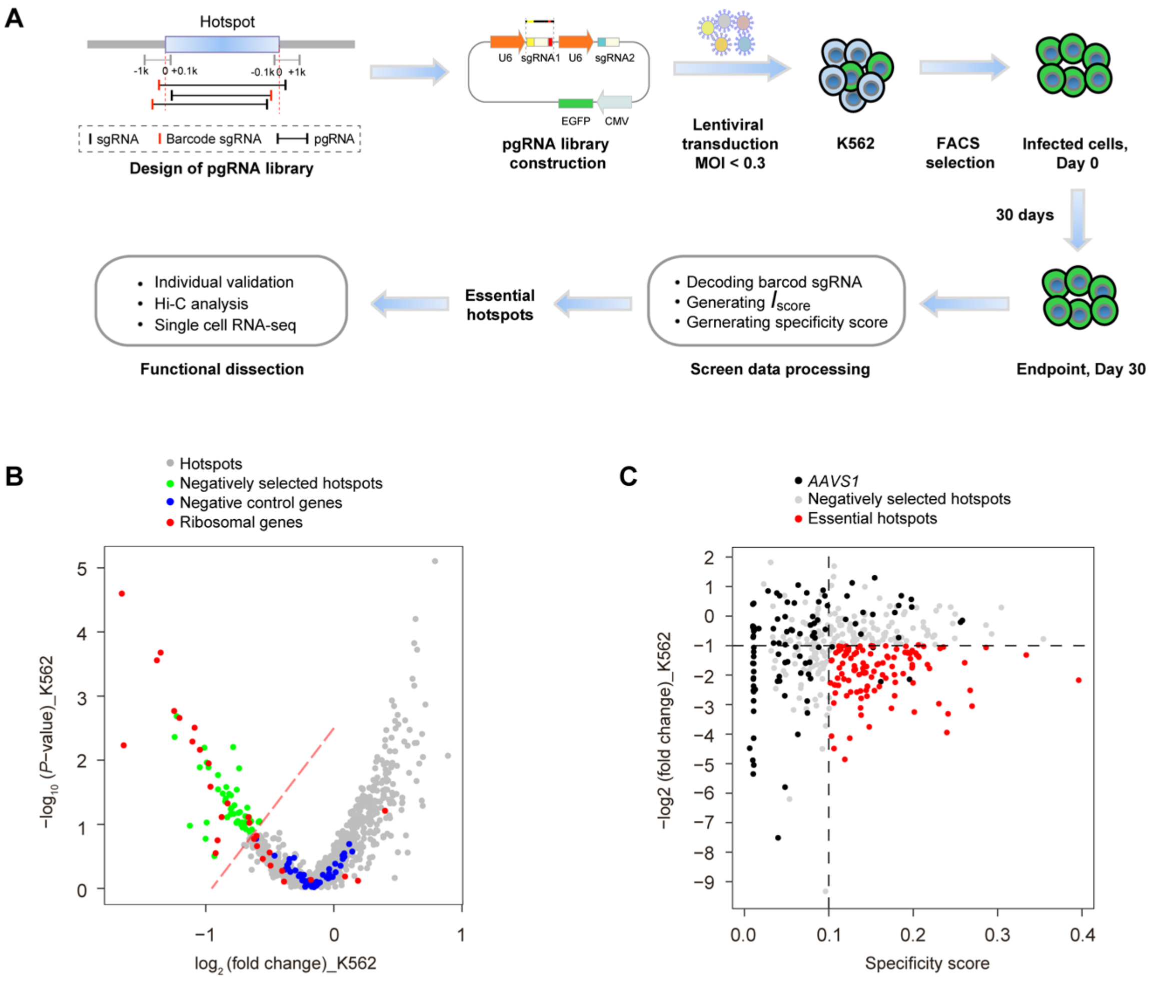
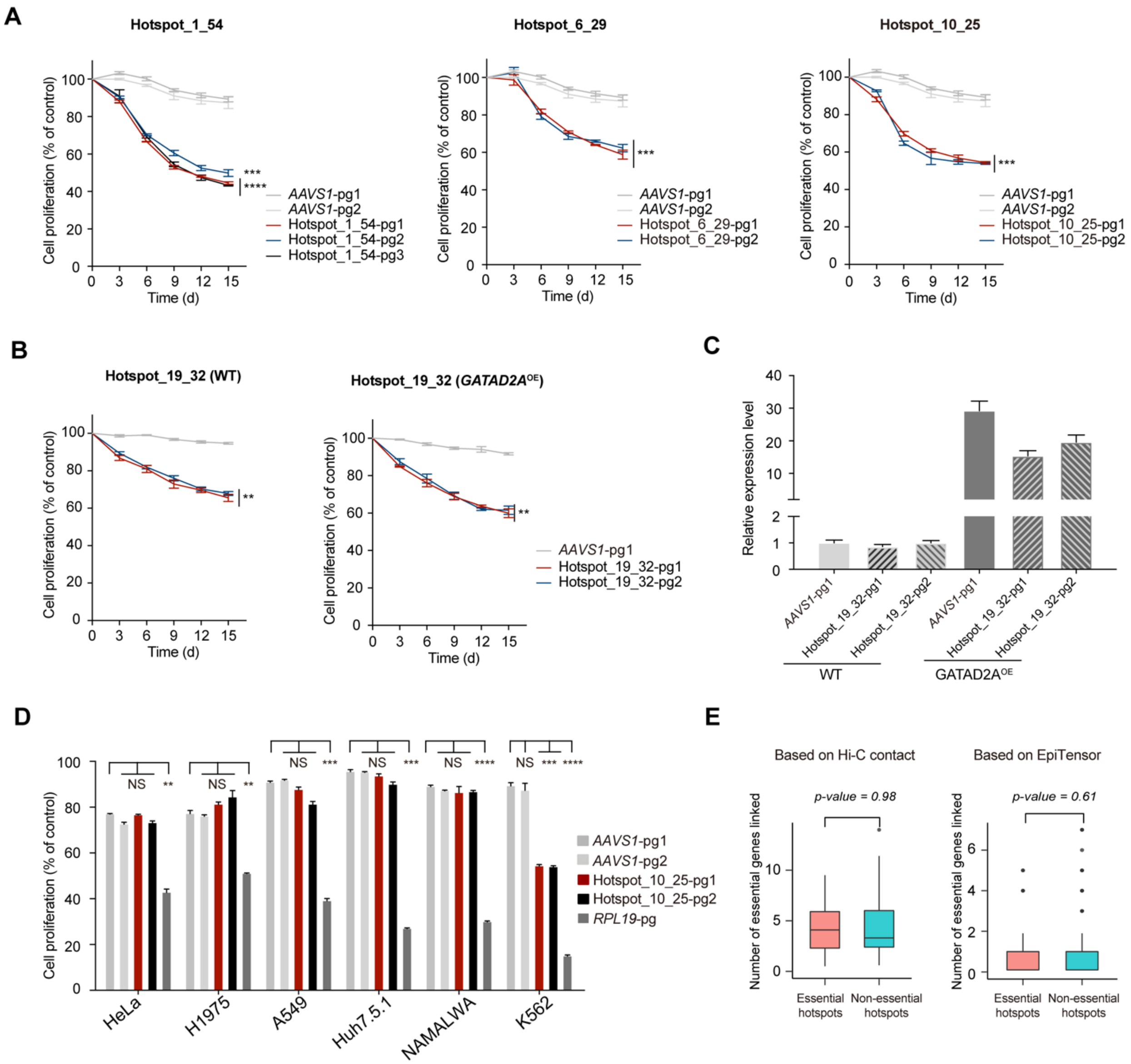


Figure 2.2. Identification of essential hotspots for cell growth and proliferation in the K562 cell line through pgRNA deletion-based CRISPR screening. (**A**) Schematic of the pgRNA library design, cloning and functional screening of selected hotspot loci. (**B**) Volcano plot of the fold change and *p-value* of hotspots in the K562 cell line. Negative control genes were generated by randomly sampling 20 *AAVS1*-targeting pgRNAs with replacement per gene, and ribosomal genes served as positive controls in the screening. The dotted red line represents a *screen score* = -2.5. (**C**) Selection of candidate essential hotspots by the fold change and specificity score of each pgRNA. These essential hits were selected under the threshold of a specificity score > 0.1 and $\log_2$(fold change) < -1.

Figure 2.3. Validation of candidate essential hotspot loci in K562 cells and multiple cell lines. (**A**) Validation of the top-ranked essential hotspot in K562 cells by cell proliferation assay. *AAVS1*-pg1 and *AAVS1*-pg2 are pgRNAs targeting the *AAVS1* locus and serve as negative controls. The asterisk (*) represents *p-value* compared with pgRNAs targeting *AAVS1*-pg1 at Day 15, which were calculated by two-tailed Student's *t*-test and adjusted for multiple comparisons by Benjamini-Hochberg procedure. Data are presented as the mean ± s.d. (n = 3 biologically independent samples). * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$; NS, not significant. (**B**) Validation of essential hotspots overlapped with the intronic region of an essential gene in K562 cells by cell proliferation assay. Left: WT K562 cells infected with pgRNAs targeting hotspot_19_32. Right: *GATAD2A*-overexpressed K562 cells infected with pgRNAs targeting hotspot_19_32. (**C**) The expression levels of *GATAD2A* in WT and *GATAD2A*-overexpressed K562 cells infected with pgRNAs targeting *AAVS1* or hotspot_19_32. (**D**) Validation of hotspots_10_25 in multiple cancer cell lines, including A549, H1975, HeLa, Huh7.5.1 and NAMALWA cell lines. Asterisk (*) represents *p-value* compared with pgRNAs targeting *AAVS1*-pg1 at Day 15, which were calculated by two-tailed Student's *t*-test and adjusted by Bonferroni correction accounting for multiple testings. * $p < 0.05$; ** $p < 0.01$; *** $p <$ 0,001; **** $p < 0.0001$; NS, not significant. (**E**) No significant difference between the numbers of essential genes contacting essential and non-essential hotspots from Hi-C or EpiTensor in K562 cells.


A
Hotspot_1_54
Hotspot_6_29
Hotspot_10_25
Cell proliferation (% of control)
Time (d)
0
3
6
9
12
15
100
80
60
40
20
***
****
AAVS1-pg1
AAVS1-pg2
Hotspot_1_54-pg1
Hotspot_1_54-pg2
Hotspot_1_54-pg3
Hotspot_6_29-pg1
Hotspot_6_29-pg2
Hotspot_10_25-pg1
Hotspot_10_25-pg2
B
Hotspot_19_32 (WT)
Hotspot_19_32 (GATAD2A^OE)
**
Hotspot_19_32-pg1
Hotspot_19_32-pg2
C
Relative expression level
40
30
20
10
2
1
0
AAVS1-pg1
Hotspot_19_32-pg1
Hotspot_19_32-pg2
WT
GATAD2A^OE
D
NS
**
***
****
HeLa
H1975
A549
Huh7 5.1
NAMALWA
K562
RPL19-pg
E
Based on Hi-C contact
Based on EpiTensor
p-value = 0.98
p-value = 0.61
Number of essential genes linked
Essential hotspots
Non-essential hotspots
10
5
6
4
2

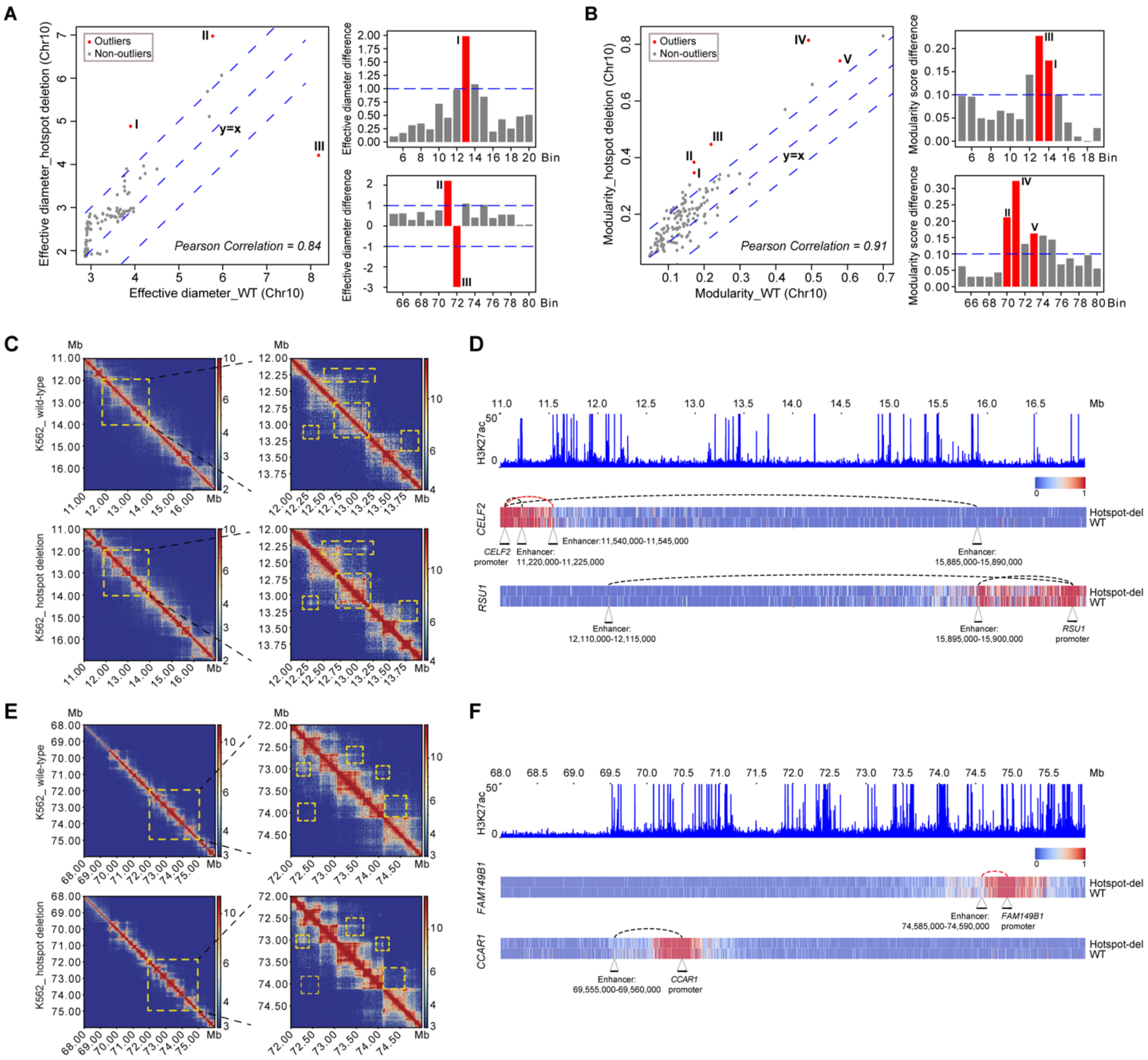


Figure 2.4. Deletion of an essential hotspot impacts broad chromatin structure. (**A-B**) The effective diameter (**A**) and modularity (**B**) before and after hotspot deletion in the sliding 5-Mb sub-networks on chr10 (left). The outliers are labeled, and their genomic locations are shown on the right. (**C**) Hi-C contact maps of chr10: 11-17 Mb at 5-kb resolution (left) and 12-14 Mb at 5-kb resolution (right), before and after hotspot deletion. (**D**) Two examples, *CELF2* and *RSU1*, for enhancer-promoter interactions altered after hotspot deletion within chr10: 11-17 Mb. (**E**) Hi-C contact maps of chr10: 68-76 Mb at 5-kb resolution (left) and chr10: 72-75 Mb at 5-kb resolution (right), before and after hotspot deletion. (**F**) Two examples, *FAM149B1* and *CCAR1*, for enhancer-promoter interactions altered after hotspot deletion within chr10: 68-76 Mb. In Figure 2.4D and 2.4F, Black dash line indicates decreased interactions in hotspot_10_25-deleted K562 cells, red dash line indicates enhanced interactions in hotspot_10_25-deleted K562 cells.

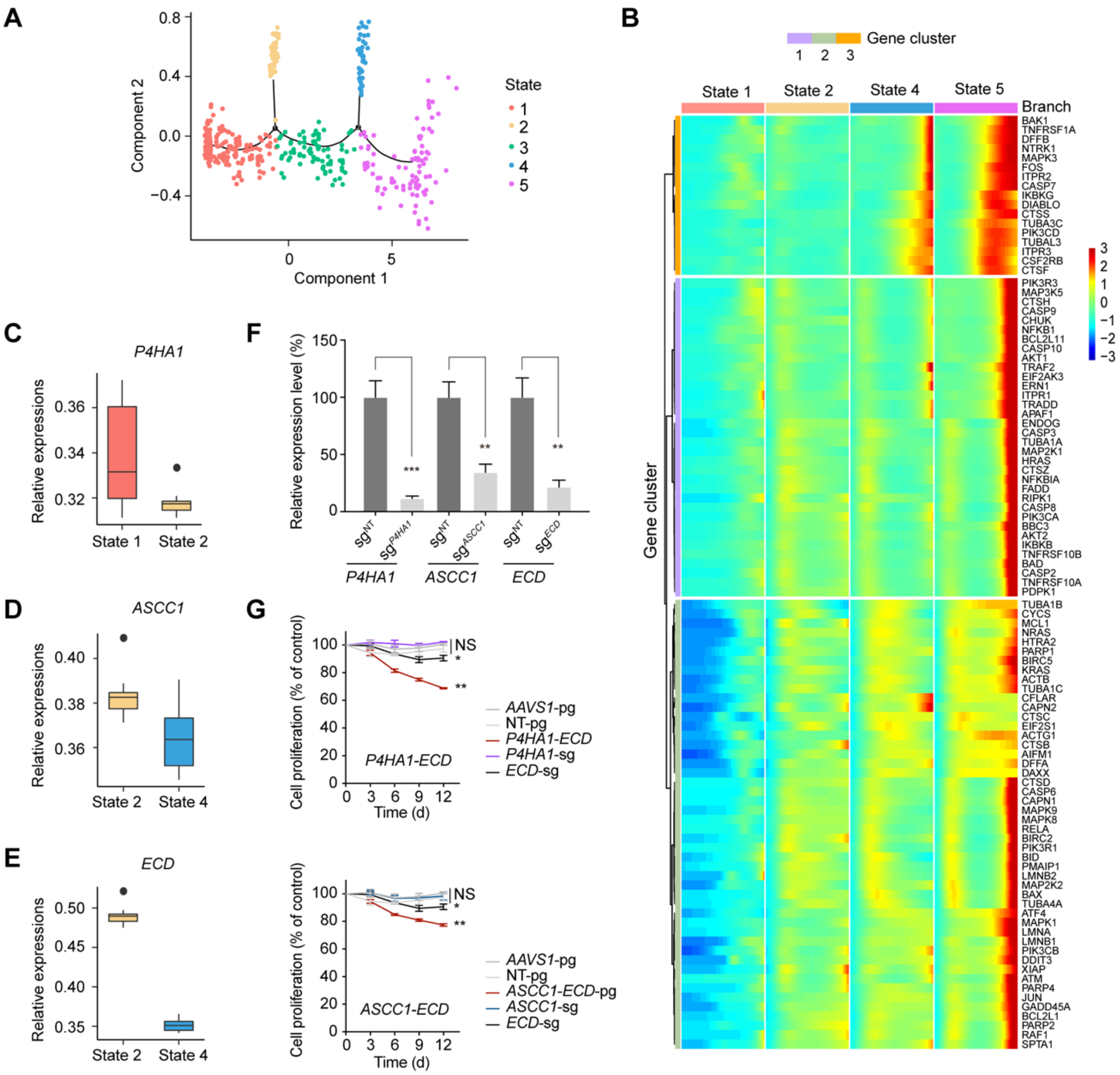


Figure 2.5. Synergistic change in gene expression after hotspot deletion. (**A**) Pseudotime clusters of hotspot_10_25-deleted and wild-type K562 cells based on apoptosis gene expression. (**B**) Global analysis of the expression levels of 99 KEGG apoptosis genes in state 1, 2, 4 and 5. Genes were clustered into 3 groups. (**C-E**) The relative expression levels of three representative downregulated genes *P4HA1, ASCC1, ECD* in different states as determined by single cell RNA-seq. (**F**) The knockdown efficiency of the indicated sgRNAs targeting each downregulated gene in K562 cells stably expressing KRAB-dCas9. The expression level of each gene was detected by real-time qPCR. sgRNA^{NT} represents the non-targeting sgRNA serving as the negative control. (**G**) Validation of the synergistic effects of two sets of gene pairs on K562 cell fitness by cell proliferation assay. Asterisk (*) represents *p-value* compared with pgRNAs targeting *AAVS1*-pg at Day 12, which were calculated by two-tailed Student's *t*-test and adjusted for multiple comparisons by Benjamini–Hochberg procedure. * $p < 0.05$; ** $p < 0.01$; NS, not significant.

