UNIVERSITY OF CALIFORNIA SAN DIEGO

Deciphering molecular mechanisms of mammalian insulators and enhancers

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Biomedical Sciences

by

Hui Huang

Committee in charge:

Professor Bing Ren, Chair
Professor Xiang-Dong Fu
Professor James T. Kadonaga
Professor Cornelis Murre
Professor Kun Zhang

2021

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

University of California San Diego

2021

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

mESC: Mouse embryonic stem cell

CBS: CTCF binding sites

SE: Super-enhancer

TAD: Topologically associating domain

FISH: Fluorescence *in situ* hybridization

FACS: Fluorescence activated cell sorting

ChIP-seq: Chromatin immunoprecipitation with sequencing

PLAC-seq: Proximity Ligation-Assisted ChIP-seq

ATAC-seq: Assays for transposase-accessible chromatin using sequencing

GWAS: Genome-wide association studies

SILAC: Stable isotope labeling by/with amino acids in cell culture

NE: Nuclear extract

CRs: Chromatin regulators

LIST OF SUPPLEMENTAL FILES

Huang_Supplementary_Table_2.1-2.3: Probe sequences for multiplexed FISH experiments in Chapter 2.

Huang_Supplementary_Table_3.1: Results from SILAC experiments in Chapter 3.

Huang_Chapter3_Supplemental_Information.

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Professor Bing Ren, my mentor, colleague, and dear friend, for his full support of my graduate research and his tireless guidance over the years. I had countless discussions with Bing on my project, whose insights and rigorousness in science are not only invaluable to my dissertation but also will be a treasure for my entire academic career. I also thank Bing for the support of my long-term development. Bing provided great opportunities to interact with leading scientists, and rich resources to learn cutting-edge technologies. Bing is also extremely supportive during my job search. I am also grateful for the numerous colorful group activities organized by Bing, which greatly enriched my research life.

I would also like to thank Professor Xiang-Dong Fu, Professor James T. Kadonaga, Professor Cornelis Murre, and Professor Kun Zhang for serving on my dissertation committee. Discussions with them greatly facilitated the development of my research. I am also grateful for the support from Professor James T. Kadonaga during my search for postdoctoral positions.

I thank Adam Jusilla and Yuanyuan Han from Ren laboratory, Quan Zhu, and Colin Kern from Center for Epigenomics at UCSD for carrying out the multiplexed imaging experiments and relentless efforts in data analyses.

I would like to thank Bogdan Bintu and Professor Xiaowei Zhuang from Harvard University for their help in designing the multiplexed FISH experiments and the constructive discussions on data analyses.

I also thank Simona Bianco, Andrea M. Chiariello, Mattia Conte and Professor Mario Nicodemi from Università di Napoli Federico II, and INFN Napoli for valuable discussions on the multiplexed FISH experiments. I thank Professor Mario Nicodemi for the kind support during my search for postdoctoral positions.

I would like to thank my colleagues, Miao Yu and Rong Hu for performing chromatin conformation capture experiments and Yanxiao Zhang for advice on data analyses. I thank Melodi Tastemel for performing western blot experiments.

I would also like to thanks Ivan Juric and Professor Ming Hu for performing allele-specific analyses of proximity assisted-ligation ChIP-seq data.

I thank the current and past members of Ren Laboratory, Chenxu Zhu, Yanxiao Zhang, James Hocker, Yang Li, Andrea Local, Yarui Diao, Jian Yan, Miao Yu, Yunjiang Qiu, Rong Hu, Jason Li, Adam Jussila, Quan Zhu, David Gorkin, Sebastian Preissl, Anthony Schmitt, Tristin Liu, Yuan Zhao, Rongxin Fang, Ramya Raviram, Naoki Kubo, Haruhico Ishi, Bin Li who are colleagues and friends of mine. Their friendship has brought me great joy both inside and outside of the laboratory. Especially, I'd like to thank the managers of the Ren laboratory, Ye Zhen, Kuan Samantha, and Bernadeth Torres, whose work brought great convenience to my research. I also thank them for organizing colorful group events and for the delicious homemade cookies and dishes Samantha and Ye shared.

Lastly, I would like to thank my parents Xingfu Huang, Guoxiu Duan, my brother, Tao Huang, and my girlfriend, Linzhen Kong. I have missed all family reunions over the past four years, which has been difficult both for them and me. Their selfless support

and love have guided me through the dark days and lent me the courage to overcome the setbacks in life. I would never have completed this journey without their full support.

Chapter 2, in full, is a reprint of the accepted manuscript in Nature Genetics 2021. Huang, H., Zhu, Q., Jussila, A. P., Han, Y., Bintu, B., Kern, C., Conte, M., Zhang, Y., Bianco, S., Chiariello, A.M., Yu, M., Hu, R., Tastemel, M., Juric, I., Hu, M., Necodemi, M., Zhuang, X., and Ren, B. CTCF mediates dosage- and sequence-context-dependent transcriptional insulation by forming local chromatin domains. Nature Genetics (in press, 2021). The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the paper published in Nature Genetics 2018. Andrea Local*, Hui Huang*, Claudio P. Albuquerque, Namit Singh, Ah Young Lee, Wei Wang, Chaochen Wang, Judy E. Hsia, Andrew K. Shiau, Kai Ge, Kevin D. Corbett, Dong Wang, Huilin Zhou, and Bing Ren. "Identification of H3K4me1-associated proteins at mammalian enhancers." Nature Genetics 50.1 (2018): 73-82. *authors contributed equally to this work. The dissertation author was one of the primary investigators and authors of this paper.

VITA

2015 Bachelor of Biological Sciences, University of Science and Technology of China

2021 Doctor of Philosophy, Biomedical Sciences, University of California San Diego

PUBLICATIONS

1. **Huang, H**., Zhu, Q., Jussila, A. P., Han, Y., Bintu, B., Kern, C., Conte, M., Zhang, Y., Bianco, S., Chiariello, A.M., Yu, M., Hu, R., Tastemel, M., Juric, I., Hu, M., Necodemi, M., Zhuang, X., and Ren, B. CTCF mediates dosage- and sequence-context-dependent transcriptional insulation by forming local chromatin domains. *Nature Genetics*, (2021).

2. Gorkin, D.U*., Barozzi, I*., Zhao, Y*., Zhang, Y*., **Huang, H***., Lee, A.Y., Li, B., Chiou, J., Wildberg, A., Ding, B., Zhang, B., Wang, M., Strattan, J.S., Davidson, J.M., Qiu, Y., Afzal, V., Akiyama, J.A., Plajzer-Frick, I., Novak, C.S., Kato, M., Garvin, T.H., Pham, Q.T., Harrington, A.N., Mannion, B.J., Lee, E.A., Fukuda-Yuzawa, Y., He, Y., Preissl, S., Chee, S., Han, J.Y., Williams, B.A., Trout, D., Amrhein, H., Yang, H., Cherry, J.M., Wang, W., Gaulton, K., Ecker, J.R., Shen, Y., Dickel, D.E., Visel, A., Pennacchio, L.A. & Ren, B. An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744-751 (2020).

3. Local, A*., **Huang, H***., Albuquerque, C.P., Singh, N., Lee, A.Y., Wang, W., Wang, C., Hsia, J.E., Shiau, A.K., Ge, K., Corbett, K.D., Wang, D., Zhou, H. & Ren, B. Identification of H3K4me1-associated proteins at mammalian enhancers. *Nat Genet* **50**, 73-82 (2018).

4. Zhu, C., Yu, M., **Huang, H**., Juric, I., Abnousi, A., Hu, R., Lucero, J., Behrens, M.M., Hu, M. & Ren, B. An ultra high-throughput method for single-cell joint analysis of open chromatin and transcriptome. *Nat Struct Mol Biol* **26**, 1063-1070 (2019).

5. Preissl, S., Fang, R., **Huang, H**., Zhao, Y., Raviram, R., Gorkin, D.U., Zhang, Y., Sos, B.C., Afzal, V., Dickel, D.E., Kuan, S., Visel, A., Pennacchio, L.A., Zhang, K. & Ren, B. Single-nucleus analysis of accessible chromatin in developing mouse

forebrain reveals cell-type-specific transcriptional regulation. *Nat Neurosci* **21**, 432-439 (2018).

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

7. Zhang, Y., Li, T., Preissl, S., Amaral, M.L., Grinstein, J.D., Farah, E.N., Destici, E., Qiu, Y., Hu, R., Lee, A.Y., Chee, S., Ma, K., Ye, Z., Zhu, Q., **Huang, H**., Fang, R., Yu, L., Izpisua Belmonte, J.C., Wu, J., Evans, S.M., Chi, N.C. & Ren, B. Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet* **51**, 1380-1388 (2019).

8. Batra, R., Stark, T.J., Clark, E., Belzile, J.P., Wheeler, E.C., Yee, B.A., **Huang, H**., Gelboin-Burkhart, C., Huelga, S.C., Aigner, S., Roberts, B.T., Bos, T.J., Sathe, S., Donohue, J.P., Rigo, F., Ares, M., Jr., Spector, D.H. & Yeo, G.W. RNA-binding protein CPEB1 remodels host and viral RNA landscapes. *Nat Struct Mol Biol* **23**, 1101-1110 (2016).

*authors contributed equally.

FIELDS OF STUDY

Major Field: Biology

Study in genetics/epigenetics, genomics/epigenomics, and gene regulation

Professor Bing Ren

ABSTRACT OF THE DISSERTATION

Deciphering molecular mechanisms of mammalian insulators and enhancers

by

Hui Huang

Doctor of Philosophy in Biomedical Sciences

University of California San Diego, 2021

Professor Bing Ren, Chair

Gene expression in animals is finely controlled by *cis*-regulatory elements in the non-coding sequences, yet the mechanisms by which they regulate transcription are not fully understood. During my graduate study, I dissected the molecular mechanisms of CTCF-mediated transcriptional insulation, explored the mechanisms of how monomethylation on histone H3 lysine 4 (H3K4me1) facilitates enhancer function, and

finally, characterized the dynamic regulatory landscape during mouse embryonic development. Chapter 1 is an overview of the (epi)genomics field. I introduce the background of my three research projects and summarize the major findings. In chapter 2, I systematically introduce my research on the mechanisms of CTCF-mediated transcriptional insulation. I investigate the context-specific insulator function of CTCF bound DNA elements using an insulator reporter assay in mouse embryonic stem cells. I demonstrate that insulation strength depends on the number of CTCF binding sites in tandem, the upstream flanking sequences, and the 9-11 zinc fingers of CTCF protein. Further, I find insulators are sufficient to create chromatin boundaries and reduce enhancer-promoter communications. In chapter 3, colleagues and I identify multiple proteins associated with H3K4me1 in the nucleus. We demonstrate that H3K4me1 facilitates the recruitment of BAF complex on active enhancers. The details of other projects of my graduate research are not included in this dissertation.

CHAPTER 1: Introduction

1.1 The significance of understanding the regulatory genome

It is once believed that finding out all the genes in the genome would reveal all the secrets of life. However, a large proportion of sequences in the genome are not protein-coding but rather regulate gene expression, including enhancers, insulators, silencers, and many other less-well characterized *cis*-regulatory elements[1-3]. Enhancers are a type of *cis*-regulatory elements that facilitate the expression of the target genes in a position-independent manner[4, 5]. By contrast, insulators can block activation signals from enhancers only when located between enhancer the target gene[6, 7]. Enhancers, insulators, and many other *cis*-regulatory elements orchestrate the spatial and temporal gene expression in higher-order organisms.

Many genetic diseases are caused by mutations in the non-coding genome[8]. Genome-wide association studies reveal that the majority of genetic variants linked with human diseases lie in non-coding genome[9, 10]. Additionally, many non-coding regulatory elements are involved in the evolution and control of development[11]. For instance, mutations in the ZRS enhancer are responsible for the loss of limbs in snakes[12]. Therefore, understanding the regulatory functions of the non-coding genome bears great value in biomedical researches.

1.2 Identification and characterization of *cis*-regulatory elements

The first few long-range regulatory elements acting on eukaryotic genes were discovered by genetic tests in cell lines and model organisms[4-7]. However, this approach is laborious and is hardly applicable to regulatory elements that control complex traits. The development of second-generation sequencing technology allows the fast and cheap surveys of thousands of genomes of individuals. Genome-wide association studies (GWAS) links genetic variations in the population to particular diseases and traits[10, 13]. In combination with expression data, genetic polymorphisms that contribute to individual differences in gene expression can be statistically identified[14]. These genetic variants in non-coding sequences mark the location of candidate *cis*-regulatory elements in the genome.

Different types of *cis*-regulatory elements are associated with specific histone modifications and transcriptional factors. Active promoters are enriched for trimethylation on Histone H3 lysine 4 (H3K4me3), whereas enhancers are devoid of H3K4me3 but enriched for monomethylation on Histone H3 lysine 4 (H3K4me1)[15]. Active enhancers are additionally marked by acetylation on Histone H3 lysine 27, which can be deposited by CBP/P300[16-18]. Insulators are featured by the binding of CCCTC factor (CTCF) in mammalian cells[19]. Epigenetic signatures can be surveyed by chromatin immunoprecipitation technologies and are commonly used to predict *cis*-regulatory elements in the genome[15, 20]. Additionally, functional elements are less tightly wrapped by histones, resulting in DNA regions that are hypersensitive to DNase I

digestion. The high-throughput measure of DNase I hypersensitive sites by sequencing (DHS-seq) has been a powerful tool to map different types of candidate *cis*-regulatory elements at the genome-wide scale[21, 22]. In 2013, Greenleaf and colleagues developed an assay for transposase-accessible chromatin using sequencing (ATAC-seq) to survey open chromatin regions[23]. Because of its simplicity and robustness, ATAC-seq has been widely used to assay *cis*-regulatory elements in different cell types and tissues. To understand how cis-regulatory elements program tissue development, colleagues and I systematically examined the *cis*-regulatory elements in mouse fetal development (not included in this dissertation)[20]. We profiled eight histone modifications and chromatin accessibility in 72 distinct tissue stages. Our data provide the most comprehensive view of the regulatory landscape in mammalian fetal development. However, animal tissues are composed of complex cell types. It is necessary to characterize cis-regulatory elements in single-cell resolution to delineate cell-type-specific regulatory programs. In recent years, dozens of high-throughput single-cell technologies have been developed using transposase-medicated cell indexing or microfluidics barcoding system[24], which greatly facilities the understanding of the cell-type-specific regulatory programs in organ development and disease progression.

Enhancers can act over large linear distances. For instance, the expression of sonic hedgehog (*Shh*) in the limb bud is activated by the ZRS enhancer located about 1Mb away[25]. It is intriguing how *cis*-regulatory elements find the target genes. In 2009, Dekker and colleagues developed a high-throughput chromosome conformation capture assay (Hi-C) to investigate the three-dimensional organization of the genome[26]. Later

on, the DNA-DNA proximity ligation step of Hi-C was performed in intact nuclei (*in situ* Hi-C), which simplified experimental procedures and enabled higher resolution[27]. Hi-C has been used to identify candidate *cis*-regulatory elements of specific genes, associate genetic risks to target genes, and detect structural variations in cancer genomes[27-29]. Hi-C can also be combined with immunoprecipitation to investigate chromatin interactions centered around the specific protein or histone modifications[30, 31]. These technologies not only facilitate the identification of candidate *cis*-regulatory elements but also provide the tools to understand how *cis*-regulatory elements are organized in 3D space.

An intriguing question in gene regulation is how distal enhancers find their targets. Thanks to Hi-C, it is now recognized that the mammalian genome is partitioned into mega-base-sized topologically associating domains (TADs)[32, 33]. TADs have been considered as basic architecture units that define the range of enhancer action. Disruption of TAD structure allows ectopic enhancer-gene interactions across TADs, leading to pathologic transcription in many diseases, including developmental disorder and cancers[34-36]. Boundaries of TADs are enriched for binding sites of the insulator protein CTCF[32]. The CTCF binding sites at TAD boundaries are predominantly positioned in convergent orientation, with the asymmetric motifs facing inward the TAD structures[27]. TADs are hypothesized to be formed through “loop extrusion”, by which cohesin complex binds to chromatin, extrudes it as a loop, and gets hindered when encounters inward-facing CTCF[37-39]. This model is supported by a large amount of evidence. Firstly, the loop extrusion model can predict changes in chromatin structure caused by genetic engineering of CTCF binding sites[39, 40]. Secondly, the global TAD

structures are eliminated upon acute depletion of CTCF and cohesin complex by auxin-inducible degron system[41, 42]. Additionally, extrusion of DNA loops by cohesin complex has been observed in *in vitro* systems[43, 44].

1.3 The mechanisms of action of mammalian insulators and enhancers

CTCF, in cooperation with the cohesin complex, creates boundaries between topologically associating domains[41, 42]. However, CTCF is also known as an insulator protein in vertebrates. It is unclear whether insulators are sufficient to establish TAD boundaries. Acute deletion of CTCF significantly weakened TAD structures, whereas only moderate changes in transcription were observed, with a similar number of genes upregulated and downregulated[42]. Furthermore, deletion of CTCF binding sites at the long non-coding RNA locus *Firre* leads to little alterations in local chromatin structure[45]. Additionally, the majority of CTCF binding sites (>80%) in the genome do not coincide with TAD boundaries[32]. To address this question, I developed an insulator reporter assay in mouse embryonic stem cells in chapter 2. I demonstrate that multiple insulator elements are sufficient to create chromatin domains *de novo* and reduce enhancer-promoter communications.

Although mostly known as an insulator protein, CTCF has been found to function as a transcription factor capable of repressing or activating gene expression in heterologous reporter assays[46, 47]. CTCF binding sites within TADs have been reported

to facilitate enhancer-promoter interactions and reduce cell-to-cell variation of gene expression[48]. It is unclear exactly how CTCF functions as insulators.

CTCF is a highly conserved DNA binding factor with eleven zinc fingers. The first genome-wide distribution of CTCF binding sites was mapped by ChIP-chip, a method based on immunoprecipitation and genome-tilling arrays[49]. Most *in vivo* CTCF binding sites share a GC-rich 20-bp motif, while the adjacent sequences are highly variable[49]. It turns out that CTCF only employs its central zinc fingers (zinc fingers 3-8) to recognize the core consensus motif[50, 51]. Peripheral zinc fingers are required for CTCF binding at different subsets of genomic locations and regulate the residence time of CTCF on chromatin[52]. Through combinatorial use of the eleven zinc fingers, CTCF is hypothesized to recognize diverse sequences, interact with distinct co-factors, and carry out various functions[53].

In chapter 2, I analyzed the insulation effects of different CTCF binding sites. I demonstrate that CTCF-mediated transcriptional insulation depends on the combinatorial effects of multiple factors, including the number of CTCF binding sites, the upstream flanking sequences, and the 9-11 peripheral zinc fingers of CTCF itself. The results provide novel insights on how insulators work in the genome. Recently, a CTCF N-terminal segment has been shown to enable loop formation by stabilizing cohesin residence on chromatin[54]. It is possible that involving CTCF 9-11 zinc fingers in DNA binding induces a conformation change that facilitates interactions between the N-terminal segment and cohesin complex, thereby enhancing chromatin loop formation.

Future work is needed to illustrate the molecular mechanisms of CTCF-dependent insulator elements.

Although H3K4me1 is a predictive mark for enhancers, whether H3K4me1 directly regulates or simply correlates with enhancer activity is poorly understood. H3K4me1 at enhancers is mainly deposited by H3K4 methyltransferase KMT2C and KMT2D (also known as MLL3 and MLL4)[55]. Catalytic deficient mutation of KMT2C and KMT2D in *Drosophila* abrogated the deposition of H3K4me1, yet the mutant exhibit little developmental defects[56]. Further, KMT2C/D null instead of catalytic dead mutant mESCs showed a significant reduction in Pol II loading and eRNA production on enhancers[57]. However, KMT2C/D null mutations are lethal to both *Drosophila* and mice. It is argued that the methyltransferases KMT2C and KMT2D are essential for enhancer function, whereas the H3K4me1 modification catalyzed by them is dispensable.

However, H3K4me1 is known to inhibit DNA methylation and block binding of H3K4me3-associated factors to enhancers such as ING1[58, 59]. It should also be noted that H3K4me1 marks primed enhancers, a state where enhancers are not yet activated but are ready to respond to stimuli. Additionally, KMT2C and KMT2D catalytic deficient flies are susceptible to environmental stress and genetic perturbations[56]. These observations highlight the possibility that H3K4me1 positively regulates enhancer function, especially for enhancers that are responsive to environmental stimuli.

In chapter 3, colleagues and I identified chromatin regulators associated with H3K4me1 by nucleosome pulldown coupled with SILAC (stable isotope labeling by amino acids in cell culture) mass spectrometry analysis[60]. We identified multiple chromatin regulators, including the BAF complex, with preferential association with

mononucleosomes bearing H3K4me1 over H3K4me3 modification. We further demonstrate that H3K4me1 recruits BAF complex to distal enhancers to facilitate transcription of target genes. Our results highlight that H3K4me1 plays an active role in the functions of *cis*-regulatory elements in mammalian cells.

Gene expression in mammalian cells is specified by complex and dynamic regulatory networks. New technologies are being developed to examine *cis*-regulatory elements at single-cell resolution and simultaneously acquire multi-dimensional information including histone/DNA modifications, genome organization, gene expression, and binding of transcription factors. Many exciting discoveries will be made in the near future.

1.4 References

1. Levine, M., Cattoglio, C. & Tjian, R. Looping back to leap forward: transcription enters a new era. *Cell* **157**, 13-25 (2014).

2. Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E., Dekker, J., Dunham, I., Elnitski, L.L., Farnham, P.J., Feingold, E.A., Gerstein, M., Giddings, M.C., Gilbert, D.M., Gingeras, T.R., Green, E.D., Guigo, R., Hubbard, T., Kent, J., Lieb, J.D., Myers, R.M., Pazin, M.J., Ren, B., Stamatoyannopoulos, J.A., Weng, Z., White, K.P. & Hardison, R.C. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* **111**, 6131-6138 (2014).

3. Hnisz, D., Day, D.S. & Young, R.A. Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell* **167**, 1188-1200 (2016).

4. Moreau, P., Hen, R., Wasylyk, B., Everett, R., Gaub, M.P. & Chambon, P. The SV40 72 base repair repeat has a striking effect on gene expression both in SV40 and other chimeric recombinants. *Nucleic Acids Res* **9**, 6047-6068 (1981).

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

6. West, A.G., Gaszner, M. & Felsenfeld, G. Insulators: many functions, many mechanisms. *Genes Dev* **16**, 271-288 (2002).

7. Geyer, P.K. & Corces, V.G. DNA position-specific repression of transcription by a Drosophila zinc finger protein. *Genes Dev* **6**, 1865-1873 (1992).

8. Valente, E.M. & Bhatia, K.P. Solving Mendelian Mysteries: The Non-coding Genome May Hold the Key. *Cell* **172**, 889-891 (2018).

9. Zhang, F. & Lupski, J.R. Non-coding genetic variants in human disease. *Hum Mol Genet* **24**, R102-110 (2015).

10. Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A.K., Canfield, T.K., Giste, E., Diegel,

M., Bates, D., Hansen, R.S., Neph, S., Sabo, P.J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S.R., Kaul, R. & Stamatoyannopoulos, J.A. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195 (2012).

11. McLean, C.Y., Reno, P.L., Pollen, A.A., Bassan, A.I., Capellini, T.D., Guenther, C., Indjeian, V.B., Lim, X., Menke, D.B., Schaar, B.T., Wenger, A.M., Bejerano, G. & Kingsley, D.M. Human-specific loss of regulatory DNA and the evolution of human-specific traits. *Nature* **471**, 216-219 (2011).

12. Kvon, E.Z., Kamneva, O.K., Melo, U.S., Barozzi, I., Osterwalder, M., Mannion, B.J., Tissieres, V., Pickle, C.S., Plajzer-Frick, I., Lee, E.A., Kato, M., Garvin, T.H., Akiyama, J.A., Afzal, V., Lopez-Rios, J., Rubin, E.M., Dickel, D.E., Pennacchio, L.A. & Visel, A. Progressive Loss of Function in a Limb Enhancer during Snake Evolution. *Cell* **167**, 633-642 e611 (2016).

13. Mills, M.C. & Rahal, C. A scientometric review of genome-wide association studies. *Commun Biol* **2**, 9 (2019).

14. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-585 (2013).

15. Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., Wang, W., Weng, Z., Green, R.D., Crawford, G.E. & Ren, B. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318 (2007).

16. Raisner, R., Kharbanda, S., Jin, L., Jeng, E., Chan, E., Merchant, M., Haverty, P.M., Bainer, R., Cheung, T., Arnott, D., Flynn, E.M., Romero, F.A., Magnuson, S. & Gascoigne, K.E. Enhancer Activity Requires CBP/P300 Bromodomain-Dependent Histone H3K27 Acetylation. *Cell Rep* **24**, 1722-1729 (2018).

17. Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A. & Wysocka, J. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283 (2011).

18. Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., Boyer, L.A., Young, R.A. & Jaenisch, R. Histone H3K27ac separates active from poised

enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936 (2010).

19. Farrell, C.M., West, A.G. & Felsenfeld, G. Conserved CTCF insulator elements flank the mouse and human beta-globin loci. *Mol Cell Biol* **22**, 3820-3831 (2002).

20. Gorkin, D.U., Barozzi, I., Zhao, Y., Zhang, Y., Huang, H., Lee, A.Y., Li, B., Chiou, J., Wildberg, A., Ding, B., Zhang, B., Wang, M., Strattan, J.S., Davidson, J.M., Qiu, Y., Afzal, V., Akiyama, J.A., Plajzer-Frick, I., Novak, C.S., Kato, M., Garvin, T.H., Pham, Q.T., Harrington, A.N., Mannion, B.J., Lee, E.A., Fukuda-Yuzawa, Y., He, Y., Preissl, S., Chee, S., Han, J.Y., Williams, B.A., Trout, D., Amrhein, H., Yang, H., Cherry, J.M., Wang, W., Gaulton, K., Ecker, J.R., Shen, Y., Dickel, D.E., Visel, A., Pennacchio, L.A. & Ren, B. An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744-751 (2020).

21. Meuleman, W., Muratov, A., Rynes, E., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Teodosiadis, A., Reynolds, A., Haugen, E., Nelson, J., Johnson, A., Frerker, M., Buckley, M., Sandstrom, R., Vierstra, J., Kaul, R. & Stamatoyannopoulos, J. Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244-251 (2020).

22. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T.K., Diegel, M., Dunn, D., Ebersol, A.K., Frum, T., Giste, E., Johnson, A.K., Johnson, E.M., Kutyavin, T., Lajoie, B., Lee, B.K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Stamatoyannopoulos, G., Iyer, V.R., Lieb, J.D., Sunyaev, S.R., Akey, J.M., Sabo, P.J., Kaul, R., Furey, T.S., Dekker, J., Crawford, G.E. & Stamatoyannopoulos, J.A. The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82 (2012).

23. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218 (2013).

24. Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nat Methods* **17**, 11-14 (2020).

25. Lettice, L.A., Heaney, S.J., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E. & de Graaff, E. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum Mol Genet* **12**, 1725-1735 (2003).

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

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

28. Javierre, B.M., Burren, O.S., Wilder, S.P., Kreuzhuber, R., Hill, S.M., Sewitz, S., Cairns, J., Wingett, S.W., Varnai, C., Thiecke, M.J., Burden, F., Farrow, S., Cutler, A.J., Rehnstrom, K., Downes, K., Grassi, L., Kostadima, M., Freire-Pritchett, P., Wang, F., Consortium, B., Stunnenberg, H.G., Todd, J.A., Zerbino, D.R., Stegle, O., Ouwehand, W.H., Frontini, M., Wallace, C., Spivakov, M. & Fraser, P. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**, 1369-1384 e1319 (2016).

29. Dixon, J.R., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V.T., Yardimci, G.G., Chakraborty, A., Bann, D.V., Wang, Y., Clark, R., Zhang, L., Yang, H., Liu, T., Iyyanki, S., An, L., Pool, C., Sasaki, T., Rivera-Mulia, J.C., Ozadam, H., Lajoie, B.R., Kaul, R., Buckley, M., Lee, K., Diegel, M., Pezic, D., Ernst, C., Hadjur, S., Odom, D.T., Stamatoyannopoulos, J.A., Broach, J.R., Hardison, R.C., Ay, F., Noble, W.S., Dekker, J., Gilbert, D.M. & Yue, F. Integrative detection and analysis of structural variation in cancer genomes. *Nat Genet* **50**, 1388-1398 (2018).

30. Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J. & Chang, H.Y. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13**, 919-922 (2016).

31. Fang, R., Yu, M., Li, G., Chee, S., Liu, T., Schmitt, A.D. & Ren, B. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res* **26**, 1345-1348 (2016).

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

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

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

35. Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schopflin, R., Kraft, K., Kempfer, R., Jerkovic, I., Chan, W.L., Spielmann, M., Timmermann, B., Wittler, L., Kurth, I., Cambiaso, P., Zuffardi, O., Houge, G., Lambie, L., Brancati, F., Pombo, A., Vingron, M., Spitz, F. & Mundlos, S. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265-269 (2016).

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

37. Fudenberg, G., Abdennur, N., Imakaev, M., Goloborodko, A. & Mirny, L.A. Emerging Evidence of Chromosome Folding by Loop Extrusion. *Cold Spring Harb Symp Quant Biol* **82**, 45-55 (2017).

38. Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N. & Mirny, L.A. Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep* **15**, 2038-2049 (2016).

39. Sanborn, A.L., Rao, S.S., Huang, S.C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K.P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E.K., Lander, E.S. & Aiden, E.L. Chromatin extrusion explains key features of loop and domain formation in wild-

type and engineered genomes. *Proc Natl Acad Sci U S A* **112**, E6456-6465 (2015).

40. Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., Lu, Y., Wu, Y., Jia, Z., Li, W., Zhang, M.Q., Ren, B., Krainer, A.R., Maniatis, T. & Wu, Q. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**, 900-910 (2015).

41. Rao, S.S.P., Huang, S.C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., Huang, X., Shamim, M.S., Shin, J., Turner, D., Ye, Z., Omer, A.D., Robinson, J.T., Schlick, T., Bernstein, B.E., Casellas, R., Lander, E.S. & Aiden, E.L. Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320 e324 (2017).

42. Nora, E.P., Goloborodko, A., Valton, A.L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A. & Bruneau, B.G. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944 e922 (2017).

43. Kim, Y., Shi, Z., Zhang, H., Finkelstein, I.J. & Yu, H. Human cohesin compacts DNA by loop extrusion. *Science* **366**, 1345-1349 (2019).

44. Davidson, I.F., Bauer, B., Goetz, D., Tang, W., Wutz, G. & Peters, J.M. DNA loop extrusion by human cohesin. *Science* **366**, 1338-1345 (2019).

45. Barutcu, A.R., Maass, P.G., Lewandowski, J.P., Weiner, C.L. & Rinn, J.L. A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. *Nat Commun* **9**, 1444 (2018).

46. Baniahmad, A., Steiner, C., Kohne, A.C. & Renkawitz, R. Modular structure of a chicken lysozyme silencer: involvement of an unusual thyroid hormone receptor binding site. *Cell* **61**, 505-514 (1990).

47. Lobanenkov, V.V., Nicolas, R.H., Adler, V.V., Paterson, H., Klenova, E.M., Polotskaja, A.V. & Goodwin, G.H. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene* **5**, 1743-1753 (1990).

48. Ren, G., Jin, W., Cui, K., Rodrigez, J., Hu, G., Zhang, Z., Larson, D.R. & Zhao, K. CTCF-Mediated Enhancer-Promoter Interaction Is a Critical Regulator of Cell-to-Cell Variation of Gene Expression. *Mol Cell* **67**, 1049-1058 e1046 (2017).

49. Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V. & Ren, B. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* **128**, 1231-1245 (2007).

50. Yin, M., Wang, J., Wang, M., Li, X., Zhang, M., Wu, Q. & Wang, Y. Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites. *Cell Res* **27**, 1365-1377 (2017).

51. Hashimoto, H., Wang, D., Horton, J.R., Zhang, X., Corces, V.G. & Cheng, X. Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. *Mol Cell* **66**, 711-720 e713 (2017).

52. Nakahashi, H., Kieffer Kwon, K.R., Resch, W., Vian, L., Dose, M., Stavreva, D., Hakim, O., Pruett, N., Nelson, S., Yamane, A., Qian, J., Dubois, W., Welsh, S., Phair, R.D., Pugh, B.F., Lobanenkov, V., Hager, G.L. & Casellas, R. A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep* **3**, 1678-1689 (2013).

53. Ohlsson, R., Lobanenkov, V. & Klenova, E. Does CTCF mediate between nuclear organization and gene expression? *Bioessays* **32**, 37-50 (2010).

54. Li, Y., Haarhuis, J.H.I., Sedeno Cacciatore, A., Oldenkamp, R., van Ruiten, M.S., Willems, L., Teunissen, H., Muir, K.W., de Wit, E., Rowland, B.D. & Panne, D. The structural basis for cohesin-CTCF-anchored loops. *Nature* **578**, 472-476 (2020).

55. Herz, H.M., Mohan, M., Garruss, A.S., Liang, K., Takahashi, Y.H., Mickey, K., Voets, O., Verrijzer, C.P. & Shilatifard, A. Enhancer-associated H3K4 monomethylation by Trithorax-related, the Drosophila homolog of mammalian Mll3/Mll4. *Genes Dev* **26**, 2604-2620 (2012).

56. Rickels, R., Herz, H.M., Sze, C.C., Cao, K., Morgan, M.A., Collings, C.K., Gause, M., Takahashi, Y.H., Wang, L., Rendleman, E.J., Marshall, S.A., Krueger, A., Bartom, E.T., Piunti, A., Smith, E.R., Abshiru, N.A., Kelleher, N.L., Dorsett, D. & Shilatifard, A. Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nat Genet* **49**, 1647-1653 (2017).

57. Dorighi, K.M., Swigut, T., Henriques, T., Bhanu, N.V., Scruggs, B.S., Nady, N., Still, C.D., 2nd, Garcia, B.A., Adelman, K. & Wysocka, J. Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Mol Cell* **66**, 568-576 e564 (2017).

58. Ooi, S.K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.P., Allis, C.D., Cheng, X. & Bestor, T.H. DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448**, 714-717 (2007).

59. Cheng, J., Blum, R., Bowman, C., Hu, D., Shilatifard, A., Shen, S. & Dynlacht, B.D. A role for H3K4 monomethylation in gene repression and partitioning of chromatin readers. *Mol Cell* **53**, 979-992 (2014).

60. Local, A., Huang, H., Albuquerque, C.P., Singh, N., Lee, A.Y., Wang, W., Wang, C., Hsia, J.E., Shiau, A.K., Ge, K., Corbett, K.D., Wang, D., Zhou, H. & Ren, B. Identification of H3K4me1-associated proteins at mammalian enhancers. *Nat Genet* **50**, 73-82 (2018).

CHAPTER 2. CTCF mediates dosage and sequence-context-dependent transcriptional insulation by forming local chromatin domains

2.1 Abstract

Insulators play a critical role in spatiotemporal gene regulation in animals. The evolutionarily conserved CCCTC-binding factor (CTCF) is required for insulator function in mammals, but not all of its binding sites act as insulators. Here, we explore the sequence requirements of CTCF-mediated transcriptional insulation using a sensitive insulator reporter in mouse embryonic stem cells (mESCs). We find that insulation potency depends on the number of CTCF binding sites in tandem. Furthermore, CTCF-mediated insulation is dependent on upstream flanking sequences at its binding sites. CTCF binding sites at topologically associating domain (TAD) boundaries are more likely to function as insulators than those outside TAD boundaries, independently of binding strength. We demonstrate that insulators form local chromatin domain boundaries and weaken enhancer-promoter contacts. Taken together, our results provide genetic, molecular, and structural evidence connecting chromatin topology to the action of insulators in the mammalian genome.

2.2 Introduction

The spatial and temporal patterns of gene expression are encoded in the genome in the form of *cis*-regulatory elements, which are categorized into promoters, enhancers, insulators, and other less-studied regulatory sequences, including repressive/silencing elements[1-3]. In animals, insulators play an essential role in cell-type-specific gene expression by protecting genes from improper regulatory signals from the neighboring chromatin environment[4]. Enhancer-blocking insulators act in a position-dependent manner in that they prevent enhancer-dependent gene activation only when placed in between the enhancer and target gene[5-7]. Insulators were initially identified in *Drosophila*, where the molecular machinery for insulation was first elucidated[4, 5, 8]. The first identified enhancer-blocking insulator in vertebrates is the 5'-HS4 element of the chicken β-globin locus[9]. Detailed analysis of this insulator led to the finding that the evolutionarily conserved zinc-finger family transcription factor CTCF, first identified as a DNA-binding protein at the chicken *c-Myc* gene promoter[10], was essential for its enhancer-blocking activity[11]. Mutations in the CTCF protein or its binding sites at insulators have since been implicated in a broad spectrum of human diseases[12-14]. In addition to its function at insulators, CTCF has also been demonstrated to play roles in transcriptional repression, gene activation, alternative splicing, and class switch recombination depending on the context of genomic locus[10, 15-19]. There are reports that CTCF binding at gene promoters could promote, instead of block, enhancer-promoter interactions[20, 21]. To date, exactly how and where CTCF mediates insulator function remains unclear.

CTCF has long been postulated to function as an organizer of three-dimensional chromosome architecture[1, 22, 23]. Genome-wide chromosome conformation capture analyses showed that the interphase chromosomes in mammalian cells are partitioned into megabase-sized TADs[24, 25], and CTCF binding sites were found at over 75% of TAD boundaries[24], suggesting a probable link between TAD boundaries and CTCF-mediated transcriptional insulation. Supporting this connection, disruption of TAD boundaries has been shown to permit ectopic enhancer-promoter contacts and aberrant gene expression, thereby leading to developmental abnormalities and cancer[16, 26]. Additionally, depletion of CTCF can lead to the weakening or disappearance of TADs[27-29]. CTCF drives TAD formation by working together with the cohesin complex to establish dynamic chromatin loops between distant CTCF binding sites, likely through a loop-extrusion process[29-39] or other mechanisms such as phase separation[40-45]. However, it is still debated whether TAD boundaries are sufficient to provide transcriptional insulation. Rapidly dissolving the global TAD structure by acute depletion of CTCF or cohesin subunits only altered transcription of a small number of genes in many different cellular contexts[27, 29, 33, 35, 37, 46]. Moreover, deletion of CTCF sites at the developmental locus *Sox9-Kcnj2* TAD boundary did not cause discernible phenotypes[47]. Furthermore, a majority of CTCF binding sites are not located at TAD boundaries, and whether these CTCF sites may function as insulators is unclear. These observations warrant an in-depth investigation of the role that CTCF and TADs play in transcriptional insulation.

To better understand where and how CTCF may mediate transcriptional insulation in the genome, we have developed an insulator reporter assay to evaluate the function of any DNA fragments in blocking enhancer-dependent transcriptional activation in mESCs. Using this system, we demonstrated that isolated single CTCF sites have weak or no insulator activity, regardless of its DNA binding strength. Instead, multiple copies of CTCF sites placed in tandem can provide a potent insulation effect. We also observed that CTCF binding sites at TAD boundaries could function as potent insulators, while the CTCF sites not located at TAD boundaries were incapable of insulating transcription. We attributed this difference in insulation activity to a sequence located 10-20 bp upstream of the CTCF core motifs, which promotes optimal insulation likely through contacts with CTCF's zinc fingers 9-11. We further discovered that insulators act by forming local TAD boundaries to reduce productive enhancer-promoter contacts, using both chromosome conformation capture assays and high-throughput multiplexed DNA fluorescence *in situ* hybridization (FISH) techniques. These results, taken together, shed light on how CTCF mediates transcriptional insulation in mammalian cells and establish a direct link between TAD boundaries and insulators.

2.3 Results

2.3.1 An insulator reporter assay in mouse embryonic stem cells

To quantitatively assay insulator activities in the context of native chromatin in cells, we engineered the *Sox2* gene locus in the F123 hybrid mESC line (*Mus musculus castaneus* × *S129/SvJae*)[48]. We and others previously showed that a super-enhancer (SE) located ~110 kb downstream of the *Sox2* gene was responsible for over 90% of its expression in the mESCs [49, 50]. We reasoned that the insulator activity of DNA elements could be measured by the reduction in *Sox2* gene expression when inserted between the *Sox2* gene and the downstream super-enhancer. To create the insulator reporter, we first tagged the two copies of the *Sox2* gene with *egfp* (CAST allele) and *mcherry* (129 allele) to quantify allelic Sox2 expression by live-cell fluorescence-activated cell sorting (FACS) (Figure 2.1a, Extended Data Figure 2.1a). Subsequently, we inserted a negative-selection fusion gene Tg(CAG-*HyTK*) flanked by a pair of heterotypic Flippase recognition sites (*Frt/F3*) between the *Sox2* gene and its downstream super-enhancer on the CAST allele (Figure 2.1a, Extended Data Figure 2.1b). As enhancer-blocking insulation is position-dependent, we created a control clone with the same replaceable cassette placed further downstream of the *Sox2* super-enhancer at equal distance on the CAST allele (Figure 2.1a, Extended Data Figure 2.1c). The Tg(CAG-*HyTK*) marker gene can be replaced by a donor sequence using the **r**ecombinase**-m**ediated **c**assette **e**xchange (RMCE) strategy (Figure 2.1b, Supplementary Figure 2.1a). By killing off unmodified mESCs with ganciclovir, we could achieve nearly 100% efficiency of marker-

free insertion (Supplementary Figure 2.1b).

As the insertion was specifically on the CAST allele, we used the 129 allele as the internal control to correct clone-to-clone variations in Sox2 expression (Figure 2.1b, Supplementary Figure 2.2a-b), which allowed quantitative comparisons of insulator activities of different CTCF binding sites (CBSs). We tested the insulation activity of a total of 11 different CBSs selected from several known TAD boundaries and chromatin loop anchors (Table 2.1). Each CBS insert was amplified from mouse or human genomic DNA by PCR and was 1-4 kb in length. Surprisingly, isolated single CBSs tested in both the forward and reverse orientations generally exhibited little or no insulator effect (Figure 2.1c). Only two of the probed CBSs in reverse orientation and four of the probed CBSs in forward orientation showed significant yet modest insulator effects (Figure 2.1c). The CBS of a canonical insulator, the HS5 sequence of the human beta-globin locus, reduced Sox2 expression by 11.0% ± 1.9% when inserted in forward orientation but had no effect in reverse orientation (Figure 2.1c, Supplementary Figure 2.2c-d). On average, individual isolated CBSs in forward and reverse orientations reduced Sox2 expression to 93.0%(±6.5%) and 97.0%(±6.0%) of parental cells with no insertion, respectively (Figure 2.1c).

2.3.2 Tandem CTCF sites enable strong transcriptional insulation

We hypothesized that multiple CBSs collectively may provide more robust insulation, since TAD boundaries are enriched for clustered CTCF binding sites[24, 51]. To

test this possibility, we constructed a series of insertion clones harboring multiple CBSs from the *Sox9-Kcnj2* TAD boundary (Extended Data Figure 2.2a). Two or more CBSs were PCR-amplified from mouse genomic DNA, ligated together and inserted in between the *Sox2* gene and super-enhancer on the CAST allele by RMCE as described above. We found that two CBSs, in forward tandem, reverse tandem, or divergent orientations, all had significantly stronger insulation effect than individual CBSs alone (Figure 2.2a). Notably, combining a weak CBS insulator with one that had a negligible insulator activity gave rise to stronger insulation than the summed effects of the two individual sites (Figure 2.2a), suggesting that CBSs could have synergistic insulation effects. Nevertheless, a weak CBS insulator did not enhance the insulator activity of a stronger CBS insulator if placed in convergent orientation (Figure 2.2a). Next, we measured the insulator activity of CBS clusters consisting of up to all four CBSs from the *Sox9-Kcnj2* TAD boundary. ChIP-seq analyses indicated that CTCF was recruited to the extra copy of the boundary sequence inserted in the *Sox2* domain (Extended Data Figure 2.2b). We found that the insulation effect became stronger as the number of CBSs increased, regardless of the orientation of CTCF motifs (Figure 2.2b, Table 2.2). Interestingly, the enhancement of insulation conferred by each additional CBS became smaller when the number of CBSs exceeds two (Extended Data Figure 2.2c). Consistent with the requirement for CTCF in transcriptional insulation, removal of the binding motifs of CTCF within the inserts completely abolished insulation effects of CBSs (Figure 2.2c). Furthermore, introducing CTCF sites downstream of the *Sox2* super-enhancer did not reduce but rather slightly increased Sox2 expression (Figure 2.2b), likely due to the insulation of interactions between the super-enhancer and further

downstream chromatin. Taken together, these results suggest that multiple CTCF binding sites arranged in tandem can function as a potent insulator due to synergistic or additive effects from individual sites.

Surprisingly, we observed that the insulator containing four CBSs was able to reduce Sox2 expression by 38.47 ± 3.16%, rather than completely blocking the *Sox2* super-enhancer activity (Figure 2.2b). The reduction of *Sox2* expression from the CAST allele was further confirmed by allele sensitive RNA-seq analysis (Extended Data Figure 2.2d-e). Interestingly, this insulator substantially increased cell-to-cell variations in Sox2 expression, evidenced by the accumulation of cells with extremely low Sox2-eGFP signals (Extended Data Figure 2.2f). Moreover, the sub-population of cells expressing ultra-low Sox2-eGFP could revert to the state of higher expression level after extended culturing, suggesting that the cell-to-cell variation of *Sox2* gene expression was a meta-stable state (Extended Data Figure 2.2g). Furthermore, CTCF insulation did not change the active chromatin state on either the *Sox2* promoter or its enhancer (Extended Data Figure 2.2h-i). Collectively, these results suggest that CBS-mediated insulation is permissive and highly dynamic.

2.3.3. CTCF-mediated insulation depends on sequence context

To better understand the sequence requirements for CTCF-mediated insulation, we synthesized insulators by concatenating multiple 139-bp genomic DNA sequences, each containing a 19-bp CTCF motif and two 60-bp flanking sequences. Each site was

selected from the aforementioned CBSs (Table 2.3-4). Consistent with the observations described above, the synthetic DNA sequences showed additive effects in transcriptional insulation (Extended Data Figure 2.3a). Additionally, ChIP-seq analyses confirmed the recruitment of CTCF and the cohesin complex to the synthetic insulators (Figure 2.3a). Interestingly, we observed that CBSs with longer flanking sequences (1 kb or longer) had stronger insulation effects than the shorter 139-bp CBSs, suggesting the existence of additional elements that could facilitate insulation (Extended Data Figure 2.3b).

Using the same approach, we also tested whether CBSs from outside of TAD boundaries could function as insulators. We selected multiple CBSs from non-TAD boundary regions in the genome, concatenated multiple 139-bp genomic sequences containing CTCF binding motifs together, and tested their insulation ability in our insulator reporter assay (Table 2.4). Surprisingly, although these non-TAD boundary CBSs displayed stronger CTCF binding than those from TAD boundaries at their original loci (Extended Data Figure 2.3c), the synthetic DNA sequences made up of six or fifteen tandemly arrayed 139-bp CBSs from non-boundary regions were unable to function as insulators, despite the presence of strong CTCF ChIP-seq signals at the insertion site (Fig 3b, Extended Data Figure 2.3d), indicating that CTCF binding alone is insufficient to bring transcriptional insulation.

To further dissect the sequence dependence of CTCF-mediated insulation, we exchanged the core motifs of 139-bp boundary CBSs with those of the synthetic CBSs

from non-boundary regions (Table 2.4). Combining boundary CBS core motifs with non-boundary adjacent sequences resulted in a much weaker insulation effect than with their original neighboring sequences of equal lengths (Figure 2.3c). In contrast, replacing adjacent sequences of non-boundary CBSs with those from boundary sites significantly strengthened their insulation effect (Figure 2.3c). However, when the adjacent sequences were scrambled or kept the same for boundary and non-boundary core motifs, their effects in insulating Sox2 expression were comparable (Figure 2.3c). Together, these results suggest that transcriptional insulation by CTCF is sequence-context-dependent, requiring DNA elements flanking the CTCF binding motif. It should be noted that ChIP-seq analysis showed that differential insulation activity of the synthetic insulators is not strictly correlated with CTCF occupancy (Extended Data Figure 2.4a-d).

To further delineate the key element in CTCF flanking sequences that promote transcriptional insulation, we tested the insulator activity of a series of synthetic CBSs with gradually decreasing flanking sequences from each side. Interestingly, strong insulation was retained at a synthetic insulator with just 20-bp flanking sequences on both sides of the core CTCF binding motifs, however, significantly reduced when the flanking sequences were shortened to 10 bp (Figure 2.3d), suggesting a critical role for the 10-20-bp flanking sequences of the core CTCF binding motif in insulation. We used the GLAM2 tool[52], a multiple sequence aligner that allows gaps and deletions among motifs, to identify a composite element in the six boundary CBSs (Extended Data Figure 2.4e). We found a central motif that matches the CTCF core motif and an upstream

motif at the same location as a previously reported element recognized by CTCF zinc fingers 9-11[53-55] (Figure 2.3e). To test whether CTCF zinc fingers 9-11 indeed contribute to transcriptional insulation, we deleted the DNA segment coding for zinc fingers 9-11 from both copies of the endogenous CTCF gene using CRISPR editing tools as previously described[37](Extended Data Figure 2.5a-c). Deletion of CTCF zinc fingers 9-11 significantly weakened insulation of the boundary CBSs but did not further reduce the insulation strength of the synthetic insulator with just 10-bp flanking sequences (Figure 2.3f, Extended Data Figure 2.5d-e). Together, these results suggest that flanking sequences of the boundary CBSs promote CTCF-mediated transcriptional insulation likely through contacts with CTCF zinc fingers 9-11. Further, ChIP-seq analysis showed that CTCF binding to CBSs with just ten-base-pair flanking sequences did not decrease significantly (Extended Data Figure 2.4a-b).

2.3.4. Insulators form TADs and weaken enhancer-promoter contacts

Previous studies suggest that the *Sox2* super-enhancer forms long-range chromatin contacts with the *Sox2* promoter[50, 56]. We hypothesized that insulators may change chromosome topology to limit enhancer-promoter communication. To test this hypothesis, we performed PLAC-seq[57] (also known as HiChIP[58]) experiments using mESC clones with various insulators inserted at the *Sox2* locus to detect promoter-centered chromatin contacts. Contact frequencies between the *Sox2* promoter and downstream super-enhancer were similar between the CAST and 129 alleles in mESCs with no insertion (Figure 2.4a). Inserting two CBSs from the *Sox9-Kcnj2* TAD boundary between the *Sox2* promoter and super-enhancer reduced the enhancer-promoter

contacts significantly (Fig.4a). Consistent with the observed dosage-dependent insulation effects, the *Sox2* enhancer-promoter contacts on the CAST allele were further reduced in cells with the insertion of four CBSs (Figure 2.4a). By contrast, placing two or four CBSs downstream of the *Sox2* super-enhancer did not reduce the *Sox2* enhancer-promoter contacts (Figure 2.4a). These results support the model that insulators act by reducing the enhancer-promoter contacts.

To further understand the effect of the insulators on local chromatin structure, we performed *in situ* Hi-C experiments[59] with mESC clones containing either two or four CBSs inserted between the *Sox2* gene and its super-enhancer on the CAST allele (Figure 2.4b-c). On the 129 allele, *Sox2* promoter and downstream super-enhancer were found to be in a single TAD (Figure 2.4b). By contrast, insertion of two CBSs between the *Sox2* gene and super-enhancer on the CAST allele created a new TAD boundary that separated the *Sox2* locus into two local chromatin domains (Figure 2.4b). Introducing four CBSs in the same location created an even stronger TAD boundary, and contacts across the new local domains were further reduced (Figure 2.4c). Additionally, we found that the inserted CBSs showed elevated levels of chromatin contacts with the CBSs located on *Sox2* promoter and super-enhancer, following the convergent rule[59] (Extended Data Figure 2.6a-d). Collectively, these results suggest that CTCF-dependent insulators create local TAD domains by forming chromatin loops between convergent CTCF binding sites.

2.3.5. Visualizing *Sox2* locus by multiplexed FISH for DNA and RNA

To directly visualize the impacts of insulators on chromatin architecture, we used the recently developed multiplexed DNA FISH imaging method to trace the chromatin conformation[60-62]. We traced the three-dimensional structure of the 210-kb genomic region (chr3: 34601078-34811078) containing the *Sox2* and super-enhancer loci across thousands of individual chromosomes at 5-kb intervals. We partitioned the 210-kb region into forty-two 5-kb segments and sequentially labeled and imaged each segment using 14 rounds of hybridization of readout probes with a three-color imaging scheme (Figure 2.5a, Extended Data Figure 2.7a-c, Supplementary Tables 1-2). The identity of the CAST allele was determined within each nucleus based on the presence of FISH signal corresponding to the 7.5-kb 4CBS insulator sequence inserted into the CAST allele that was absent in the 129 allele (Fig.5a, Extended Data Figure 2.7d).

We first carried out chromatin tracing experiments with the mESC clone containing an insertion of the 4CBS insulator between the *Sox2* gene and the downstream super-enhancer on the CAST allele. We obtained chromatin tracing data from 571 cells where both CAST and 129 alleles were robustly discerned (**Methods**). Consistent with results from Hi-C (Figure 2.4c), the median spatial distance matrix for the 129 allele showed a single TAD harboring both the *Sox2* and super-enhancer loci, whereas the spatial distance matrix for the CAST allele showed two TADs with a new boundary formed at the insertion site separating the *Sox2* and super-enhancer loci (Figure 2.5b-c; Extended Data Figure 2.8a-c). Accordingly, individual CAST chromosomes were more likely to form a boundary at the 4CBS insertion (Figure 2.5d-e). Moreover, the level of insulation between the two sub-regions to either side of the

inserted 4CBS, containing the *Sox2* promoter and the super-enhancer was statistically significantly enhanced on the CAST alleles (Figure 2.5f).

As controls, we also performed chromatin tracing experiments with one mESC line where all CTCF binding motifs of the insertion were removed, and another cell line where the insertion was at an equal distance further downstream of the *Sox2* super-enhancer. We obtained chromatin tracing data on both CAST and 129 alleles from 659 and 784 cells of the two cell lines, respectively. Based on FACS analyses, neither control insert reduced *Sox2* expression on the CAST allele (Extended Data Figure 2.8d). Consistently, no local chromatin domain boundary was visible between the *Sox2* and super-enhancer loci, and spatial insulation between the *Sox2* gene and the super-enhancer was indistinguishable between the CAST and 129 alleles (Extended Data Figure 2.8e-j). Interestingly, the distances between regions across the insulator were increased on the CAST allele compared to the 129 allele ,whereas mutant CBS inserted at the same location did not increase the distance between regions across the insertion (Extended Data Figure 2.9a-b). In contrast, the 4CBS insulator inserted downstream of the *Sox2* super-enhancer appeared to promote segregation of the *Sox2* domain from downstream chromatin, which may explain the slightly increased Sox2 expression in this clone (Extended Data Figure 2.9c).

Surprisingly, although the 4CBS insulator substantially reduced *Sox2* expression and the contact frequency between *Sox2* and its super-enhancer, the median spatial distance between *Sox2* super-enhancer and promoter only mildly increased on the

CAST alleles (282 nm) compared to the 129 alleles (264 nm) (Wilcoxon rank sum test, $P = 0.066$) (Figure 2.5g). We hypothesized that only on a small fraction of chromosomes the *Sox2* super-enhancer was in physical proximity with the *Sox2* promoter to engage in productive transcription, and insertion of an insulator on the CAST allele could reduce this fraction of engaged *Sox2* enhancer-promoter configuration selectively on the CAST allele. To test this hypothesis, we quantified the fraction of CAST alleles that showed a spatial distance between the *Sox2* promoter and the super-enhancer shorter than a particular threshold and compared it to that of the 129 alleles in the same cells. Indeed, in the mESCs where the 4CBS insulator was inserted between the *Sox2* gene and super-enhancer on the CAST allele, the ratio between the fraction of CAST alleles with spatially proximal enhancer-promoter pairs and the fraction of 129 alleles with spatially proximal enhancer-promoter pairs was much smaller than 1, at a spatial distance threshold of 150 nm, and the ratio increased gradually to 1 at a spatial distance threshold of ~300 nm (Figure 2.5h). By contrast, no reduction of this ratio was observed at a shorter spatial threshold in mESC clones where CTCF motifs were deleted from the insulator, or when the insulator sequence was inserted downstream of the *Sox2* super-enhancer(Figure 2.5h).