2.8 Supplementary Figures

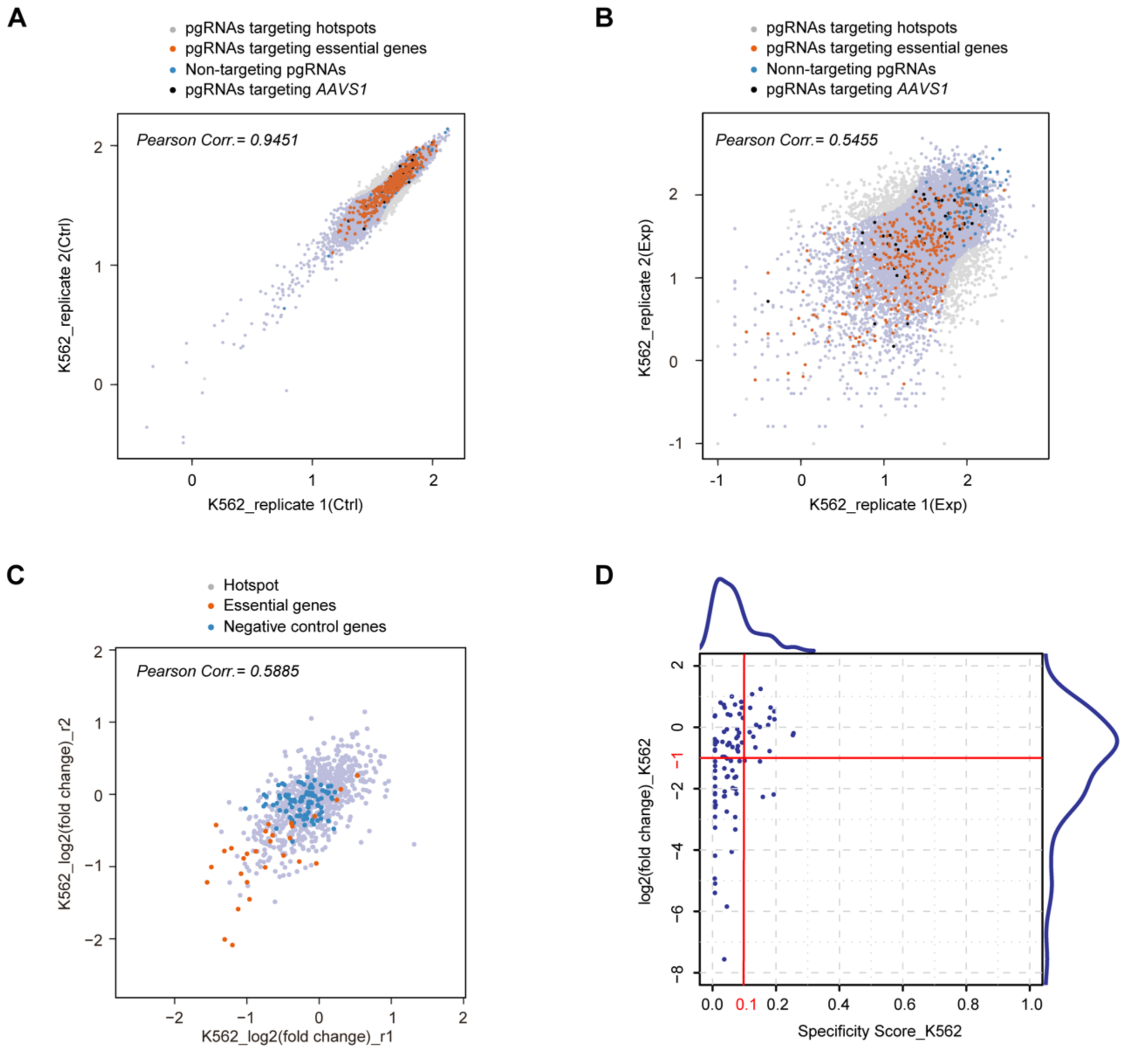


Figure S2.1. The correlations between replicates in the functional screening for essential hotspots in the K562 cell line. (**A-B**) Scatter plots of normalized pgRNA read counts of the hotspot libraries from the Day-0 control samples (Ctrl) (**A**) and Day-30 experimental samples (Exp) (**B**) in the K562 cell line. The light violet dots represent the pgRNAs targeting hotspots that passed the filter, and the light gray dots represent the pgRNAs that are filtered out. The *Pearson* correlation coefficients (*Pearson corr.*) of the two biologically independent replicates of each sample are also presented. (**C**) Scatter plots of pgRNA fold changes between the two biological replicates in K562 cells. (**D**) The distribution of pgRNAs targeting *AAVS1* with different $\log_2$(fold change) and specificity scores.

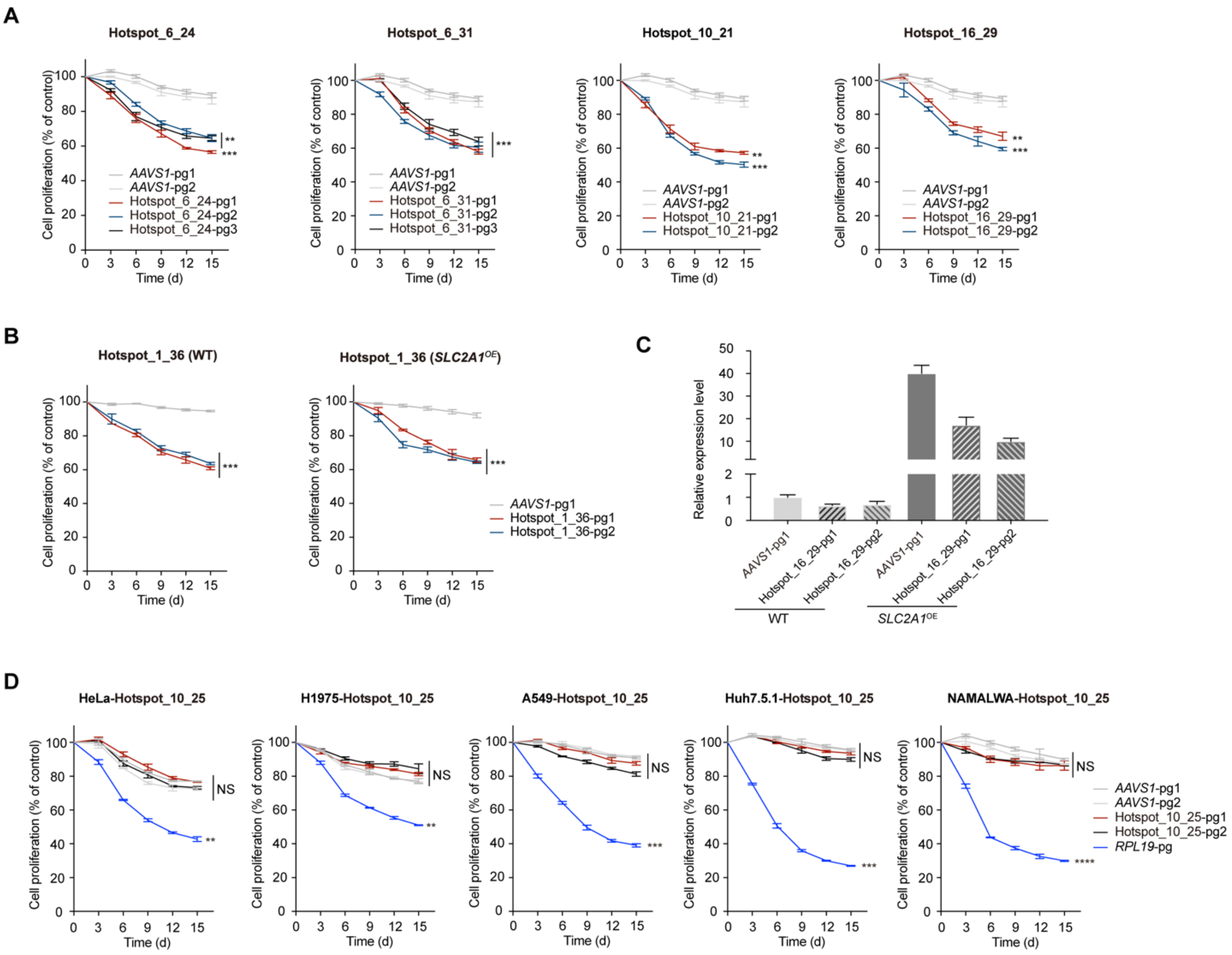
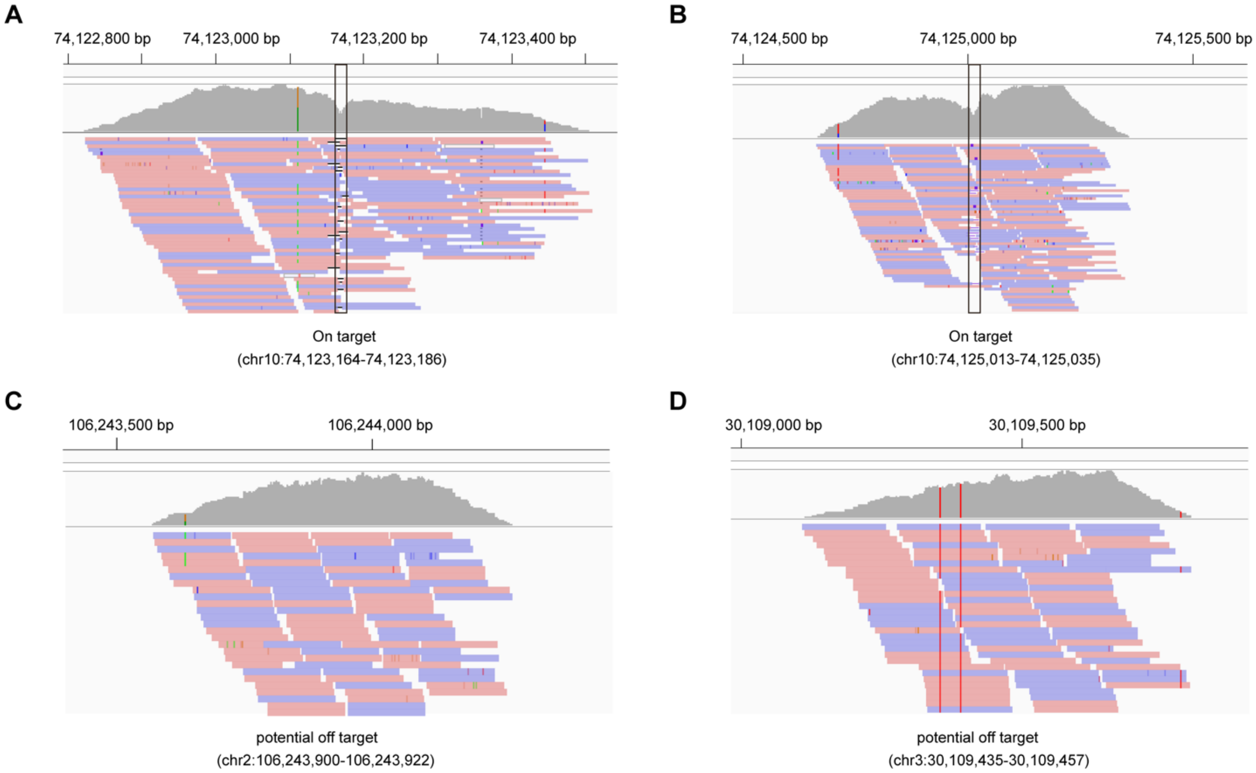


Figure S2.2. Validation of essential hotspots in K562 cells and multiple cancer cell lines as assessed by fragment deletion. (**A**) Validation of essential hotspots not overlapping with essential protein-coding genes in K562 cells. (**B**) Validation of essential hotspots overlapping with the promoter region of an essential gene in K562 cells by cell proliferation assay. Left: WT K562 cells infected with pgRNAs targeting hotspot_1_36. Right: *SLC2A1*-overexpressing K562 cells infected with pgRNAs targeting hotspot_1_36. The method for determining the effect of each hotspot on cell growth or proliferation was the same as that described in Figure 2.3A. Asterisks represent *p-values* compared with *AAVS1*_pg1 at Day 15, which were calculated by the same method as described in Figure 2.3A. (**C**) The expression levels of *SLC2A1* in WT and *SLC2A1*-overexpressing K562 cells transduced with pgRNAs targeting *AAVS1* or hotspot_1_36. (**D**) Validation of hotspot_19_32 in multiple cancer cell lines, including A549, H1975, HeLa, Huh7.5.1 and NAMALWA cell lines. The asterisks (*) represents the *p-values* compared with pgRNAs targeting *AAVS1*-pg1 at Day 15, which were calculated by the same method as described in Figure 2.3D. In Figure S2.2A-B and D, data are presented as the means ± s.d.s (n = 3 biologically independent samples). * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$; NS, not significant.



On target
(chr10:74,123,164-74,123,186)

On target
(chr10:74,125,013-74,125,035)

potential off target
(chr2:106,243,900-106,243,922)

potential off target
(chr3:30,109,435-30,109,457)

Figure S2.3. WGS analysis of the off-target effects of the pgRNA targeting hotspot_10_25 in the genome browser view. (A-B) Pileup reads for on-target loci chr10: 74,123,164-74,123,186 (**A**, one sgRNA target site) and chr10: 74,125,013-74,125,035 (**B**, another sgRNA target site) with clear cleavages as indicated in the frames. **(C-D)** Pileup reads for two examples of putative off-target loci chr10: 106,243,900-106,243,922 (**C**) and chr10: 30,109,435-30,109,457 (**D**) without any clear cleavages, indicating no off-target effects.

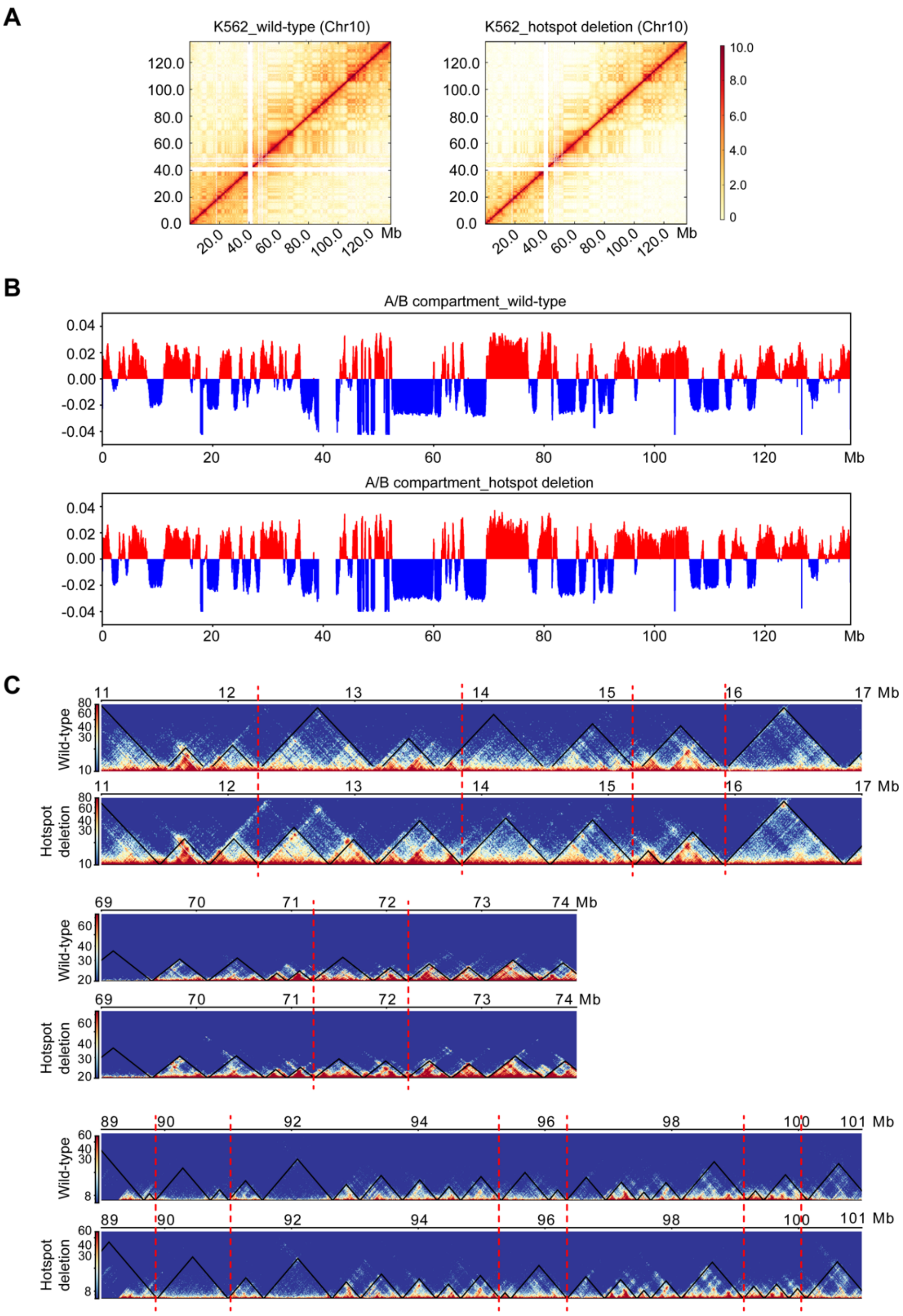


Figure S2.4. Hi-C contact maps, A/B compartments and TADs before and after hotspot_10_25 deletion. (**A**) Hi-C contact maps of the entire chr10 at 100-kb resolution. **(B)** A/B compartments at 50-kb resolution for wild-type and hotspot_10_25-deleted K562 cells in chr10. **(C)** TADs merge/split at 10-kb resolution for wild-type and hotspot_10_25-deleted K562 cells in chr10.

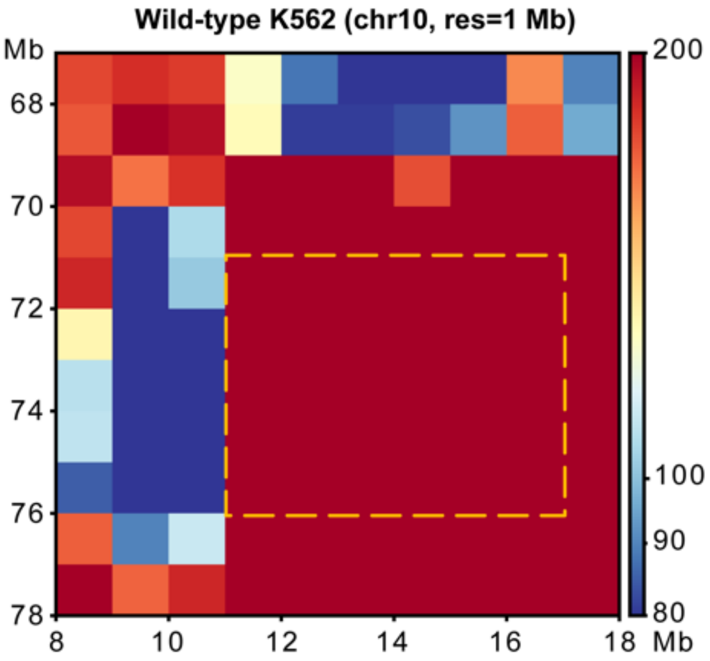


Figure S2.5. Contact map (chr10) of wild-type K562 at a 1-Mb resolution. The yellow box highlights the strong contact between chr10: 71-76 Mb and chr10: 11-17 Mb.

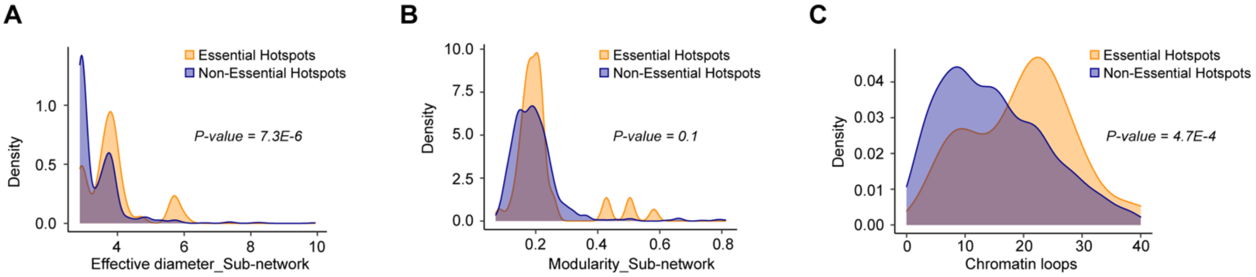


Figure S2.6. Distributions of the effective diameter (A), modularity score (B) and chromatin loops (C) in the 5-Mbp regions around the essential hotspots and non-essential hotspots.

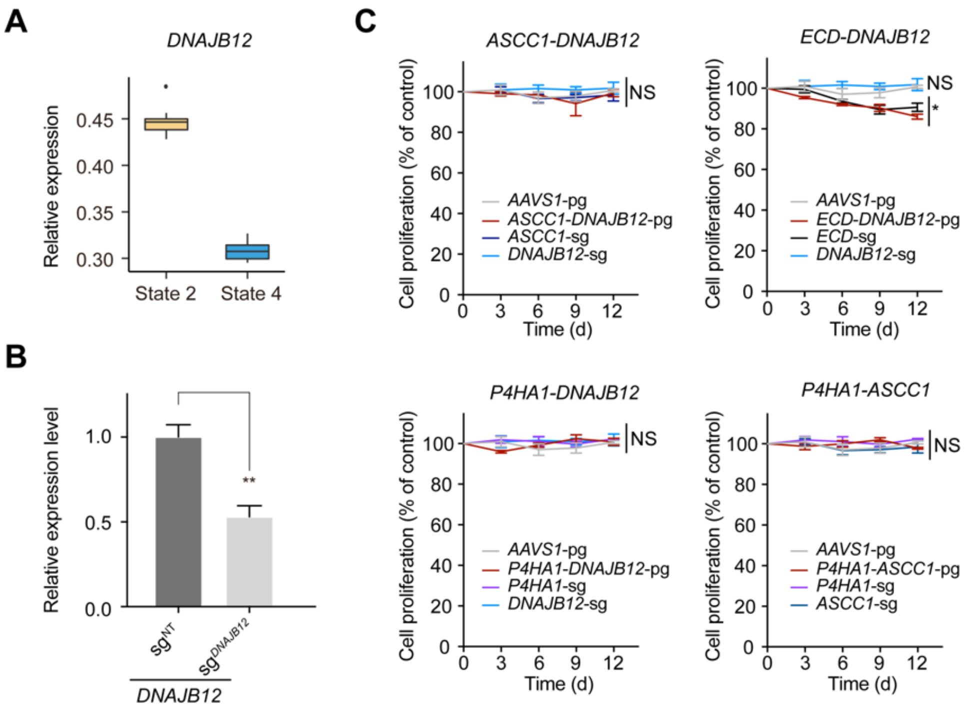


Figure S2.7. The growth effects of gene pairs that showed downregulated expression according to single-cell RNA-seq after hotspot_10_25 deletion in K562 cells. (**A**) The relative expression level of *DNAJB12* (from state 2 to state 4) in hotspot_10_25-deleted K562 cells from the single-cell RNA-seq results. (**B**) The knockdown efficiency of sgRNA targeting *DNAJB12* in K562 cells stably expressing KRAB-dCas9 protein. The expression levels of *DNAJB12* in K562 cells infected with non-targeting sgRNA and *DNAJB12*-targeted sgRNA were detected by real-time qPCR. sgRNA^{NT} represents the non-targeting sgRNA (serving as the negative control). (**C**) Validation of the gene pairs through a cell proliferation assay. The asterisks (*) represent the *p-values* compared with pgRNAs targeting *AAVS1*-pg at day 15, which were calculated by the same method as described in Figure 2.4F. Data are presented as the means ± s.d.s (n = 3 biologically independent samples). * $p < 0.05$; NS, not significant.