To further study how insulators affect enhancer-promoter spatial proximity and enhancer-dependent transcriptional activation at single-cell resolution, we simultaneously probed the chromatin structure with multiplexed DNA FISH and the transcripts at the *Sox2* locus with single-molecule RNA FISH[61, 63]. We first hybridized three sets of RNA-FISH probes targeting *Sox2*, *egfp*, and *mcherry* each with a unique

readout sequence to distinguish the transcripts made from the two *Sox2* chromosome copies in each cell (Supplementary Table 2.3). We then performed multiplexed DNA FISH with the same cells to trace the local chromatin configuration. The *Sox2* chromatin loci that spatially overlapped with nascent *Sox2* transcripts were designated as transcriptionally bursting loci, and the remaining *Sox2* loci without a coincident transcript were regarded to be in resting state (Extended Data Figure 2.10a). Consistent with the RNA-seq analysis described above (Extended Data Figure 2.2d-e), the frequency of detecting the nascent *Sox2* transcripts on the CAST allele was substantially lower than that of the 129 allele in the 4CBS clone (Extended Data Figure 2.10b). By contrast, the frequency of detecting nascent *Sox2* transcripts on the CAST allele was slightly higher than the 129 allele in the control cells in which the CBS insulator was inserted downstream of the *Sox2* enhancer (Extended Data Figure 2.10b). Consistent with previous studies[61, 64], we found that nascent *Sox2* transcripts were detected across a wide range of spatial distances between the *Sox2* enhancer and promoter, although the median enhancer-promoter distances at the *Sox2* gene with coincident nascent transcripts were slightly but significantly shorter than those on the resting loci (Extended Data Figure 2.10c-d). However, the fraction of the *Sox2* gene with coincident nascent transcripts on the CAST allele in the 4CBS clone was consistently lower than that on 129 allele even though the spatial distances between the enhancer and promoter are similar (Figure 2.5i). By contrast, the fraction of the *Sox2* genes with coincident nascent transcripts was comparable between the two alleles when the 4CBS was inserted downstream of the *Sox2* enhancer (Figure 2.5j). These results, taken together, suggest that enhancer proximity is positively correlated to transcriptional activity at target gene in

general; however, itself alone is not sensitive enough to differentiate transcriptional states. In summary, our results suggest that CTCF-insulators decrease the frequency of transcription bursting at *Sox2* when inserted between the enhancer and promoter, likely by establishing local chromatin domain boundaries that weaken productive communications between spatially close enhancer and promoter (Figure 2.5k).

2.4 Discussion

The sequence-specific DNA binding protein CTCF plays a role in both chromatin organization and transcriptional insulation, but exactly how chromatin topology is related to transcriptional insulation remains to be understood. In this study, we developed an experimental system using mESCs to quantify the enhancer-blocking activity of insulators in the native chromatin context at the *Sox2* locus. The well-defined distal enhancer of *Sox2* gene activation afforded an excellent opportunity to quantify the effects of insulator insertions on local chromatin structure and transcription in *cis*. We determined the insulator activity of a number of CTCF binding sites either alone or in various combinations, and demonstrated that potent insulation was rendered by two or more clustered CTCF binding sites. Importantly, we found that CTCF binding alone was insufficient to confer insulation activity; rather, sequences immediately adjacent to CTCF binding motifs were required for potent insulator function. Consistent with this observation, CTCF binding sites within TAD boundaries are more likely to function as insulators than those not located at TAD boundaries, regardless of the strength of their binding by CTCF. Finally, using two complementary approaches to profile chromatin architecture, we showed that CTCF likely mediates transcriptional insulation by creating local chromatin domain boundaries and reducing the frequency of productive enhancer-promoter contacts. Our results, therefore, provide mechanistic insights into the link between TAD boundaries that are enriched for CTCF binding sites and CTCF-mediated transcriptional insulation.

We demonstrated that several factors may be involved in CTCF-mediated

transcriptional insulation in mammalian cells. First, a single CBS has weak insulation effects, varies depending on the orientation of the CTCF motif. The orientation bias is likely due to a pair of convergent CBSs located on the *Sox2* promoter and enhancer. The CBS insertion is predicted to loop with the enhancer CBS in a forward orientation[59]. A loop formed with the enhancer may block enhancer activity more efficiently than one formed with the promoter. Given that the inserted insulator is closer to the *Sox2* enhancer, where CTCF binding is stronger, it is also possible that looping with CBS on the *Sox2* super-enhancer is more efficient, thereby, favoring insulation by forward-orientated CBSs.

Second, we found that multiple CBSs taken from TAD boundaries exert potent transcriptional insulation activities. Our finding is consistent with a recent study of the mouse *Pcdh* clusters reporting that insertion of tandem CTCF sites could block enhancers from activating proximal genes[65]. These observations with CTCF insulators are different from the *Drosophila* gypsy insulator, which was ineffective in blocking enhancer activity when two tandem copies were tested[66, 67].

Third and more importantly, through sequence swapping experiments, we showed that sequences immediately adjacent to CTCF binding motifs were necessary for enhancer-blocking function. We further found an upstream element in the flanking sequences of CTCF binding motifs to be crucial for transcriptional insulation. Previous studies reported an upstream motif that stabilizes CTCF binding via interactions with the 9-11 zinc fingers of CTCF[53-55, 68]. We speculate that CTCF zinc fingers 9-11 may

promote transcriptional insulation by inducing a tertiary structure on insulators that stabilizes CTCF-cohesin interactions, thereby blocking the loop extrusion process that facilitates long-range enhancer-promoter contacts. It is noteworthy that deleting zinc fingers 9-11 did not fully abolish insulation of the boundary CBSs, suggesting the involvement of additional factors in transcriptional insulation.

Our study also relates the chromatin structure involving enhancer-promoter contacts, as revealed by various 3C-based and microscopy-based experiments, to enhancer-dependent transcription. From both the 3C and imaging experiments, we found that the insertion of multiple CBS sites in tandem, with the appropriate flanking sequences, induced the formation of a TAD boundary at the insertion site and reduced interactions between the enhancer and the promoter. Spatial proximity between an enhancer and a promoter has been thought to be positively correlated with enhancer-dependent activation in general. However, recent studies have also shown that spatial proximity is not strictly correlated with transcriptional activation[69], and is a poor predictor of transcriptional activity in live cells[64]. We showed that transcriptional activities of *Sox2* promoter, measured by the frequency of nascent transcripts detected at the gene locus in a population, could be reduced by insulators with only modest changes to the enhancer-promoter proximity. These studies, together, highlight that enhancer-promoter proximity is just one of the many elements regulating transcriptional activity in mammalian cells. The point-to-point spatial distances between the enhancer and promoter does not fully reflect the chromatin structure of the entire locus. Simultaneous imaging of chromatin and transcripts indicated that *Sox2* transcription could take place

in chromosomes showing a broad range of spatial distances between the enhancer and *Sox2* promoter[61]. One possibility is that the *Sox2* super-enhancer forms a phase-separated environment[70], where the *Sox2* gene needs not be very close to its enhancer to be activated. Another possibility is that the temporal duration of Sox2 enhancer-promoter interaction is relatively short compared to a transcriptional bursting cycle, which would make it difficult to capture the two events simultaneously in fixed cells using FISH. Finally, transcription is not likely to happen immediately after enhancer-promoter contacts[71]. The lagging between these two events could also explain the lack of strict correlation between enhancer-promoter proximity and transcriptional bursting in live cells[64].

In summary, our results suggest that CTCF sites in the genome are not all equivalent to each other, and CTCF-mediated insulation depends on both dosage and upstream flanking sequences. Our findings explain why CBSs at TAD boundaries are more likely to act as transcriptional insulators than those outside TAD boundaries. One potential limitation of the current study is that the insulation effects of CBSs were tested only in the *Sox2* locus. Future experiments will be needed to demonstrate whether observations made from the *Sox2* locus can be generalized to other gene loci in the mammalian genome.

2.5 Figures

Figure 2.1: A sensitive insulator reporter assay measures the insulation activity of different CTCF binding sites at the *Sox2* locus in mouse ES cells. a, Left, the regulatory landscape of the *Sox2* locus in mESCs. Orientations of CTCF sites are indicated on the top of the signal tracks; Right, genetic constructs of mESC lines. Boxed *Sox2* in blue represents *Sox2-p2a-egfp in situ* fusion gene, boxed *Sox2* in orange represents *Sox2-p2a-mcherry in situ* fusion gene. The hygromycin phosphotransferase-thymidine kinase fusion gene *HyTK* is flanked by Flippase recognition sites *FRT* and *F3*. **b,** Experimental scheme to insert a test sequence into the *Sox2* locus by recombinase-mediated cassette exchange (RMCE). The Flippase expression plasmid and donor plasmid containing the test sequence were co-electroporated into cells. The orientation of the insert was controlled by the positions of the Not1 and Sbf1 restriction enzyme sites. Mouse ESC clones containing the insert were picked, genotyped, and allelic Sox2 expression was measured by FACS. **c,** A bar graph shows the normalized Sox2-eGFP expression of the no insertion clone (n = 8), different CBS insertion clones (n = 3; For Sox9_CBS1 in the forward orientation, n = 2.) and downstream insertion controls (n = 27). Each dot represents an independently picked colony. One-way analysis of variance with Bonferroni's multiple comparisons test. Data are mean ± sd. The exact *P* values for each comparison are listed in Table 2.6. ns $P > 0.05$, *$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$, ****$P \leq 0.0001$.

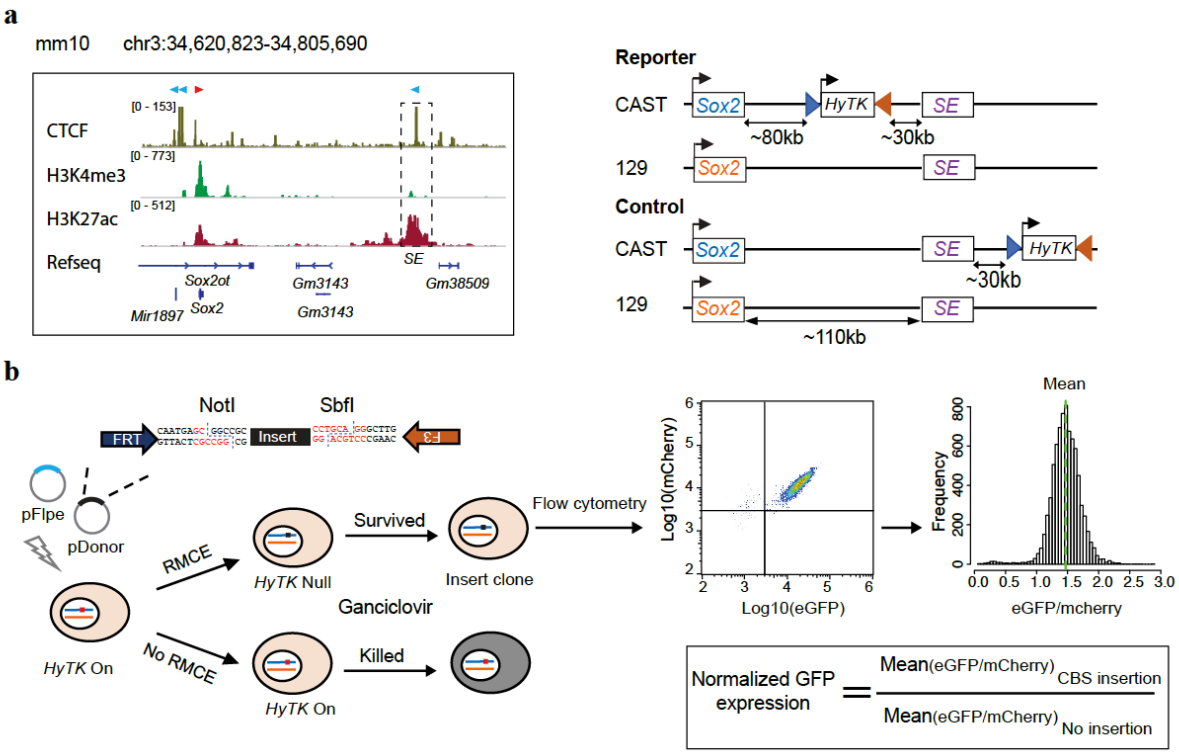

a
mm10
chr3:34,620,823-34,805,690
CTCF
H3K4me3
H3K27ac
Refseq
[0 - 153]
[0 - 773]
[0 - 512]
SE
Sox2ot
Gm3143
Gm38509
Mir1897
Sox2
Gm3143
Reporter
CAST
Sox2
HyTK
SE
~80kb
~30kb
129
Sox2
SE
Control
CAST
Sox2
SE
HyTK
~30kb
129
Sox2
SE
~110kb
b
NotI
SbfI
FRT
Insert
F3
pFlpe
pDonor
RMCE
HyTK Null
Survived
Insert clone
Flow cytometry
Ganciclovir
HyTK On
No RMCE
Killed
HyTK On
Log10(mCherry)
Log10(eGFP)
Mean
Frequency
eGFP/mcherry
Normalized GFP expression
Mean(eGFP/mCherry) CBS insertion
Mean(eGFP/mCherry) No insertion
c
Sox2
CBS
SE
~30kb
Normalized Sox2-eGFP expression
No insertion
HS5_CBS
Tal1_CBS
Lmo2_CBS
Pax3_CBS1
Pax3_CBS2
Pax3_CBS3
Pax3_CBS4
Sox9_CBS3
Sox9_CBS4
Sox9_CBS2
Sox9_CBS1
Downstream controls
Reverse orientation
Forward orientation

Figure 2.2: Multiple CTCF sites in tandem enable strong transcriptional insulation. **a,** Bar graphs showing insulation effects of two combined CBSs from the *Sox9-Kcnj2* TAD boundary (no insertion n = 8, for convergent group, n = 7; insertion clones n = 3, for Sox9 CBS1+2 in convergent group n = 4). Individual CBS sequences were combined by PCR to create two-CBS insertions. Arrows indicate the motif orientation of each individual CBS. Every insertion construct was created by an independent RMCE experiment. **b**, A bar graph shows insulation effects of multiple CBS from the *Sox9-Kcnj2* TAD boundary. Individual or combined CBS sequences were PCR cloned from mouse genomic DNA. Every insertion construct was created by an independent RMCE experiment (0 CBS, n = 8; 1 CBS inside, n = 23; 2 CBS inside, n = 18; 3 CBS inside, n = 13; 4 CBS inside, n = 5; 1 CBS downstream, n = 15; 2 CBS downstream, n = 8; 3 CBS downstream, n = 4; 4 CBS downstream, n = 6.). **c,** A bar graph shows insulation effects of λ DNA (n = 3), a combined two-CBS sequence, *Sox9* CBS1+2 (n = 3), and *Sox9* CBS1+2 Δcore motifs, which is the same two-CBS sequence but with the two19-bp CTCF core motifs deleted (n = 3). Inserts were comparable in length (~4 kb). Data are mean ± sd. *P* values were determined by one-way analysis of variance with Bonferroni's multiple comparisons test.

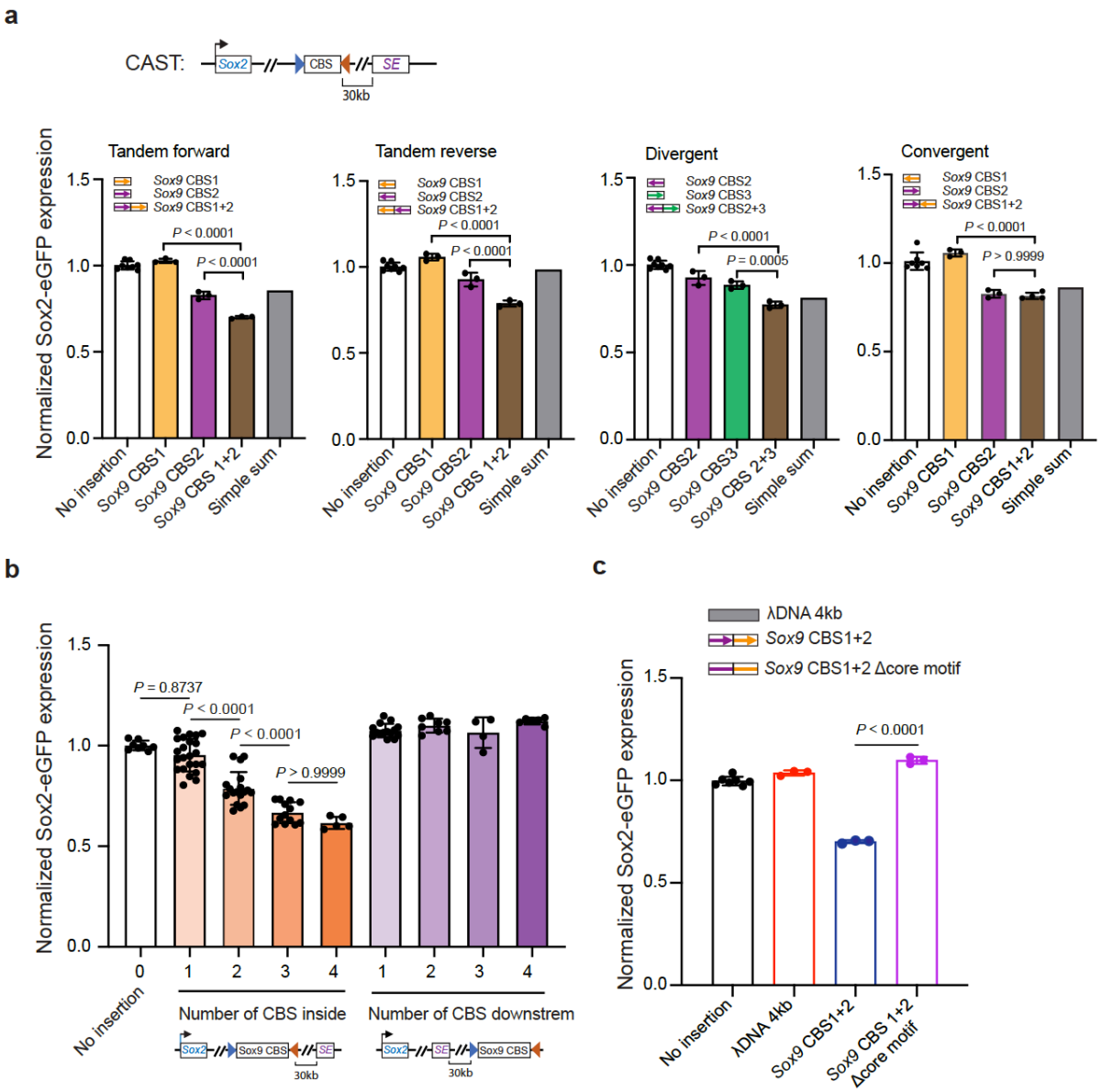

a
CAST:
Sox2
CBS
SE
30kb
Tandem forward
Tandem reverse
Divergent
Convergent
Normalized Sox2-eGFP expression
Sox9 CBS1
Sox9 CBS2
Sox9 CBS1+2
Sox9 CBS3
Sox9 CBS2+3
P < 0.0001
P = 0.0005
P > 0.9999
No insertion
Sox9 CBS1
Sox9 CBS2
Sox9 CBS 1+2
Sox9 CBS 2+3
Simple sum
b
P = 0.8737
P > 0.9999
Number of CBS inside
Number of CBS downstrem
Sox9 CBS
c
λDNA 4kb
Sox9 CBS1+2 Δcore motif
Sox9 CBS 1+2 Δcore motif

Figure 2.3: Synthetic insulators reveal sequence requirements for CTCF-mediated enhancer-blocking. a, ChIP-seq of CTCF and Rad21. The "Bd syn-6" mES clone contains the insertion of six 139-bp boundary CBSs (*Sox9*_CBS1-4, *Pax3*_CBS3 and *HS5*_CBS, Tables 3-4.) between *Sox2* and its super-enhancer. Sequencing reads from the insertion clone were aligned to a customized mm10 genome that included the inserted sequence at the target location. Motif orientations of nearby CBS and inserted CBS were indicated on the top of signal tracks. The *Sox2* super-enhancer is highlighted in the orange box. **b,** A bar plot shows insulation effects of synthetic sequences containing tandemly arrayed 139-bp CBSs from boundary and non-boundary regions. For each synthetic sequence, six insertion clones were picked with three of them in forward orientation and the other three in reverse orientation. **c,** A bar plot shows insulation effects of recombined tandemly arrayed 139-bp CBSs. CBS core motifs of boundary and non-boundary sites were combined with either their native adjacent sequences, scrambled adjacent sequences, or exchanged adjacent sequences with each other (n = 3). Each test sequence contains six tandemly arrayed 139-bp CBSs. The order of the six CBS core motifs was kept the same. **d,** A bar plot shows the insulation effect of the "Bd syn-6" sequence with truncated adjacent sequences. All insertions were between the *Sox2* promoter and super-enhancer (n = 3; for 50 bp adj, n = 2.). **e,** A composite motif discovered in the six boundary CBSs tested. Each CBS consists of a 19-bp core motif and 20-bp adjacent sequences on both sides. The motif was searched by the GLAM2 program of the MEME suite. **f,** A bar plot shows the impact of CTCF zinc fingers 9-11 deletion on insulation effects of boundary CBSs containing sixty-base-pair adjacent sequences and ten-base-pair adjacent sequences (No insertion, n = 4; for WT with 60 bp adj, n = 4; for ΔZF9-11 with 60 bp adj, n = 9; for WT with 10 bp adj, n = 5; for ΔZF9-11 with 10 bp adj, n = 10.). *P* values were determined by one-way analysis of variance with Bonferroni's multiple comparisons test. Data are mean ± sd.

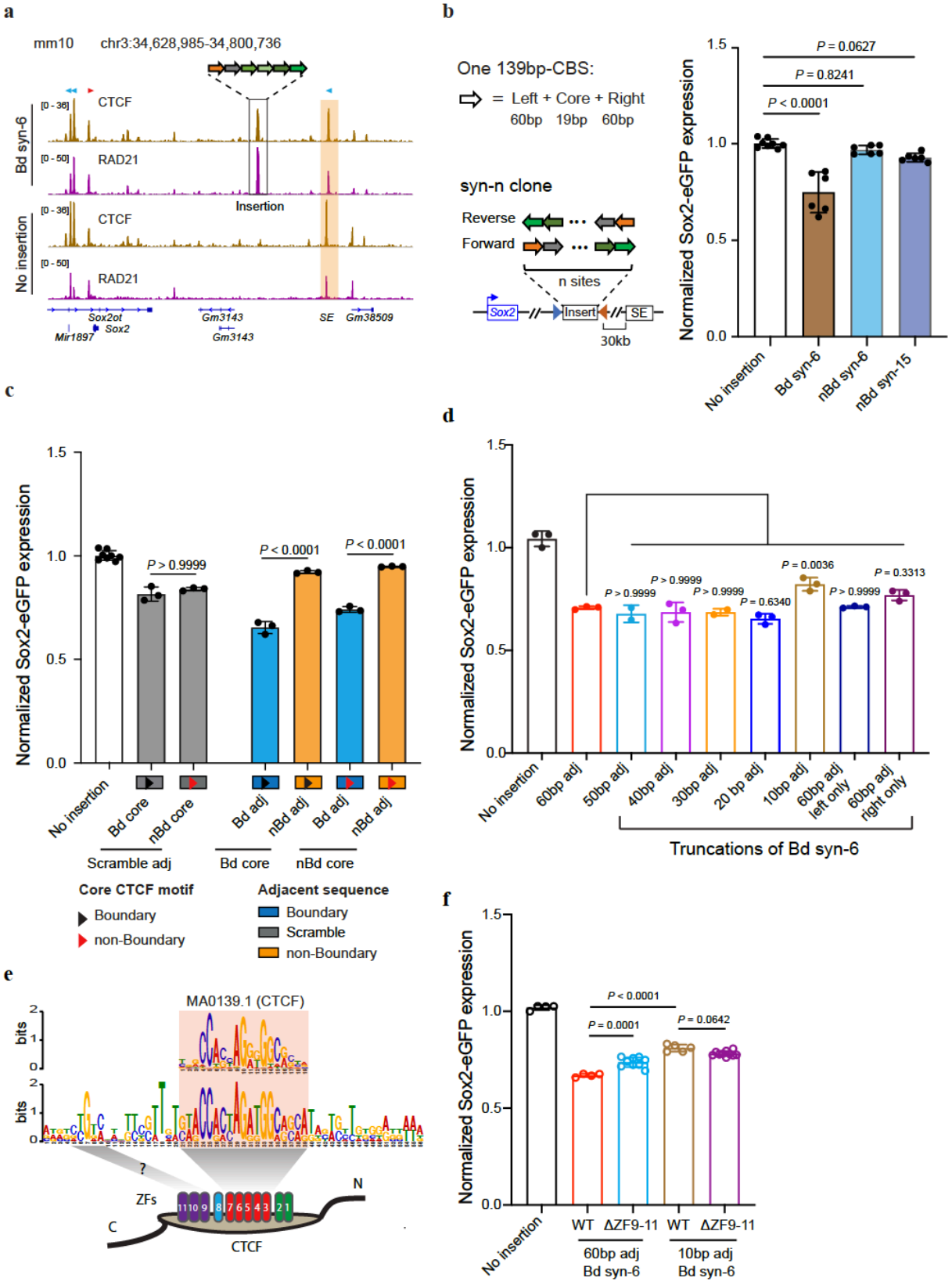

a
mm10
chr3:34,628,985-34,800,736
Bd syn-6
No insertion
CTCF
RAD21
Insertion
Sox2ot
Sox2
Mir1897
Gm3143
SE
Gm38509
b
One 139bp-CBS:
= Left + Core + Right
60bp 19bp 60bp
syn-n clone
Reverse
Forward
n sites
Sox2
Insert
SE
30kb
Normalized Sox2-eGFP expression
P = 0.0627
P = 0.8241
P < 0.0001
No insertion
Bd syn-6
nBd syn-6
nBd syn-15
c
Normalized Sox2-eGFP expression
P > 0.9999
P < 0.0001
P < 0.0001
No insertion
Bd core
nBd core
Bd adj
nBd adj
Bd adj
nBd adj
Scramble adj
Bd core
nBd core
Core CTCF motif
Boundary
non-Boundary
Adjacent sequence
Boundary
Scramble
non-Boundary
d
Normalized Sox2-eGFP expression
No insertion
60bp adj
50bp adj
40bp adj
30bp adj
20 bp adj
10bp adj
60bp adj left only
60bp adj right only
P > 0.9999
P > 0.9999
P > 0.9999
P = 0.6340
P = 0.0036
P > 0.9999
P = 0.3313
Truncations of Bd syn-6
e
MA0139.1 (CTCF)
bits
ZFs
C
N
CTCF
f
Normalized Sox2-eGFP expression
P < 0.0001
P = 0.0001
P = 0.0642
No insertion
WT
ΔZF9-11
WT
ΔZF9-11
60bp adj Bd syn-6
10bp adj Bd syn-6

Figure 2.4: Enhancer-blocking insulator forms local chromatin domains and reduces *Sox2* enhancer-promoter chromatin contacts. a, Allelic chromatin contacts from PLAC-seq data are shown at the viewpoint of the *Sox2* promoter (n = 2, replicates were merged). PLAC-seq experiments were carried out using a monoclonal antibody (Millipore, 04-745) against H3K4me3. Sequencing reads were mapped to the mm10 reference genome and split to CAST and 129 allele based on the haplotypes of parental strains. DNA fragments connecting the promoter and each of the surrounding 10-kb bins were counted. Contact frequency was normalized by the total *cis* contacts of the *Sox2* promoter for each allele, interactions within the 10-kb *Sox2* promoter bin were not shown. Arrows indicate the insertion location of CBSs. Fisher exact tests of *Sox2* enhancer-promoter contacts of the two alleles were performed (Two sided tests, ns $P > 0.05$, ***$P = 4.91 \times 10^{-4}$, ****$P = 5.34 \times 10^{-5}$). Right, insertion construct matching each clone on the left. The CBS clusters were obtained from the *Sox9-Kcnj2* TAD boundary by PCR. **b-c,** Allelic Hi-C contact map at *Sox2* locus. Mouse ESCs with the insertion of two CBSs or four CBSs from the *Sox9-Kcnj2* TAD boundary in the CAST allele were used for the experiments. Hi-C reads were mapped to the mm10 reference genome and split to CAST and 129 allele based on the haplotypes of parental strains. Allele-specific contact matrix was normalized by K-R matrix balancing. Top right, no insertion allele (129); Bottom left, insertion allele from the same cells (CAST). Bottom**,** allelic directionality index (DI) score of Hi-C interaction frequency (n = 2, replicates were merged).