2.9 Supplementary Tables

Table S2.1. Summary of WGS analysis for hotspot_10_25 deletion in K562 cells

Cell Line	K562-WT	K562-del
Genome	hg19	hg19
Genome coverage	42x	26x
Mapping rate	94.39%	84.9%
SNV+INDEL (compared to hg19)	4.7 M	4.05 M
SNV+INDEL (compared to hg19) confirmation rate	86.20%	
Variations in putative off-target loci (compared to WT)	N/A	0

Table S2.2. Hi-C comparison between the wild-type and hotspot_10_25-deleted K562 cells

Chromosome	Compared.pairs	off-diagonal contacts with *p* value < 0.05	Percentage
chr1	3746079	8700	0.2322%
chr2	3415305	21238	0.6218%
chr3	2497235	20972	0.8398%
chr4	2526220	16836	0.6665%
chr5	2579300	6523	0.2529%
chr6	2230346	10784	0.4835%
chr7	2439371	15266	0.6258%
chr8	2077913	11813	0.5685%
chr9	945967	7	0.0007%
chr10	1655689	1287	0.0777%
chr11	1823573	13085	0.7175%
chr12	1660129	13771	0.8295%
chr13	951182	6723	0.7068%
chr14	916781	3969	0.4329%
chr15	884917	3029	0.3423%
chr16	922993	3601	0.3901%
chr17	765104	4354	0.5691%
chr18	974682	9741	0.9994%
chr19	583452	3101	0.5315%
chr20	675116	6192	0.9172%
chr21	384661	272	0.0707%
chr22	290993	878	0.3017%
chrX	1476247	9356	0.6338%

Table S2.3. Comparison of chromatin loops before/after hotspot deletion

Chromosome	Disappeared in hotspot deletion	Newly formed in hotspot deletion	Identified in both before/after hotspot deletion
chr1	266	1096	315
chr10	91	428	90
chr11	142	595	151
chr12	100	488	102
chr13	21	143	18
chr14	51	276	35
chr15	86	278	69
chr16	78	294	72
chr17	121	496	132
chr18	30	217	25
chr19	82	439	126
chr2	161	679	145
chr20	63	274	69
chr21	27	108	16
chr22	35	209	51
chr3	118	553	131
chr4	80	336	87
chr5	81	475	117
chr6	144	475	117
chr7	100	482	141
chr8	89	414	69
chr9	106	445	97
chrX	58	289	42

Table S2.4. Comparison of chromatin loops before/after hotspot deletion in chr10: 11-17 Mb and chr10: 68-76 Mb

Genomic Locus	Disappeared in hotspot deletion	Newly formed in hotspot deletion	Identified in both before/after hotspot deletion
Chr10: 11-17 Mb	4	22	7
Chr10: 68-76 Mb	15	61	15

2.10 References

1. R. Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Li, L. Schwarzfischer, D. Glatz, J. Raithel, B. Lilje, N. Rapin, F. O. Bagger, M. Jørgensen, P. R. Andersen, N. Bertin, O. Rackham, A. M. Burroughs, J. K. Baillie, Y. Ishizu, Y. Shimizu, E. Furuhata, S. Maeda, Y. Negishi, C. J. Mungall, T. F. Meehan, T. Lassmann, M. Itoh, H. Kawaji, N. Kondo, J. Kawai, A. Lennartsson, C. O. Daub, P. Heutink, D. A. Hume, T. H. Jensen, H. Suzuki, Y. Hayashizaki, F. Müller, A. R. R. Forrest, P. Carninci, M. Rehli, A. Sandelin, An atlas of active enhancers across human cell types and tissues. *Nature*. **507**, 455–461 (2014).

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

3. N. D. Tippens, A. Vihervaara, J. T. Lis, Enhancer transcription: what, where, when, and why? *Genes Dev.* **32**, 1–3 (2018).

4. J. L. Plank, A. Dean, Enhancer function: mechanistic and genome-wide insights come together. *Mol. Cell*. **55**, 5–14 (2014).

5. J. van Arensbergen, B. van Steensel, H. J. Bussemaker, In search of the determinants of enhancer-promoter interaction specificity. *Trends Cell Biol.* **24**, 695–702 (2014).

6. J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, B. E. Bernstein, Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*. **473**, 43–49 (2011).

7. I. Chepelev, G. Wei, D. Wangsa, Q. Tang, K. Zhao, Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. *Cell Res.* **22**, 490–503 (2012).

8. J. Fitz, T. Neumann, M. Steininger, E.-M. Wiedemann, A. C. Garcia, A. Athanasiadis, U. E. Schoeberl, R. Pavri, Spt5-mediated enhancer transcription directly couples enhancer activation with physical promoter interaction. *Nat. Genet.* **52**, 505–515 (2020).

9. A. Pombo, N. Dillon, Three-dimensional genome architecture: players and mechanisms. *Nat. Rev. Mol. Cell Biol.* **16**, 245–257 (2015).

10. H. K. Long, S. L. Prescott, J. Wysocka, Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell*. **167**, 1170–1187 (2016).

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

12. Y. Ghavi-Helm, F. A. Klein, T. Pakozdi, L. Ciglar, D. Noordermeer, W. Huber, E. E. M. Furlong, Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*. **512**, 96–100 (2014).

13. A. Sanyal, B. R. Lajoie, G. Jain, J. Dekker, The long-range interaction landscape of gene promoters. *Nature*. **489**, 109–113 (2012).

14. S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, E. L. Aiden, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. **159**, 1665–1680 (2014).

15. A. S. Hansen, I. Pustova, C. Cattoglio, R. Tjian, X. Darzacq, CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *Elife*. **6** (2017), doi:10.7554/eLife.25776.

16. Y. Shen, F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, B. Ren, A map of the cis-regulatory sequences in the mouse genome. *Nature*. **488**, 116–120 (2012).

17. R. Siersbæk, J. G. S. Madsen, B. M. Javierre, R. Nielsen, E. K. Bagge, J. Cairns, S. W. Wingett, S. Traynor, M. Spivakov, P. Fraser, S. Mandrup, Dynamic Rewiring of Promoter-Anchored Chromatin Loops during Adipocyte Differentiation. *Mol. Cell*. **66**, 420–435.e5 (2017).

18. P. Morcillo, C. Rosen, M. K. Baylies, D. Dorsett, Chip, a widely expressed chromosomal protein required for segmentation and activity of a remote wing margin enhancer in Drosophila. *Genes Dev.* **11**, 2729–2740 (1997).

19. W. Deng, J. Lee, H. Wang, J. Miller, A. Reik, P. D. Gregory, A. Dean, G. A. Blobel, Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*. **149**, 1233–1244 (2012).

20. S. Kong, D. Bohl, C. Li, D. Tuan, Transcription of the HS2 enhancer toward a cis-linked gene is independent of the orientation, position, and distance of the enhancer relative to the gene. *Mol. Cell. Biol.* **17**, 3955–3965 (1997).

21. C. Rickman, W. A. Bickmore, Transcription. Flashing a light on the spatial organization of transcription. *Science*. **341** (2013), pp. 621–622.

22. S. S. Teves, L. An, A. S. Hansen, L. Xie, X. Darzacq, R. Tjian, A dynamic mode of mitotic bookmarking by transcription factors. *Elife*. **5** (2016), doi:10.7554/eLife.22280.

23. Z. Liu, W. R. Legant, B.-C. Chen, L. Li, J. B. Grimm, L. D. Lavis, E. Betzig, R. Tjian, 3D imaging of Sox2 enhancer clusters in embryonic stem cells. *Elife*. **3**, e04236 (2014).

24. S. Heinz, L. Texari, M. G. B. Hayes, M. Urbanowski, M. W. Chang, N. Givarkes, A. Rialdi, K. M. White, R. A. Albrecht, L. Pache, I. Marazzi, A. García-Sastre, M. L. Shaw, C. Benner, Transcription Elongation Can Affect Genome 3D Structure. *Cell*. **174**, 1522–1536.e22 (2018).

25. Y. Zhu, Z. Chen, K. Zhang, M. Wang, D. Medovoy, J. W. Whitaker, B. Ding, N. Li, L. Zheng, W. Wang, Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.* **7**, 10812 (2016).

26. Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis, Integrative analysis of 111 reference human epigenomes. *Nature*. **518**, 317–330 (2015).

27. S. Zhu, W. Li, J. Liu, C.-H. Chen, Q. Liao, P. Xu, H. Xu, T. Xiao, Z. Cao, J. Peng, P. Yuan, M. Brown, X. S. Liu, W. Wei, Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR-Cas9 library. *Nat. Biotechnol.* **34**, 1279–1286 (2016).

28. A. R. Perez, Y. Pritykin, J. A. Vidigal, S. Chhangawala, L. Zamparo, C. S. Leslie, A. Ventura, GuideScan software for improved single and paired CRISPR guide RNA design. *Nat. Biotechnol.* **35**, 347–349 (2017).

29. N. C. Durand, M. S. Shamim, I. Machol, S. S. P. Rao, M. H. Huntley, E. S. Lander, E. L. Aiden, Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst*. **3**, 95–98 (2016).

30. F. Ramírez, V. Bhardwaj, L. Arrigoni, K. C. Lam, B. A. Grüning, J. Villaveces, B. Habermann, A. Akhtar, T. Manke, High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, 189 (2018).

31. J. Wolff, V. Bhardwaj, S. Nothjunge, G. Richard, G. Renschler, R. Gilsbach, T. Manke, R. Backofen, F. Ramírez, B. A. Grüning, Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **46**, W11–W16 (2018).

32. K. C. Akdemir, L. Chin, HiCPlotter integrates genomic data with interaction matrices. *Genome Biol.* **16**, 198 (2015).

33. D. Lin, J. Sanders, W. S. Noble, HiCRep.py : Fast comparison of Hi-C contact matrices in Python. *Bioinformatics* (2021), doi:10.1093/bioinformatics/btab097.

34. T. Yang, F. Zhang, G. G. Yardımcı, F. Song, R. C. Hardison, W. S. Noble, F. Yue, Q. Li, HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. *Genome Res.* **27**, 1939–1949 (2017).

35. J. C. Stansfield, K. G. Cresswell, V. I. Vladimirov, M. G. Dozmorov, HiCcompare: an R-package for joint normalization and comparison of HI-C datasets. *BMC Bioinformatics*. **19** (2018), , doi:10.1186/s12859-018-2288-x.

36. D. Kim, S. Bae, J. Park, E. Kim, S. Kim, H. R. Yu, J. Hwang, J.-I. Kim, J.-S. Kim, *Nat. Methods*, in press.

37. C. Smith, A. Gore, W. Yan, L. Abalde-Atristain, Z. Li, C. He, Y. Wang, R. A. Brodsky, K. Zhang, L. Cheng, Z. Ye, Whole-genome sequencing analysis reveals high specificity of CRISPR/Cas9 and TALEN-based genome editing in human iPSCs. *Cell Stem Cell*. **15**, 12–13 (2014).

38. B. Ding, Y. Liu, Z. Liu, L. Zheng, P. Xu, Z. Chen, P. Wu, Y. Zhao, Q. Pan, Y. Guo, W. Wei, W. Wang, Noncoding loci without epigenomic signals can be essential for maintaining global chromatin organization and cell viability. *Sci Adv*. **7**, eabi6020 (2021).

39. M. Pertea, D. Kim, G. M. Pertea, J. T. Leek, S. L. Salzberg, Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).

40. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).

41. D. Kim, B. Langmead, S. L. Salzberg, HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*. **12**, 357–360 (2015).

42. M. Pertea, G. M. Pertea, C. M. Antonescu, T.-C. Chang, J. T. Mendell, S. L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

43. E. Z. Macosko, A. Basu, R. Satija, J. Nemesh, K. Shekhar, M. Goldman, I. Tirosh, A. R. Bialas, N. Kamitaki, E. M. Martersteck, J. J. Trombetta, D. A. Weitz, J. R. Sanes, A. K. Shalek, A. Regev, S. A. McCarroll, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*. **161**, 1202–1214 (2015).

44. M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: new

perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).

45. X. Qiu, A. Hill, J. Packer, D. Lin, Y.-A. Ma, C. Trapnell, Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods*. **14**, 309–315 (2017).

46. C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, J. L. Rinn, The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32**, 381–386 (2014).

47. W. Haynes, Wilcoxon Rank Sum Test. *Encyclopedia of Systems Biology* (2013), pp. 2354–2355.

48. M. A. Horlbeck, L. A. Gilbert, J. E. Villalta, B. Adamson, R. A. Pak, Y. Chen, A. P. Fields, C. Y. Park, J. E. Corn, M. Kampmann, J. S. Weissman, Compact and highly active next-generation libraries for CRISPR-mediated gene repression and activation. *Elife*. **5** (2016), doi:10.7554/eLife.19760.

49. A. R. Benson, D. F. Gleich, J. Leskovec, Higher-order organization of complex networks. *Science*. **353**, 163–166 (2016).

50. W. W. R. Ball, H. S. M. Coxeter, *Mathematical Recreations and Essays* (Courier Corporation, 1987).

51. R. Cohen, S. Havlin, *Complex Networks: Structure, Robustness and Function* (Cambridge University Press, 2010).

52. C. Grabow, S. Grosskinsky, J. Kurths, M. Timme, Collective relaxation dynamics of small-world networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **91**, 052815 (2015).

53. S. J. Liu, M. A. Horlbeck, S. W. Cho, H. S. Birk, M. Malatesta, D. He, F. J. Attenello, J. E. Villalta, M. Y. Cho, Y. Chen, M. A. Mandegar, M. P. Olvera, L. A. Gilbert, B. R. Conklin, H. Y. Chang, J. S. Weissman, D. A. Lim, CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science*. **355** (2017), doi:10.1126/science.aah7111.

54. Y. Liu, Z. Cao, Y. Wang, Y. Guo, P. Xu, P. Yuan, Z. Liu, Y. He, W. Wei, Genome-wide screening for functional long noncoding RNAs in human cells by Cas9 targeting of splice sites. *Nat. Biotechnol.* (2018), doi:10.1038/nbt.4283.

55. D. M. Munoz, P. J. Cassiani, L. Li, E. Billy, J. M. Korn, M. D. Jones, J. Golji, D. A. Ruddy, K. Yu, G. McAllister, A. DeWeck, D. Abramowski, J. Wan, M. D. Shirley, S. Y. Neshat, D. Rakiec, R. de Beaumont, O. Weber, A. Kauffmann, E. R. McDonald, N. Keen, F. Hofmann, W. R. Sellers, T. Schmelzle, F. Stegmeier, M. R. Schlabach, CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities

but Generate False-Positive Hits for Highly Amplified Genomic Regions. *Cancer Discovery*. **6** (2016), pp. 900–913.

56. A. J. Aguirre, R. M. Meyers, B. A. Weir, F. Vazquez, C.-Z. Zhang, U. Ben-David, A. Cook, G. Ha, W. F. Harrington, M. B. Doshi, M. Kost-Alimova, S. Gill, H. Xu, L. D. Ali, G. Jiang, S. Pantel, Y. Lee, A. Goodale, A. D. Cherniack, C. Oh, G. Kryukov, G. S. Cowley, L. A. Garraway, K. Stegmaier, C. W. Roberts, T. R. Golub, M. Meyerson, D. E. Root, A. Tsherniak, W. C. Hahn, Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. *Cancer Discov.* **6**, 914–929 (2016).

57. D. W. Morgens, M. Wainberg, E. A. Boyle, O. Ursu, C. L. Araya, C. K. Tsui, M. S. Haney, G. T. Hess, K. Han, E. E. Jeng, A. Li, M. P. Snyder, W. J. Greenleaf, A. Kundaje, M. C. Bassik, Genome-scale measurement of off-target activity using Cas9 toxicity in high-throughput screens. *Nat. Commun.* **8**, 15178 (2017).

58. J. Tycko, M. Wainberg, G. K. Marinov, O. Ursu, G. T. Hess, B. K. Ego, Aradhana, A. Li, A. Truong, A. E. Trevino, K. Spees, D. Yao, I. M. Kaplow, P. G. Greenside, D. W. Morgens, D. H. Phanstiel, M. P. Snyder, L. Bintu, W. J. Greenleaf, A. Kundaje, M. C. Bassik, Identification and mitigation of pervasive off-target activity in CRISPR-Cas9 screens for essential non-coding elements: Supplementary Information, , doi:10.1101/520569.

59. The ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature*. **489**, 57–74 (2012).

60. S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, E. L. Aiden, A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*. **162** (2015), pp. 687–688.

Chapter 3. Regulation associated modules reflect 3D genome modularity associated with chromatin activity

3.1 Abstract

The 3D genome has been shown to be organized into modules including topologically associating domains (TADs) and compartments that are primarily defined by spatial contacts from Hi-C or other experiments. There exists a gap to investigate whether and how the spatial modularity of the chromatin is related to the functional modularity resulting from the chromatin activity. Increasing evidence shows a tight interplay between histone modifications and 3D chromatin organization. As the histone modifications reflect the chromatin activity, it is tempting to infer the spatial modularity of the genome directly from the histone modification patterns, which would establish the connection between the spatial and functional modularity of the genome. However, uncovering the 3D genomic modules using histone modifications has not been well explored. Here, we report that the histone modifications show a modular pattern (referred to as regulation associated modules, RAMs) that reflects the spatial modularity of the chromatin structure. We found that enhancer-promoter interactions and extrachromosomal DNAs (ecDNAs) occur more often within the same RAMs than within the same TADs, indicating stronger insulation of the RAM boundaries and a modularization of the 3D genome at a scale better aligned with the chromatin activity. Consistently, compared to the TAD boundaries, in silico predictions showed that deletions of RAM boundaries perturb the chromatin structure more severely and somatic variants in the cancer samples are more enriched in the RAM boundaries. These observations suggest that RAMs reflect a modular organization of the

3D genome at a scale better aligned with chromatin activity, providing a bridge connecting the structural and functional modularity of the genome.

3.2 Introduction

Histone modifications are critical to shape the chromatin structure and regulate gene expression(*1*, *2*). Active marks such as H3K27ac and H3K4me3 open up chromatin to allow access of transcription factors (TFs) and transcription machinery to promoters or enhancers. Repressive marks such as H3K9me3 and H3K27me3 condensate chromatin to block TF binding and suppress gene expression. DNA marked by active and repressive histone modifications form euchromatin and heterochromatin that are distinct on the compactness. These observations suggest that histone modifications have an important impact on organizing the regional and global 3D genome.

Accumulating evidence has revealed the association of histone modifications with the topologically associating domains (TADs)(*3*–*6*) and compartments(*7*, *8*) derived from the Hi-C contact maps showing plaid patterns. TADs represent genomic domains forming dense internal contacts but fewer contacts with neighboring regions. The TAD boundaries are demarcated with CTCF sites or active transcribed DNA sequences. The Hi-C data also shows that the 3D genome is partitioned into transcriptionally active (compartment A) and suppressed (compartment B) compartments. Active and repressive histone marks are enriched, but do not exclusively appear, in the A and B compartments, respectively(*7*, *8*). Computational models have shown that histone modification signals are predictive of Hi-C contacts particularly for enhancer-promoter interactions(*9*), TAD boundaries(*10*) and compartments(*11*). Histone modifications are tightly associated with transcriptional activity(*12*–*14*) while transcription and proteins involved in transcriptional regulation including RNA polymerase and TFs have been shown to contribute to compartmentation and active promoters and enhancers tend to form clusters in the nucleus(*15–18*).

Despite the mechanisms underlying the interplay between histone modifications and chromatin organization remain elusive, histone modifications can indicate the spatial organization of the genome as readout signals for regulatory modules. However, the current analysis has been limited to associating histone marks to Hi-C derived TADs and compartments. An unfilled gap is to use histone modifications to directly elucidate the modular organization of the 3D genome. We propose here to define the spatial module of the genome organization resulting from the chromatin activities reflected by histone modifications.

We found that the frequency profiles of the H3K27ac peaks present a modular structure (referred to as regulation associated modules, RAMs). A large number of these modules are shared across cell types and can be independently derived using other active histone marks, including H3K4me3 and H3K4me1. We uncovered several lines of evidence to support the hypothesis that the RAMs are spatial modules resulting from functional activities: the enhancer-promoter interactions dominantly occur within RAMs; the extrachromosomal DNAs (ecDNAs) tend to be originated from the same RAMs rather than split in multiple RAMs; RAMs are resistant to cohesin degradation. These properties of RAMs distinguish them from TADs and compartments. Furthermore, deletion of the RAM boundaries is predicted to alter the chromatin organization more significantly than the deletion of TAD boundaries. Consistently, the somatic genetic variations in cancer patients are enriched in RAM boundaries, suggesting a possible mechanism of tumorigenesis involved in altering the chromatin modules.

3.3 Materials and Methods

3.3.1 Regulatory associated modules (RAM) identification

Data Source. The 93 normal and 19 cancer samples with the called H3K27ac narrow peaks in hg19 were downloaded from Roadmap Epigenomics portal (https://egg2.wustl.edu/roadmap/web_portal/)(*19*) and ENCODE portal (https://www.encodeproject.org/). **Table S3.1** lists the samples used in this study.