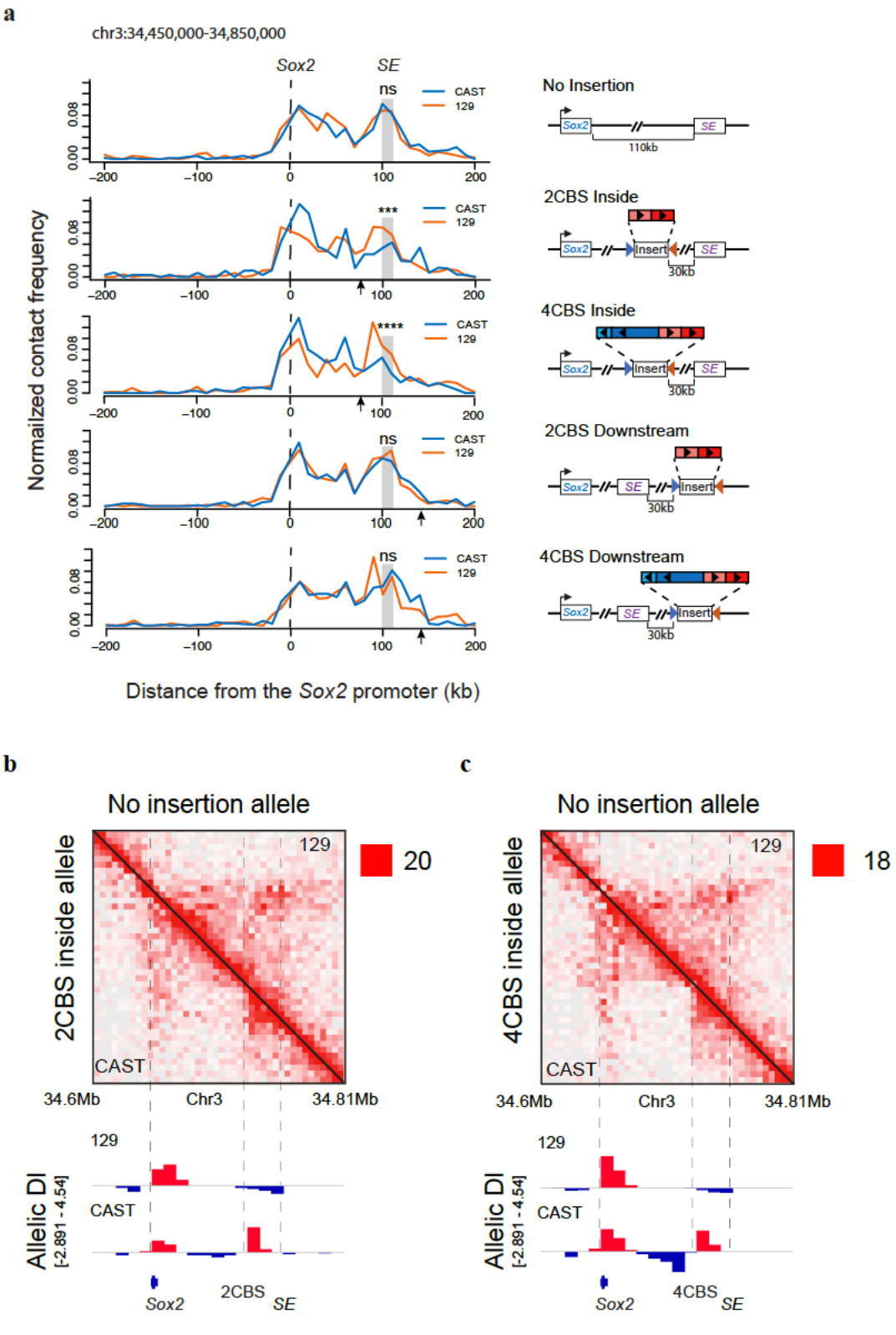
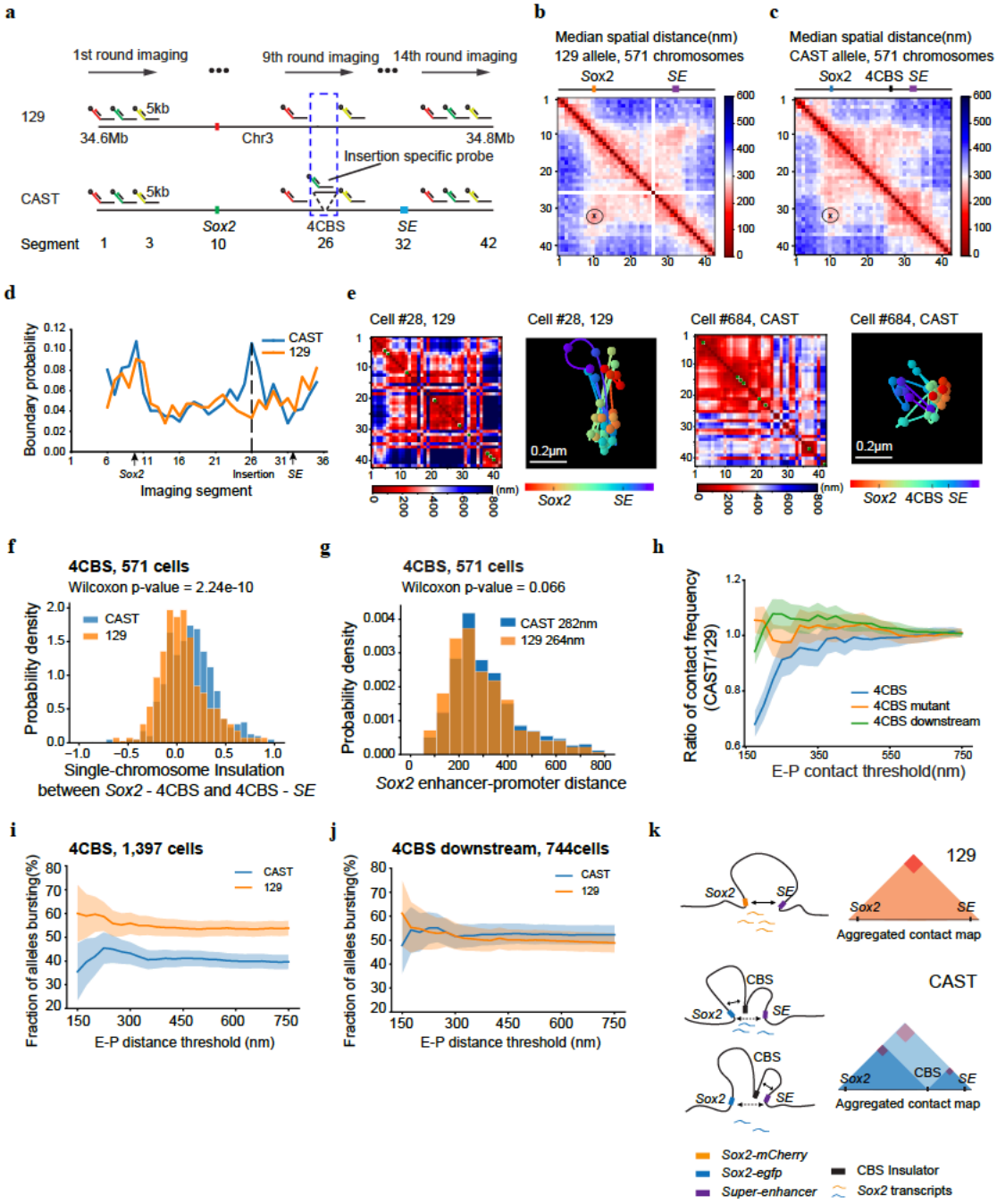
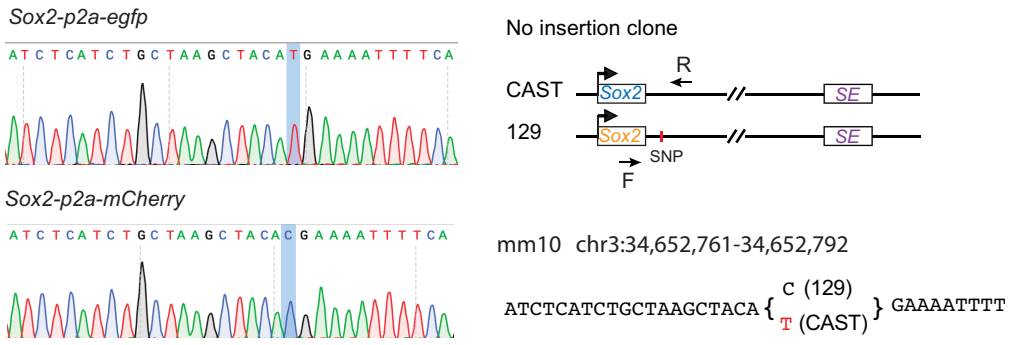

a
chr3:34,450,000-34,850,000
Sox2
SE
ns
***
****
CAST
129
Normalized contact frequency
Distance from the Sox2 promoter (kb)
No Insertion
110kb
2CBS Inside
4CBS Inside
2CBS Downstream
4CBS Downstream
Insert
30kb
b
No insertion allele
2CBS inside allele
129
CAST
20
34.6Mb
Chr3
34.81Mb
Allelic DI
[-2.891 - 4.54]
2CBS
c
No insertion allele
4CBS inside allele
18
4CBS

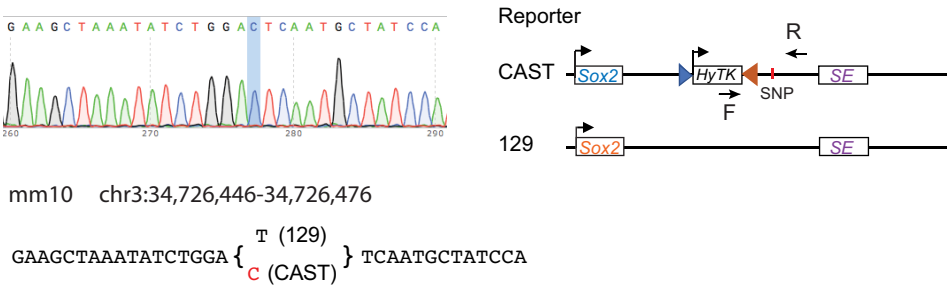
Figure 2.5: Effects of an enhancer-blocking insulator on chromatin topology and transcription revealed by multiplexed FISH. a, Scheme of the chromatin tracing experiments targeting the 210-kb *Sox2* region (chr3: 34601078-34811078). **b-c,** Median spatial-distance matrix for 129 (**b**) and CAST (**c**) chromosomes. **d,** The probability of each segment to be a single-chromosome domain boundary for the two alleles in **b-c**. The 26^{th} segment on the CAST allele is the 4CBS insertion. **e,** Exemplary single-chromosome structures of the imaged *Sox2* locus of CAST and 129 alleles. Green pixels on the interpolated matrices indicate missing values in the displayed examples of chromatin traces. **f,** The distribution of single-chromosome insulation scores for each of the alleles between *Sox2* promoter – 4CBS insertion (segments 10-25) and 4CBS insertion – *Sox2* enhancer (segments 26-33). Two-sided Wilcoxon rank-sum test was performed. **g,** The distribution of *Sox2* enhancer-promoter distance for the CAST and 129 chromosomes in **b-c.** Two-sided Wilcoxon rank-sum test was performed. **h**, The ratio of *Sox2* enhancer-promoter contact frequency of CAST chromosomes to that of 129 chromosomes. The distribution of contact frequency ratio (CAST/129) of the "4CBS" (n = 571 cells) clone is significantly different from that of the "4CBS mutant" (n = 659 cells) and "4CBS downstream" (n = 784 cells) clone, with *P* values of two-sided Kolmogorov–Smirnov tests equal to 6.38×10^{-5} and 1.09×10^{-9}, respectively. Shadow indicates the 95% confidence interval based upon binomial distribution. **i-j,** The bursting frequency of the *Sox2* gene on CAST and 129 chromosomes. (i) the 4CBS clone (n = 1,397 cells), (j) the control clone with 4CBS inserted downstream of the *Sox2* super-enhancer (n = 744 cells). Shadow indicates the 95% confidence interval based upon binomial distribution. **k,** A model of the *Sox2* locus on the two alleles. On the 129 allele, the super-enhancer interacts with the *Sox2* promoter and activates transcription of the *Sox2* gene. On the CAST allele, the CBS insulators can interact with both the *Sox2* promoter and the super-enhancer, resulting in fewer productive enhancer-promoter contacts.


a
1st round imaging
9th round imaging
14th round imaging
129
5kb
34.6Mb
Chr3
34.8Mb
Insertion specific probe
CAST
Sox2
4CBS
SE
Segment
b
Median spatial distance(nm)
129 allele, 571 chromosomes
c
Median spatial distance(nm)
CAST allele, 571 chromosomes
d
Boundary probability
CAST
129
Imaging segment
e
Cell #28, 129
Cell #684, CAST
0.2μm
f
4CBS, 571 cells
Wilcoxon p-value = 2.24e-10
Probability density
Single-chromosome Insulation
between Sox2 - 4CBS and 4CBS - SE
g
4CBS, 571 cells
Wilcoxon p-value = 0.066
CAST 282nm
129 264nm
Sox2 enhancer-promoter distance
h
Ratio of contact frequency
(CAST/129)
4CBS
4CBS mutant
4CBS downstream
E-P contact threshold(nm)
i
4CBS, 1,397 cells
Fraction of alleles bursting(%)
E-P distance threshold (nm)
j
4CBS downstream, 744cells
k
129
Aggregated contact map
CAST
CBS
Sox2-mCherry
Sox2-egfp
Super-enhancer
CBS Insulator
Sox2 transcripts

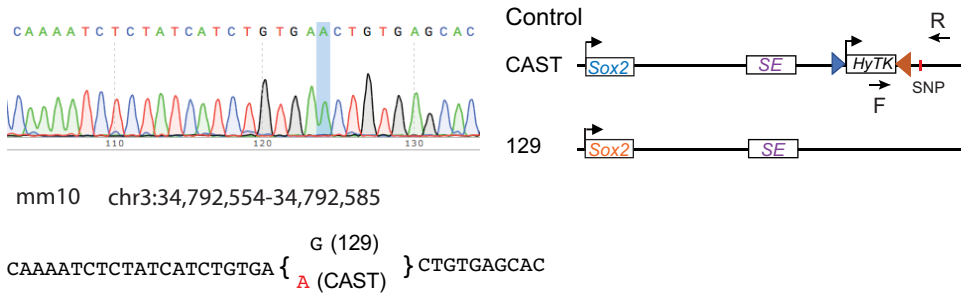
a



b

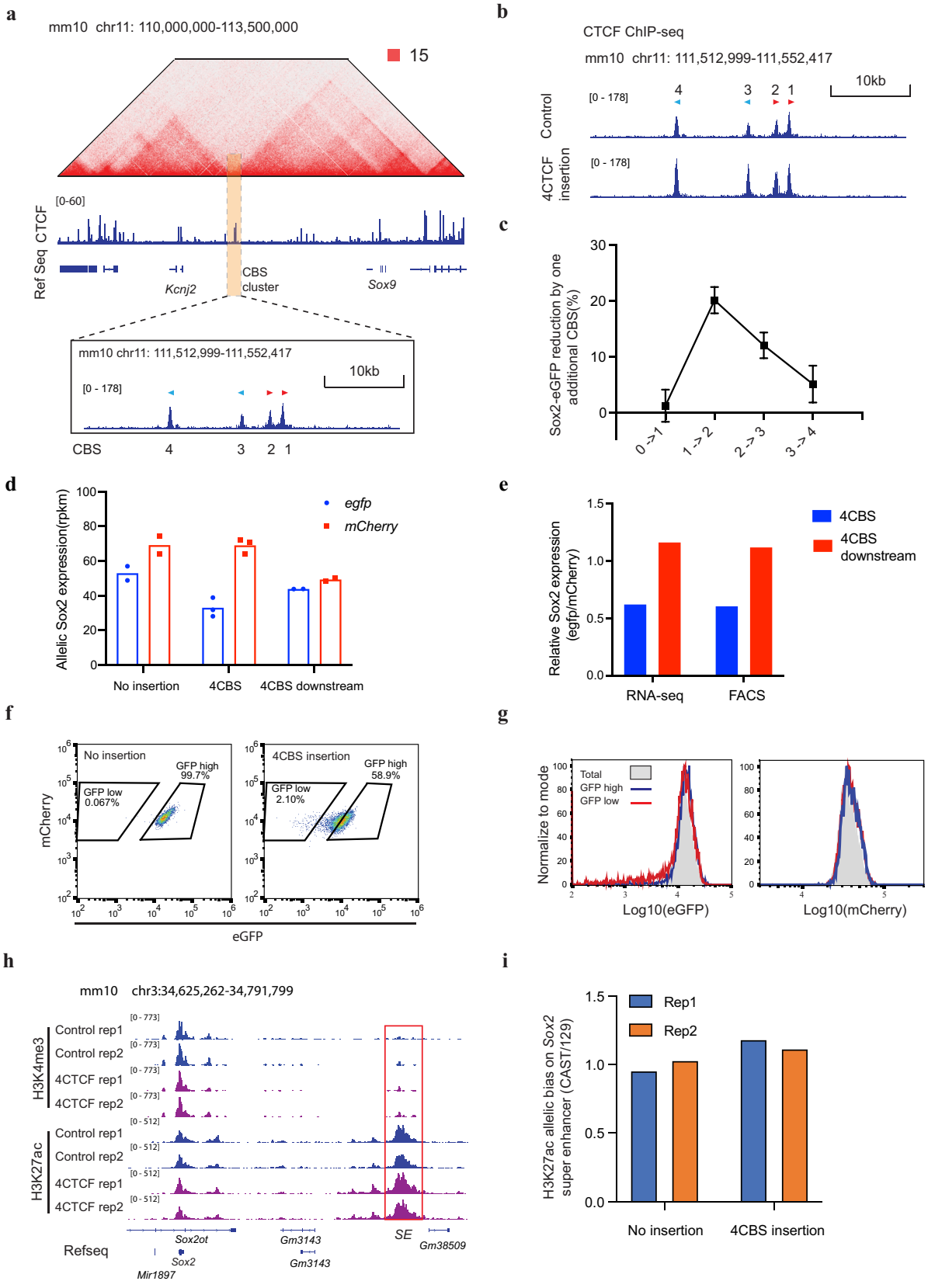


c

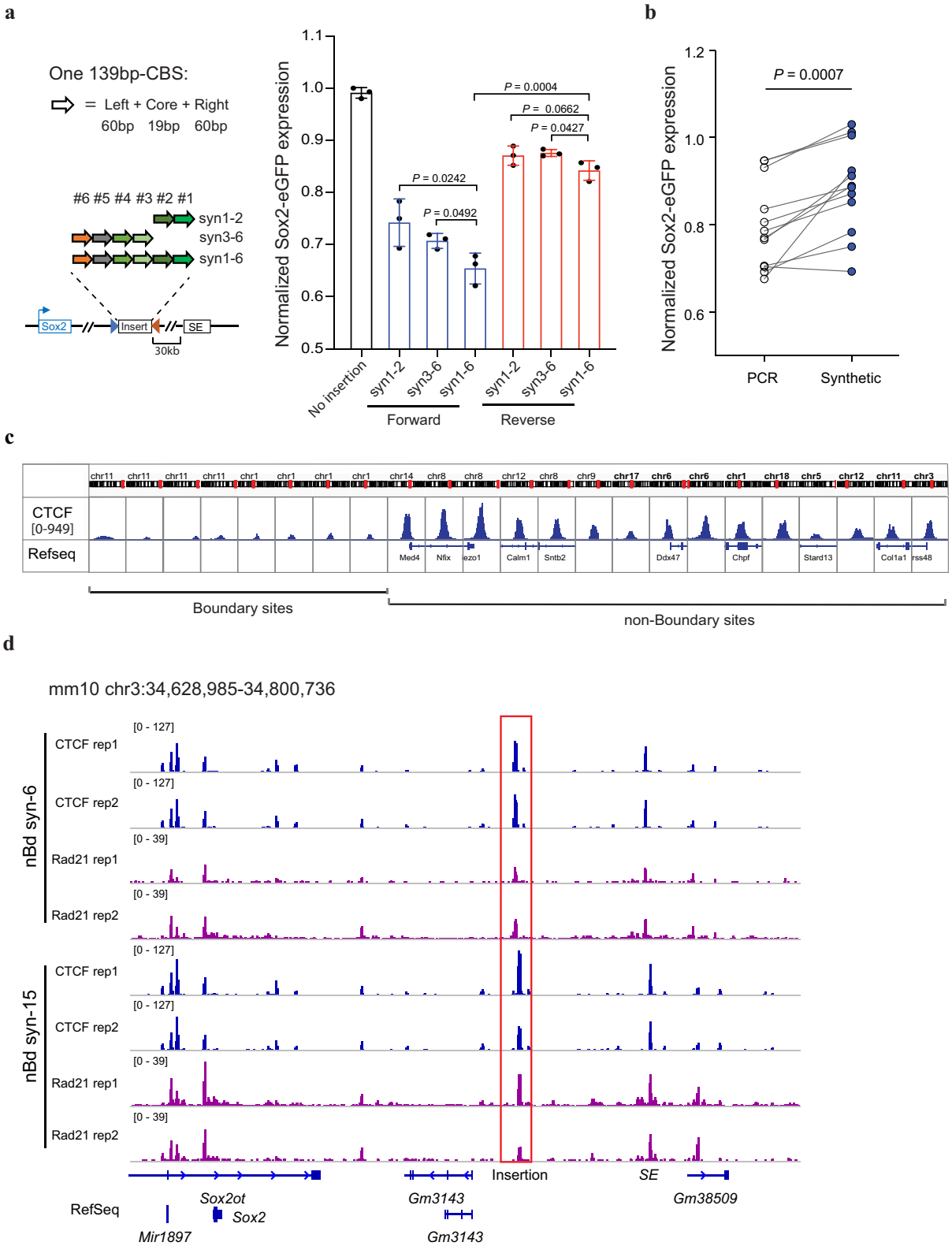


Extended Data Figure 2.1: Genotyping mESC reporter cell lines. a, Genotyping *egfp* and *mcherry* labeled *Sox2* gene. Left, Sanger sequencing results for allele-specific PCR products. Allele-specific SNP is highlighted. Right, the construct of the clone and the SNP information used to distinguish the two alleles. The reverse primer was common, while the forward primer was allele-specific, matching with *egfp* and *mcherry* sequence, respectively. **b-c,** Genotyping the Insulator reporter and control cell lines. Left, Sanger sequencing and SNP information. Right, Construct of the clone and positions of PCR primers. The forward primer is specific to the inserted *HyTK* gene. **b,** insulator reporter cell line. **c,** Insulator control cell line.

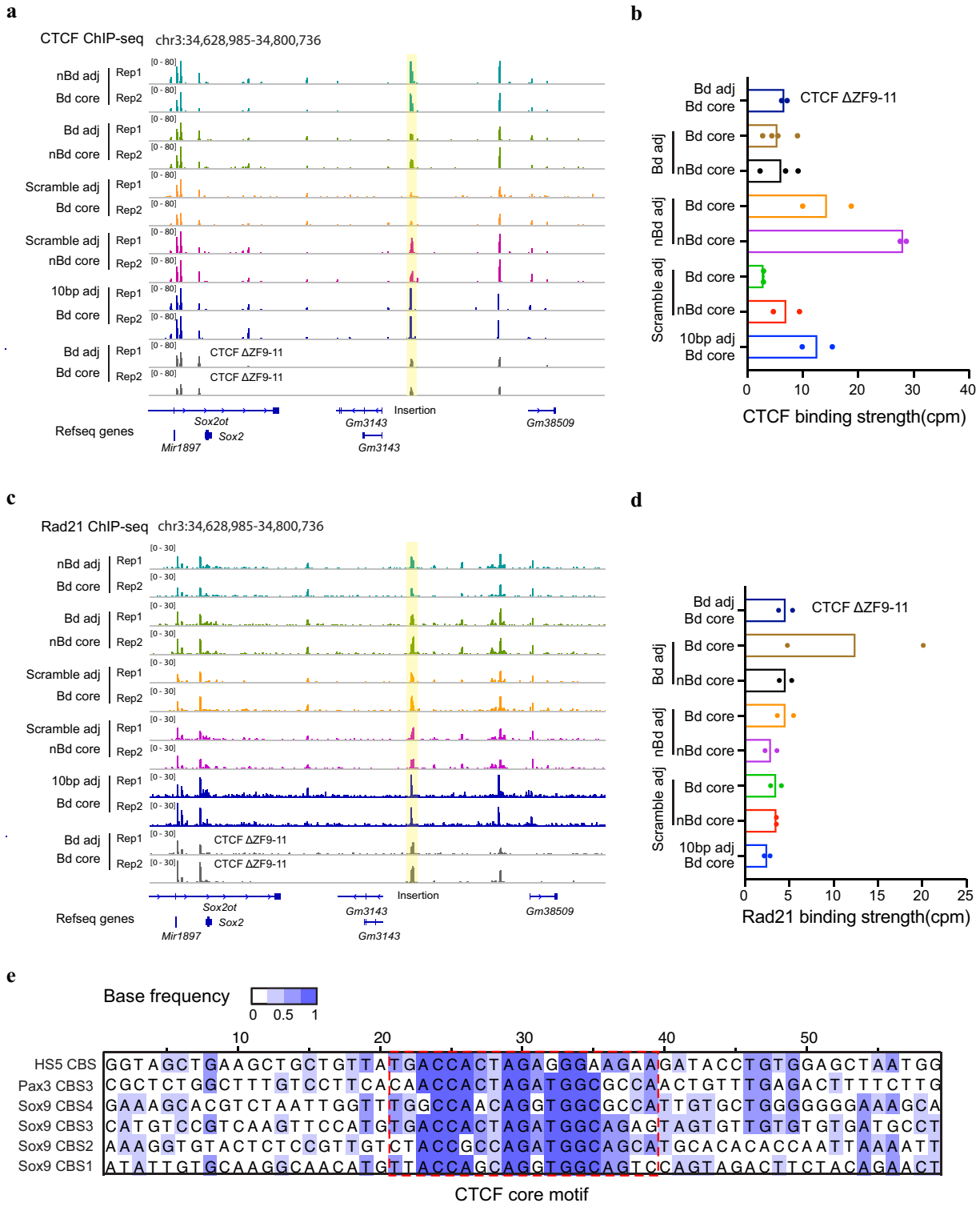
Extended Data Figure 2.2: Insulation features of CBSs from the *Sox9-Kcnj2* TAD boundary. a, Hi-C contact map of the *Sox9-Kcnj2* locus in mouse ES cells. Zoom in view shows the four CTCF binding sites cloned for insulator activity test. **b,** ChIP-seq of CTCF in the no insertion clone and the clone with an extra copy of the four *Sox9-Kcnj2* TAD boundary CBS inserted inside the *Sox2* domain. **c,** Reduction in Sox2-eGFP expression by one additional CBS. The comparison was between the clones presented in **Figure 2.2b**. (0 CBS, n = 8; 1 CBS inside, n = 23; 2 CBS inside, n = 18; 3 CBS inside, n = 13; 4 CBS inside, n = 5; Data are mean ± sd). **d,** Allele-specific *Sox2* expression in the no insertion clone (n = 2), the 4CBS clone (n = 3) , and the 4CBS downstream clone (n = 2) as measured by RNA-seq. *Sox2* expression from the CAST and 129 allele was represented by normalized read counts (rpkm) of the tagged *egfp* and *mcherry* gene, respectively. **e,** Relative *Sox2* expression in the 4CBS and the 4CBS downstream clone in **d** measured by RNA-seq and FACS. The *Sox2* expression from the e*gfp* allele was first normalized to the *mcherry* allele, then compared to the no insertion clone. **f,** FACS profiling of the no insertion clone and the 4CBS clone. **g,** FACS profiling of GFP^{low}, GFP^{high} sub-populations, and the unsort total population of the 4CBS insertion clone in **f** after extended culturing for 8 days. Left, GFP signal, right, mCherry signal from the same cells. **h,** ChIP-seq of H3K4me3 and H3K27ac in the no insertion clone and the 4CBS clone (n = 2). **i,** Allelic quantification of H3K27ac signal on the *Sox2* super-enhancer of clones in **h**. H3K27ac ChIP-seq reads on the *Sox2* super-enhancer were normalized by the total reads mapped to chromosome 3 for each allele.