RAM identification in individual samples. We calculated the H3K27ac narrow-peak density using a sliding window of step size equal to 10kb, 50kb, 100kb, 250kb and 500kb respectively and 500kb flanking size of each window in every sample. The H3K27ac narrow-peak profiles were then smoothed by a local polynomial regression fitting(*20*). The RAM boundaries (valley/minima on the smoothing curves) and peaks (summit/maxima on the smoothing curve) in the smoothed curves were then detected using the "findpeaks" function in R package "pracma".

Consensus-RAM (cRAM) identification. We first identified RAM boundaries in the 93 normal or 19 cancer samples using different step sizes (10kb, 50kb, 100kb, 250kb and 500kb) and then counted the percentage of a genomic region identified as RAM boundary in the 93 normal or 19 cancer samples. A genomic region with occurring percentage >=25% was considered as a consensus-RAM (cRAM) boundary in the normal or cancer samples. We merged cRAM boundaries if they are located <250kb apart from each other and required cRAMs have size >250kb.

3.3.2 Hi-C data analysis

Hi-C processing. The Hi-C data for the wildtype K562, GM12878, A549, IMR90, NHEK, HUVEC, HMEC and the HCT116 cell lines were downloaded from GEO (GSE63525) and the ENCODE portal (ENCSR662QKG and GSE104333). All the raw fastq files were aligned to hg19 genome and then processed using Juicer with the default settings(*7*, *21*). The contact reads in a given cell line were further normalized by vanilla coverage (VC) normalization using the Juicer pipeline. The significance for a given fragment contact was computed by Poisson distribution with VC-normalized expected contact reads versus the VC-normalized observed contact reads. We then used HiCExplorer(*22–24*) and HiCplotter(*25*) software to visualize the Hi-C data.

A/B compartment. We performed A/B compartment analysis at 250kb resolution. The eigenvectors for each chromosome in all the cell lines involved in the Hi-C data analysis were extracted from the VC normalized Hi-C counts processed by the Juicer pipeline with the default parameters(*26*). POLR2A ChIP-seq data were obtained from the ENCODE(*27*) portal (https://www.encodeproject.org/). To determine A or B compartment, we calculated the correlation between the first eigenvector of each chromosome and the Pol II peak density(*28*). As there was no Pol II ChIP-seq data available for HMEC, we used TSS density in the hg19 genome to assign A/B compartments in HMEC.

Topological associated domains (TAD). To identify topological associated domains (TAD), we applied the insulation score method(*29*) to the Hi-C data at 50kb resolution in the K562, GM12878, A549, IMR90, NHEK, HUVEC, HMEC and HCT116 cell lines. The HFF cell line TADs were downloaded from the 4DN website (accession number 4DNFIMROE6N4). The H1 cell line TADs were downloaded from ref(*4*).

3.3.3 Lamina-associated domains (LADs) data processing

The Lamin-B1 signal data in the K562, HCT116, H1, HAP1, RPE-hTERT, HFFc6 cell lines generated by the DamID technique were obtained from the 4DN portal(*30*). We downloaded the mean of the replicates for each cell line. The Lamin-B1 signals at 50kb resolution for the K562, HCT116, H1, HAP1, RPE-hTERT, HFFc6 cell lines were lifted over from hg38 to hg19.

3.3.4 Cohesin degradation analysis.

The H3K27ac ChIP-seq data in the untreated HCT-116 RAD21-mAC cells and HCT-116 RAD21-mAC cells treated for 6 hours with IAA were downloaded from GEO (GSE104888). We processed the H3K27ac data same as ref(*31*). In brief, we aligned the raw data to the hg19 human genome using the BWA software(*32*), and then deduplicated the reads using PicardTools. The narrow peaks were called by comparing the associated input data using MACS2(*33*). All the parameters were set to the defaults.

3.3.5 Enrichment of upregulated genes in enhancer-promoter pairs occurring in the same RAM of K562 but split in the background cell type by Hypergeometric Test.

The hypergeometric test was employed to measure the significance of the upregulated genes involved in the enhancer-promoter pairs occurring in the same RAM in K562 (foreground cell type) but in different RAMs of the background cell type. The population size N was the overall genes involved in the K562 and the compared cell type. The population success size M was the number of all upregulated genes in the K562 compared to the background cell type. The sampling size n was the number of the genes involved in the enhancer-promoter pairs occurring in the same RAM of K562 but in different RAMs of the background cell type, and the sampling success size m was the upregulated genes involved in the enhancer-promoter pairs occurring in the same RAM of K562 but in different RAMs of the background cell type. Enrichment was considered significant if p-value<0.05.

3.3.6 Enrichment of the somatic variants in cancer cRAM boundaries compared to the TAD boundaries assessed by Two-sample Proportion Tests.

The two-sample proportion test null hypothesis was to test the equal proportion of the number of the somatic variants relative to the genome coverage (in base pair) in the cancer cRAM boundaries and TAD boundaries. The two proportions were calculated separately by the number of the somatic variants divided by the boundary length (in base pair) for cancer cRAM boundaries and TAD boundaries. Enrichment was considered significant if p-value<0.05.

3.3.7 Somatic mutations and structural variation analysis for cancer patients from PCAWG

The consensus somatic SNV and indels were downloaded from PCAWG. (https://dcc.icgc.org/releases/PCAWG)(*34*, *35*). The VCF files were transformed to bed files by BEDOPS vcf2bed tools(*36*). The number of somatic SNV and indels overlapping with the RAM and TAD boundaries were then counted.

3.3.8 Motif analysis

The motif analysis was done using the Homer pipeline(*37*) with default parameters. The motif occurrence was called using FIMO(*38*) with p-value <=1E-4.

3.4 Results

3.4.1 Regulation associated modules (RAMs) detected by the histone modification peaks

We analyzed the density profile of H3K27ac peaks (i.e. peak count in a sliding window) from chromatin immunoprecipitation assays with sequencing (ChIP-Seq) experiments as using the peak density instead of read count density can better remove the noise from the background signals. We downloaded ChIP-seq data of 93 normal and 19 cancer samples from Roadmap Epigenomics Project (http://www.roadmapepigenomics.org/)(*19*) and ENCODE portal (https://www.encodeproject.org/)(*39*) (**Table S3.1**). Using a sliding window (a fixed flanking size of 500kbp and step size varying from 10kbp to 500kbp), we computed the H3K27ac peak densities in the linear genome. Regardless of the step size, the H3K27ac peak densities were not evenly distributed and showed a modular pattern (**Figure 3.1A**). The active marks of H3K4me1, H3K4me3 and H3K36me3 showed similar peak density profiles to H3K27ac in the 93 samples, indicated by high Pearson correlations between them, whereas the repressive marks of H3K27me3 and H3K9me3 had less consistent patterns (**Figure 3.1B**). Given the highly correlated active mark patterns, we focused on analyzing the H3K27ac signals as the other active marks show similar modular structure.

At a given step size, we identified the valley or minima of the H3K27ac peak profile that was smoothed using local polynomial fit in each chromosome and in each cell type (see **Materials and Methods**). These valleys demarcated the boundaries of the modular domains (called Regulation Associated Domains or RAMs). We varied the step size from

10kbp to 500kbp and fixed the window size to 500kbp. It is not surprising that, with the increasing step size, the RAM size increased and a higher percentage of RAMs were shared between samples (**Figure S3.1, Figure S3.2**). We observed that the number of common RAMs in all the chromosomes reached a plateau at 250kbp step size in both normal and cancer samples, which indicates the identified RAMs are most conserved across diverse cell types (**Figure S3.3A-B**). We thus used this step size of 250kbp for the remaining analyses. A RAM boundary is called consensus RAM (cRAM) boundary if it is shared by >25% of the samples. This way, 711 cRAMs were detected in the normal samples and 771 cRAMs in the cancer samples (see **Materials and Methods**). On average, 60% of the RAMs in a cell type are consensus (referred to as cRAMs) and the remaining cell-type specific (**Figure S3.3C**).

One example of the identified RAMs in chr12 of the GM12878 cell by IGV software(*40*) is shown in **Figure 3.1C-E**. Obviously, the RAM boundaries have lower signals of the active histone marks (H3K27ac, H3K4me1, H3K4me3, H3K36me3) and higher repressive marks (H3K9me3 and H3K27me3) compared to the within RAM regions. Consistently, they tend to align with the B compartment or subcompartments (B1, B2, B3). Furthermore, by counting the number of the 3D contacting neighbors for each locus using the 10kb resolution Hi-C data in GM12878 (contacts with log(P-value) <=-10), we found that the RAM boundaries tend to harbor many 3D contacts, indicated an enrichment with densely packed DNA sequences forming many spatial contacts (**Figure 3.1C**).

3.4.2 Characterization of the consensus regulation associated modules (cRAMs)

As cRAMs are largely shared between diverse cell lines, we further characterized them. Among the cell lines that have both active and repressive marks (all are normal cells), as expected, we found that the cRAM boundaries have lower peak density of the active marks (H3K27ac, H3K4me3, H3K4me1 and H3K36me3) and slightly higher peak densities of repressive marks (H3K27me3 and H3K9me3) than the cRAM regions (see examples in GM12878 and HUVEC cell lines in **Figure 3.2A**). To quantify the difference of the histone modifications among the cRAM boundaries and the non-boundary regions, we counted the peak density using a sliding window, and compared the histone modifications enrichment across 93 normal samples. The P-value < 0.05 from the Wilcoxon Rank Sum test indicated that cRAM boundaries have significantly lower active marks and higher repressive marks than the non-boundaries of cRAMs (**Figure 3.2B-G**). Furthermore, using the available 10kb-resolution Hi-C data in the K562, GM12878, A549, IMR90, NHEK, HUVEC, HMEC and HCT116 cell lines, we found that the cRAM boundaries have significantly more Hi-C contacts (intrachromosomal contacts with log(P-value)<=-10) compared to the whole genome (**Figure 3.2H**), which is consistent with the genome browser view for any RAM boundary in **Figure 3.1C**. These observations suggested that the cRAM boundaries are formed by densely packed DNA sequences harboring many 3D contacts.

We next investigated how RAMs are related to the previously identified chromatin modules. First, the median size of cRAMs (~3.3Mbp) is larger than TADs (~600kbp) and one RAM often spans across multiple TADs (**Figure 3.2I**). Second, using the Hi-C data, we identified the A/B compartments at 250kb resolution (see **Materials and Methods**). We calculated the percentage of the A and B compartments in each cRAM (250kb bin size) across the cell types. While the A-compartments account around 50%-75% in each of the cRAM, a single cRAM is largely composed by a mixture of A and B compartments, indicating a distinction between cRAMs and compartments (**Figure 3.2J**). Consistently, the cRAM boundaries are enriched with B-compartment but also with a significant portion of A compartments (**Figure 3.2J**). Third, we checked the Lamin-B1 signals for the cRAM boundaries. Lamin-B1 is a scaffolding component of the nuclear envelope(*41*, *42*). A positive signal for Lamin-B1 suggests a close distance to the nuclear lamina, which could be used to define lamina associated domain (LAD). When aligning the cRAM boundaries with the Lamin-B1 signals (see **Materials and Methods**), we found on average 69% of the cRAM boundaries overlapping with Lamin-B1 signals across the cell types and meanwhile on average 62.7% of the LADs identified from each cell type overlapping with cRAM boundaries (**Figure 3.2K**), indicating that LADs and cRAMs are also different. Taken together, the cRAM boundaries are formed by densely packed DNA sequences; while they are enriched with B compartment and Lamin-B1 signals, RAMs are clearly distinct from the previously reported domain structures such as TADs, LADs and A/B compartments.

3.4.3 RAMs are functional units

If RAMs are functional modules, we reason that the majority of the promoter-enhancer interactions should occur within the same RAMs. We downloaded 970 high-confidence promoter-enhancer interactions in the K562 cell line that were experimentally validated in ref(*43*, *44*) and 95% of them are located within the same RAMs, compared to 75% of them in the same TADs(*4*, *6*, *8*, *45*) (**Figure 3.3A-B**). Two examples of promoter-enhancer interactions are shown in **Figure 3.3G**: the enhancer-promoter interactions of *STEAP1B* and *VGF* are across multiple TADs but within the same K562 RAMs marked by continuous strong H3K27ac peaks. This observation suggests that RAMs may represent regulatory modules and RAM boundaries insulate promoter-enhancer contacts across RAMs at a scale more appropriate than TADs to capture functional modularity of chromatin activity.

We further investigated how the modularity defined by RAMs affects gene expression. To this end, we examined whether the enhancer-promoter pairs located within the same K562 RAMs but separated by RAM boundaries in other cells would specifically impact gene expression in K562. When comparing K562 to the normal cell line GM12878, we found 885 K562 enhancer-promoter interactions were within the same RAMs in both cell lines and 39 only in K562 (**Figure 3.3C**). The majority of the genes regulated by the 39 enhancer-promoter interactions are upregulated in K562 compared to in GM12878 (P-value=0.0002 by Hypergeometric Test, see **Materials and Methods**) (**Figure 3.3D**), indicating that the RAM organization facilitates promoter-enhancer interactions to activate gene expression. For example, the *RAB31* promoter interacts with an enhancer that is located within the same RAM in K562 but in a RAM boundary in GM12878 where the enhancer would be silenced in GM12878; the *PRELID2* promoter and its interacting enhancer are located within the same K562 RAM but reside in a GM12878 RAM boundary indicating suppression in GM12878 (**Figure 3.3H**). In fact, the

RAB31 and *PRELID2* normalized expression levels are 34.29 and 5.37 folds higher in K562 than in GM12878, respectively. All the upregulated gene expressions involved in enhancer-promoter interactions that occurred in the same RAM in K562 but in different GM12878 RAMs are shown in **Table S3.2**. We had a similar observation by comparing K562 and HEPG2: while the majority of the promoter-enhancer pairs are intra-RAM in both cell lines, 28 are only occurred in the same RAM in K562 but in different HEPG2 RAMs (**Figure 3.3E**) and the corresponding genes have higher expressions in K562 than in HEPG2 (P-value=0.0006 by Hypergeometric Test) (**Figure 3.3F, Table S3.3**). Furthermore, we examined the K562 enhancer-promoter pairs in K562-specific RAMs and the cancer consensus RAMs (cancer cRAMs). 750 pairs were identified as intra-RAM interactions in both K562 RAMs and cancer cRAMs (**Figure S3.4A**). 77 genes were involved in the 174 pairs that are only intra-RAM interactions in K562, and 20 out of the 77 genes were detected as K562 specifically highly expressed genes across 92 cancer cell lines (hypergeometric test P-value=0.03) documented in the Harmonizome database(*46*) (**Figure S3.4B**). These observations further illustrated that RAMs represent a modularity directly associated with functional activity of the chromatin.

3.4.4 RAMs are insensitive to cohesin degradation

Previous studies showed that cohesin degradation would disrupt loop domains and topological associated domains (TADs) but largely not change histone modifications and gene expression(*47*, *48*). Therefore, RAMs are not expected to be affected by cohesin degradation. For confirmation, we identified RAMs using H3K27ac data in the HCT-116 RAD21-mAC cells untreated and treated for 6 hours with IAA. The RAM patterns for each chromosome were highly correlated between treated and untreated cells (**Figure S3.5A**) and the recall rate for the RAM boundaries was 0.9 on average for all the chromosomes (**Figure S3.5B**). This observation indicates that the RAM formation is independent from cohesin, distinguishing RAMs from TADs and loop domains.

3.4.5 Extrachromosomal DNA (ecDNA) from cancer patients majorly originated from intra-RAM

Circular extrachromosomal DNAs (ecDNAs) are prevalent in tumors and their length ranges from 100kbp to megabases, and the genes encoded in the ecDNAs are often amplified in cancers(*49–51*). We reason that, if RAMs are functional modules, ecDNAs would form within RAMs because RAM boundaries are highly condensed DNAs that would restrain the transcription of genes residing in ecDNAs. To test this hypothesis, we downloaded the ecDNAs identified from cancer patients(*52*), and filtered the ecDNAs corresponding to the median size of the cancer cRAM length (2.5Mb), i.e. only ecDNAs with size <2.5Mb were kept (i.e. 78% of all the ecDNAs). We found that 98% of 2459 ecDNAs were located within the individual RAMs. As a comparison, we performed the same analysis on TADs. We took the conserved TADs defined in the Dixon et al. study(*4*) and only kept the ecDNAs shorter than 880kbp (68% of all the ecDNAs), the median size of the TAD length. We found that 86% of the 2150 ecDNAs were within individual TADs (**Figure 3.3J, 3.3K**). Because the ecDNAs were filtered to have comparable length with the RAM and TAD sizes, respectively, this lower intra-domain percentage for TAD compared to RAM is not due to the larger size of RAMs. Furthermore, GREAT analysis(*53*) (http://great.stanford.edu/public/html/) on the ecDNAs that fall into intra-cancer cRAMs but split in TADs revealed that they are highly enriched in “positive regulation of DNA replication” with a P-value of 5.1641E-19. There are 12 genes involved in this pathway: *ATF1, BMP5, BMP6, EGFR, FGFR1, GLI2, IGF1, IL6, JUN, KITLG, PDGFA, PDGFRA*, which are known important for cell proliferation and cancer pathogenesis. For example, *EGFR* is a driver of tumorigenesis(*54*). Deregulation of the oncogenic FGFR signaling has been frequently observed in multiple types of cancers(*55*). The PDGF mediated signaling has been reported to be involved in the cell proliferation and invasion(*56*). The observations that ecDNAs tend to originate from intra-RAMs, which suggests that RAM is a functional module.

3.4.6 Deletion of the cRAM boundaries are predicted to alter the 3D chromatin structures

To systematically examine the impact of deleting cRAM boundaries to the chromatin structure, we resorted to computational predictions using a deep learning model ORCA(*57*) (https://github.com/jzhoulab/orca) as it is prohibitive to perform hundreds of Hi-C experiments with sufficient resolution. We took the ORCA model pre-trained on the high resolution Hi-C and Micro-C data in H1-hESC and HFF cell lines to predict 3D chromatin architecture from kilobase to whole-chromosome scale using DNA sequences. It also provided perturbation predictions if certain sequences were targeted. cRAM boundaries shared between cancer and normal samples are apparently important, therefore we selected all 418 of them that are located at least 16Mb away from centromere to predict their impacts on Hi-C contacts if deleted. As a comparison, we also included 298 H1-hESC and 187 HFF TAD boundaries (length of the TAD boundaries >=

100kbp; nonoverlap with selected cRAM boundaries; away from centromere at least 16Mb) in the computational perturbations screening. Considering that the cRAM boundaries are often larger than the TAD boundaries, we only deleted the center 100kbp of cRAM and TAD boundaries to avoid bias introduced by deletion size.

To measure the similarity between the deletion and wildtype Hi-C contact matrices, we calculated the Pearson correlation between them. Compared to the TAD deletions, deleting cRAM boundaries obviously resulted in lower correlation coefficients, indicating larger chromatin alterations, at the highest resolutions the ORCA model could predict (4kb and 8kb resolutions with Wilcoxon Rank Sum test P-values of 2.4E-11 and 2.6E-7, respectively) (**Figure 3.4B**). Deletion of the cRAM boundary (*chr10:115,940,000-116,040,000,* in hg38) on Hi-C contacts in HFF and H1-hESC cells is shown as an example (**Figure 3.4A, Figure S3.6**). The 3D contacts are severely weakened by deleting the cRAM boundary in both cell types.

3.4.7 Somatic genetic variations enriched in regulation associated modules boundaries

If RAMs are functional modules important for regulating functional activities, we reason that somatic mutations in cancers may target the RAM boundaries to disrupt the modular organization of chromatin leading to aberrant regulation of gene expression and resulting tumorigenesis. The PCAWG study revealed consensus mutations and variations from thousands of cancer patients including ~20 millions of somatic single nucleotide variations (SNVs) and ~1.08 millions of indels (https://dcc.icgc.org/releases)(*58*). We found that, while cancer cRAM boundaries cover ~21.6% of the genome, they host 25.9% of somatic SNVs and 23.4% somatic indels. As a comparison, the conserved TAD boundaries(*4*) covering 6.5% genome containing 5.2% somatic SNVs and 6% somatic indels (**Figure 3.4C**). The cancer cRAM boundaries are significantly enriched with both somatic SNVs and indels compared to the TAD boundaries (P-value<1E-5 from Two-sample Proportion tests, see **Materials and Methods**), indicating a stronger association with cancer mutations.

To elucidate the sequence features associated with cRAM boundaries and investigate how the somatic mutations change such features, we performed motif analysis on the cRAM boundaries using Homer(*37*). We focused on the cRAMs that are common in cancer (normal) but not in normal (cancer) samples as cancer (normal)-specific cRAMs, as they represent changed modularity between cancer and normal samples. By comparing cancer-specific and normal-specific cRAMs, we found 73 and 74 motifs enriched only in cancer and normal specific cRAM boundaries, respectively (example motifs shown in **Figure 3.4D, 3.4F**). We employed FIMO(*38*) to identify the occurrences of the enriched motifs that counted for 25.8% and 30.1% in base pairs, respectively, in the cancer and normal specific cRAM boundaries (**Figure 3.4E, 3.4G**). We next mapped the PCAWG somatic SNVs and indels onto the cancer and normal specific cRAM boundaries. While somatic SNVs do not show a preferred occurrences within the enriched motifs (24.9% and 29.9% for cancer and normal specific cRAM boundaries, respectively), the somatic indels overlapping with the cancer/normal-specific cRAM boundaries preferentially hit the enriched motifs (52.3% and 63% for cancer and normal specific cRAM boundaries, respectively), more than two fold by chance, in the altered cRAM boundaries between normal and cancer samples. We speculate that the enriched motifs in the cancer and normal-specific cRAM boundaries may respectively facilitate disruption and formation of cRAM boundaries in the normal samples. Two examples of these motifs overlapping with indels are shown in **Figure 3.4H-I**. We identified genes within 2.5kb from the enriched motifs overlapping with somatic indels and analyzed the enriched pathways using g:Profiler(*59*). The KEGG over-represented pathways are shown in **Figure 3.4J,** and the gene ontology molecular functions are shown in **Figure S3.7**. Furthermore, we downloaded the normalized gene expressions of the TCGA and GTEx samples from Expression Atlas (https://www.ebi.ac.uk/gxa/home) and identified differentially expressed genes (DEGs) (P-value<=0.05 by the Wilcoxon Rank Sum tests between cancer and normal samples). The top ranked pathways associated with cancer and cell proliferation are highly enriched with the DEGs, such as *PIK3AP1, LABM3, AKT1, MYB* in "PI3K-AKT pathway" (32%) and "pathways in cancer" (20%) (**Figure 3.4K**). These observations suggested that the motifs specifically enriched in the formation of cancer cRAM boundaries and disruption of normal cRAM boundaries are close to genes important for tumorigenesis, cell survival and cell proliferation. Somatic indels can severely alter these motifs and may contribute to the cRAM boundary change, affecting the expressions of the nearby genes.