a
mm10 chr11: 110,000,000-113,500,000
15
[0-60]
Ref Seq CTCF
Kcnj2
CBS cluster
Sox9
mm10 chr11: 111,512,999-111,552,417
10kb
[0 - 178]
CBS 4 3 2 1
b
CTCF ChIP-seq
mm10 chr11: 111,512,999-111,552,417
10kb
[0 - 178]
4 3 2 1
Control
4CTCF insertion
c
Sox2-eGFP reduction by one additional CBS(%)
0->1
1->2
2->3
3->4
d
Allelic Sox2 expression(rpkm)
egfp
mCherry
No insertion
4CBS
4CBS downstream
e
Relative Sox2 expression (egfp/mCherry)
4CBS
4CBS downstream
RNA-seq
FACS
f
No insertion
GFP high 99.7%
GFP low 0.067%
4CBS insertion
GFP high 58.9%
GFP low 2.10%
mCherry
eGFP
g
Normalize to mode
Total
GFP high
GFP low
Log10(eGFP)
Log10(mCherry)
h
mm10 chr3:34,625,262-34,791,799
H3K4me3
Control rep1
Control rep2
4CTCF rep1
4CTCF rep2
H3K27ac
Control rep1
Control rep2
4CTCF rep1
4CTCF rep2
Refseq
Sox2ot
Sox2
Mir1897
Gm3143
Gm3143
SE
Gm38509
i
H3K27ac allelic bias on Sox2 super enhancer (CAST/129)
Rep1
Rep2
No insertion
4CBS insertion

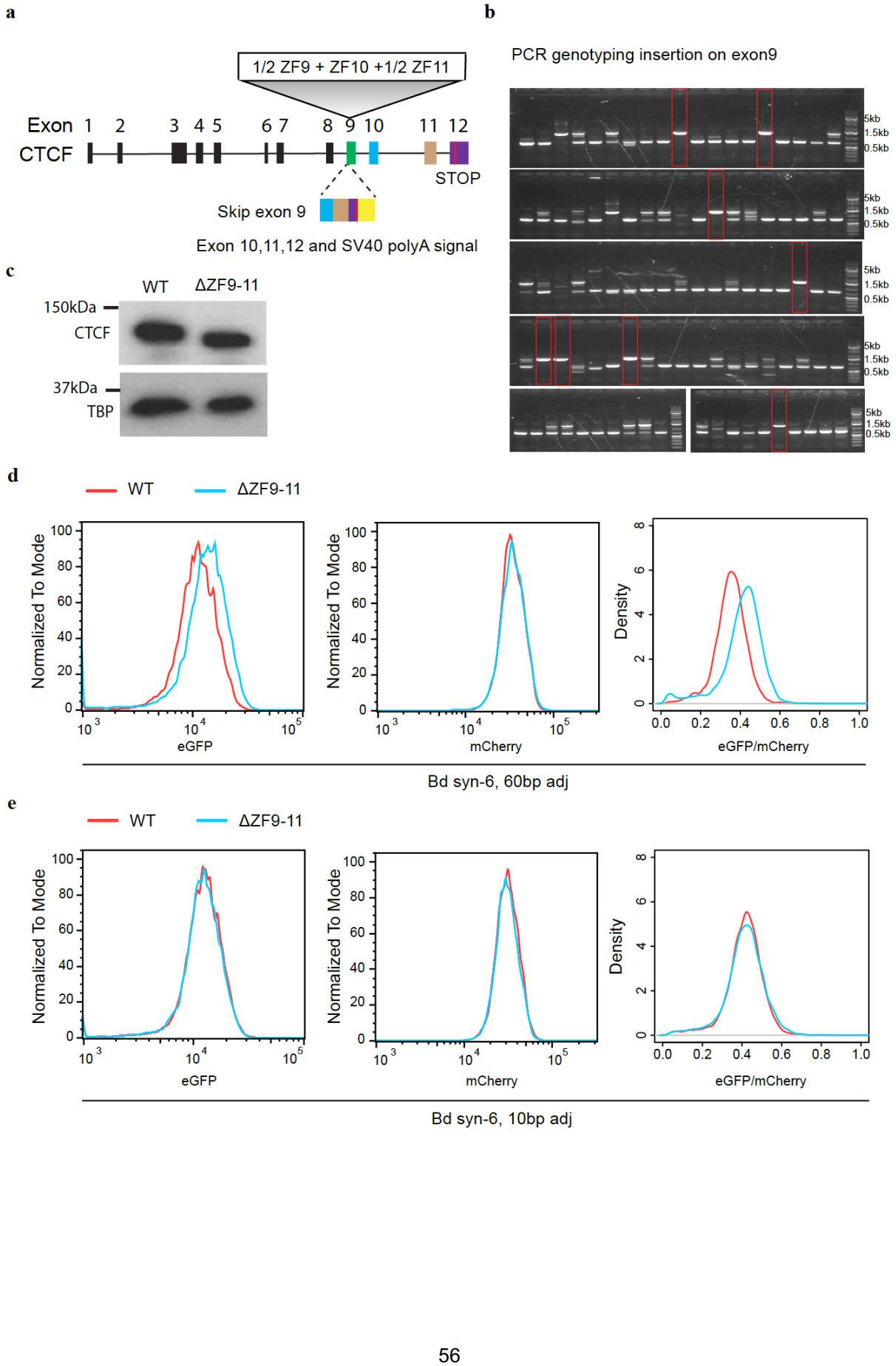
Extended Data Figure 2.3: Insulation effects of synthetic CTCF binding sites. a, Additive insulation by synthetic CBS from boundary regions. Left top, compositions of one 139bp-CBS that was synthesized; Left bottom, tandemly arrayed 139bp-CBSs tested for insulator activity. Right, normalized Sox2-eGFP expression of clones with the tandemly arrayed 139bp-CBSs inserted between the *Sox2* gene and its super-enhancer. Blue, CBS core motifs were in forward orientation; Red, CBS core motifs were in reverse orientation. Insertions were on the CAST allele only. n = 3, unpaired t-test, two-tailed. Data are mean ± sd. **b,** Insulation effects of PCR cloned large size CBSs (1-4 kb) and the synthesized 139bp-CBSs that contain the same CTCF motifs. (n = 12, paired t-test, two-tailed, ***P = 0.0007.). **c,** CTCF binding strength at selected boundary sites and non-boundary sites in mouse ES cells. ChIP-seq signals of CTCF are shown in 2-kb window. **d,** ChIP-seq of CTCF and Rad21 in clones with the insertion of six (nBd-syn6) or fifteen (nBd-syn15) 139-bp CBSs obtained from non-boundary regions. ChIP-seq reads were mapped to a customized mm10 genome that included the inserted sequence at the target site. Insertion position is highlighted in the red box.


a
One 139bp-CBS:
= Left + Core + Right
60bp 19bp 60bp
#6 #5 #4 #3 #2 #1
syn1-2
syn3-6
syn1-6
Sox2
Insert
SE
30kb
Normalized Sox2-eGFP expression
P = 0.0004
P = 0.0662
P = 0.0427
P = 0.0242
P = 0.0492
No insertion
syn1-2
syn3-6
syn1-6
syn1-2
syn3-6
syn1-6
Forward
Reverse
b
Normalized Sox2-eGFP expression
P = 0.0007
PCR
Synthetic
c
CTCF [0-949]
Refseq
Med4
Nfix
ezo1
Calm1
Sntb2
Ddx47
Chpf
Stard13
Col1a1
rss48
Boundary sites
non-Boundary sites
d
mm10 chr3:34,628,985-34,800,736
nBd syn-6
nBd syn-15
CTCF rep1
CTCF rep2
Rad21 rep1
Rad21 rep2
[0 - 127]
[0 - 39]
Insertion
SE
RefSeq
Sox2ot
Sox2
Mir1897
Gm3143
Gm3143
Gm38509

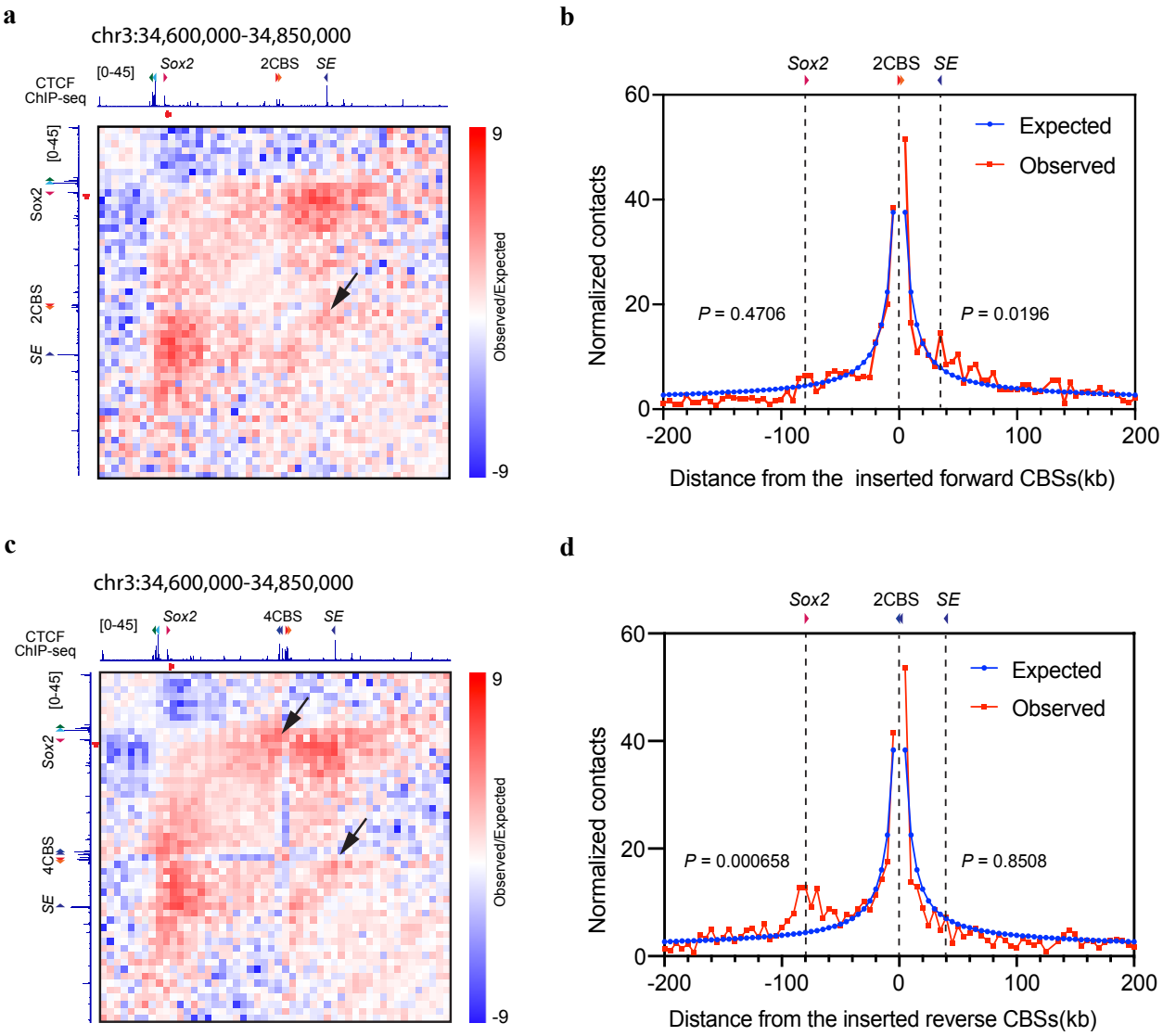
Extended Data Figure 2.4: ChIP-seq analysis of CTCF and cohesin binding at the synthetic insulators in various insulator reporter clones. a, ChIP-seq signal tracks of CTCF in clones with the insertion of different synthetic CBS variants (n = 2). Each insertion consists of six CBSs that were tandemly arrayed in forward orientation (Table 2.4). The insertion location is highlighted in the yellow box. **b,** CTCF binding strength (counts per million uniquely mapped reads) at the insertion location in the clones in (**a**). For each clone, ChIP-seq reads were mapped to a specific customized genome that contains the corresponding insertion in the *Sox2* locus (n = 2; for bd core with bd adj, n = 4; for nbd core with bd adj n = 3.). **c,** ChIP-seq signal tracks of Rad21 in the same clones in **(a)** (n = 2). The insertion location is highlighted in the yellow box. **d,** Rad21 binding strength (counts per million uniquely mapped reads) at the insertion location in the clones in (**c**). For each clone, ChIP-seq reads were mapped to a specific customized genome that contains the corresponding insertion in the *Sox2* locus (n = 2). **e,** Sequence alignment of the six boundary CBSs. Each CBS consists of 19-bp core motif plus 20-bp adjacent sequences on both sides. The color indicates the base frequency at each position. The CTCF motifs are highlighted in the red box.

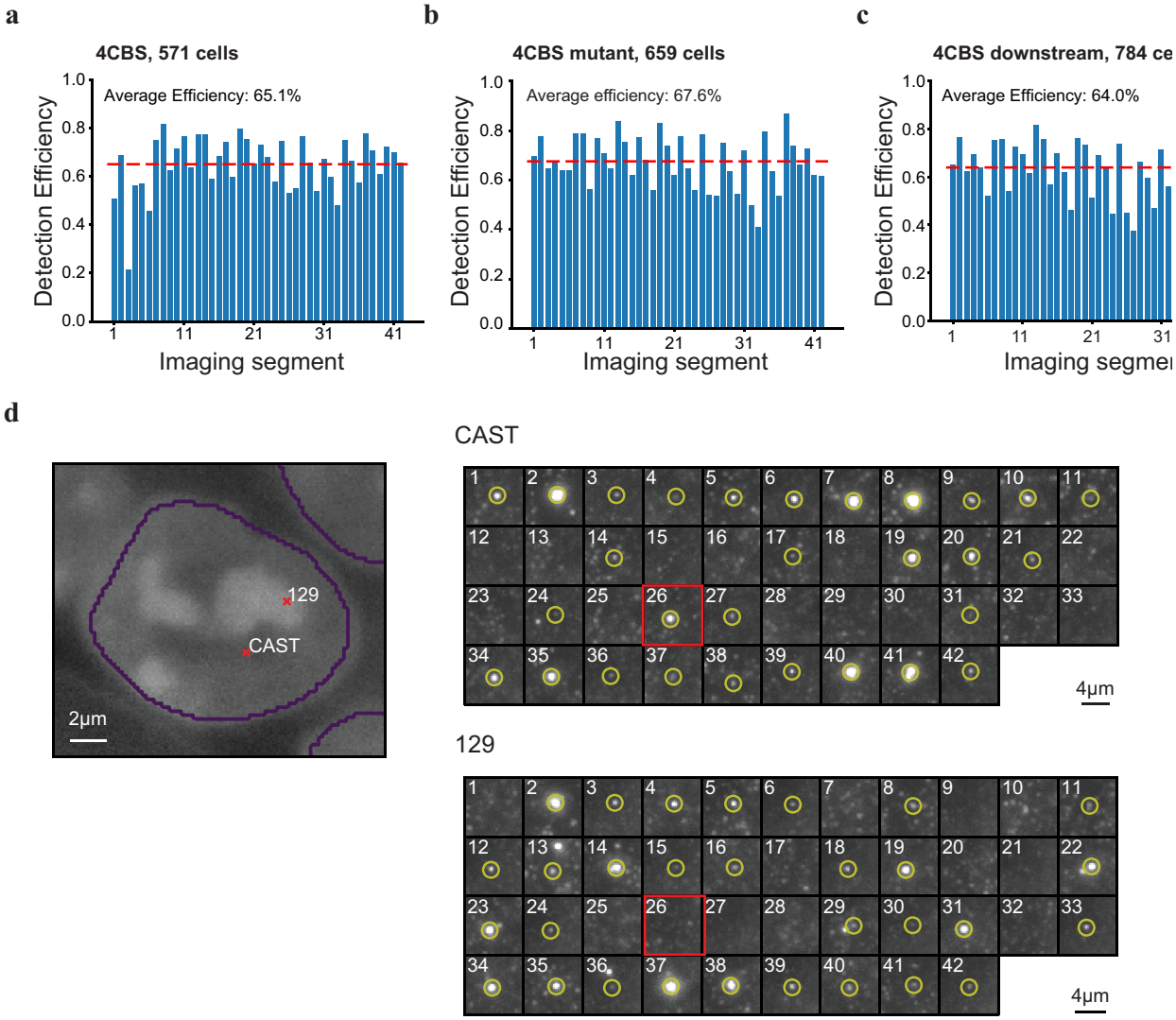

a
CTCF ChIP-seq chr3:34,628,985-34,800,736
nBd adj Rep1
Bd core Rep2
Bd adj Rep1
nBd core Rep2
Scramble adj Rep1
Bd core Rep2
Scramble adj Rep1
nBd core Rep2
10bp adj Rep1
Bd core Rep2
Bd adj Rep1
Bd core Rep2
CTCF ΔZF9-11
CTCF ΔZF9-11
Insertion
Refseq genes
Sox2ot
Sox2
Mir1897
Gm3143
Gm3143
Gm38509
b
CTCF ΔZF9-11
Bd adj
Bd core
Bd adj
Bd core
nBd core
nBd adj
Bd core
nBd core
Scramble adj
Bd core
nBd core
10bp adj
Bd core
CTCF binding strength(cpm)
c
Rad21 ChIP-seq chr3:34,628,985-34,800,736
d
Rad21 binding strength(cpm)
e
Base frequency
HS5 CBS GGTAGCTGAAGCTGCTGTTATGACCACTAGAGGGAAGAAGATACCTGTGGAGCTAATGG
Pax3 CBS3 CGCTCTGGCTTTGTCCTTCACAACCACTAGATGGCGCCAACTGTTTGAGACTTTTCTTG
Sox9 CBS4 GAAAGCACGTCTAATTGGTTTGGCCAACAGGTGGCGCCATTGTGCTGGGGGGGAAAGCA
Sox9 CBS3 CATGTCCGTCAAGTTCCATGTGACCACTAGATGGCAGAGTAGTGTTGTGTGATGCCT
Sox9 CBS2 AAAGGTGTACTCTCCGTTGTCTACCGCCAGATGGCAGCATGCACACACCAATTAAAATT
Sox9 CBS1 ATATTGTGCAAGGCAACATGTTACCAGCAGGTGGCAGTCCAGTAGACTTCTACAGAACT
CTCF core motif

Extended Data Figure 2.5: Impact of CTCF ZF-9-11 deletion on transcriptional insulation by a synthetic insulator. a, A schematic shows the experimental design to delete ZF9-11 of CTCF in mESCs. The exon 10, 11, partial of exon12, and an SV40 polyA signal were inserted into exon 9, resulting in the skip of exon9 in mRNA of the CTCF gene. **b,** PCR genotyping CTCF ΔZF9-11 mutant colonies. Genotyping primers spanned the insertion in exon 9. Highlighted in red boxes were homozygous mutant colonies evidenced by a single large-sized PCR fragment. PCR products from the homozygous mutant clones were further confirmed by Sanger sequencing. Genotyping of the homozygous mutants was repeated once with similar results. **c,** Western blot of CTCF in wild type and an exemplary CTCF ΔZF9-11 mutant clone. The primary antibody was the same one used for ChIP-seq (catalog: ab70303, lot GR3281212-7). Bottom, TBP loading control (primary antibody: sc-421, lot #B0304). Western blot was repeated once with similar results. **d,** Impact of CTCF zinc fingers 9-11 deletion on insulation effects of boundary CBSs with sixty-base-pair adjacent sequences. Left, eGFP profile of exemplary clones expressing wild-type and mutant CTCF protein; middle, mCherry profile of the same cells; right, normalized eGFP signal (eGFP/mCherry) of the wild-type and mutant clones. **e,** Impact of CTCF zinc fingers 9-11 deletion on insulation effects of boundary CBSs with ten-base-pair adjacent sequences. Left, eGFP profile of exemplary clones expressing wild-type and mutant CTCF protein; middle, mCherry profile of the same cells; right, normalized eGFP signal (eGFP/mCherry) of the wild-type and mutant clones.


a
1/2 ZF9 + ZF10 +1/2 ZF11
Exon 1 2 3 4 5 6 7 8 9 10 11 12
CTCF
STOP
Skip exon 9
Exon 10,11,12 and SV40 polyA signal
b
PCR genotyping insertion on exon9
5kb
1.5kb
0.5kb
c
WT
ΔZF9-11
150kDa
CTCF
37kDa
TBP
d
WT
ΔZF9-11
Normalized To Mode
eGFP
mCherry
Density
eGFP/mCherry
Bd syn-6, 60bp adj
e
WT
ΔZF9-11
Normalized To Mode
eGFP
mCherry
Density
eGFP/mCherry
Bd syn-6, 10bp adj

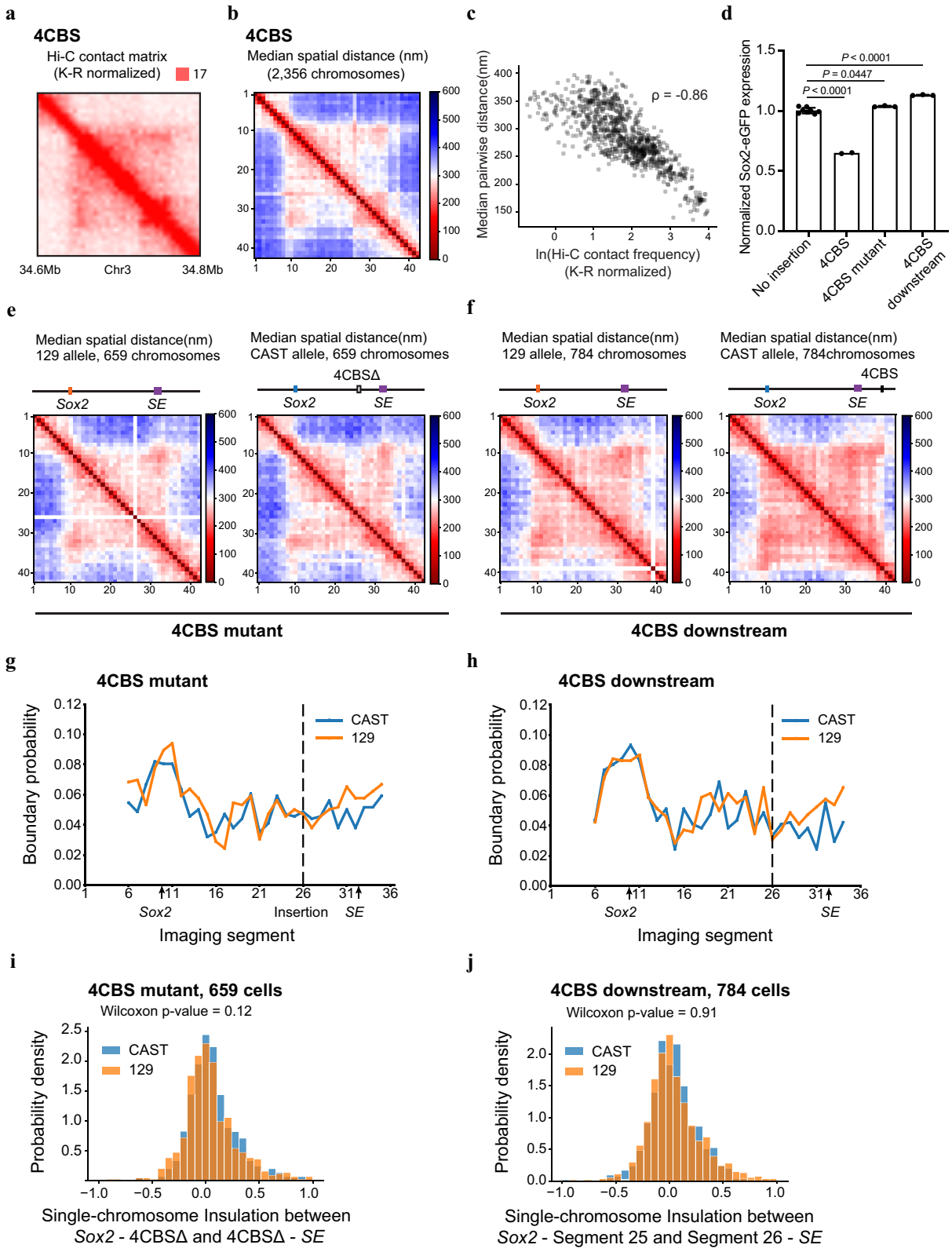
Extended Data Figure 2.6: Chromatin contacts at inserted CBSs. a, K-R normalized Hi-C matrix (Observed/Expected) in the clone with the two forward *Sox9-Kcnj2* TAD boundary CBSs inserted between the *Sox2* promoter and super-enhancer (n = 2, replicates were merged). Hi-C reads were mapped to a customized chromosome 3 containing the insertion of the two forward CBSs. ChIP-seq signal of CTCF and orientations of the inserted CBSs and CBSs around the *Sox2* promoter and super-enhancer were shown. The black arrow indicates the interactions between the inserted CBSs and the CBS on the *Sox2* super-enhancer. **b,** Virtual 4C derived from Hi-C contacts in (**a)** at the viewpoint of the two inserted CBSs. Contacts were counted in each 5kb-bin. Contacts between the inserted CBSs and the *Sox2* promoter or super-enhancer were compared to expected values, two-sided Poisson test. **c,** K-R normalized Hi-C matrix (Observed/Expected) in the clone with the four *Sox9-Kcnj2* TAD boundary CBSs inserted between the *Sox2* promoter and super-enhancer (n = 2, replicates were merged). Hi-C reads were mapped to a customized chromosome 3 containing the insertion of the four CBSs. ChIP-seq signal of CTCF and orientations of the inserted CBSs and CBSs around the *Sox2* promoter and super-enhancer were shown. The black arrows indicate the interactions between the inserted CBSs and the CBS on the *Sox2* promoter and super-enhancer. **d,** Virtual 4C derived from Hi-C contacts in (**c)** at the viewpoint of the two reverse-orientated CBSs inserted between the *Sox2* promoter and super-enhancer. Contacts were counted in each 5kb-bin. Contacts between the two reverse CBSs and the *Sox2* promoter or super-enhancer were compared to expected values, two-sided Poisson test.