3.5 Discussion

In this study, we analyzed the peak density profiles of histone modifications data and found they show modular patterns. These modules are clearly defined by active marks such as H3K27ac, H3K4me1 and H3K4me3, indicating their association with functional activity of the genome, and are thus termed regulation associated modules (RAMs). While TADs and compartments are identified from the 3D contacts measured by Hi-C, RAMs are delineated by histone modifications that are directly related to chromatin accessibility and gene expression. We showed that RAMs are obviously distinct from TADs, compartments and LADs although some RAM boundaries do overlap with TAD, compartment or LAD boundaries.

By surveying 93 normal and 19 cancer samples, we found the following evidence to support that RAMs are spatial modules resulting from functional activities. First, we observed that on average 60% of the RAMs (i.e. consensus RAMs) are largely shared across samples, while some of them are sample specific. Compared to TADs, consensus RAMs host higher percent of experimentally confirmed promoter-enhancer contacts (i.e. within the same RAMs), suggesting RAMs represent a modularization of the genome at a scale better aligned with transcriptional regulation. Second, ecDNAs detected from cancer patients tend to originate from the same RAMs rather than across multiple RAMs, supporting the insulation effect of RAM boundaries. Third, deletion of the cRAM boundaries would result in more severe chromatin alteration than the TAD boundaries based on in silico predictions of Hi-C contacts, suggesting the importance of cRAM boundaries in maintaining the chromatin structure. Fourth, cRAM boundaries are also more enriched with somatic genetic variants of SNVs and indels than the TAD boundaries. In particular, the somatic indels tend to disrupt the motifs specifically enriched in cancer or normal specific cRAM boundaries, suggesting a possible mechanism of tumorigenesis involved in altering the chromatin modularity.

To investigate the mechanisms underlying the RAM formation, we found that the RAMs are separated by densely packed DNA regions (as shown by their large number of Hi-C contacts) enriched with repressive histone modifications and lacking open chromatin, active histone marks or transcriptional events. Furthermore, unlike TADs, RAMs are insensitive to cohesin degradation. Taken together, these observations clearly show that RAMs are distinct from loop domains and TADs. RAMs are also different from lamina associated domains (LADs) defined by measuring the intermediate filament protein *LMNB1* localization. The LADs are formed through interactions between chromatin and lamina, and they are located at the periphery of the genome. The RAM boundaries are demarcated by densely packed DNAs and many RAM boundaries are not located in regions interacting with lamina or overlapping with TAD boundaries, and thus the mechanism underlying these RAM boundary formation should be different from other chromatin modules including TADs, LADs and compartments.

Many studies (such as in ref(*60*, *61*)) have shown that multivalent cations such as calcium, magnesium, and manganese can reduce the electrostatic repulsions between the DNA chains and induce DNA condensation. Furthermore, these cations may bind to specific DNA sequences(*62*) and affect nucleosome positioning(*63*, *64*). Therefore, a possible mechanism can be that genomic DNAs become densely packed around cations such as Ca^{2+}, Mg^{2+} and Mn^{2+} to form RAM boundaries even if they are not marked by H3K9me3 or interacting with lamina. Proteins such as calcium binding proteins that carry many cations or their interacting partners may recognize specific DNA sequences such as those motifs enriched in the cancer or normal specific cRMA boundaries to facilitate locus-specific localization of cations. Interestingly, the most enriched molecular function of the genes close to (<2.5kbp) the enriched motifs overlapping with somatic indels in the cancer or normal specific cRAM boundaries is calcium ion binding (**Figure S3.7**), and ~30% of them are differentially expressed in cancer and normal samples (**Table S3.4**), implying a possible feedback mechanism. This hypothesis and the mechanistic details are awaiting for future studies.

3.6 Acknowledgements

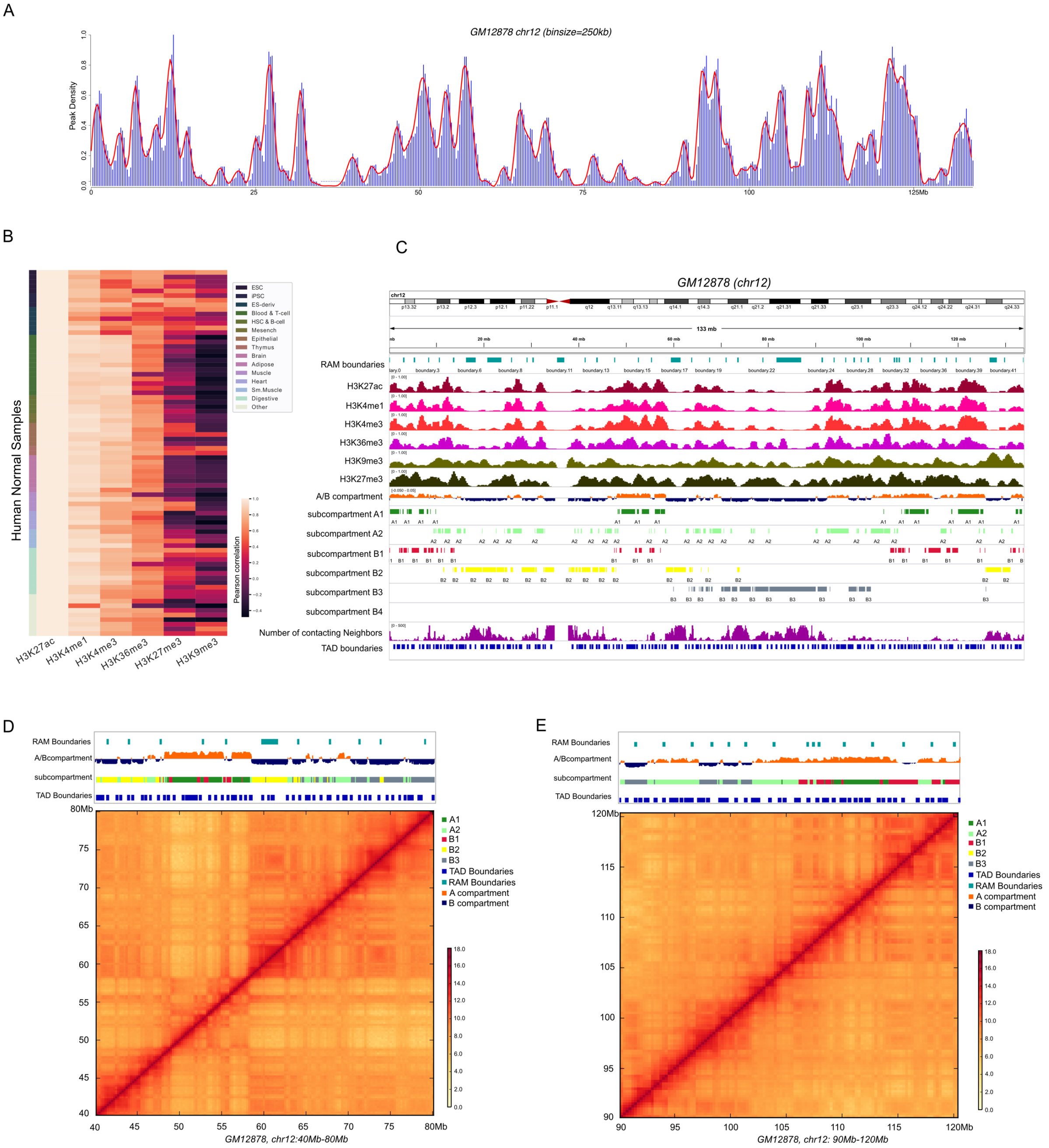
Chapter 3, in full, is a reformatted reprint of the material as it appears in “Regulation associated modules reflect 3D genome modularity associated with chromatin activity.” Lina Zheng; Wei Wang. bioRxiv, 2022. https://doi.org/10.1101/2022.03.02.482718. In submission. The dissertation author was the primary investigator and author of this paper.

We are grateful to the labs of Drs. Bas van Steensel, Erez Lieberman Aiden and Marion Cremer for data sharing.

W.W. conceived and supervised the project. L.Z. performed the bioinformatics analysis. L.Z. and W.W. wrote the manuscripts.

3.7 Figures

Figure 3.1. Regulation associated module (RAM) identification. (A) H3K27ac peaks density of chr12 in GM12878 (binsize=250kb) **(B)** Pearson correlation between histone modification marks (H3K27ac, H3K4me1, H3K4me3, H3K36me3, H3K27me3, and H3K9me3) for the Roadmap samples. **(C)** Examples of histone modifications, A/B compartments, subcompartments, number of the 3D contacts, TAD boundaries and RAM boundaries in chr12 for GM12878. **(D)** The zoom-in genomic view for chr12:40Mb-80Mb and **(E)** The zoom-in genomic view for chr12:90Mb-120Mb in hg19.


A
GM12878 chr12 (binsize=250kb)
Peak Density
0.0
0.2
0.4
0.6
0.8
1.0
0
25
50
75
100
125Mb
B
Human Normal Samples
H3K27ac
H3K4me1
H3K4me3
H3K36me3
H3K27me3
H3K9me3
ESC
iPSC
ES-deriv
Blood & T-cell
HSC & B-cell
Mesench
Epithelial
Thymus
Brain
Adipose
Muscle
Heart
Sm.Muscle
Digestive
Other
Pearson correlation
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
C
GM12878 (chr12)
chr12
133 mb
RAM boundaries
H3K27ac
H3K4me1
H3K4me3
H3K36me3
H3K9me3
H3K27me3
A/B compartment
subcompartment A1
subcompartment A2
subcompartment B1
subcompartment B2
subcompartment B3
subcompartment B4
Number of contacting Neighbors
TAD boundaries
D
RAM Boundaries
A/Bcompartment
subcompartment
TAD Boundaries
80Mb
75
70
65
60
55
50
45
40
40
45
50
55
60
65
70
75
80Mb
GM12878, chr12:40Mb-80Mb
A1
A2
B1
B2
B3
TAD Boundaries
RAM Boundaries
A compartment
B compartment
18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
E
RAM Boundaries
A/Bcompartment
subcompartment
TAD Boundaries
120Mb
115
110
105
100
95
90
90
95
100
105
110
115
120Mb
GM12878, chr12: 90Mb-120Mb
A1
A2
B1
B2
B3
TAD Boundaries
RAM Boundaries
A compartment
B compartment
18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0

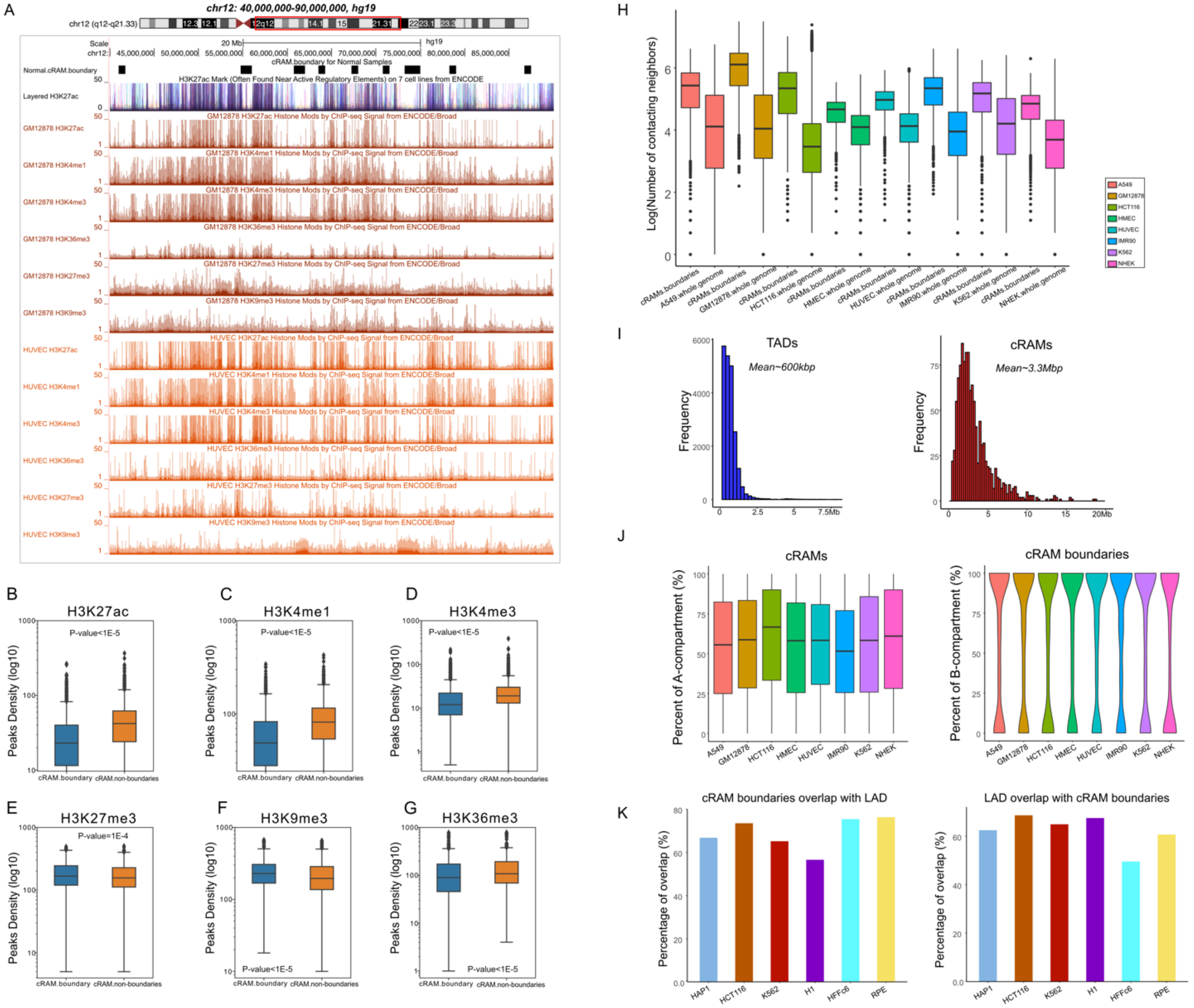


Figure 3.2. Characterization of the cRAMs and boundaries. (**A**) Genome browser examples of normal cRAM boundaries and the histone modifications in genomic region chr12:40Mb-90Mb in hg19 for GM12878 and the HUVEC cells. (**B-G**) The genome-wide enrichment of the histone modification marks. (**B**) H3K27ac, (**C**) H3K4me1, (**D**) H3K4me3, (**E**) H3K27me3, (**F**) H3K9me3, and (**G**) H3K36me3 in cRAM boundaries and non-boundaries (**H**) The contacting neighbors distribution of cRAM boundaries and whole genome locus in the 3D contact network in a diverse of the cell types. (**I**) Sizes of the TADs and cRAMs. (**J**) cRAM boundaries distribution over A/B compartments. (**K**) cRAM boundaries distribution over LaminB1 signals (LAD).

Figure 3.3. RAM is a functional unit. (A) K562 enhancer-promoter pairs distribution over K562 RAMs. **(B)** K562 enhancer-promoter pairs distribution over K562 TADs. **(C)** K562 enhancer-promoter pairs distribution over K562 RAMs and GM12878 RAMs. (**D**) Genes regulated by the 39 enhancer-promoter interactions only within K562 RAMs tend to have higher expressions in K562 compared to GM12878. **(E)** K562 enhancer-promoter pairs distribution over K562 RAMs and HEPG2 RAMs. (**F**) Genes regulated by the 28 enhancer-promoter interactions only within K562 RAMs tend to have higher expressions in K562 compared to HEPG2. (**G**) Examples of K562 enhancer-promoter pairs relative to K562 TAD and RAM boundaries. (**H**) Examples of K562 enhancer-promoter pairs relative to K562 and GM12878 RAM boundaries. (**I**) ecDNA distribution over TADs. (**J**) ecDNA distribution over cancer cRAMs.

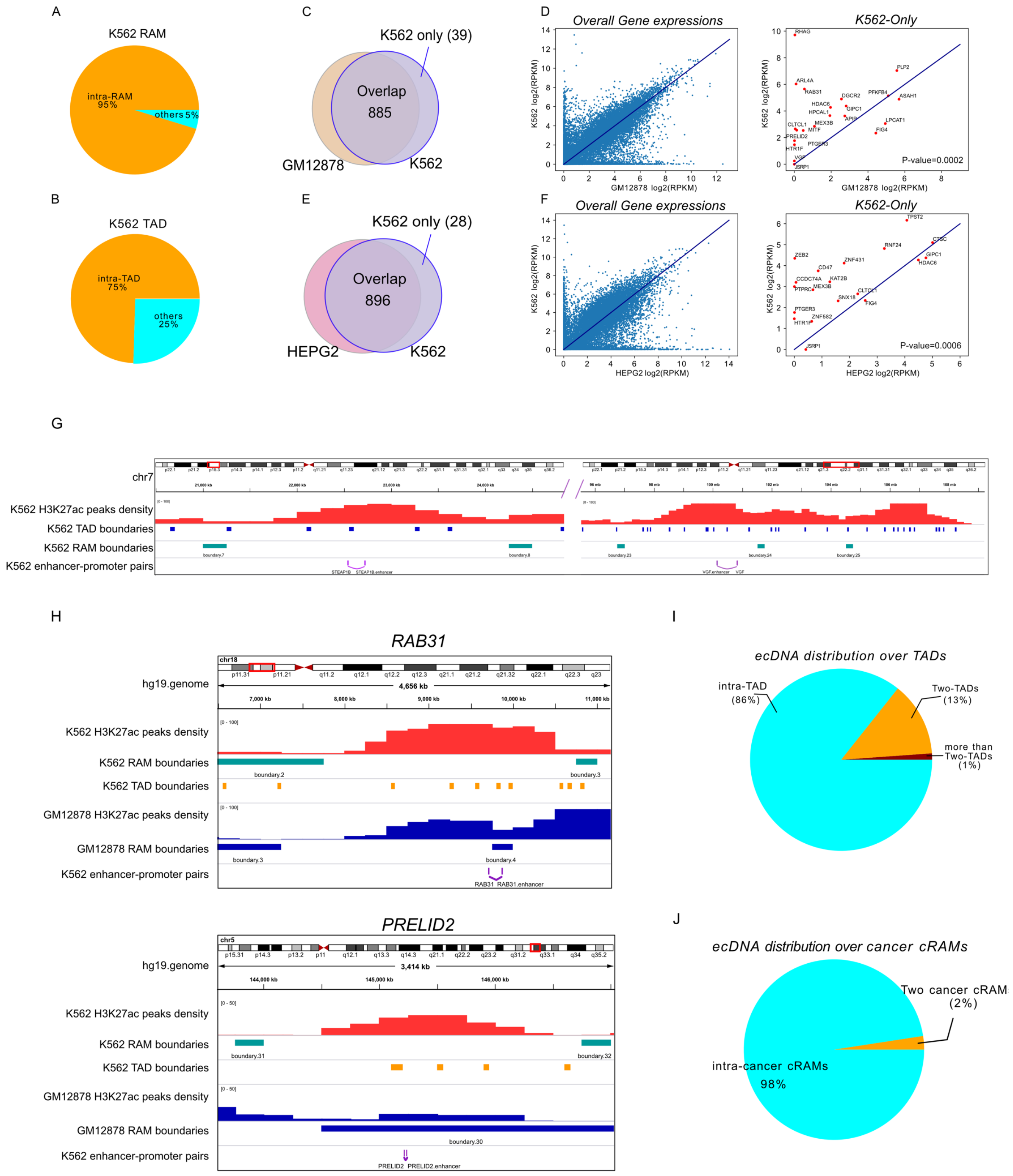
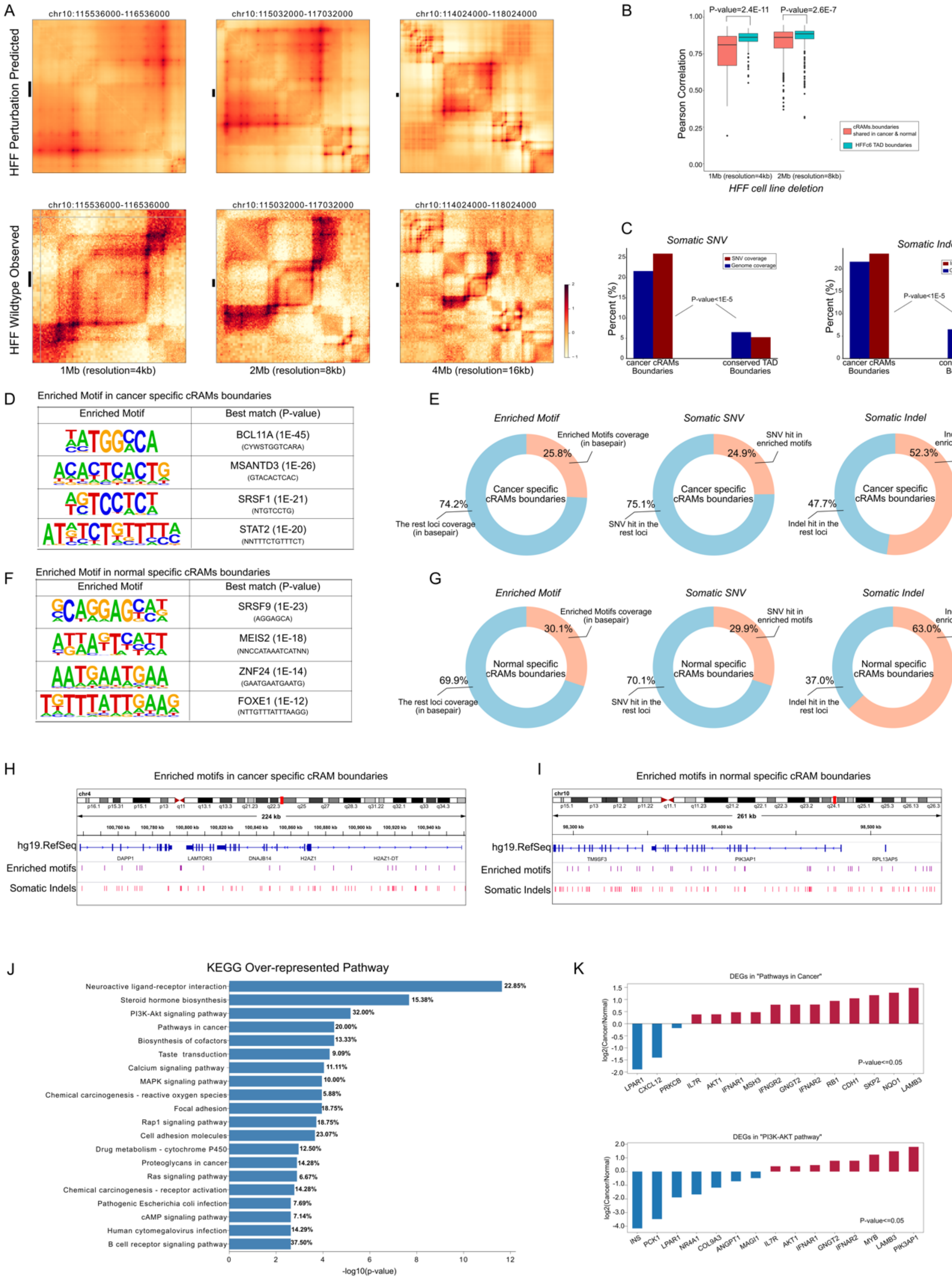