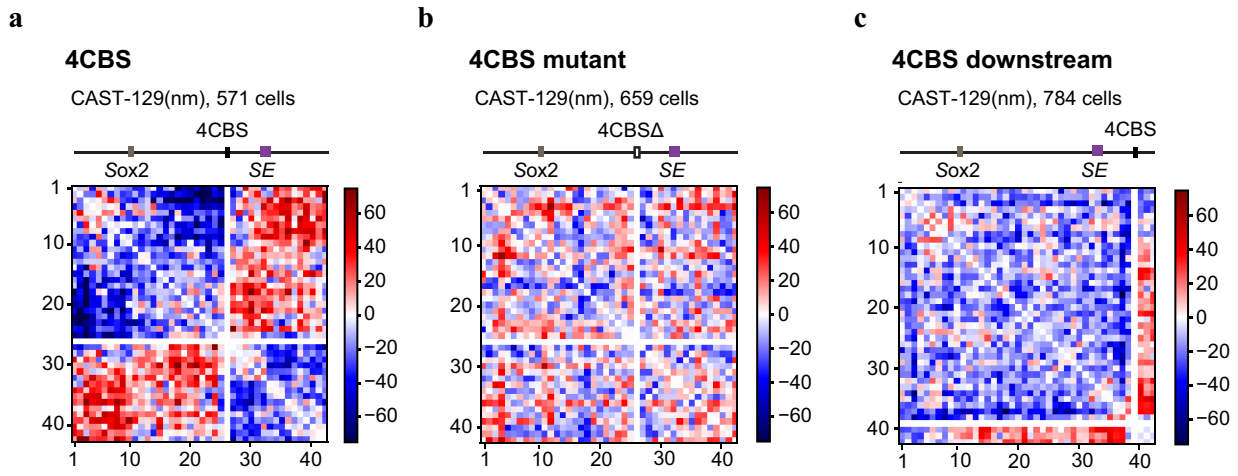

a
chr3:34,600,000-34,850,000
CTCF ChIP-seq
[0-45]
Sox2
2CBS
SE
Observed/Expected
9
-9
b
Sox2
2CBS
SE
Expected
Observed
Normalized contacts
P = 0.4706
P = 0.0196
Distance from the inserted forward CBSs(kb)
c
chr3:34,600,000-34,850,000
4CBS
d
P = 0.000658
P = 0.8508
Distance from the inserted reverse CBSs(kb)



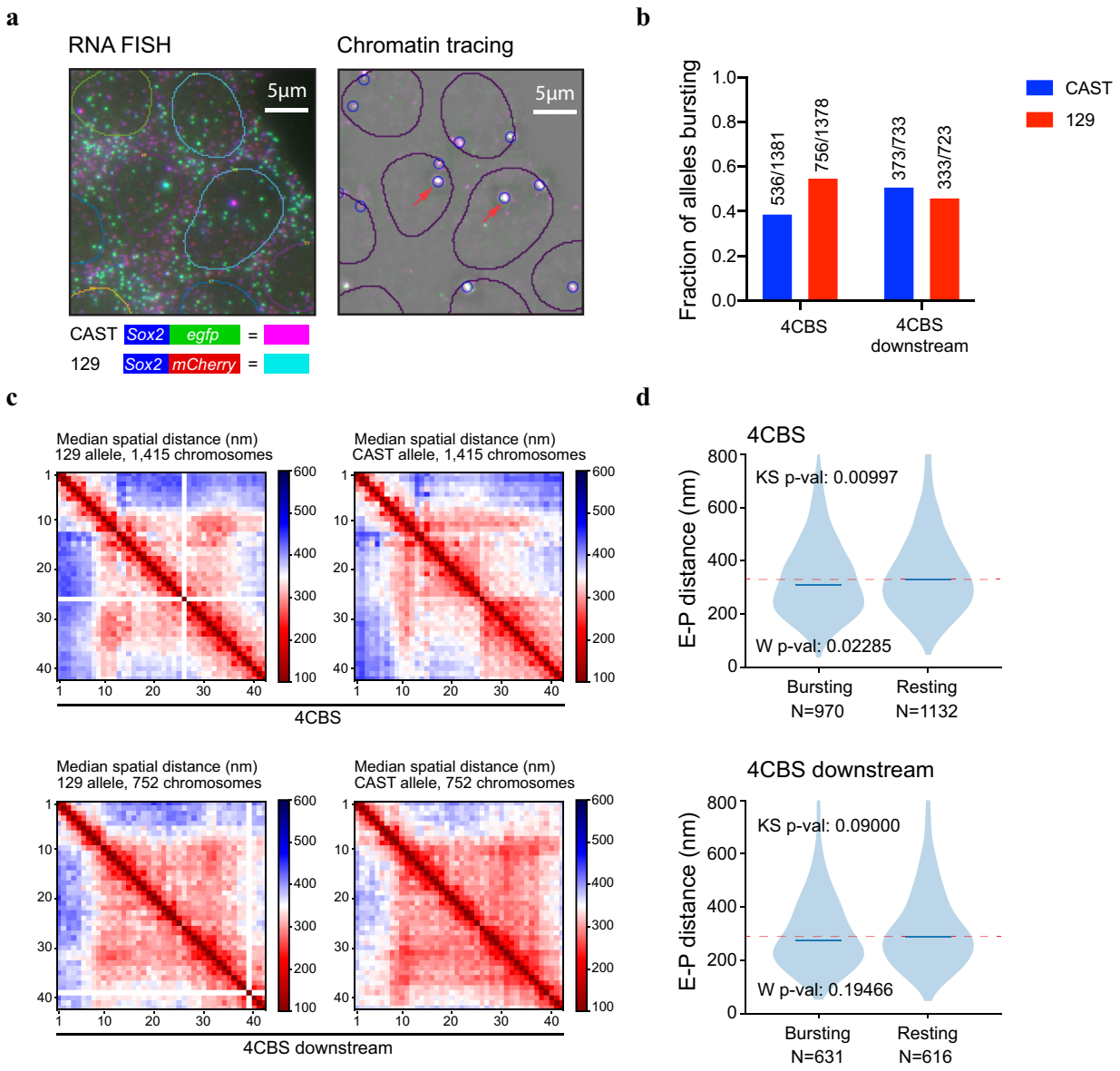
Extended Data Figure 2.7: Allele classification by multiplexed DNA FISH. a-c, Bar plots showing detect efficiency of the 42 segments of chromatin tracing experiments in the "4CBS" clone **(a)**, the "4CBS mutant" clone **(b)**, and the "4CBS downstream" clone **(c)**. Detect efficiency of each segment was calculated as the fraction of chromosomes that showed a positive fluorescence signal at the specific imaging round. **d,** Exemplary images of allele classification. Left, nuclei segmentation and the positions of CAST and 129 allele in the nucleus. Right, images of the forty-two 5-kb segments (chr3:34,601,078-34,811,078) of the CAST and 129 allele. The hybridization probes of the 26th segment (highlighted in the red box) specifically targeted the 4CBS sequence. The chromosome positive for the 26th segment (inserted 4CBS) was classified as CAST allele, the negative chromosome in the same cell was classified as 129 allele. Cells with both chromosomes positive or both chromosomes negative for the 26th segment were discarded.

Extended Data Figure 2.8: Spatial organization of the *Sox2* locus in engineered mESCs. a, Bulk Hi-C contact matrix (K-R normalized) of the *Sox2* locus in the 4CBS clone. **b,** Median pairwise distance of the same *Sox2* region measured by chromatin tracing experiment in the same clone in **a,** CAST and 129 chromosomes were combined. **c,** Correlation between the Hi-C contact frequency matrix (**a**) and median distance matrix(**b**). **d,** Normalized Sox2-eGFP expression in the no insertion clone(n = 8), the "4CBS" clone (same cells in **a-b**, n = 2), and two insertion controls, "4CBS mutant" (n = 3) and "4CBS downstream" (n = 3). One-way analysis of variance with Bonferroni's multiple comparisons test. Data are mean ± sd. **e-f,** Median spatial-distance matrix for the 210kb *Sox2* region (chr3: 34601078-34811078) of 129 (left) and CAST (right) chromosomes of the "4CBS mutant" clone**(e)** and the "4CBS downstream clone"**(f)**. The 26th segment was imaged by 4CBS specific probes in **e**. Similarly, the 38th segment was imaged by 4CBS specific probes in **f**. **g-h,** The probability of forming single-chromosome domain boundaries at each segment for the two alleles of the "4CBS mutant" clone (**g**), and the "4CBS downstream" clone (**h**). **i,** The distribution of single-chromosome insulation scores for each of the alleles between *Sox2* promoter – 4CBSΔ insertion (segments 10-25) and 4CBSΔ insertion – *Sox2* super-enhancer (segments 26-33), respectively. Two-sided Wilcoxon rank-sum test was performed. **j,** The distribution of single-chromosome insulation scores for each of the alleles between the same two domains (segment 10-25 and segment 26-33) in (**i**) for the "4CBS downstream" clone. Insulation score was calculated in the same way as in (**i**). Two-sided Wilcoxon rank-sum test was performed.

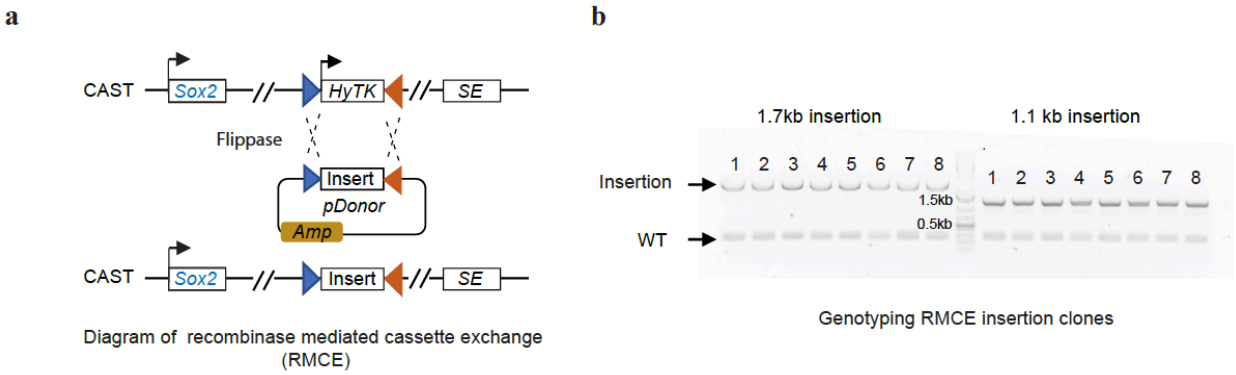

a
4CBS
Hi-C contact matrix
(K-R normalized)
17
34.6Mb
Chr3
34.8Mb
b
4CBS
Median spatial distance (nm)
(2,356 chromosomes)
c
Median pairwise distance(nm)
ρ = -0.86
ln(Hi-C contact frequency)
(K-R normalized)
d
Normalized Sox2-eGFP expression
P < 0.0001
P = 0.0447
P < 0.0001
No insertion
4CBS
4CBS mutant
4CBS downstream
e
Median spatial distance(nm)
129 allele, 659 chromosomes
Median spatial distance(nm)
CAST allele, 659 chromosomes
4CBSΔ
Sox2
SE
4CBS mutant
f
Median spatial distance(nm)
129 allele, 784 chromosomes
Median spatial distance(nm)
CAST allele, 784chromosomes
4CBS
4CBS downstream
g
4CBS mutant
Boundary probability
CAST
129
Sox2
Insertion
SE
Imaging segment
h
4CBS downstream
Boundary probability
CAST
129
Sox2
SE
Imaging segment
i
4CBS mutant, 659 cells
Wilcoxon p-value = 0.12
Probability density
CAST
129
Single-chromosome Insulation between
Sox2 - 4CBSΔ and 4CBSΔ - SE
j
4CBS downstream, 784 cells
Wilcoxon p-value = 0.91
Probability density
CAST
129
Single-chromosome Insulation between
Sox2 - Segment 25 and Segment 26 - SE



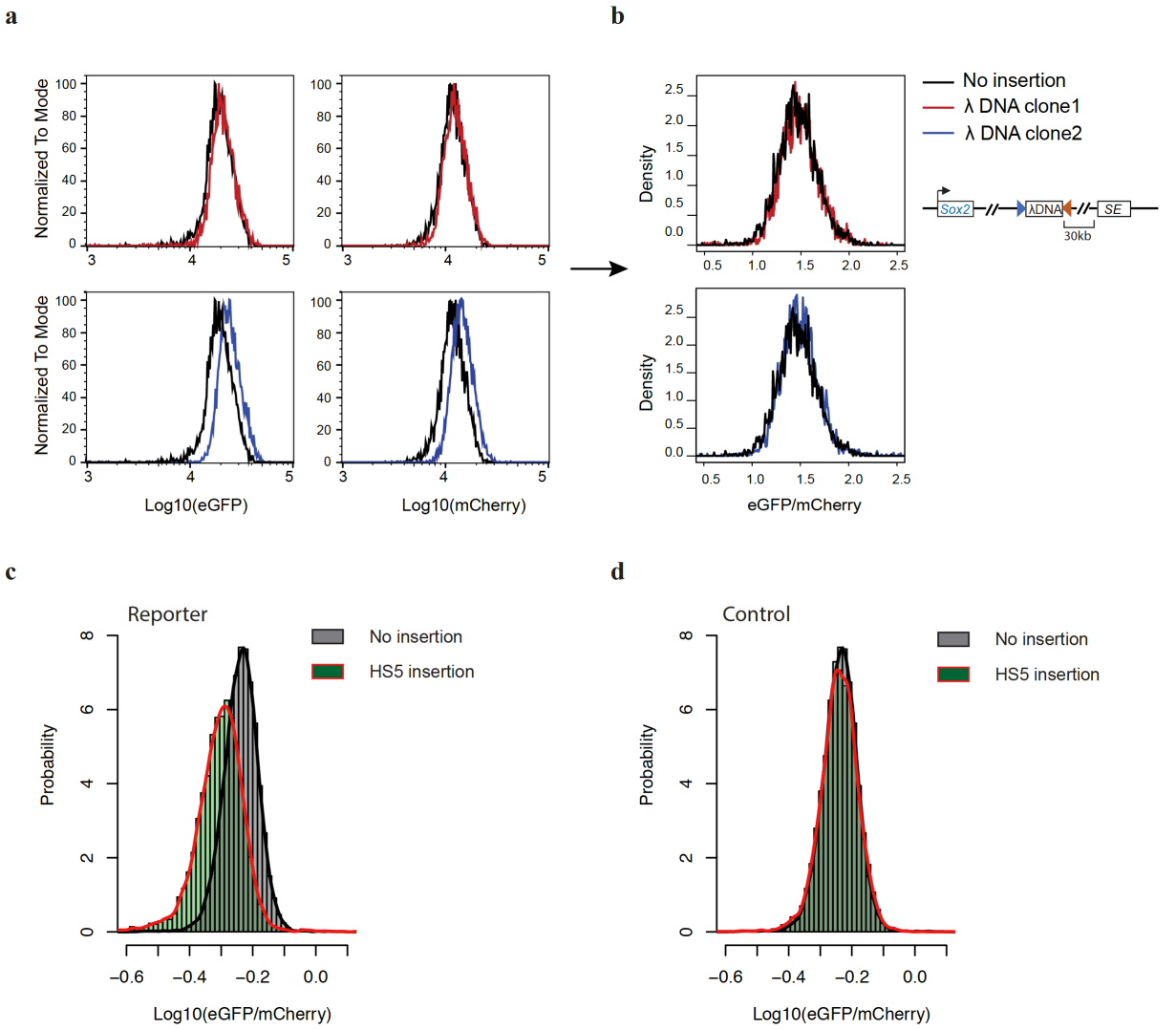
Extended Data Figure 2.9: Allele differences in median spatial distance. a-c, Difference of the median distance matrices between the CAST and 129 allele of the "4CBS" clone (**a**), the "4CBS mutant" clone (**b**) and the "4CBS downstream" clone(**c**).



Extended Data Figure 2.10: Imaging of both nascent transcripts and chromatin structure at the *Sox2* locus. a, Example images of RNA FISH(left) and chromatin tracing(right) in the same cells. Circles indicate individual nuclei. Transcripts from the CAST allele are indicated by dots in purple pseudo color. Transcripts from the 129 allele are indicated by dots in cyan pseudo color. Arrows highlight examples of bursting alleles. **b,** Bursting frequencies of the CAST and 129 allele in the 4CBS clone and the control clone with 4CBS inserted downstream of the *Sox2* SE. The numbers of bursting chromosomes and total chromosomes are indicated on the top of each bar. **c,** Median spatial distance matrices of the CAST and 129 allele in the 4CBS and the control clone. Multiplexed DNA FISH experiments were performed in the same cells following the RNA FISH experiments. **d,** Enhancer-promoter distances of the bursting and resting chromosomes in the 4CBS and the 4CBS downstream clone. A two-sided KS test between the distributions and a two-sided Wilcoxon test were performed.



Supplementary Figure 2.1: Efficiency of insertion by recombinase-mediated cassette exchange. a, Diagram of recombinase-mediated cassette exchange (RMCE) in the insulator reporter cell line. Flippase expression plasmid and the donor plasmid carrying the insertion sequence were co-electroporated into cells. The replacement only happens on the CAST allele. **b,** Genotyping insertion clones of λDNA fragments generated by RMCE. PCR primers were designed from genomic locations that spanned the insertion position. Top band, insertion fragments; Bottom band, PCR products from the no insertion allele.



Supplementary Figure 2.2: Normalization of Sox2 expression.
a-b, FACS profiles of two clones with the insertion of the same λDNA fragment. **a,** Histograms showing eGFP and mCherry signals of the two clones; **b,** Density plots of normalized signal (eGFP/mCherry) of cells from the two clones. For every cell, the ratio of eGFP signal over mCherry signal was calculated. **c,** A histogram shows the normalized Sox2-eGFP expression of cells with the human β-globin HS5 insulator inserted between the Sox2 gene and its super-enhancer. The CTCF motif of the HS5 insulator was in forward orientation and Sox2-eGFP expression was reduced by 14.1%. **d,** A histogram shows the normalized Sox2-eGFP of cells with the human β-globin HS5 insulator inserted downstream of the Sox2 super-enhancer. The CTCF motif of the HS5 insulator was in forward orientation.

2.6 Tables

Table 2.1: Genomic coordinates of individual CTCF binding sites tested.

CBS	Cloning method	Chr	Start	End	Size(bp)	Genome build
Sox9 CBS4	PCR	chr11	111,523,291	111,524,273	983	mm10
Sox9 CBS3	PCR	chr11	111,531,104	111,533,964	2,861	mm10
Sox9 CBS2	PCR	chr11	111,533,965	111,536,560	2,596	mm10
Sox9 CBS1	PCR	chr11	111,536,561	111,538,959	2,399	mm10
STIL1 CBS	PCR	chr1	47,727,792	47,729,368	1,577	hg19
HS5 CBS	PCR	chr11	5,310,702	5,313,343	2642	hg19
LMO2 CBS	PCR	chr11	34,008,004	34,011,199	3196	hg19
PAX3 CBS1	PCR	chr1	77,942,190	77,944,770	2581	mm10
PAX3 CBS2	PCR	chr1	77,970,975	77,974,415	3441	mm10
PAX3 CBS3	PCR	chr1	77,976,869	77,980,896	4028	mm10
PAX3 CBS4	PCR	chr1	78,006,962	78,010,814	3853	mm10

Table 2.2: Combinations of Sox9-Kcnj2 TAD boundary CBSs.

Clone	# of CBS	Orientation	Position	Normalized Sox2-eGFP expression
No insertion control_tech_rep1	0	NA	NA	0.98
No insertion control_tech_rep2	0	NA	NA	1
No insertion control_tech_rep3	0	NA	NA	0.993
No insertion control_tech_rep4	0	NA	NA	1.004
No insertion control_tech_rep5	0	NA	NA	1.038
No insertion control_tech_rep6	0	NA	NA	0.989
No insertion control_tech_rep7	0	NA	NA	0.972
No insertion control_tech_rep8	0	NA	NA	1.034
mSox9-CBS4-pos-clone1	1	<	between Sox2 promoter and SE	1.053
mSox9-CBS4-pos-clone2	1	<	between Sox2 promoter and SE	1.057
mSox9-CBS4-pos_clone3	1	<	between Sox2 promoter and SE	1.012
mSox9-CBS2-neg-clone1	1	<	between Sox2 promoter and SE	0.931
mSox9-CBS2-neg-clone2	1	<	between Sox2 promoter and SE	0.884
mSox9-CBS2-neg-clone3	1	<	between Sox2 promoter and SE	0.964
mSox9-CBS2-pos-clone1	1	>	between Sox2 promoter and SE	0.848
mSox9-CBS2-pos-clone2	1	>	between Sox2 promoter and SE	0.828
mSox9-CBS2-pos-clone3	1	>	between Sox2 promoter and SE	0.804
mSox9-CBS1-neg-clone1	1	<	between Sox2 promoter and SE	1.075

Table 2.2: Combinations of Sox9-Kcnj2 TAD boundary CBSs, continued.

Clone	# of CBS	Orientation	Position	Normalized Sox2-eGFP expression
mSox9-CBS1-neg-clone2	1	<	between Sox2 promoter and SE	1.06
mSox9-CBS1-neg-clone3	1	<	between Sox2 promoter and SE	1.036
mSox9-CBS1-pos-clone2	1	>	between Sox2 promoter and SE	1.02
mSox9-CBS1-pos-clone1	1	>	between Sox2 promoter and SE	1.042
mSox9-CBS4-neg-clone1	1	>	between Sox2 promoter and SE	0.908
mSox9-CBS4-neg-clone2	1	>	between Sox2 promoter and SE	0.991
mSox9-CBS4-neg-clone3	1	>	between Sox2 promoter and SE	0.972
mSox9-CBS3-neg-clone1	1	>	between Sox2 promoter and SE	0.908
mSox9-CBS3-neg-clone2	1	>	between Sox2 promoter and SE	0.882
mSox9-CBS3-neg-clone3	1	>	between Sox2 promoter and SE	0.865
mSox9-CBS3-pos-clone1	1	<	between Sox2 promoter and SE	0.949
mSox9-CBS3-pos-clone2	1	<	between Sox2 promoter and SE	0.941
mSox9-CBS3-pos-clone3	1	<	between Sox2 promoter and SE	0.908
mSox9-CBS23-pos-clone1	2	< >	between Sox2 promoter and SE	0.778
mSox9-CBS23-pos-clone2	2	< >	between Sox2 promoter and SE	0.767
mSox9-CBS23-pos-clone3	2	< >	between Sox2 promoter and SE	0.754
mSox9-CBS23-neg-clone1	2	< >	between Sox2 promoter and SE	0.792
mSox9-CBS23-neg-clone2	2	< >	between Sox2 promoter and SE	0.755

Table 2.2: Combinations of Sox9-Kcnj2 TAD boundary CBSs, continued.

Clone	# of CBS	Orientation	Position	Normalized Sox2-eGFP expression
mSox9-CBS23-neg-clone3	2	< >	between Sox2 promoter and SE	0.77
mSox9-CBS34-neg_clone2	2	> >	between Sox2 promoter and SE	0.836
mSox9-CBS34-neg_clone3	2	> >	between Sox2 promoter and SE	0.676
mSox9-CBS34-neg_clone5	2	> >	between Sox2 promoter and SE	0.766
mSox9-CBS34-pos_clone1	2	< <	between Sox2 promoter and SE	0.947
mSox9-CBS34-pos_clone2	2	< <	between Sox2 promoter and SE	0.947
mSox9-CBS34-pos_clone3	2	< <	between Sox2 promoter and SE	0.931
mSox9_CBS12_neg_clone1	2	< <	between Sox2 promoter and SE	0.77
mSox9_CBS12_neg_clone2	2	< <	between Sox2 promoter and SE	0.786
mSox9_CBS12_neg_clone3	2	< <	between Sox2 promoter and SE	0.806
mSox9_CBS12_pos_clone1	2	> >	between Sox2 promoter and SE	0.704
mSox9_CBS12_pos_clone2	2	> >	between Sox2 promoter and SE	0.706
mSox9_CBS12_pos_clone3	2	> >	between Sox2 promoter and SE	0.692
mSox9_CBS123_neg_clone1	3	< < >	between Sox2 promoter and SE	0.729
mSox9_CBS123_neg_clone2	3	< < >	between Sox2 promoter and SE	0.734
mSox9_CBS123_neg_clone3	3	< < >	between Sox2 promoter and SE	0.72
mSox9_CBS234_neg_clone2	3	< > >	between Sox2 promoter and SE	0.61
mSox9_CBS234_neg_clone3	3	< > >	between Sox2 promoter and SE	0.638

Table 2.2: Combinations of Sox9-Kcnj2 TAD boundary CBSs, continued.

Clone	# of CBS	Orientation	Position	Normalized Sox2-eGFP expression
mSox9_CBS234_neg_clone4	3	< > >	between Sox2 promoter and SE	0.616
mSox9-CBS123-pos-clone1	3	< > >	between Sox2 promoter and SE	0.709
mSox9-CBS123-pos-clone2	3	< > >	between Sox2 promoter and SE	0.687
mSox9-CBS123-pos-clone3	3	< > >	between Sox2 promoter and SE	0.734
mSox9-CBS123-pos-clone4	3	< > >	between Sox2 promoter and SE	0.655
mSox9_CBS234_pos_clone2	3	< < >	between Sox2 promoter and SE	0.609
mSox9_CBS234_pos_clone4	3	< < >	between Sox2 promoter and SE	0.631
mSox9_CBS234_pos_clone5	3	< < >	between Sox2 promoter and SE	0.607
mSox9-CBS1234-pos-clone1	4	< < > >	between Sox2 promoter and SE	0.652
mSox9-CBS1234-pos-clone3	4	< < > >	between Sox2 promoter and SE	0.645
mSox9-CBS1234-neg-clone1	4	< < > >	between Sox2 promoter and SE	0.584
mSox9-CBS1234-neg-clone2	4	< < > >	between Sox2 promoter and SE	0.6
mSox9-CBS1234-neg-clone3	4	< < > >	between Sox2 promoter and SE	0.602
SE-Sox9-CBS2-neg-clone1	1	<	Downstream of Sox2 SE	1.078
SE-Sox9-CBS1-pos-clone1	1	>	Downstream of Sox2 SE	1.088

Table 2.2: Combinations of Sox9-Kcnj2 TAD boundary CBSs, continued

Clone	# of CBS	Orientation	Position	Normalized Sox2-eGFP expression
SE-Sox9-CBS4_pos_clone1	1	<	Downstream of Sox2 SE	1.051
SE-Sox9-CBS4_pos_clone2	1	<	Downstream of Sox2 SE	1.066
SE-Sox9-CBS4_pos_clone4	1	<	Downstream of Sox2 SE	1.062
SE-Sox9-CBS2-neg-clone2	1	<	Downstream of Sox2 SE	1.128
SE-Sox9-CBS2-neg-clone3	1	<	Downstream of Sox2 SE	1.07
SE-Sox9-CBS2-pos-clone1	1	>	Downstream of Sox2 SE	1.148
SE-Sox9-CBS2-pos-clone2	1	>	Downstream of Sox2 SE	1.092
SE-Sox9-CBS2-pos-clone3	1	>	Downstream of Sox2 SE	1.113
SE-Sox9-CBS1-neg-clone1	1	<	Downstream of Sox2 SE	1.031
SE-Sox9-CBS1-neg-clone3	1	<	Downstream of Sox2 SE	1.061
SE-Sox9-CBS1-pos-clone1	1	>	Downstream of Sox2 SE	1.066
SE-Sox9-CBS1-pos-clone2	1	>	Downstream of Sox2 SE	1.03
SE-Sox9-CBS1-pos-clone3	1	>	Downstream of Sox2 SE	1.048
SE-Sox9-CBS23-pos-clone1	2	< >	Downstream of Sox2 SE	1.094
SE-Sox9-CBS23-neg-clone1	2	< >	Downstream of Sox2 SE	1.117
SE-Sox9-CBS12-pos-clone1	2	> >	Downstream of Sox2 SE	1.149
SE-Sox9-CBS12-pos-clone2	2	> >	Downstream of Sox2 SE	1.136

Table 2.2: Combinations of Sox9-Kcnj2 TAD boundary CBSs, continued

Clone	# of CBS	Orientation	Position	Normalized Sox2-eGFP expression
SE-Sox9-CBS12-pos-clone1-p4	2	> >	Downstream of Sox2 SE	1.114
SE-Sox9-CBS12-neg-clone3	2	< <	Downstream of Sox2 SE	1.073
SE-Sox9-CBS12-neg-clone1	2	< <	Downstream of Sox2 SE	1.05
SE-Sox9-CBS12-neg-clone2	2	< <	Downstream of Sox2 SE	1.071
SE-Sox9-CBS123-pos-clone1	3	< > >	Downstream of Sox2 SE	1.13
SE_Sox9_CBS234_pos_clone2	3	< < >	Downstream of Sox2 SE	0.966
SE_Sox9_CBS234_pos_clone3	3	< < >	Downstream of Sox2 SE	1.119
SE_Sox9_CBS234_pos_clone4	3	< < >	Downstream of Sox2 SE	1.047
SE-Sox9-CBS1234-pos-clone1	4	< < > >	Downstream of Sox2 SE	1.134
SE-Sox9-CBS1234-pos-clone2	4	< < > >	Downstream of Sox2 SE	1.128
SE-Sox9-CBS1234-pos-clone3	4	< < > >	Downstream of Sox2 SE	1.127
SE-Sox9-CBS1234-neg-clone2	4	< < > >	Downstream of Sox2 SE	1.093
SE-Sox9-CBS1234-neg-clone3	4	< < > >	Downstream of Sox2 SE	1.116
SE-Sox9-CBS1234-neg-clone5	4	< < > >	Downstream of Sox2 SE	1.139

Table 2.3: Genomic coordinates of synthetic 139bp-CBSs.

CBS	Cloning method	Chr	Start	End	Size	Genome
HS5	synthetic	chr1	77978819	77978957	139bp	hg19
Pax3 CBS3	synthetic	chr11	111523608	111523746	139bp	mm10
Sox9 CBS4	synthetic	chr11	111523608	111523746	139bp	mm10
Sox9 CBS3	synthetic	chr11	111532592	111532730	139bp	mm10
Sox9 CBS2	synthetic	chr11	111536158	111536296	139bp	mm10
Sox9 CBS1	synthetic	chr11	111537779	111537917	139bp	mm10
nBd1	synthetic	chr14	73509834	73509972	139bp	mm10
nBd2	synthetic	chr8	84737673	84737811	139bp	mm10
nBd3	synthetic	chr8	122551650	122551788	139bp	mm10
nBd4	synthetic	chr12	100201933	100202071	139bp	mm10
nBd5	synthetic	chr8	107002599	107002737	139bp	mm10
nBd6	synthetic	chr9	77596465	77596603	139bp	mm10
nBd7	synthetic	chr17	49282601	49282739	139bp	mm10
nBd8	synthetic	chr6	135011479	135011617	139bp	mm10
nBd9	synthetic	chr6	3183938	3184076	139bp	mm10
nBd10	synthetic	chr1	75477514	75477652	139bp	mm10
nBd11	synthetic	chr18	35939577	35939715	139bp	mm10
nBd12	synthetic	chr5	151102500	151102638	139bp	mm10
nBd13	synthetic	chr12	32292017	32292155	139bp	mm10
nBd14	synthetic	chr11	94937112	94937250	139bp	mm10
nBd15	synthetic	chr3	86002612	86002750	139bp	mm10

Table 2.4: Sequences of tandemly arrayed synthetic 139bp-CBSs.

Construct	Sequence
Tandemly arrayed six 139-bp boundary CBSs	TTATATTTCCTGACCTATATCTGGCAGGACTCTTTAGAGAGGTAGCTGA AGCTGCTGTTATGACCACTAGAGGGAAGAAGATACCTGTGGAGCTAAT GGTCCAAGATGGTGGAGCCCCAAGCAAGGAAGTTGTTAAGGATTCCAG AAGTGTGTGGGCTTACTAGCCTGGAATCGGTTCGCGCTCTGGCTTTGT CCTTCACAACCACTAGATGGCGCCAACTGTTTGAGACTTTTCTTGGTGC CTGACACTGTCTGGTTTCCAAACCCCCAAAAATGAAAGTGGCTTATTTC AGGTGGACCAGAACACCCAAAACAAATGAAAGCACGTCTAATTGGTTTG GCCAACAGGTGGCGCCATTGTGCTGGGGGGGAAAGCATGCCCTCCCT TATGGGAAGTTGTAGGCAATATTCTGGGTGATTCACAGCATTAGAATAA AATTTAAACATGGCTAAATAACATGTCCGTCAAGTTCCATGTGACCACTA GATGGCAGAGTAGTGTTGTGTGTGATGCCTGCAAGTTTTACCCCTAAAT TATTTGCCCgtacgtgtgtatAACTATCAAGATGTATTTGCCGATTGAACTGAA AACAAGAAAAGGTGTACTCTCCGTTGTCTACCGCCAGATGGCAGCATG CACACACCAATTAAAATTTGTTCCTCTAGAGGCTAAAAGTTATCCATTGG AAACAACTGTCAACAGACACTTTGGACGTTTTAGATCAGGAAAGGAAAA TATTGTGCAAGGCAACATGTTACCAGCAGGTGGCAGTCCAGTAGACTT CTACAGAACTGTGGCTGAAAAGGAGAAAGACACACAGCCTCATGAAGC TA
Boundary core motifs + scramble adajacent sequences	CGGATATAAACAGAGAAATGGCTGATTATACCTATTCGTATAGTATCGA TAAATAGCCTCTGACCACTAGAGGGAAGAAGCGGAGCCTTATGCCATA CTTGTCTGCGGAGCACTCTAGTAATGCATATGGTCCACAGGAGAGATC ACTGACCAATCTATCTGAACGGCAACCTTGTATCGTACTGGAGCTTGGA GAGATCAACCACTAGATGGCGCCAAACTACAATGCCGCTTACAACCTCT CTGTCGTCGCTGACGTTTATAGTCTAGTCTCATTATAATTGTACGCTATT GAGGCATTGACTAATACCGGAACATCTGAAATGAACTAATCTATATGGC CAACAGGTGGCGCCACGACAGAAACCAAGTGCACCTACCAAATCTCTT TAGTGTAAGTTCTGACTAATTCGTACTGCTTTCGTGTGCCTCTAATGGC TCGTTAGATAGTCTAGCCGCTGGTAAACACTCCATGACTGACCACTAGA TGGCAGAGCTCGGCTCTCCATTGATACTACGGCGATTGTTGGAGAGCC AGCAGCGATTGCAAATGTCAGAATTATCGCGGCAATGACAACGAAGCA ACATCTCGGGTCTTGCCTAACCAAGGTCTACACTACCGCCAGATGGCA GCATGTTGATATAGTGAATCACTGAACCCAGTGCCACACAATGGAATGT CCTTAATTCTGGCAGGGTGTACTCAGTTCTATAAACGAGCTATTAAATAT GAGATCCGTAGATTGAAGGGTAACTTACCAGCAGGTGGCAGTCAGAAT TTGCCTGGATGCAAGACGGACAGCTAGGTATCCTAAGTATAGTTGCGA ACGTCCG

Table 2.4: Sequences of tandemly arrayed synthetic 139bp-CBSs, continued.

Construct	Sequence
Boundary core motifs + non-boundary adajacent sequences	TTCAGAGGGAACCCTGTCTGCGGAACTTCTAGGCCAGCACCCCACGGG CTCGCCCTACTGTGACCACTAGAGGGAAGAAAGTGTTAACTCGCTGCGC CACCTAGCTACCTTGTTGGGGAGCGCAACGTGAAGCTCCTTGCCACCCC AACCACCAAGGCCTAGCCAGCGTGGCCGCCAGTGCCACAGTTAGGGAC TCTGACAACCACTAGATGGCGCCACGCCATCACGGCTTGGGGAACAAA GCGAAAGTTGCTCCAACCACGAGGGGCGCAGTTGCGCTTCTTGTCCGC GGCATTGGGCTGGGTCTCTTGGATGAAATCAGCAACCGCAGTGTTTATT GGCCAACAGGTGGCGCCATTCCTTCTCTGCCTGCTGCTGCAGCTGCTG CTGAGGGTACTTTAACCACTAGAGGCGCCAAGTTCATAATGTCTGCCAC CAGGGGCACATCCACTGTGGAACAAAGCAACCAGCAGGGACTGACCAC TAGATGGCAGAGCCAGAGTGGAAATGAGGCAACCAACTCCTGAATGTTG GACAGACATTTTCAAGCTTTTGAGGTAGGTTGAGAAACTCGCGATAATTC CCAGCTAGTCTTTACAGATGTTCTTCTAGGCTACTACCGCCAGATGGCA GCACAGGCTGCTTCAATTTCCTAGGACATTTTCCCCACTAGGGGCGCTG TAAGTTATGGCGAGTGGTGGGGTAATAGTAAAAGTCCCAACACATGCGA GGCAGCACACAGCCAATAACTTTTCTTACCAGCAGGTGGCAGTCTAGTG CCCCTTCTGGCCACTGCAGGTACTGCATGCATGTGATGCGGATACACAT AGAAGG
Tandemly arrayed six 139-bp non-boundary CBSs	TTCAGAGGGAACCCTGTCTGCGGAACTTCTAGGCCAGCACCCCACGGG CTCGCCCTACTGTGACCGGCAGAGGGCAGCTAGTGTTAACTCGCTGCG CCACCTAGCTACCTTGTTGGGGAGCGCAACGTGAAGCTCCTTGCCACCC CAACCACCAAGGCCTAGCCAGCGTGGCCGCCAGTGCCACAGTTAGGGA CTCTGACAGCCCCCAGAGGGCGCTGCGCCATCACGGCTTGGGGAACAA AGCGAAAGTTGCTCCAACCACGAGGGGCGCAGTTGCGCTTCTTGTCCG CGGCATTGGGCTGGGTCTCTTGGATGAAATCAGCAACCGCAGTGTTTAT GGCGCAGTAGGTGGCGCTCTTCCTTCTCTGCCTGCTGCTGCAGCTGCT GCTGAGGGTACTTTAACCACTAGAGGCGCCAAGTTCATAATGTCTGCCA CCAGGGGCACATCCACTGTGGAACAAAGCAACCAGCAGGGACTCGCCA CTAGAGGGAGCGCCCAGAGTGGAAATGAGGCAACCAACTCCTGAATGT TGGACAGACATTTTCAAGCTTTTGAGGTAGGTTGAGAAACTCGCGATAAT TCCCAGCTAGTCTTTACAGATGTTCTTCTAGGCTATCGCCACTAGGGGG CAGGGCAGGCTGCTTCAATTTCCTAGGACATTTTCCCCACTAGGGGCGC TGTAAGTTATGGCGAGTGGTGGGGTAATAGTAAAAGTCCCAACACATGC GAGGCAGCACACAGCCAATAACTTTTCCGGTCGCTAGGGGGAGATCTA GTGCCCCTTCTGGCCACTGCAGGTACTGCATGCATGTGATGCGGATACA CATAGAAGG

Table 2.4: Sequences of tandemly arrayed synthetic 139bp-CBSs, continued.

Construct	Sequence
non-Boundary core motifs + scramble adajacent sequences	TTCAGACCGTCCTTTAATTTCCCTTGCATATATGTTGCGTTTCTTCGACCT TCTAACCGCTGACCGGCAGAGGGCAGCTACCCTTAGGACGGAGACAGA TCCACGTTCTTACCCGTGCCACCGTTGGCAGCGGGATCGCACCGACCT GCGTTCGGCATTGTGGGCAGAGTGAAGTATTGGCAAACGTTAAGTGCC GAACCAGCCCCCAGAGGGCGCTGTAGATCTGACCTAACGGTAAGAGAG TTTCATAATACGTCCAGCCGCACGCGCAGGGTACATGACTCAAACAGAG TACATCCTGCCCGCGTTTCGCATGAATCAAGTTGGAGGTTATGGAGGGC GCAGTAGGTGGCGCTCCATAGTAACATGTGGACGGCCAGTGGTCGGTT GCTACACGCCTGCCGCAACGTTGAAGGTTGGTGTCTCGTATTCCTCTTG GAGATCGAGGAAATGTTTCACGACCAAGGGAAAGGTCGCTCGCCACTA GAGGGAGCGCCCTACGGAATAGATTTGCGTTACTGCCTGCATAAGGAGT CCGGTGTAGCCAAGGACGAAGCAAATTATAGCCGTACAGACCCTAATCT CATGTCATATCACGCGACTAGCCTCTGCTTAATCGCCACTAGGGGGCAG GGTTTCTGTGCTCAAGTTGGTTGGTCCGCCCGAGCGGTGCTGCCGATTA GGACCATCAAATGCGCGCCATCTCTGAGCAGGTGGGCCGACGAGACAC TGTCCCTGATTTCTCCGCTACTAATCGGTCGCTAGGGGGAGATCAGCAC TCACGGCGCAATACCAGCACAGCCCAGTCTCGCCGGAACGCTGGTCAG CATACGA
non-Boundary core motifs +boundary adajacent sequences	TTATATTTCCTGACCTATATCTGGCAGGACTCTTTAGAGAGGTAGCTGAA GCTGCTGTTATGACCGGCAGAGGGCAGCTGATACCTGTGGAGCTAATG GTCCAAGATGGTGGAGCCCCAAGCAAGGAAGTTGTTAAGGATTCCAGAA GTGTGTGGGCTTACTAGCCTGGAATCGGTTCGCGCTCTGGCTTTGTCCT TCACAGCCCCCAGAGGGCGCTGACTGTTTGAGACTTTTCTTGGTGCCTG ACACTGTCTGGTTTCCAAACCCCCAAAAATGAAAGTGGCTTATTTCAGGT GGACCAGAACACCCAAAACAAATGAAAGCACGTCTAATTGGTTGGCGCA GTAGGTGGCGCTCTTGTGCTGGGGGGGAAAGCATGCCCTCCCTTATGG GAAGTTGTAGGCAATATTCTGGGTGATTCACAGCATTAGAATAAAATTTA AACATGGCTAAATAACATGTCCGTCAAGTTCCATGTCGCCACTAGAGGG AGCGCTAGTGTTGTGTGTGATGCCTGCAAGTTTTACCCCTAAATTATTTG CCCgtacgtgtgtatAACTATCAAGATGTATTTGCCGATTGAACTGAAAACAAG AAAAGGTGTACTCTCCGTTGTTCGCCACTAGGGGGCAGGGTGCACACA CCAATTAAAATTTGTTCCTCTAGAGGCTAAAAGTTATCCATTGGAAACAA CTGTCAACAGACACTTTGGACGTTTTAGATCAGGAAAGGAAAATATTGTG CAAGGCAACATGCGGTCGCTAGGGGGAGATCCAGTAGACTTCTACAGA ACTGTGGCTGAAAAGGAGAAAGACACACAGCCTCATGAAGCTA

Table 2.4: Sequences of tandemly arrayed synthetic 139bp-CBSs, continued.

Construct	Sequence
Tandemly arrayed fifteen 139-bp non-boundary CBSs	ATCTCTCATCCTGGGTCCGGTTGGAGCATGTCAGATTGGAACGAGTCAC AAGCCCTCCTTTGCCCTGTAGATGGCGCTAGAACTGTAGTGCACCTCCA GGCACATAGGCGGCGCTCTAGCCGGACCTCTCCGTCCCGACTTCAGAG GGAACCCTGTCTGCGGAACTTCTAGGCCAGCACCCCACGGGCTCGCCC TACTGTGACCGGCAGAGGGCAGCTAGTGTTAACTCGCTGCGCCACCTA GCTACCTTGTTGGGGAGCGCAACGTGAAGCTCCTTGCCACCCCAACCA CCAAGGCCTAGCCAGCGTGGCCGCCAGTGCCACAGTTAGGGACTCTGA CAGCCCCCAGAGGGCGCTGCGCCATCACGGCTTGGGGAACAAAGCGA AAGTTGCTCCAACCACGAGGGGCGCAGTTGCGCTTCTTGTCCGCGGCA TTGGGCTGGGTCTCTTGGATGAAATCAGCAACCGCAGTGTTTATGGCGC AGTAGGTGGCGCTCTTCCTTCTCTGCCTGCTGCTGCAGCTGCTGCTGAG GGTACTTTAACCACTAGAGGCGCCAAGTTCATAATGTCTGCCACCAGGG GCACATCCACTGTGGAACAAAGCAACCAGCAGGGACTCGCCACTAGAG GGAGCGCCCAGAGTGGAAATGAGGCAACCAACTCCTGAATGTTGGACA GACATTTTCAAGCTTTTGAGGTAGGTTGAGAAACTCGCGATAATTCCCAG CTAGTCTTTACAGATGTTCTTCTAGGCTATCGCCACTAGGGGGCAGGGC AGGCTGCTTCAATTTCCTAGGACATTTTCCCCACTAGGGGCGCTGTAAG TTATGGCGAGTGGTGGGGTAATAGTAAAAGTCCCAACACATGCGAGGCA GCACACAGCCAATAACTTTTCCGGTCGCTAGGGGGAGATCTAGTGCCCC TTCTGGCCACTGCAGGTACTGCATGCATGTGATGCGGATACACATAGAA GGGGCTTCCAGCCATTGCTGTTCTTCCAAGTCCCTCCTGTGTGGGGCGC TGCCACGCCAGACCCGCCAGCAGGAGGCGTTCAGAACCTACACTCTAG GAAAGTAATCGCCTACATTTCCCAGCACGCCTTGCGCCTCGGCCACATT GGTGTTAAACTCCTCTAGTATATGTTAAAAGTGAGCTTTTGAATTCAGTCA GGAACTCCCAGTAGGGGGCGCTCTGTGCCCTAGTTAGGTGAAAGTGAG ACTCAAAGATTTGTTGGGCCACAGTTTCTACTTTGGCGCTTCAGTATAGG TGGCATCAGGCACTAGGAAAAACCAGTCAAAGTCATCGCCGTGTTGCTC CAGCAGGTGGCGCAGCGCCAGGTGCAGGTGTCCGATGGGCCTCTCTTC GCCCAGTGCCACCACCGCCATGCCTGATTTGATTGTCCTCACTTCCACC CTCTAGGCCTGAGCACCGACAGGCCGTAGAGCTAGCGAAGGCCAGGAG GGGGCGCACGAACAGTCGCCCGGAGGTCACAGAGCAGGGTGCGGCCA GAGTGGAAGGCCAAGGGGTGAAATCTTATTAATCTTAAAGACATATTTAA TTTTAATTGTGTGTGTGAGAATGTGCATGTGGAGGCCAGAAGAGGGCGC CAGATCCCCCTGAGCTTGGGTTCCAGATGGCTGTGAGCTGCCCAGTACA GGTGCTGAGGATTTGATAAACTTGAAGAGCAGGGAGTGTGATCTAAGGC GCCTCGTTTTAAAATACCTCAAAATTACCACAAGGTGGCGCTCTCGGGA ATTTGACGCTCCAGACAATCCTGCTTCCCTGAAGTCCCGCTTTGTAGAA GGGTGCAAGTCTACAAATAGTAATGATGAGTTAACCACGCCCACCGCTC CAGGACCTGATCAAGGCAACCACTAGGTGGCGGGCAAAACTAAGCAAT GCAGGGGCGGACCGACTTTTACTCGGCCTGGGTCACTTCTTTCAAAGCC CCCACCTTCGGCAAAAACAAAACAAAAAGAAACAACAAAACTCCACAGT GGCTCCCCCAAGACAGTAGGGGGCGCTGCACAGTCCCGCAGCCGCTC GGCTTAAATCCCTTGCAGTCCGGCCCAGGCGAAGGGGAAAA

Table 2.5: PCR primers for CBS cloning and genotyping.

Primer name	Sequence
HS5_pos_F	gTGAGCGGCCGCGCTTAAATCAGGCACAAGCTTAGG
HS5_pos_R	ctAGCCCTGCAGGACCAAGCCAATGTTCCTCTCTATG
HS5_neg_F	ctAGCCCTGCAGGGCTTAAATCAGGCACAAGCTTAGG
HS5_neg_R	gTGAGCGGCCGCACCAAGCCAATGTTCCTCTCTATG
LMO2_F_pos	gTGAGCGGCCGCAAGGCAACAACATAACCTAGCTATAA
LMO2_R_pos	ctAGCCCTGCAGGCCTACAGGAGAGACATCTGCACG
PAX3_cbs1_F_pos	gTGAGCGGCCGCttttcaggtgtgttggggtgtcc
PAX3-cbs1_R_pos	ctAGCCCTGCAGGGAATGTTCAGTCTGCATTCTGGAGC
PAX3_cbs2_F_pos	gTGAGCGGCCGCtggaactcacaaactgagatgcctc
PAX3_cbs2_R_pos	ctAGCCCTGCAGGAATTAGAACAGGCTCACCTCTCTGT
PAX3_cbs3_F_pos	gTGAGCGGCCGCCTCATCGTGCCTTCTGCTGTGA
PAX3_cbs3-R_pos	ctAGCCCTGCAGGTCCCAGTGCCAGTTCACTCTTTG
PAX3_cbs4_F_pos	gTGAGCGGCCGCATCCAGAATGGGAGCATATTGTAGG
PAX3_cbs4_R_pos	ctAGCCCTGCAGGGGTGTGTGTGTGAAGATTTCACAGT
SOX9-cbs1_F_pos	gTGAGCGGCCGCGCATTTTGGACTGTGAGTTATTGGC
SOX9_cbs1_R_pos	ctAGCCCTGCAGGAGAAGACATTGGAATCCGGTTACCA
SOX9_cbs1&2_F_pos	gTGAGCGGCCGCctctgtggttggcaattcagtctct
SOX9_cbs1&2_R_pos	ctAGCCCTGCAGGGGAGGTGAAGGGTCTCTGCTCT
SOX9_cbs3_F_pos	gTGAGCGGCCGCTGTTCCATGCTTGGCTGATCTTCT
SOX9_cbs3_R_pos	ctAGCCCTGCAGGaagtgggtggtactatttcctaggc
LMO2_F_neg	ctAGCCCTGCAGGAAGGCAACAACATAACCTAGCTATAA
LMO2_R_neg	gTGAGCGGCCGCCCTACAGGAGAGACATCTGCACG
PAX3_cbs1_F_neg	ctAGCCCTGCAGGttttcaggtgtgttggggtgtcc
PAX3-cbs1_R_neg	gTGAGCGGCCGCGAATGTTCAGTCTGCATTCTGGAGC
PAX3_cbs2_F_neg	ctAGCCCTGCAGGtggaactcacaaactgagatgcctc
PAX3_cbs2_R_neg	gTGAGCGGCCGCAATTAGAACAGGCTCACCTCTCTGT
PAX3_cbs3_F_neg	ctAGCCCTGCAGGCTCATCGTGCCTTCTGCTGTGA
PAX3-cbs3_R_neg	gTGAGCGGCCGCTCCCAGTGCCAGTTCACTCTTTG
PAX3_cbs4_F_neg	ctAGCCCTGCAGGATCCAGAATGGGAGCATATTGTAGG
PAX3_cbs4_R_neg	gTGAGCGGCCGCGGTGTGTGTGTGAAGATTTCACAGT
SOX9-cbs1_F_neg	ctAGCCCTGCAGGGCATTTTGGACTGTGAGTTATTGGC
SOX9_cbs1_R_neg	gTGAGCGGCCGCAGAAGACATTGGAATCCGGTTACCA
SOX9_cbs1&2_F_neg	ctAGCCCTGCAGGctctgtggttggcaattcagtctct
SOX9_cbs1&2_R_neg	gTGAGCGGCCGCGGAGGTGAAGGGTCTCTGCTCT
SOX9_cbs3_F_neg	ctAGCCCTGCAGGTGTTCCATGCTTGGCTGATCTTCT
SOX9_cbs3_R_neg	gTGAGCGGCCGCaagtgggtggtactatttcctaggc
SOX9_cbs1_F_pos	gTGAGCGGCCGCATAGCAAACACATCTGGGAAACAGCG
SOX9_cbs1_R_pos	ctAGCCCTGCAGGCGCTGTTTCCCAGATGTGTTTGCTAT
SOX9_cbs2_F_pos	gTGAGCGGCCGCGAAGTTCACATCTGTCTGCTGCC

Table 2.5: PCR primers for CBS cloning and genotyping, continued.

Primer name	Sequence
SOX9_cbs2_R_pos	ctAGCCCTGCAGGGGCAGCAGACAGATGTGAACTTC
SOX9_cbs1_F_neg	ctAGCCCTGCAGGATAGCAAACACATCTGGGAAACAGCG
SOX9_cbs1_R_neg	gTGAGCGGCCGCCGCTGTTTCCCAGATGTGTTTGCTAT
SOX9_cbs2_F_neg	ctAGCCCTGCAGGGAAGTTCACATCTGTCTGCTGCC
SOX9_cbs2_R_neg	gTGAGCGGCCGCGGCAGCAGACAGATGTGAACTTC
SOX9_cbs4_F_pos	gTGAGCGGCCGCCCAAAGTCTATGACATTTCAGTCAACCA
SOX9_cbs4_F_neg	ctAGCCCTGCAGGCCAAAGTCTATGACATTTCAGTCAACCA
SOX9_cbs4_pos_R	CTAGCCCTGCAGGTCTGAAATCTACCACAGAGATGGAACAC
SOX9_cbs4_neg_R	GTGAGCGGCCGCTCTGAAATCTACCACAGAGATGGAACAC
lmdDNA_4k_F	gTGAGCGGCCGCGACCATCACCGTGTATGAAG
lmdDNA_4k_R	ctAGCCCTGCAGGTCGCACTTGCTCAAATGCTG
GT_sox2_mcherry_F	CGTGGAACAGTACGAACGCG
GT_sox2_egfp_F	GTCCTGCTGGAGTTCGTGAC
GT_sox2_common_R	AGAACGCTCGGCGCGTCTACTT
GT_hytk_between_E-P_F	GGAGCTCACCGATTATGTGC
GT_hytk_between_E-P_R	GAACTTCGGATCCACTGAAAACA
GT_hytk_downstream_SE_F	GGATGGTCCAGACCCACGTC
GT_hytk_downstream_SE_R	AGATGCTCTGTCGGTCACTG
GT_insertion_betweenE-P_F	GGAGACAAGAGATGTCAGGAG
GT_insertion_betweenE-P_R	TCCGCAAGCAAATAGCTCCATTC
GT_insertion_downstream_F	CATCGGCAATGAGTGTGTGTCA
GT_insertion_downstream_R	GTGATCTCCAGAGTATACGCATGTC
sg_for_9-11_mut	AAGCAGTTCTGTTTCAGGAG.
GT_ΔZF9-11_F	GACTATTGCTTCCTGAGTGC
GT_ΔZF9-11_R	TTGGAACAGACAAAGGCAGC

Table 2.6: One-way ANOVA analysis of insulation effects by single CBSs.

Bonferroni's multiple comparisons test	Mean Diff.	95.00% CI of diff.	Significant?	Summary	Adjusted P Value
No insertion vs. Downstream controls	-0.05153	-0.09317 to -0.009884	Yes	**	0.0043
No insertion vs. HS5_CBS_reverse	-0.02942	-0.09945 to 0.04062	No	ns	>0.9999
No insertion vs. TAL1_CBS_reverse	0.03658	-0.03345 to 0.1066	No	ns	>0.9999
No insertion vs. LMO2_CBS_reverse	-0.01075	-0.08079 to 0.05929	No	ns	>0.9999
No insertion vs. Pax3_CBS1_reverse	0.1379	0.06788 to 0.2080	Yes	****	<0.0001
No insertion vs. Pax3_CBS2_reverse	0.04625	-0.02379 to 0.1163	No	ns	0.9071
No insertion vs. Pax3_CBS3_reverse	0.05325	-0.01679 to 0.1233	No	ns	0.4195
No insertion vs. Pax3_CBS4_reverse	0.04392	-0.02612 to 0.1140	No	ns	>0.9999
No insertion vs. Sox9_CBS3_reverse	0.06858	-0.001455 to 0.1386	No	ns	0.061
No insertion vs. Sox9_CBS4_reverse	-0.03942	-0.1095 to 0.03062	No	ns	>0.9999
No insertion vs. Sox9_CBS2_reverse	0.07492	0.004879 to 0.1450	Yes	*	0.0252
No insertion vs. Sox9_CBS1_reverse	-0.05575	-0.1258 to 0.01429	No	ns	0.3131
No insertion vs. HS5_CBS_forward	0.1099	0.03988 to 0.1800	Yes	****	<0.0001
No insertion vs. TAL1_CBS_forward	0.02825	-0.04179 to 0.09829	No	ns	>0.9999
No insertion vs. LMO2_CBS_forward	0.04492	-0.02512 to 0.1150	No	ns	>0.9999
No insertion vs. Pax3_CBS1_forward	-0.001417	-0.07145 to 0.06862	No	ns	>0.9999
No insertion vs. Pax3_CBS2_forward	0.06192	-0.008121 to 0.1320	No	ns	0.1466
No insertion vs. Pax3_CBS3_forward	0.1549	0.08488 to 0.2250	Yes	****	<0.0001
No insertion vs. Pax3_CBS4_forward	0.04325	-0.02679 to 0.1133	No	ns	>0.9999
No insertion vs. Sox9_CBS3_forward	0.1163	0.04621 to 0.1863	Yes	****	<0.0001
No insertion vs. Sox9_CBS4_forward	0.04425	-0.02579 to 0.1143	No	ns	>0.9999
No insertion vs. Sox9_CBS2_forward	0.1746	0.1045 to 0.2446	Yes	****	<0.0001
No insertion vs. Sox9_CBS1_forward	-0.02975	-0.1115 to 0.05204	No	ns	>0.9999

2.7 Methods

Cell culture

The hybrid F123 mESC line (F1 *Mus musculus castaneus* × S129/SvJae, maternal 129/Sv, paternal CAST) was from Dr. Rudolf Jaenisch's laboratory at the Whitehead Institute at MIT. The wild type F123 mESC line and engineered clones were maintained in feeder-free, serum-free 2i conditions (1 μM PD03259010, 3 μM CHIR99021, 2 mM glutamine, 0.15 μM Monothioglycerol, 1,000 U/ml LIF). The growth medium was changed every day. Cells were dissociated by Accutase (AT104) and passaged onto 0.2% gelatin-coated plates every 2-3 days.