A
K562 RAM
intra-RAM 95%
others 5%
B
K562 TAD
intra-TAD 75%
others 25%
C
K562 only (39)
Overlap 885
GM12878
K562
E
K562 only (28)
Overlap 896
HEPG2
K562
D
Overall Gene expressions
K562 log2(RPKM)
GM12878 log2(RPKM)
K562-Only
RHAG
PLP2
ARL4A
RAB31
DGCR2
PFKFB4
ASAH1
HDAC6
GIPC1
HPCAL1
APIP
CLTCL1
MEX3B
MITF
LPCAT1
FIG4
PRELID2
PTGER3
HTR1F
VGF
JSRP1
P-value=0.0002
F
Overall Gene expressions
K562 log2(RPKM)
HEPG2 log2(RPKM)
K562-Only
TPST2
CTSC
RNF24
ZEB2
GIPC1
ZNF431
HDAC6
CD47
CCDC74A
KAT2B
PTPRC
MEX3B
CLTCL1
SNX18
FIG4
PTGER3
ZNF582
HTR1F
JSRP1
P-value=0.0006
G
chr7
K562 H3K27ac peaks density
K562 TAD boundaries
K562 RAM boundaries
K562 enhancer-promoter pairs
21,000 kb
22,000 kb
23,000 kb
24,000 kb
96 mb
98 mb
100 mb
102 mb
104 mb
106 mb
108 mb
[0 - 100]
boundary.7
boundary.8
boundary.23
boundary.24
boundary.25
STEAP1B
STEAP1B.enhancer
VGF.enhancer
VGF
H
RAB31
chr18
hg19.genome
4,656 kb
7,000 kb
8,000 kb
9,000 kb
10,000 kb
11,000 kb
K562 H3K27ac peaks density
[0 - 100]
K562 RAM boundaries
boundary.2
boundary.3
K562 TAD boundaries
GM12878 H3K27ac peaks density
[0 - 100]
GM12878 RAM boundaries
boundary.3
boundary.4
K562 enhancer-promoter pairs
RAB31
RAB31.enhancer
PRELID2
chr5
hg19.genome
3,414 kb
144,000 kb
145,000 kb
146,000 kb
K562 H3K27ac peaks density
[0 - 50]
K562 RAM boundaries
boundary.31
boundary.32
K562 TAD boundaries
GM12878 H3K27ac peaks density
[0 - 50]
GM12878 RAM boundaries
boundary.30
K562 enhancer-promoter pairs
PRELID2
PRELID2.enhancer
I
ecDNA distribution over TADs
intra-TAD (86%)
Two-TADs (13%)
more than Two-TADs (1%)
J
ecDNA distribution over cancer cRAMs
Two cancer cRAM (2%)
intra-cancer cRAMs 98%

Figure 3.4. The association of the cRAM boundaries with the 3D chromatin structure and cancer somatic variants. (A) An example of the Hi-C contact change upon deletion of the cRAM boundary (chr10:115,940,000-116,040,000 in hg38) in HFF cells predicted by a deep learning model ORCA. (B) Pearson correlations between the predicted Hi-C contacts before and after cRAM boundary and TAD boundary deletion in HFF cells. A lower correlation indicates a larger perturbation to the wildtype chromatin structure upon deletion. (**C**) Somatic SNV and indels enrichment in cancer cRAM boundaries and TAD boundaries. Genome coverage: the total base pairs of the cancer cRAM boundaries or TAD boundaries in the whole genome; SNV coverage: the percentage of the SNVs in the cancer cRAM boundaries or TAD boundaries in the whole genome; INDEL coverage: the percentage of the indels in cancer cRAM boundaries or TAD boundaries in the whole genome. (**D**) Examples of the enriched motifs in the cancer specific cRAM boundaries. (**E**) Overlaps of somatic SNVs and indels with the enriched motifs in the cancer specific cRAM boundaries (**F**) Examples of the enriched motifs in the normal specific cRAM boundaries. (**G**) Overlaps of somatic SNVs and indels with the enriched motifs in the normal specific cRAM boundaries (**H**) The zoom-in genomic view for the enriched motifs with somatic indels in cancer specific cRAM boundaries (**I**) The zoom-in genomic view for the enriched motifs with somatic indels in normal specific cRAM boundaries (**J**) KEGG over-represented pathways for the genes within 2.5kb from the enriched motifs overlapping with somatic indels (**K**) Differentially expressed genes in KEGG “Pathways in cancer” and “PI3K-AKT pathway”.


A
HFF Perturbation Predicted
HFF Wildtype Observed
chr10:115536000-116536000
chr10:115032000-117032000
chr10:114024000-118024000
1Mb (resolution=4kb)
2Mb (resolution=8kb)
4Mb (resolution=16kb)
B
P-value=2.4E-11
P-value=2.6E-7
Pearson Correlation
1.00
0.75
0.50
0.25
0.00
cRAMs.boundaries shared in cancer & normal
HFFc6 TAD boundaries
1Mb (resolution=4kb)
2Mb (resolution=8kb)
HFF cell line deletion
C
Somatic SNV
SNV coverage
Genome coverage
Percent (%)
P-value<1E-5
cancer cRAMs Boundaries
conserved TAD Boundaries
Somatic Indel
Indel coverage
Genome coverage
P-value<1E-5
D
Enriched Motif in cancer specific cRAMs boundaries
Enriched Motif
Best match (P-value)
BCL11A (1E-45)
(CYWSTGGTCARA)
MSANTD3 (1E-26)
(GTACACTCAC)
SRSF1 (1E-21)
(NTGTCCTG)
STAT2 (1E-20)
(NNTTTCTGTTTCT)
E
Enriched Motif
Enriched Motifs coverage (in basepair)
25.8%
74.2%
The rest loci coverage (in basepair)
Cancer specific cRAMs boundaries
Somatic SNV
SNV hit in enriched motifs
24.9%
75.1%
SNV hit in the rest loci
Somatic Indel
Indel hit in enriched motifs
52.3%
47.7%
Indel hit in the rest loci
F
Enriched Motif in normal specific cRAMs boundaries
Enriched Motif
Best match (P-value)
SRSF9 (1E-23)
(AGGAGCA)
MEIS2 (1E-18)
(NNCCATAAATCATNN)
ZNF24 (1E-14)
(GAATGAATGAATG)
FOXE1 (1E-12)
(NTTGTTTATTTAAGG)
G
Enriched Motif
Enriched Motifs coverage (in basepair)
30.1%
69.9%
The rest loci coverage (in basepair)
Normal specific cRAMs boundaries
Somatic SNV
SNV hit in enriched motifs
29.9%
70.1%
SNV hit in the rest loci
Somatic Indel
Indel hit in enriched motifs
63.0%
37.0%
Indel hit in the rest loci
H
Enriched motifs in cancer specific cRAM boundaries
chr4
224 kb
hg19.RefSeq
DAPP1
LAMTOR3
DNAJB14
H2AZ1
H2AZ1-DT
Enriched motifs
Somatic Indels
I
Enriched motifs in normal specific cRAM boundaries
chr10
261 kb
hg19.RefSeq
TM9SF3
PIK3AP1
RPL13AP5
Enriched motifs
Somatic Indels
J
KEGG Over-represented Pathway
Neuroactive ligand-receptor interaction 22.85%
Steroid hormone biosynthesis 15.38%
PI3K-Akt signaling pathway 32.00%
Pathways in cancer 20.00%
Biosynthesis of cofactors 13.33%
Taste transduction 9.09%
Calcium signaling pathway 11.11%
MAPK signaling pathway 10.00%
Chemical carcinogenesis - reactive oxygen species 5.88%
Focal adhesion 18.75%
Rap1 signaling pathway 18.75%
Cell adhesion molecules 23.07%
Drug metabolism - cytochrome P450 12.50%
Proteoglycans in cancer 14.28%
Ras signaling pathway 6.67%
Chemical carcinogenesis - receptor activation 14.28%
Pathogenic Escherichia coli infection 7.69%
cAMP signaling pathway 7.14%
Human cytomegalovirus infection 14.29%
B cell receptor signaling pathway 37.50%
0
2
4
6
8
10
12
-log10(p-value)
K
DEGs in "Pathways in Cancer"
log2(Cancer/Normal)
P-value<=0.05
LPAR1
CXCL12
PRKCB
IL7R
AKT1
IFNAR1
MSH3
IFNGR2
GNGT2
IFNAR2
RB1
CDH1
SKP2
NQO1
LAMB3
DEGs in "PI3K-AKT pathway"
log2(Cancer/Normal)
P-value<=0.05
INS
PCK1
LPAR1
NR4A1
COL9A3
ANGPT1
MAGI1
IL7R
AKT1
IFNAR1
GNGT2
IFNAR2
MYB
LAMB3
PIK3AP1

3.8 Supplementary Figures

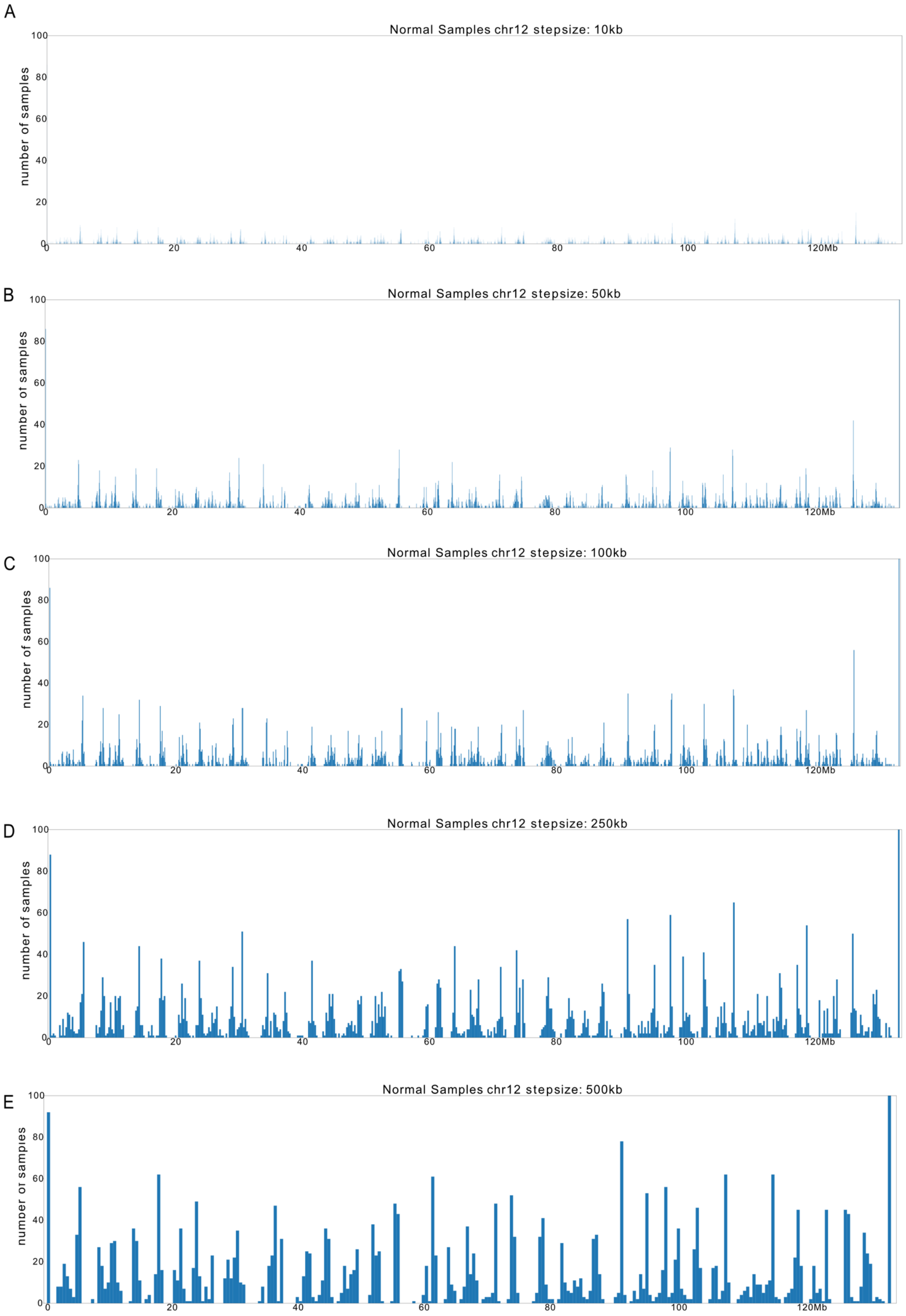


Figure S3.1. The frequency of the consensus RAM boundaries in the 93 normal samples. The step sizes are (**A**) 10kb, (**B**) 50kb, (**C**)100kb, (**D**) 250kb, (**E**) 500kb.

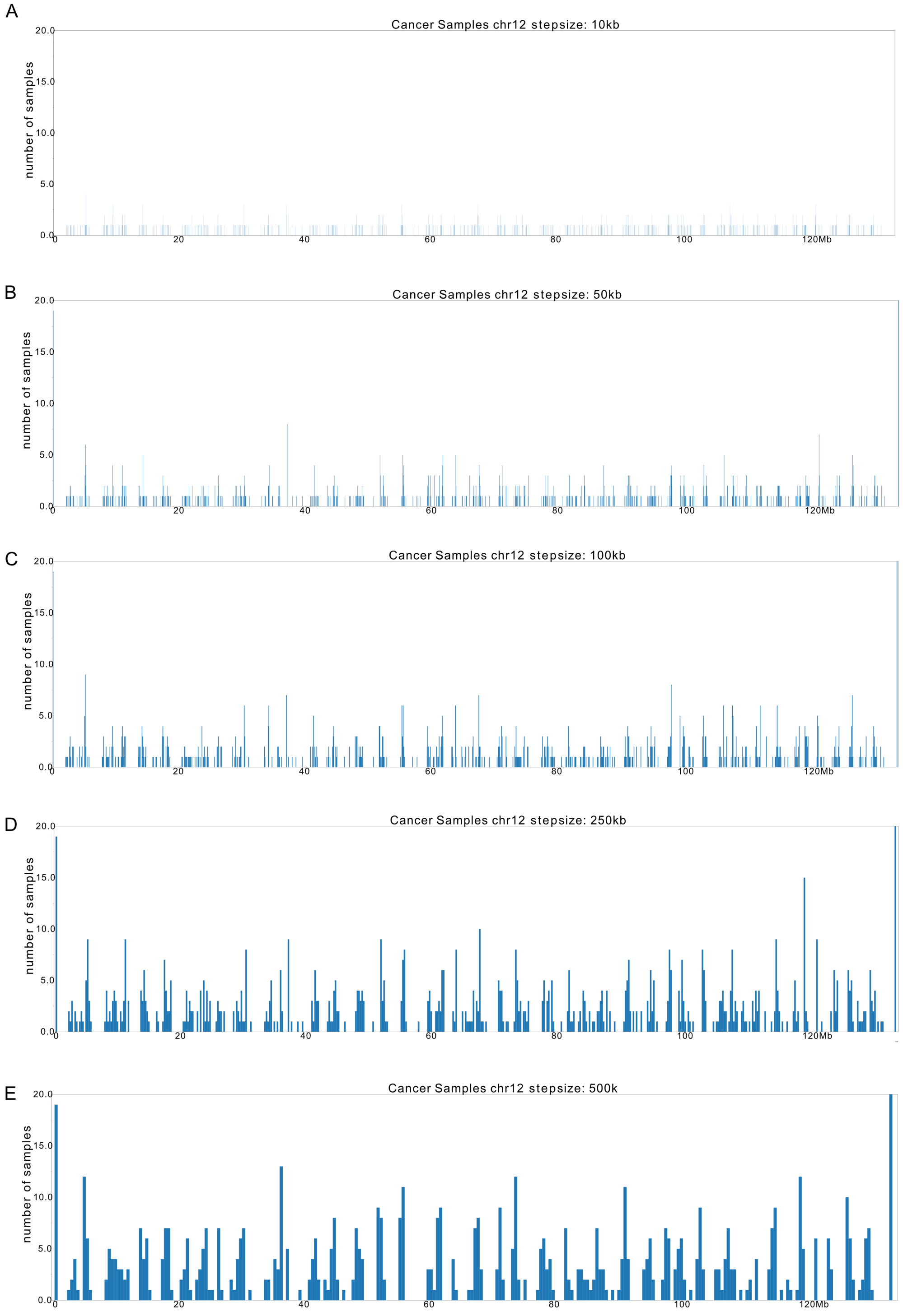


Figure S3.2. The frequency of the consensus RAM boundaries in the 19 cancer samples. The step sizes are (**A**) 10kb, (**B**) 50kb, (**C**)100kb, (**D**) 250kb, (**E**) 500kb.

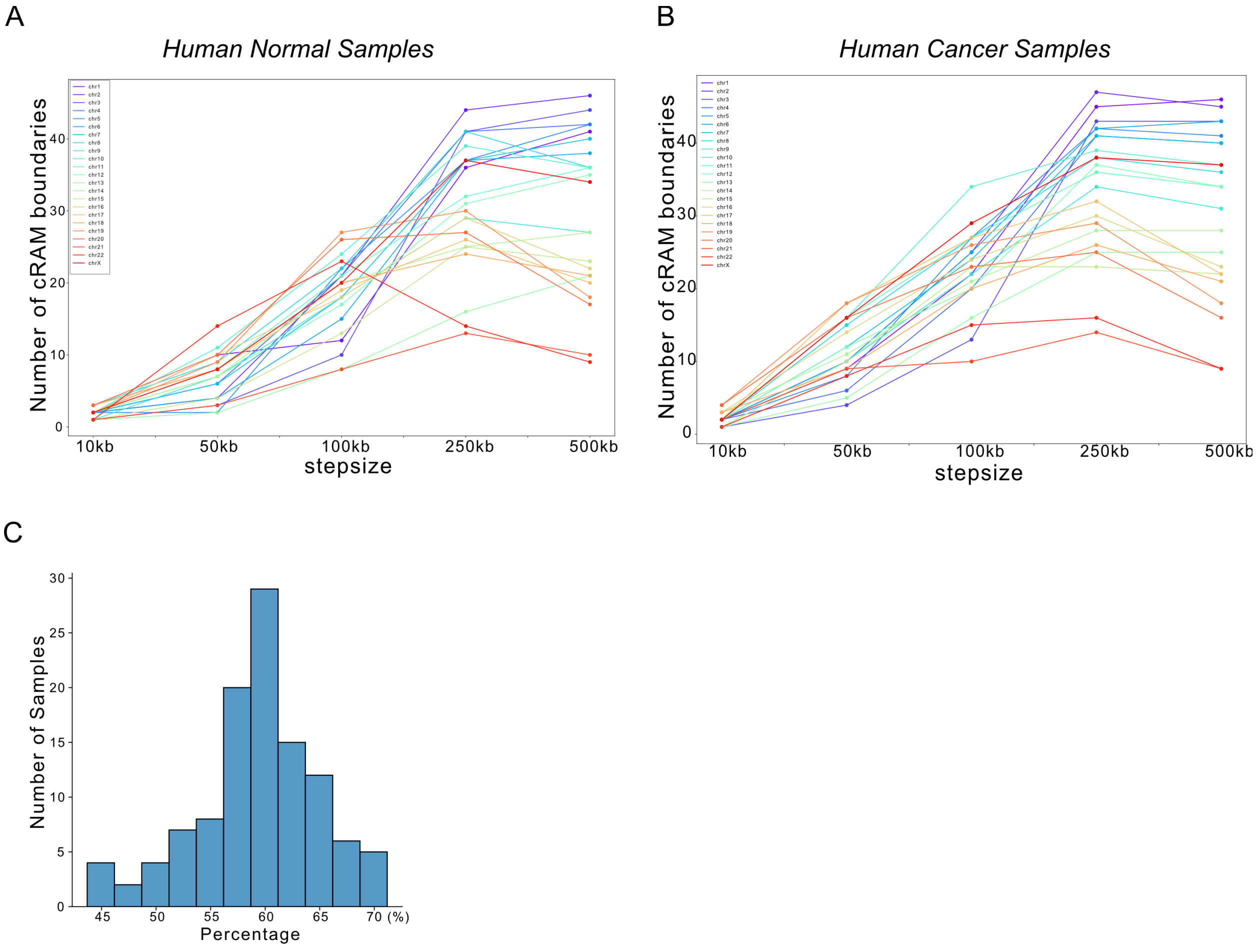


Figure S3.3. Consensus RAMs. The number of consensus RAMs (i.e. RAMs shared between samples) using different step sizes in **(A)** normal and **(B)** cancer samples. (**C**) The percentage of the consensus RAMs among all the RAMs in each sample.

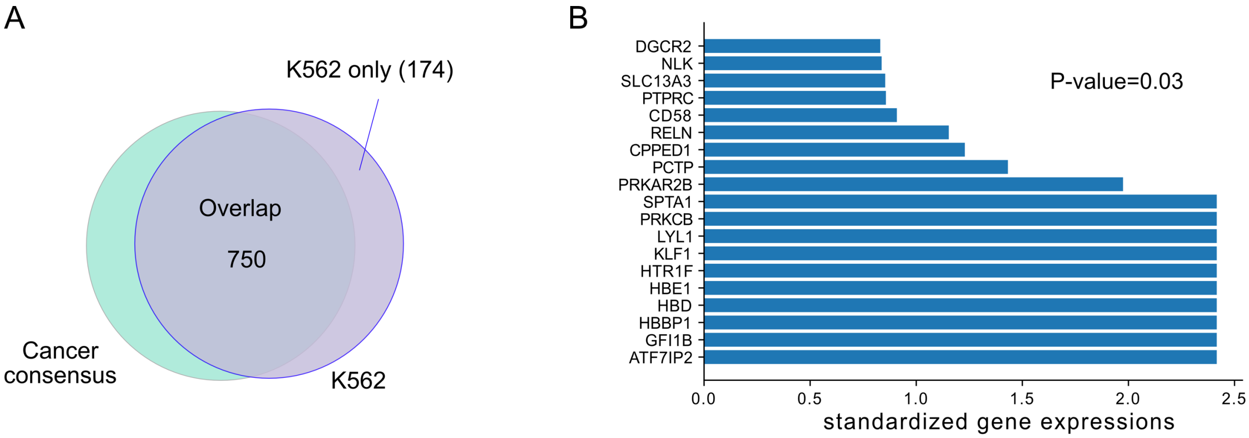


Figure S3.4. The K562 enhancer-promoter interactions over K562 RAMs and cancer cRAMs. (A) K562 enhancer-promoter pairs distribution over K562 RAMs and cancer cRAMs. (**B**) 20 highly expressed genes in the enhancer-promoter pairs uniquely observed in K562 RAMs.

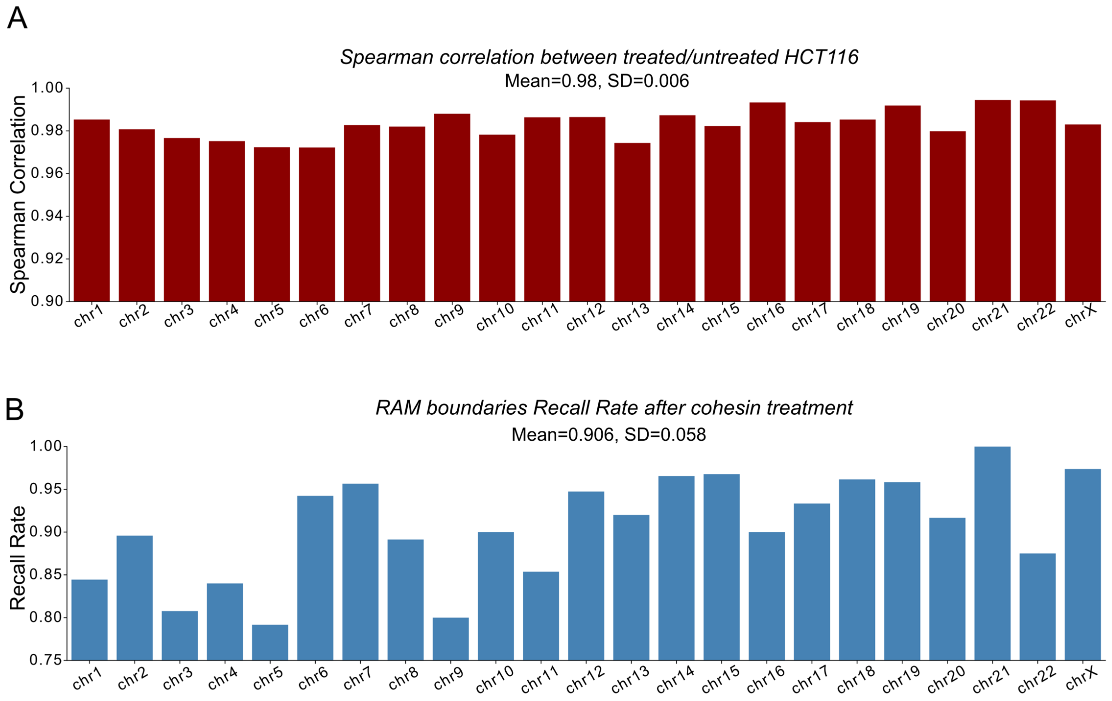


Figure S3.5. RAMs are resistant to cohesion degradation. (A) Spearman correlation of the RAMs between the treated and untreated HCT116 cells (**B**) RAM boundaries recall rate after cohesin treatment for HCT116 cells.

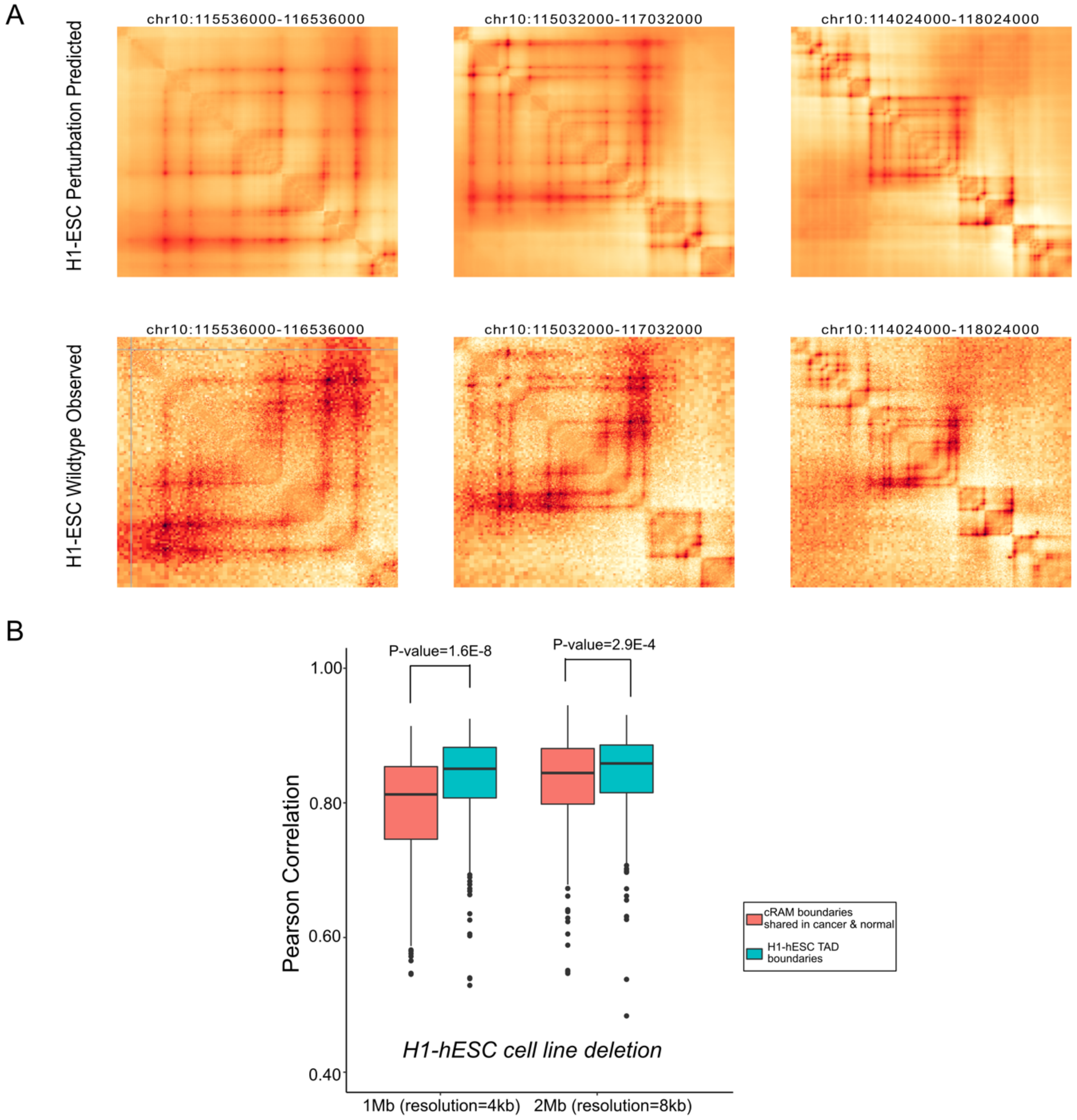


Figure S3.6. Predicted chromatin structure change upon deletion of cRAM boundaries and TAD boundaries in H1-hESC cell line. (A) An example of the Hi-C contact change upon deletion of the cRAM boundary (chr10:115,940,000-116,040,000 in hg38) in H1-hESC cells predicted by a deep learning model ORCA. (B) Pearson correlations between the predicted Hi-C contacts before and after cRAM boundary and TAD boundary deletion in H1-hESC cells. A lower correlation indicates a larger perturbation to the wildtype chromatin structure upon deletion.

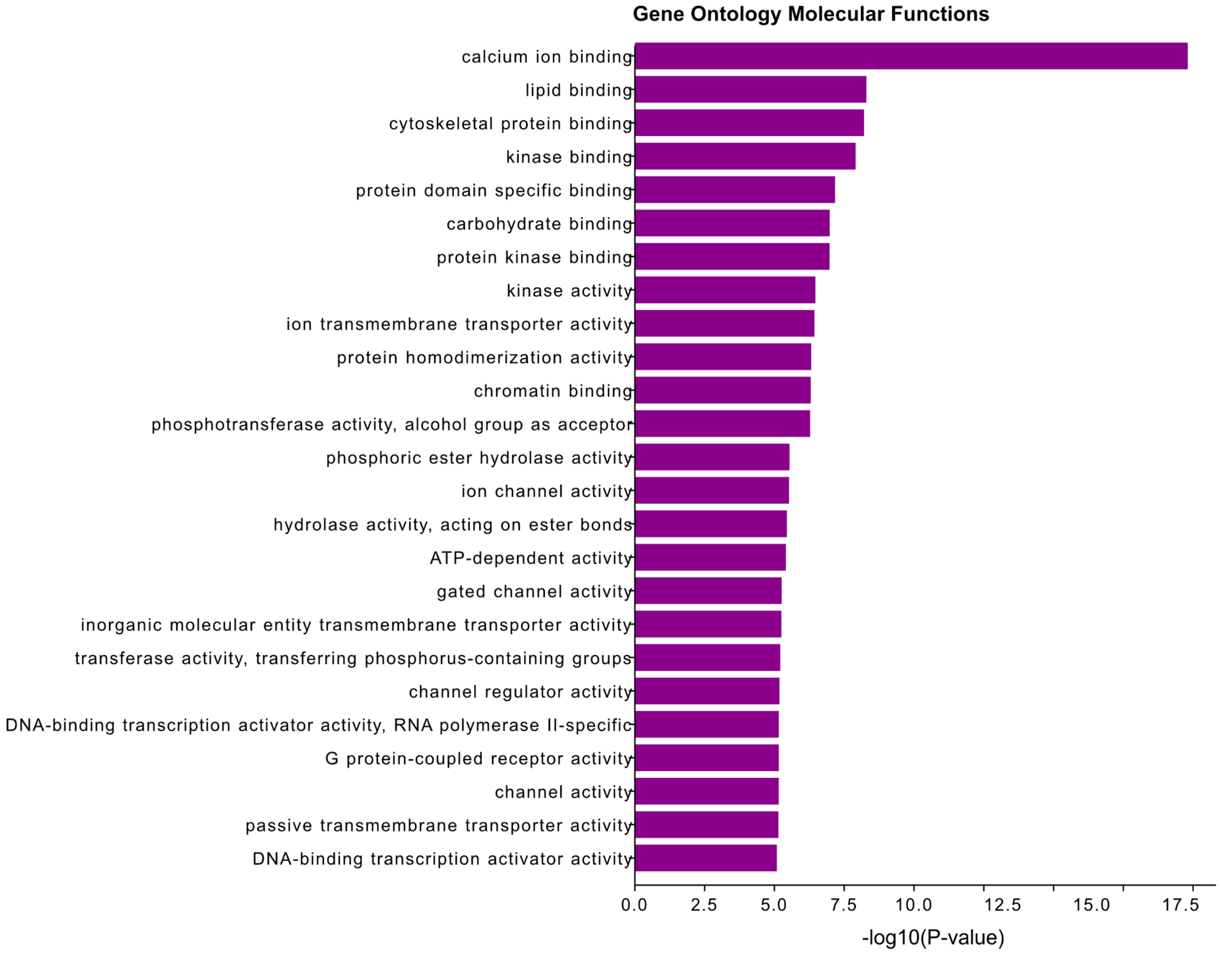


Figure S3.7. Gene Ontology Molecular Functions of the genes within 2.5kb from the enriched motifs overlapping with somatic indels

3.9 Supplementary Tables

Table S3.1. Collected ChIP-seq samples from ROADMAP and ENCODE.

Data ID	Source	Tissue/SampleType	Details
E017	Roadmap	IMR90	IMR90 fetal lung fibroblasts Cell Line
E008	Roadmap	ESC	H9 Cells
E015	Roadmap	ESC	HUES6 Cells
E014	Roadmap	ESC	HUES48 Cells
E016	Roadmap	ESC	HUES64 Cells
E003	Roadmap	ESC	H1 Cells
E020	Roadmap	iPSC	iPS-20b Cells
E019	Roadmap	iPSC	iPS-18 Cells
E021	Roadmap	iPSC	iPS DF 6.9 Cells
E022	Roadmap	iPSC	iPS DF 19.11 Cells
E007	Roadmap	ES-deriv	H1 Derived Neuronal Progenitor Cultured Cells
E013	Roadmap	ES-deriv	hESC Derived CD56+ Mesoderm Cultured Cells
E012	Roadmap	ES-deriv	hESC Derived CD56+ Ectoderm Cultured Cells
E011	Roadmap	ES-deriv	hESC Derived CD184+ Endoderm Cultured Cells
E004	Roadmap	ES-deriv	H1 BMP4 Derived Mesendoderm Cultured Cells
E005	Roadmap	ES-deriv	H1 BMP4 Derived Trophoblast Cultured Cells
E006	Roadmap	ES-deriv	H1 Derived Mesenchymal Stem Cells
E062	Roadmap	Blood&Tcell	Primary mononuclear cells from peripheral blood
E034	Roadmap	Blood&Tcell	Primary T cells from peripheral blood
E045	Roadmap	Blood&Tcell	Primary T cells effector/memory enriched from peripheral blood

Table S3.1. Collected ChIP-seq samples from ROADMAP and ENCODE, continued

Data ID	Source	Tissue/SampleType	Details
E044	Roadmap	Blood&Tcell	Primary T regulatory cells from peripheral blood
E043	Roadmap	Blood&Tcell	Primary T helper cells from peripheral blood
E039	Roadmap	Blood&Tcell	Primary T helper naive cells from peripheral blood
E041	Roadmap	Blood&Tcell	Primary T helper cells PMA-I stimulated
E042	Roadmap	Blood&Tcell	Primary T helper 17 cells PMA-I stimulated
E040	Roadmap	Blood&Tcell	Primary T helper memory cells from peripheral blood 1
E037	Roadmap	Blood&Tcell	Primary T helper memory cells from peripheral blood 2
E048	Roadmap	Blood&Tcell	Primary T CD8+ memory cells from peripheral blood
E038	Roadmap	Blood&Tcell	Primary T helper naive cells from peripheral blood
E047	Roadmap	Blood&Tcell	Primary T CD8+ naive cells from peripheral blood
E029	Roadmap	HSC&Bcell	Primary monocytes from peripheral blood
E050	Roadmap	HSC&Bcell	Primary hematopoietic stem cells G-CSF-mobilized Female
E032	Roadmap	HSC&Bcell	Primary B cells from peripheral blood
E046	Roadmap	HSC&Bcell	Primary Natural Killer cells from peripheral blood
E026	Roadmap	Mesench	Bone Marrow Derived Cultured Mesenchymal Stem Cells
E049	Roadmap	Mesench	Mesenchymal Stem Cell Derived Chondrocyte Cultured Cells
E055	Roadmap	Epithelial	Foreskin Fibroblast Primary Cells skin01
E056	Roadmap	Epithelial	Foreskin Fibroblast Primary Cells skin02

Table S3.1. Collected ChIP-seq samples from ROADMAP and ENCODE, continued

Data ID	Source	Tissue/SampleType	Details
E059	Roadmap	Epithelial	Foreskin Melanocyte Primary Cells skin01
E061	Roadmap	Epithelial	Foreskin Melanocyte Primary Cells skin03
E058	Roadmap	Epithelial	Foreskin Keratinocyte Primary Cells skin03
E112	Roadmap	Thymus	Thymus
E093	Roadmap	Thymus	Fetal Thymus
E071	Roadmap	Brain	Brain Hippocampus Middle
E074	Roadmap	Brain	Brain Substantia Nigra
E068	Roadmap	Brain	Brain Anterior Caudate
E069	Roadmap	Brain	Brain Cingulate Gyrus
E072	Roadmap	Brain	Brain Inferior Temporal Lobe
E067	Roadmap	Brain	Brain Angular Gyrus
E073	Roadmap	Brain	Brain_Dorsolateral_Prefrontal_Cortex
E063	Roadmap	Adipose	Adipose Nuclei
E100	Roadmap	Muscle	Psoas Muscle
E108	Roadmap	Muscle	Skeletal Muscle Female
E089	Roadmap	Muscle	Fetal Muscle Trunk
E090	Roadmap	Muscle	Fetal Muscle Leg
E104	Roadmap	Heart	Right Atrium
E095	Roadmap	Heart	Left Ventricle
E105	Roadmap	Heart	Right Ventricle
E065	Roadmap	Heart	Aorta
E078	Roadmap	Sm. Muscle	Duodenum Smooth Muscle
E076	Roadmap	Sm. Muscle	Colon Smooth Muscle
E103	Roadmap	Sm. Muscle	Rectal Smooth Muscle
E111	Roadmap	Sm. Muscle	Stomach Smooth Muscle
E092	Roadmap	Digestive	Fetal Stomach
E085	Roadmap	Digestive	Fetal Intestine Small
E084	Roadmap	Digestive	Fetal Intestine Large
E109	Roadmap	Digestive	Small Intestine
E106	Roadmap	Digestive	Sigmoid Colon

Table S3.1. Collected ChIP-seq samples from ROADMAP and ENCODE, continued

Data ID	Source	Tissue/SampleType	Details
E075	Roadmap	Digestive	Colonic Mucosa
E101	Roadmap	Digestive	Rectal Mucosa Donor 29
E102	Roadmap	Digestive	Rectal Mucosa Donor 31
E079	Roadmap	Digestive	Esophagus
E094	Roadmap	Digestive	Gastric
E099	Roadmap	PLCNT.AMN	Placenta Amnion
E097	Roadmap	OVRY	Ovary
E087	Roadmap	PANC.ISLT	Pancreatic Islets
E080	Roadmap	ADRL.GLND.FET	Fetal Adrenal Gland
E091	Roadmap	PLCNT.FET	Placenta
E066	Roadmap	LIV.ADLT	Liver
E098	Roadmap	PANC	Pancreas
E096	Roadmap	LNG	Lung
E113	Roadmap	SPLN	Spleen
E116	Roadmap	BLD.GM12878	GM12878 Lymphoblastoid Cells
E119	Roadmap	BRST.HMEC	HMEC Mammary Epithelial Primary Cells
E120	Roadmap	MUS.HSMM	HSMM Skeletal Muscle Myoblasts Cells
E121	Roadmap	MUS.HSMMT	HSMM cell derived Skeletal Muscle Myotubes Cells
E122	Roadmap	VAS.HUVEC	HUVEC Umbilical Vein Endothelial Primary Cells
E124	Roadmap	BLD.CD14.MONO	Monocytes-CD14+ RO01746 Primary Cells
E125	Roadmap	BRN.NHA	NH-A Astrocytes Primary Cells
E126	Roadmap	SKIN.NHDFAD	NHDF-Ad Adult Dermal Fibroblast Primary Cells
E127	Roadmap	SKIN.NHEK	NHEK-Epidermal Keratinocyte Primary Cells
E128	Roadmap	LNG.NHLF	NHLF Lung Fibroblast Primary Cells
E129	Roadmap	BONE.OSTEO	Osteoblast Primary Cells
E114	Roadmap	LNG.A549.ETOH002.CNCR	A549 EtOH 0.02pct Lung Carcinoma Cell Line

Table S3.1. Collected ChIP-seq samples from ROADMAP and ENCODE, continued

Data ID	Source	Tissue/SampleType	Details
E115	Roadmap	BLD.DND41.CNCR	Dnd41 TCell Leukemia Cell Line
E117	Roadmap	CRVX.HELAS3.CNCR	HeLa-S3 Cervical Carcinoma Cell Line
E118	Roadmap	LIV.HEPG2.CNCR	HepG2 Hepatocellular Carcinoma Cell Line
E123	Roadmap	BLD.K562.CNCR	K562 Leukemia Cells
ENCFF409EFR	Encode	SK-N-MC	neuroblastoma
ENCFF575WAS	Encode	HCT116	colon
ENCFF209VEY	Encode	PC-3	prostatic
ENCFF831KZM	Encode	MCF-7	Breast
ENCFF787ITI	Encode	OCI-LY3	non-Hodgkin.lymphoma
ENCFF161GCD	Encode	ACC112	Adenoid.cystic.carcinoma
ENCFF468GKP	Encode	SK-N-SH	neuroblastoma
ENCFF137AXJ	Encode	VCaP	prostate
ENCFF629BRY	Encode	Panc1	pancreatic
ENCFF159JKE	Encode	C4-2B	HPV.cervical
ENCFF623PRE	Encode	22Rv1	prostate
ENCFF279PSG	Encode	OCI-LY1	non-Hodgkin.lymphoma
ENCFF262PTI	Encode	A673	rhabdomyosarcoma
ENCFF152UAP	Encode	PC-9	lung

Table S3.2. Upregulated gene expressions involved in enhancer-promoter interactions occurred in the same RAM in K562 but in different GM12878 RAMs.

Genes	log2(GM12878)	log2(K562)	Difference log2(K562/GM12878)	Foldchange
PTGER3	0.01	1.77	1.76	3.38
CLTCL1	0.05	2.65	2.60	6.07
DGCR2	2.56	4.90	2.34	5.06
HDAC6	1.96	4.27	2.31	4.95
PLP2	5.57	7.04	1.47	2.77
RHAG	0.03	9.71	9.68	820.61
PFKFB4	5.10	5.14	0.04	1.03
HPCAL1	1.93	3.65	1.72	3.29
ARL4A	0.09	6.03	5.93	61.15
GIPC1	2.81	4.37	1.56	2.94
VGF	0.00	0.24	0.24	1.18
APIP	2.74	3.62	0.88	1.84
RAB31	0.55	5.65	5.10	34.29
HTR1F	0.01	1.47	1.46	2.75
MEX3B	1.09	2.85	1.76	3.39
PRELID2	0.13	2.56	2.42	5.37
MITF	0.48	2.53	2.05	4.14

Table S3.3. Upregulated gene expressions involved in enhancer-promoter interactions occurred in the same RAM in K562 but in different HEPG2 RAMs.

Genes	log2(HEPG2)	log2(K562)	Difference log2(K562/HEPG2)	Foldchange
ZNF582	0.63	1.34	0.71	1.64
PTGER3	0.01	1.77	1.76	3.39
CLTCL1	2.30	2.65	0.36	1.28
PTPRC	0.00	3.00	3.00	7.99
RNF24	3.26	4.82	1.56	2.95
CTSC	5.01	5.10	0.10	1.07
KAT2B	1.29	3.23	1.94	3.85
TPST2	4.07	6.17	2.09	4.27
CCDC74A	0.06	3.20	3.14	8.83
ZEB2	0.01	4.35	4.34	20.20
SNX18	1.59	2.32	0.74	1.66
HTR1F	0.00	1.47	1.47	2.76
MEX3B	0.68	2.85	2.17	4.51
ZNF431	1.80	4.12	2.32	4.98
CD47	0.86	3.75	2.89	7.40

Table S3.4. DEGs annotated with Gene Ontology Molecular Function of “Calcium Ion Binding”.

Gene	log2(Mean of Cancer)	log2(Mean of Normal)	log2(Cancer/Normal)	Pvalue
AIF1L	7.86	25.48	-1.70	1.40E-02
CABP1	0.16	4.40	-4.79	6.66E-06
CDH1	49.64	24.00	1.05	4.46E-03
CDH3	16.24	5.80	1.49	9.34E-03
DGKB	0.23	1.16	-2.34	1.62E-03
FAT3	0.31	0.56	-0.86	9.49E-03
HMCN2	0.24	1.43	-2.54	4.03E-03
ITSN1	1.64	2.70	-0.72	4.11E-04
LRP1B	0.20	0.57	-1.49	1.45E-03
MAN1C1	2.14	3.60	-0.75	7.23E-04
MASP1	1.31	3.35	-1.36	5.79E-05
MCTP1	0.64	1.39	-1.12	4.30E-02
MYL3	0.41	8.36	-4.33	2.63E-09
NCS1	7.01	37.53	-2.42	2.37E-03
NDUFAB1	30.63	22.33	0.46	2.93E-05
PCDHA10	0.06	0.17	-1.39	1.09E-02
PCDHA2	0.03	0.08	-1.18	3.57E-02
PCDHA3	0.06	0.14	-1.25	3.82E-03
PCDHA6	0.01	0.10	-2.69	2.64E-03
RYR2	0.13	1.73	-3.78	7.88E-07
SELL	10.52	17.24	-0.71	1.12E-04
SVEP1	0.84	3.08	-1.87	2.39E-04
SYT2	0.07	1.16	-4.12	9.57E-07
VSNL1	2.05	26.51	-3.69	3.34E-02
WDR49	0.10	0.35	-1.86	2.87E-02

3.10 References

1. T. Kouzarides, Chromatin modifications and their function. *Cell*. **128**, 693–705 (2007).

2. B. Li, M. Carey, J. L. Workman, The role of chromatin during transcription. *Cell*. **128**, 707–719 (2007).

3. E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Blüthgen, J. Dekker, E. Heard, Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*. **485** (2012), pp. 381–385.

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

5. T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, G. Cavalli, Three-dimensional folding and functional organization principles of the Drosophila genome. *Cell*. **148**, 458–472 (2012).

6. E. Crane, Q. Bian, R. P. McCord, B. R. Lajoie, B. S. Wheeler, E. J. Ralston, S. Uzawa, J. Dekker, B. J. Meyer, Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature*. **523** (2015), pp. 240–244.

7. S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, E. L. Aiden, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*. **159**, 1665–1680 (2014).

8. E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, J. Dekker, Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*. **326** (2009), pp. 289–293.

9. Y. Zhu, Z. Chen, K. Zhang, M. Wang, D. Medovoy, J. W. Whitaker, B. Ding, N. Li, L. Zheng, W. Wang, Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.* **7**, 10812 (2016).

10. J. Huang, E. Marco, L. Pinello, G.-C. Yuan, Predicting chromatin organization using histone marks. *Genome Biol.* **16**, 162 (2015).

11. H. Ashoor, X. Chen, W. Rosikiewicz, J. Wang, A. Cheng, P. Wang, Y. Ruan, S. Li, Graph embedding and unsupervised learning predict genomic sub-compartments from HiC chromatin interaction data. *Nat. Commun.* **11**, 1173 (2020).

12. A. J. Bannister, T. Kouzarides, Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–395 (2011).

13. M. Vermeulen, H. C. Eberl, F. Matarese, H. Marks, S. Denissov, F. Butter, K. K. Lee, J. V. Olsen, A. A. Hyman, H. G. Stunnenberg, M. Mann, Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell*. **142**, 967–980 (2010).

14. T. Bartke, M. Vermeulen, B. Xhemalce, S. C. Robson, M. Mann, T. Kouzarides, Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell*. **143**, 470–484 (2010).

15. Z. Liu, W. R. Legant, B.-C. Chen, L. Li, J. B. Grimm, L. D. Lavis, E. Betzig, R. Tjian, 3D imaging of Sox2 enhancer clusters in embryonic stem cells. *Elife*. **3**, e04236 (2014).

16. I. I. Cisse, I. Izeddin, S. Z. Causse, L. Boudarene, A. Senecal, L. Muresan, C. Dugast-Darzacq, B. Hajj, M. Dahan, X. Darzacq, Real-time dynamics of RNA polymerase II clustering in live human cells. *Science*. **341**, 664–667 (2013).

17. W.-K. Cho, J.-H. Spille, M. Hecht, C. Lee, C. Li, V. Grube, I. I. Cisse, Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*. **361**, 412–415 (2018).

18. A. Pancholi, T. Klingberg, W. Zhang, R. Prizak, I. Mamontova, A. Noa, M. Sobucki, A. Y. Kobitski, G. U. Nienhaus, V. Zaburdaev, L. Hilbert, RNA polymerase II clusters form in line with surface condensation on regulatory chromatin. *Mol. Syst. Biol.* **17**, e10272 (2021).

19. Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y.-C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K.-H. Farh, S. Feizi, R. Karlic, A.-R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. M. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L.-H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, M. Kellis, Integrative analysis of 111 reference human epigenomes. *Nature*. **518**, 317–330 (2015).

20. W. S. Cleveland, E. Grosse, Computational methods for local regression. *Statistics and Computing*. **1** (1991), pp. 47–62.

21. N. C. Durand, M. S. Shamim, I. Machol, S. S. P. Rao, M. H. Huntley, E. S. Lander, E. L. Aiden, Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst*. **3**, 95–98 (2016).

22. J. Wolff, L. Rabbani, R. Gilsbach, G. Richard, T. Manke, R. Backofen, B. A. Grüning, Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **48**, W177–W184 (2020).

23. F. Ramírez, V. Bhardwaj, L. Arrigoni, K. C. Lam, B. A. Grüning, J. Villaveces, B. Habermann, A. Akhtar, T. Manke, High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, 189 (2018).

24. J. Wolff, V. Bhardwaj, S. Nothjunge, G. Richard, G. Renschler, R. Gilsbach, T. Manke, R. Backofen, F. Ramírez, B. A. Grüning, Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **46**, W11–W16 (2018).

25. K. C. Akdemir, L. Chin, HiCPlotter integrates genomic data with interaction matrices. *Genome Biol.* **16**, 198 (2015).

26. N. C. Durand, M. S. Shamim, I. Machol, S. S. P. Rao, M. H. Huntley, E. S. Lander, E. L. Aiden, Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst*. **3**, 95–98 (2016).

27. K. Y. Yip, C. Cheng, N. Bhardwaj, J. B. Brown, J. Leng, A. Kundaje, J. Rozowsky, E. Birney, P. Bickel, M. Snyder, M. Gerstein, Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13**, R48 (2012).

28. R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, L. Chen, Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.* **30**, 90–98 (2011).

29. E. Crane, Q. Bian, R. P. McCord, B. R. Lajoie, B. S. Wheeler, E. J. Ralston, S. Uzawa, J. Dekker, B. J. Meyer, Condensin-Driven Remodeling of X-Chromosome Topology during Dosage Compensation. *Nature*. **523**, 240 (2015).

30. T. van Schaik, M. Vos, D. Peric-Hupkes, P. Hn Celie, B. van Steensel, Cell cycle dynamics of lamina-associated DNA. *EMBO Rep.* **21**, e50636 (2020).

31. S. S. P. Rao, S.-C. Huang, B. G. St Hilair, J. M. Engreitz, E. M. Perez, K.-R. Kieffer-Kwon, A. L. Sanborn, S. E. Johnstone, I. D. Bochkov, X. Huang, M. S. Shamim, A. D. Omer, B. E. Bernstein, R. Casellas, E. S. Lander, E. L. Aiden, Cohesin loss eliminates all loop domains, leading to links among superenhancers and downregulation of nearby genes, , doi:10.1101/139782.

32. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler

transform. *Bioinformatics*. **26**, 589–595 (2010).

33. T. Liu, Use model-based Analysis of ChIP-Seq (MACS) to analyze short reads generated by sequencing protein-DNA interactions in embryonic stem cells. *Methods Mol. Biol.* **1150**, 81–95 (2014).

34. Pan-cancer analysis of whole genomes. *Nature*. **578**, 82–93 (2020).

35. W. Jiao, G. Atwal, P. Polak, R. Karlic, E. Cuppen, PCAWG Tumor Subtypes and Clinical Translation Working Group, A. Danyi, J. de Ridder, C. van Herpen, M. P. Lolkema, N. Steeghs, G. Getz, Q. Morris, L. D. Stein, PCAWG Consortium, A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.* **11**, 728 (2020).

36. S. Neph, M. S. Kuehn, A. P. Reynolds, E. Haugen, R. E. Thurman, A. K. Johnson, E. Rynes, M. T. Maurano, J. Vierstra, S. Thomas, R. Sandstrom, R. Humbert, J. A. Stamatoyannopoulos, BEDOPS: high-performance genomic feature operations. *Bioinformatics*. **28**, 1919–1920 (2012).

37. S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, C. K. Glass, Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*. **38** (2010), pp. 576–589.

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

39. C. A. Sloan, E. T. Chan, J. M. Davidson, V. S. Malladi, J. S. Strattan, B. C. Hitz, I. Gabdank, A. K. Narayanan, M. Ho, B. T. Lee, L. D. Rowe, T. R. Dreszer, G. Roe, N. R. Podduturi, F. Tanaka, E. L. Hong, J. M. Cherry, ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44**, D726–32 (2016).

40. J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, J. P. Mesirov, Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).

41. R. Foisner, L. Gerace, Integral membrane proteins of the nuclear envelope interact with lamins and chromosomes, and binding is modulated by mitotic phosphorylation. *Cell*. **73**, 1267–1279 (1993).

42. K. L. Wydner, J. A. McNeil, F. Lin, H. J. Worman, J. B. Lawrence, Chromosomal assignment of human nuclear envelope protein genes LMNA, LMNB1, and LBR by fluorescence in situ hybridization. *Genomics*. **32**, 474–478 (1996).

43. M. Gasperini, A. J. Hill, J. L. McFaline-Figueroa, B. Martin, S. Kim, M. D. Zhang, D. Jackson, A. Leith, J. Schreiber, W. S. Noble, C. Trapnell, N. Ahituv, J. Shendure, A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*. **176**, 1516 (2019).

44. C. P. Fulco, J. Nasser, T. R. Jones, G. Munson, D. T. Bergman, V. Subramanian, S. R. Grossman, R. Anyoha, B. R. Doughty, T. A. Patwardhan, T. H. Nguyen, M. Kane, E. M. Perez, N. C. Durand, C. A. Lareau, E. K. Stamenova, E. L. Aiden, E. S. Lander, J. M. Engreitz, Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics*. **51** (2019), pp. 1664–1669.

45. A.-L. Valton, J. Dekker, TAD disruption as oncogenic driver. *Curr. Opin. Genet. Dev.* **36**, 34–40 (2016).

46. A. D. Rouillard, G. W. Gundersen, N. F. Fernandez, Z. Wang, C. D. Monteiro, M. G. McDermott, A. Ma'ayan, The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* . **2016** (2016), doi:10.1093/database/baw100.

47. M. Cremer, K. Brandstetter, A. Maiser, S. S. P. Rao, V. J. Schmid, M. Guirao-Ortiz, N. Mitra, S. Mamberti, K. N. Klein, D. M. Gilbert, H. Leonhardt, M. C. Cardoso, E. L. Aiden, H. Harz, T. Cremer, Cohesin depleted cells rebuild functional nuclear compartments after endomitosis. *Nat. Commun.* **11**, 6146 (2020).

48. S. S. P. Rao, S.-C. Huang, B. Glenn St Hilaire, J. M. Engreitz, E. M. Perez, K.-R. Kieffer-Kwon, A. L. Sanborn, S. E. Johnstone, G. D. Bascom, I. D. Bochkov, X. Huang, M. S. Shamim, J. Shin, D. Turner, Z. Ye, A. D. Omer, J. T. Robinson, T. Schlick, B. E. Bernstein, R. Casellas, E. S. Lander, E. L. Aiden, Cohesin Loss Eliminates All Loop Domains. *Cell*. **171**, 305–320.e24 (2017).

49. K. M. Turner, V. Deshpande, D. Beyter, T. Koga, J. Rusert, C. Lee, B. Li, K. Arden, B. Ren, D. A. Nathanson, H. I. Kornblum, M. D. Taylor, S. Kaushal, W. K. Cavenee, R. Wechsler-Reya, F. B. Furnari, S. R. Vandenberg, P. N. Rao, G. M. Wahl, V. Bafna, P. S. Mischel, Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature*. **543**, 122–125 (2017).

50. R. G. W. Verhaak, V. Bafna, P. S. Mischel, Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. *Nat. Rev. Cancer*. **19**, 283–288 (2019).

51. A. C. deCarvalho, H. Kim, L. M. Poisson, M. E. Winn, C. Mueller, D. Cherba, J. Koeman, S. Seth, A. Protopopov, M. Felicella, S. Zheng, A. Multani, Y. Jiang, J. Zhang, D.-H. Nam, E. F. Petricoin, L. Chin, T. Mikkelsen, R. G. W. Verhaak, Discordant inheritance of chromosomal and extrachromosomal DNA elements contributes to dynamic disease evolution in glioblastoma. *Nat. Genet.* **50**, 708–717 (2018).

52. H. Kim, N.-P. Nguyen, K. Turner, S. Wu, A. D. Gujar, J. Luebeck, J. Liu, V. Deshpande, U. Rajkumar, S. Namburi, S. B. Amin, E. Yi, F. Menghi, J. H. Schulte, A. G. Henssen, H. Y. Chang, C. R. Beck, P. S. Mischel, V. Bafna, R. G. W. Verhaak, Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat. Genet.* **52**, 891–897 (2020).

53. C. Y. McLean, D. Bristor, M. Hiller, S. L. Clarke, B. T. Schaar, C. B. Lowe, A. M. Wenger, G. Bejerano, GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).

54. S. Sigismund, D. Avanzato, L. Lanzetti, Emerging functions of the EGFR in cancer. *Mol. Oncol.* **12**, 3–20 (2018).

55. I. S. Babina, N. C. Turner, Advances and challenges in targeting FGFR signalling in cancer. *Nat. Rev. Cancer*. **17**, 318–332 (2017).

56. M. Sahraei, L. D. Roy, J. M. Curry, T. L. Teresa, S. Nath, D. Besmer, A. Kidiyoor, R. Dalia, S. J. Gendler, P. Mukherjee, MUC1 regulates PDGFA expression during pancreatic cancer progression. *Oncogene*. **31**, 4935–4945 (2012).

57. J. Zhou, Sequence-based modeling of genome 3D architecture from kilobase to chromosome-scale. *bioRxiv* (2021), p. 2021.05.19.444847.

58. T. I. P.-C. A. of W. G. Consortium, The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, Pan-cancer analysis of whole genomes. *Nature*. **578** (2020), pp. 82–93.

59. U. Raudvere, L. Kolberg, I. Kuzmin, T. Arak, P. Adler, H. Peterson, J. Vilo, g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).

60. V. A. Bloomfield, Condensation of DNA by multivalent cations: considerations on mechanism. *Biopolymers*. **31**, 1471–1481 (1991).

61. I. Koltover, K. Wagner, C. R. Safinya, DNA condensation in two dimensions. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 14046–14051 (2000).

62. A. Srivastava, R. Timsina, S. Heo, S. W. Dewage, S. Kirmizialtin, X. Qiu, Structure-guided DNA-DNA attraction mediated by divalent cations. *Nucleic Acids Res.* **48**, 7018–7026 (2020).

63. C. A. Davey, T. J. Richmond, DNA-dependent divalent cation binding in the nucleosome core particle. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 11169–11174 (2002).

64. Z. Yang, J. J. Hayes, The divalent cations Ca2+ and Mg2+ play specific roles in stabilizing histone-DNA interactions within nucleosomes that are partially redundant with the core histone tail domains. *Biochemistry*. **50**, 9973–9981 (2011).

Chapter 4. Concluding Remarks

Using the scale-free network on the 3D contacts data from Hi-C, we have identified a few non-coding DNA regions that could form many 3D contacts with other regions as “hubs”. The following CRISPR-Cas9 experiment, Hi-C, and the single-cell RNA-seq analysis have further characterized the impacts on the global 3D chromatin structures and cell fitness upon the hub loci deletion, particularly for the hubs without epigenetic signals. Moreover, we modeled the enhancer-promoter interactions by a small-world network, and then identified dozens of active enhancers and promoters harboring many 3D contacts as “hotspots”. Upon the hotspot deletion by the CRISPR-Cas9 pgRNA genome editing system, broad 3D chromatin organization alterations beyond enhancer-promoter interactions and gene expression changes were observed from the Hi-C and the single-cell RNAseq analysis. Both the hubs and hotspots are strongly associated with genetic variants in cancer cells, which provide new insights into pathology. Although there is a long way to understand the underlying mechanisms of the hubs and hotspots contributing to maintaining the 3D chromatin organization, this is the first time to illuminate the structural importance of the regulatory elements and even non-coding regions without epigenetic signals in 3D genome architecture.

Furthermore, the genome-wide regulation associated modules (“RAMs”) have been identified by investigating epigenetic histone marks across cell types. Pieces of evidence collected from histone modifications, enhancer-promoter interactions, extrachromosomal DNAs (ecDNAs), Hi-C data and LAD signals supported that RAMs are spatial modules and better aligned with the chromatin function. The characterizations of

RAMs tell apart them from compartments and TADs. Compared to the TADs, RAMs are insensitive to loss of cohesin and are predicted to have a more significant impact on chromatin organization through deep learning models. In addition, the cancer somatic indels are highly enriched in RAM boundaries, which highlights the importance of RAMs in human diseases. Although the mechanisms of RAM formation are still elusive, the RAMs provide opportunities to better understand the relationship between the structural and functional modularity of the genome.

Together, this dissertation provides new insights into the 3D chromatin organization and function. Our findings navigate future efforts in deciphering the mysteries of the 3D genome.