Genetic engineering of the *Sox2* locus

Tagging of the *Sox2* gene with fluorescence reporter was performed by CRISPR-Cas9-mediated homologous recombination. Specifically, a guide RNA expression plasmid (pX330, addgene #42230) targeting the 3' of the *Sox2* gene, together with *egfp* and *mcherry* donor plasmids were co-electroporated into wild-type F123 cells by Neon transfection system (MPK1096). Cells were recovered for 2 days, then $eGFP^+$ $mCherry^+$ cells were sorted by FACS and seeded onto a new 0.2% gelatin-coated 60-mm dish. 5 days later, a second round of FACS was performed to enrich $eGFP^+$ $mCherry^+$ cells. 500-1,000 double-positive single cells were seeded onto a new 60-mm dish and single colonies were picked manually another 5 days later. Allele-specific genotyping of *Sox2* was performed with primers spanning CAST/129 SNPs. A clone with the CAST allele *Sox2* gene fused with *egfp* and 129 allele *Sox2* gene fused with *mcherry* was selected

as the parental clone. Subsequently, the *HyTK* fusion gene was integrated into the CAST allele of the parental clone by CRISPR-Cas9 editing. Specifically, electroporated cells were recovered for 2 days and then cultured in growth media containing 200 μg/ml hygromycin for 7 days. Survived cells were dissociated into single cells and seeded at the density of 500-1,000 cells per 60-mm dish. 5 days later, colonies were manually picked and genotyped with primers spanning CAST/129 SNPs. Genotyping primers were synthesized by IDT (**Table 2.5**).

Donor plasmids cloning for RMCE

The donor vector was adapted from the pUC19 plasmid. Two heterotypic Flippase recognition sites FRT/F3, as well as NotI and SbfI restriction enzyme recognition sites, were added into pUC19 plasmid by PCR. The donor vector was then digested with the enzyme cocktail of NotI-HF (neb, R3642S), SbfI-HF(neb, R3189S), and rSAP(neb, M0371S) for 4 h at 37 °C. Individual CTCF binding sites were PCR amplified from mouse or human genomic DNA. PCR primers contain overhang sequences of NotI and SbfI sites to specify CTCF motif orientation. PCR products were purified by gel-electrophoresis, digested, and ligated into the donor vector. Ligation products were transformed into Stbl3 chemically competent cells. Positive clones were screened by PCR and plasmids were extracted using QIAGEN plasmid plus midi kit (cat 12943) and validated by Sanger sequencing.

Marker-free insertion in mESCs by RMCE

A Flippase expression plasmid(pFlpe) (addgene #13787) and a donor plasmid(pDonor) were co-electroporated into 0.1 million insulator reporter or control cells at the ratio of 1:4 (pFlpe: pDonor = 1 μg :4 μg). Cells were recovered for two days and cultured in growth media containing 2 μM ganciclovir for another 5 days. Surviving cells were dissociated into single-cell suspension and seeded at the density of 500-1,000 cells per 60-mm dish. Five days later, six colonies were picked for PCR genotyping. Genomic DNA was then extracted by QIAGEN DNeasy Blood & Tissue Kits (#69506, #69581). For each insert, three independent clones were randomly picked for FACS analysis and subsequent studies. Individual CTCF binding sites were combined by PCR to create CBS clusters. Specifically, the 4CBS cluster from the *Sox9-Kcnj2* TAD boundary was consisted of genomic sequences from chr11:111,523,291-111,524,273, chr11:111,531,104-111,533,964, and chr11:111,535,307-111,538,959. PCR primers were synthesized by IDT (**Table 2.5**).

Deleting 9-11 zinc fingers of CTCF in mESCs

Deletion of CTCF zinc fingers 9-11 was achieved by CRISPR-Cas9-mediated homologous recombination as previously described[37]. Briefly, coding sequences of exon 10-12 of the *Ctcf* gene, together with an SV40 polyA signal were inserted into exon9 of the *Ctcf* gene *in situ*, resulting in only the 1-8 zinc fingers of the CTCF protein being functional. About 0.15 million cells were transfected with a mixture of guide RNA expressing plasmid (Px330, 1 μg), homologous recombination repair plasmid (4 μg), and a co-electroporation marker (0.1 μg, puromycin resistant). After two days' recovery, cells were treated with 1 μg/ml puromycin for another three days. Surviving cells were

suspended into single cells and seed at the density of 500-1,500 cells per 10-cm Petri dish. Five days later, single colonies were manually picked and genotyped by PCR.

FACS data acquisition and analysis

Cells were treated by Accutase (#AT104) at 37°C for 5-7 min and resuspended into single cells with 2 ml warm 2i/LIF medium. Cells were then spun down at 1,000 rpm for 4 min and washed twice with 5 ml PBS. Cell pellets were resuspended into single cells with 1 ml PBS and filtered through the 35-μm strainer cap of a FACS tube (SKU: FSC-9005). Then, cells were sorted by Sony sorter SH800 (Cell Sorter Software 2.1.5) in analysis mode using a 130-μm chip. For each insertion clone, both GFP and mCherry signals were recorded for 10,000 cells. Cells were first gated by SSCA-FSCA for live cells, then by FSA-FSH for singlets using FlowJo 10.0.7r2. Fluorescence signals of cells passed gating were exported in csv files and analyzed in R 3.6.0. Specifically, the GFP signal is normalized by mCherry signal from the same cell. For each insertion clone, the normalized Sox2-eGFP expression was calculated as:

$$\mathbf{Mean}\left(\frac{eGFP}{mCherry}\right)_{\text{Insertion}} / \mathbf{Mean}\left(\frac{eGFP}{mCherry}\right)_{\text{no insertion}}$$

To better estimate instrument variability in FACS sorting, we used replicates of the no insertion clone in all experiments as controls when testing the significance of insulation effects of the inserted DNA elements.

ChIP-seq

Cells were dissociated into single cells and cross-linked by 1% formaldehyde in PBS for 15 min at room temperature. Cross-linking was then quenched by 0.125 M

glycine and cells were washed twice with 5 ml cold PBS. Permeabilized nuclei were prepared with Covaris truChIP Chromatin Shearing Kit (PN520154) following the manufacturer's instructions. 1-3 million nuclei were sonicated in 130 µl microtube by Covaris M220 instrument (Power, 75W; Duty factor, 10%; Cycle per bust, 200; Time, 10 min; Temperature, 7°C.). Sonicated chromatin was diluted with 1× Shearing Buffer into a total volume of 1 ml and spun down at 15,000 rpm at 4°C to remove cell debris. 5 µg antibodies were added to the supernatant and incubated overnight at 4°C with gentle rotation (CTCF, ab70303, lot GR3281212-6,7,8; RAD21, ab992, lot GR3253930-8, GR3310168-11; H3K4me3, Millipore, 04-745, lot 3243412; H3K27ac, Active Motif, 39685, lot 33417016.). Chromatin was pulled down by protein G Sepharose beads (GE, 17061801) and washed three times with RIPA buffer(10 mM Tris pH 8.0, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% Sodium Deoxycholate), twice with high-salt RIPA buffer (10 mM Tris pH 8.0, 300 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% SDS, 0.1% Sodium Deoxycholate), once with LiCl buffer (10 mM Tris pH 8.0, 250 mM LiCl, 1 mM EDTA, 0.5% IGEPAL CA-630, 0.1% Sodium Deoxycholate), and twice with TE buffer (10 mM Tris, pH 8.0; 0.1 mM EDTA). Washed chromatin was reverse crosslinked overnight with 2 µl proteinase K (P8107S, NEB) at 65 °C (1%SDS, 10 mM Tris, pH 8.0, 0.1 mM EDTA), column purified and subjected to end repair, A-tailing, adapter ligation, and PCR amplification. Final libraries were purified by SPRI beads (0.8:1) and quantified with Qubit HS dsDNA kit (Q32854).

RNA-seq

Total RNA from cells was extracted using the TRIzol Plus RNA purification kit (Thermo Fisher Scientific, Catalog: 12183555). RNA-seq libraries were prepared from 4 μg total RNA using the Illumina TruSeq Stranded mRNA Library Prep Kit Set A (RS-122-2101; Illumina) or Set B (RS-122-2102; Illumina). RNA-seq libraries were sequenced on illumine Next-seq 550 and Hi-seq4000 platforms (75-bp paired ends).

PLAC-seq/HiChIP

Proximity Ligation ChIP-sequencing (PLAC-seq) (also known as HiChIP) libraries were prepared as previously described[57, 58] with minor modifications. In brief, 2-3 million cells were crosslinked for 15 minutes at room temperature with 1% methanol-free formaldehyde and quenched for 5 minutes at room temperature with 0.2 M glycine. The crosslinked cells were lysed in 300 μl Hi-C lysis buffer (10 mM Tris-HCl, pH 8.0, 10 mM NaCl, 0.2% IPEGAL CA-630) for 15 minutes on ice and then washed once with 500 μl lysis buffer (2,500×g for 5 minutes). Subsequently, cells were resuspended in 50 μl 0.5% SDS and incubated for 10 min at 62°C then quenched by 160 μl 1.56% Triton X-100 for 15 min at 37°C. Then, 25 μl of 10× NEBuffer 2 and 100 U MboI were added to digest chromatin for 2 hours at 37°C with shaking (1,000 rpm). Digested fragments were biotin-labeled and subsequently ligated by T4 DNA ligase buffer (NEB) for 2 hours at 23°C with 300 rpm gentle rotation. Chromatin was sheared and washed as described in ChIP-seq. Dynabeads (M-280 Sheep anti-Rabbit IgG, catalog: 11203D) coated with 5 μg H3K4me3 antibodies (Millipore, 04-745, lot 3243412) were used for immunoprecipitation. Pulled down chromatin was treated with 10 μg RNase A for 1 hour at 37°C, reverse-crosslinked by 20 μg proteinase K at 65°C for 2 hours, then purified

with Zymo DNA Clean & Concentrator-5 kit. Ligation junctions were enriched by 25 µl myOne T1 Streptavidin Dynabeads. Libraries were prepared using QIAseq Ultralow Input Library Kit (Qiagen, #180492). Final libraries were size selected with SPRI beads (0.5:1 and 1:1), quantified, and submitted for paired-end sequencing.

Hi-C

Cells were processed in the same way as in PLAC-seq before chromatin shearing steps. Briefly, nuclei after the ligation step were digested by 50 µl of proteinase K (20 mg/ml) for 30 min at 55 °C. DNA was then purified by ethanol precipitation and resuspended in 130 µl 10 mM Tris-HCl (pH 8.0). Purified DNA was sonicated by Covaris M220 instrument with the following parameters: Duty cycle, 10%; Power, 50; Cycles/burst, 200; Time, 70 seconds. DNA fragments smaller than 300 bp were removed by Ampure XP bead-based dual size selection (0.55:1 and 0.75:1). Biotin-labeled free DNA ends were cleaned up by end-repair reaction and ligation junctions were enriched by Streptavidin Dynabeads as described in PLAC-seq. Ligation junctions were then purified and subjected to A-tailing, adapter ligation, and PCR amplification. Final libraries were purified by 0.75× Ampure XP beads, quantified, and submitted for pair-end sequencing.

Multiplexed FISH imaging for chromatin tracing

Glass coverslips were treated by poly-L-lycine for 30 min at 37°C. Then, glass coverslips were washed twice with 5ml PBS and treated with 0.2% gelatin for another 20 min at 37°C. 2.5 million mESCs were seeded in a 6-cm plastic dish containing the

treated glass coverslip. After 20 hours, cells were cross-linked by 4% paraformaldehyde and followed by chromatin tracing experiments as described in a previous publication[60]. Briefly, the entire 210-kb *Sox2* region was labeled by a library of primary Oligopaint probes[60, 61]. Each primary probe consists of a unique 42-nucleotide readout sequence that is specific for each 5 kb DNA segment. Next, secondary readout probes complementary to the readout sequences on the primary probes were added to the cells. Lastly, fluorophore-labeled common imaging probes complementary to the secondary probes were added to the cells to allow three-dimensional diffraction-limited imaging of individual DNA segments. After each round of imaging, the fluorescence signal was extinguished by using both TCEP [tris(2-carboxyethyl) phosphine] cleavage at a concentration of 50 μM in 2× SSC and high power photobleaching. The process was repeated until all DNA segments were labeled and imaged. We performed three-color imaging by using three secondary readout imaging probes that were conjugated with Cy3, Cy5, and Alexa 750, respectively. In this case, three consecutive 5-kb chromatin segments were labeled by each round of imaging. A pool of 42 oligonucleotide probe sets was designed to scan the 210-kb *Sox2* locus with each set covering a 5-kb DNA region.

Multiplexed RNA and DNA FISH imaging at the *Sox2* locus

The dual-modality FISH imaging was performed as recently described[63]. Briefly, the sample was prepared as in the “Multiplexed FISH imaging for chromatin tracing” except that after cells were cross-linked by 4% paraformaldehyde, the sample was hybridized with oligonucleotide probes (Supplementary Table 2.3) targeting the *Sox2*,

egfp, and *mcherry* transcripts followed by imaging (final concentration 100 nM). Immediately after the RNA FISH imaging was completed, the sample was washed with 50% formamide to remove residual fluorescence readout probes and crosslinked again with 4% paraformaldehyde before multiplexed FISH imaging for chromatin tracing.

2.8 Data Analysis

ChIP-seq

Sequenced reads were aligned to reference mouse genome mm10 using bowtie2 (version 2.2.9). Unmapped reads and PCR duplicates were removed. For clones with the insertion of synthetic CTCF binding sites, reads were aligned to a customized mm10 reference genome that includes the inserted sequence. Mapping pipeline is available at http://renlab.sdsc.edu/huh025/chipseq-PE/. Signal tracks were generated with the command “bamCoverage (version 3.3.1) –normlizingRPKM -bs 50 --smoothLength 150“. Peaks were called by macs2 (version 2.1.1.20160309) with default parameters.

RNA-seq

The RNA-seq alignment and quantification pipeline is available at https://github.com/ren-lab/rnaseq-pipeline. Briefly, reads were aligned to mm10 (GRCm38) and GENCODE GTF version M25 with rnaSTAR[72] (version 020201). Particularly, we created two extra chromosomes for the two tagged *Sox2* alleles. PCR duplicates were removed using Picard. Reads uniquely mapped to *egfp* and *mcherry*

sequences were counted using samtools. *Sox2* expression from the CAST and 129 allele was quantified by RPKM values of the *egfp* and *mcherry* gene, respectively.

PLAC-seq

To resolve allele-specific interactions, we created the VCF files containing SNPs with respect to the mm10 reference genome for parental strain CAST/EiJ and 129SV/Jae. Specifically, whole-genome sequencing reads from the two strains were mapped to mm10, deduplicated, and called SNPs using bcftools. We removed heterozygous SNP calls and those with sequencing depth less than 5 and quality less than 30 and further removed SNPs that were present in both strains. We used a modified mapping procedure from WASP[73] pipeline (version 0.3.4) to detect allele-specific contacts. Since WASP pipeline ignores indels, we further removed all reads which map to within 50 base pairs from the nearest indel. We modified the original WAPS mapping procedure by replacing the bowtie2 alignment tool with bwa-mem and integrated MAPS[74] feather post-filtering pipeline to resolve the chimeric reads. Analysis pipeline is available at https://github.com/ijuric/Sox2AllelicAnalysis.

Hi-C

To process Hi-C data we used our in-house pipeline available at https://github.com/ren-lab/hic-pipeline. Briefly, Hi-C reads were aligned to mm10 using BWA-MEM (version 0.7.12-r1039) for each read separately and then paired. For chimeric reads, only 5' end-mapped locations were kept. Duplicated read pairs mapped to the same location were removed to leave only one unique read pair. The output bam

files were transformed into juicer file format for visualization in Juicebox 1.11.08. Contact matrices were normalized using the Knight–Ruiz matrix balancing method[75]. Directionality Index (DI) score for each sample was generated at 50-kb resolution and 2-Mb window (40 bins) as described in a previous work[24]. Haplotype phasing was performed using the obtained CAST/129 VCF file. This created two contact matrices corresponding to 'Cast allele' and '129 allele' for each Hi-C library. For each phased haplotype of chromosome 3, the DI score was generated at 10-kb resolution and 50-kb window (5 bins).

Chromatin tracing data processing

Custom software was used to obtain images of chromatin architecture as described previously[60] with minor modifications. The software identifies centroid positions of each 5-kb chromatin segment using diffraction-limited z-stack images acquired by epifluorescence microscopy. Chromosome locations were first identified via the segmentation of the nuclei in each field of view using a convolutional neural network (CNN). The segmentation masks were then applied to limit the chromosome candidates to the two most likely clusters of fluorescence spots presented in each nucleus. We then selected the two spots that showed the strongest averaged fluorescence signal over all imaging rounds as the two alleles for each nucleus. To avoid selecting the same chromosome, we also required the two spots to be separated by at least 10 pixels (1.08 μm). The algorithm then utilized the identified chromosome locations to select candidate spots of the imaged 5-kb chromatin segments in every round of imaging. A Gaussian fitting algorithm was then used to fit both the signal of each of the candidate segments

and the fiducial beads. The chromatic aberration, flat-field, and drift correction algorithms were adopted from the published work[60].

The candidate spot of each segment was then further evaluated for their likelihood to be accepted or rejected as estimated by an expectation-maximization (EM) algorithm. The EM algorithm computes a score based upon a product of three terms, brightness of the spot, the proximity of the spot to the estimated chromosome centroid position, and the proximity of the spot to a moving average localization of the candidates selected in the previous five rounds of imaging, of each candidate spot of a segment. The EM algorithm selected the highest scoring candidate spot for each chromosome segment in each round of imaging, while all remaining candidate spots were not considered in subsequent analyses.

The misidentification rate was computed as the percentage of fluorescence spots among the top discarded candidate spots which had scores above the EM score threshold that we chose. Finally, only chromosomes that contained accepted segments with a score above the selected threshold across at least ~50% of imaging rounds (22/42 rounds) were kept for further analysis. The detection efficiency of each segment for each experiment was computed as the fraction of segments with accepted candidate spots based upon the above procedure. We only kept cells in which one and only one chromosome was detected positive for the insertion. In addition, we required the signal of the insertion to be greater than 1/2 of the median value of all segments and at least two times stronger than the signal from the other allele. In this way, the misclassification

of the two alleles is estimated to be less than 5%. Insulation score was calculated for each chromosome as the natural log of the ratio of median distance between loci across domains and median distance between loci within domains. *Sox2* enhancer-promoter distance was calculated by median pairwise Euclidean distances between the genomic locations of the *Sox2* gene (9^{th} - 11^{th} region) and its enhancer (30^{th} - 32^{nd} region) for every chromosome.

2.9 Data and Code Availability

All next-generation sequencing data are available under GEO accession GSE153403. Raw images of multiplexed FISH experiments and raw FACS data of specific mESC colonies are available upon request.

Multiplexed FISH data and code for analyses can be found on Github at https://github.com/epigen-UCSD/huang-natgen2021.

2.10 Author Contributions

B.R. conceived the study. H.H. and B.R. supervised the study. H.H. performed insulator assays and related analysis. R.H. and M.Y. performed PLAC-seq/HiChIP and Hi-C experiments. I.J. and M.H. analyzed PLAC-seq data. M.T. performed western blot experiments. Y.Z. performed Hi-C analysis. Q.Z. and Y.H. performed chromatin tracing experiments with help from B.B. and X.Z.. A.P.J., B.B., C.K., M.C., S.B., A.M.C. and M.N. analyzed chromatin tracing data. The manuscript was written by H.H. and B.R. with input from all co-authors.

2.11 Competing Interests

Bing Ren is a co-founder and consultant for Arima Genomics, Inc. and co-founder of Epigenome Technologies. Xiaowei Zhuang is a co-founder and consultant for Vizgen, Inc. The remaining authors declare no competing interests.

2.12 Funding Resources

We are grateful for comments from members of the Ren laboratory. This study was supported by funding from the Ludwig Institute for Cancer Research and NIH (U54 DK107977, to B.R., M.H. and M.N., and 3U54DK107977-05S1 to B.R.). X.Z. is a Howard Hughes Medical Institute Investigator.

2.13 Acknowledgements

Chapter 2, in full, is a reprint of the accepted manuscript in *Nature Genetics* 2021. Huang, H., Zhu, Q., Jussila, A. P., Han, Y., Bintu, B., Kern, C., Conte, M., Zhang, Y., Bianco, S., Chiariello, A.M., Yu, M., Hu, R., Tastemel, M., Juric, I., Hu, M., Necodemi, M., Zhuang, X., and Ren, B. CTCF mediates dosage- and sequence-context-dependent transcriptional insulation by forming local chromatin domains. *Nature Genetics* (in press, 2021). The dissertation author was the primary investigator and author of this paper.

2.14 References

1. Hnisz, D., Day, D.S. & Young, R.A. Insulated Neighborhoods: Structural and Functional Units of Mammalian Gene Control. *Cell* **167**, 1188-1200 (2016).

2. Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E., Dekker, J., Dunham, I., Elnitski, L.L., Farnham, P.J., Feingold, E.A., Gerstein, M., Giddings, M.C., Gilbert, D.M., Gingeras, T.R., Green, E.D., Guigo, R., Hubbard, T., Kent, J., Lieb, J.D., Myers, R.M., Pazin, M.J., Ren, B., Stamatoyannopoulos, J.A., Weng, Z., White, K.P. & Hardison, R.C. Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* **111**, 6131-6138 (2014).

3. Levine, M., Cattoglio, C. & Tjian, R. Looping back to leap forward: transcription enters a new era. *Cell* **157**, 13-25 (2014).

4. West, A.G., Gaszner, M. & Felsenfeld, G. Insulators: many functions, many mechanisms. *Genes Dev* **16**, 271-288 (2002).

5. Geyer, P.K. & Corces, V.G. DNA position-specific repression of transcription by a Drosophila zinc finger protein. *Genes Dev* **6**, 1865-1873 (1992).

6. Recillas-Targa, F., Bell, A.C. & Felsenfeld, G. Positional enhancer-blocking activity of the chicken beta-globin insulator in transiently transfected cells. *Proc Natl Acad Sci U S A* **96**, 14354-14359 (1999).

7. Stief, A., Winter, D.M., Stratling, W.H. & Sippel, A.E. A nuclear DNA attachment element mediates elevated and position-independent gene activity. *Nature* **341**, 343-345 (1989).

8. Gurudatta, B.V. & Corces, V.G. Chromatin insulators: lessons from the fly. *Brief Funct Genomic Proteomic* **8**, 276-282 (2009).

9. Chung, J.H., Bell, A.C. & Felsenfeld, G. Characterization of the chicken beta-globin insulator. *Proc Natl Acad Sci U S A* **94**, 575-580 (1997).

10. Lobanenkov, V.V., Nicolas, R.H., Adler, V.V., Paterson, H., Klenova, E.M., Polotskaja, A.V. & Goodwin, G.H. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene* **5**, 1743-1753 (1990).

11. Bell, A.C. & Felsenfeld, G. Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature* **405**, 482-485 (2000).

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

13. Katainen, R., Dave, K., Pitkanen, E., Palin, K., Kivioja, T., Valimaki, N., Gylfe, A.E., Ristolainen, H., Hanninen, U.A., Cajuso, T., Kondelin, J., Tanskanen, T., Mecklin, J.P., Jarvinen, H., Renkonen-Sinisalo, L., Lepisto, A., Kaasinen, E., Kilpivaara, O., Tuupanen, S., Enge, M., Taipale, J. & Aaltonen, L.A. CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet* **47**, 818-821 (2015).

14. Ohlsson, R., Renkawitz, R. & Lobanenkov, V. CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet* **17**, 520-527 (2001).

15. Filippova, G.N., Fagerlie, S., Klenova, E.M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P.E., Collins, S.J. & Lobanenkov, V.V. An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol Cell Biol* **16**, 2802-2813 (1996).

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

17. Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R. & Oberdoerffer, S. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74-79 (2011).

18. Vostrov, A.A. & Quitschke, W.W. The zinc finger protein CTCF binds to the APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation. *J Biol Chem* **272**, 33353-33359 (1997).

19. Zhang, X., Zhang, Y., Ba, Z., Kyritsis, N., Casellas, R. & Alt, F.W. Fundamental roles of chromatin loop extrusion in antibody class switching. *Nature* **575**, 385-389 (2019).

20. Guo, Y., Monahan, K., Wu, H., Gertz, J., Varley, K.E., Li, W., Myers, R.M., Maniatis, T. & Wu, Q. CTCF/cohesin-mediated DNA looping is required for protocadherin alpha promoter choice. *Proc Natl Acad Sci U S A* **109**, 21081-21086 (2012).

21. Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., Lu, Y., Wu, Y., Jia, Z., Li, W., Zhang, M.Q., Ren, B., Krainer, A.R., Maniatis, T. & Wu, Q. CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. *Cell* **162**, 900-910 (2015).

22. Ghirlando, R. & Felsenfeld, G. CTCF: making the right connections. *Genes Dev* **30**, 881-891 (2016).

23. Phillips-Cremins, J.E. & Corces, V.G. Chromatin insulators: linking genome organization to cellular function. *Mol Cell* **50**, 461-474 (2013).

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

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

26. Franke, M., Ibrahim, D.M., Andrey, G., Schwarzer, W., Heinrich, V., Schopflin, R., Kraft, K., Kempfer, R., Jerkovic, I., Chan, W.L., Spielmann, M., Timmermann, B., Wittler, L., Kurth, I., Cambiaso, P., Zuffardi, O., Houge, G., Lambie, L., Brancati, F., Pombo, A., Vingron, M., Spitz, F. & Mundlos, S. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265-269 (2016).

27. Nora, E.P., Goloborodko, A., Valton, A.L., Gibcus, J.H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L.A. & Bruneau, B.G. Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944 e922 (2017).

28. Luppino, J.M., Park, D.S., Nguyen, S.C., Lan, Y., Xu, Z., Yunker, R. & Joyce, E.F. Cohesin promotes stochastic domain intermingling to ensure proper regulation of boundary-proximal genes. *Nat Genet* (2020).

29. Wutz, G., Varnai, C., Nagasaka, K., Cisneros, D.A., Stocsits, R.R., Tang, W., Schoenfelder, S., Jessberger, G., Muhar, M., Hossain, M.J., Walther, N., Koch, B., Kueblbeck, M., Ellenberg, J., Zuber, J., Fraser, P. & Peters, J.M. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J* **36**, 3573-3599 (2017).

30. Alipour, E. & Marko, J.F. Self-organization of domain structures by DNA-loop-extruding enzymes. *Nucleic Acids Res* **40**, 11202-11212 (2012).

31. Davidson, I.F., Bauer, B., Goetz, D., Tang, W., Wutz, G. & Peters, J.M. DNA loop extrusion by human cohesin. *Science* **366**, 1338-1345 (2019).

32. Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N. & Mirny, L.A. Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep* **15**, 2038-2049 (2016).

33. Haarhuis, J.H.I., van der Weide, R.H., Blomen, V.A., Yanez-Cuna, J.O., Amendola, M., van Ruiten, M.S., Krijger, P.H.L., Teunissen, H., Medema, R.H., van Steensel, B., Brummelkamp, T.R., de Wit, E. & Rowland, B.D. The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell* **169**, 693-707 e614 (2017).

34. Kim, Y., Shi, Z., Zhang, H., Finkelstein, I.J. & Yu, H. Human cohesin compacts DNA by loop extrusion. *Science* **366**, 1345-1349 (2019).

35. Rao, S.S.P., Huang, S.C., Glenn St Hilaire, B., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.R., Sanborn, A.L., Johnstone, S.E., Bascom, G.D., Bochkov, I.D., Huang, X., Shamim, M.S., Shin, J., Turner, D., Ye, Z., Omer, A.D., Robinson, J.T., Schlick, T., Bernstein, B.E., Casellas, R., Lander, E.S. & Aiden, E.L. Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320 e324 (2017).

36. Sanborn, A.L., Rao, S.S., Huang, S.C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K.P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E.K., Lander, E.S. & Aiden, E.L. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A* **112**, E6456-6465 (2015).

37. Vian, L., Pekowska, A., Rao, S.S.P., Kieffer-Kwon, K.R., Jung, S., Baranello, L., Huang, S.C., El Khattabi, L., Dose, M., Pruett, N., Sanborn, A.L., Canela, A., Maman, Y., Oksanen, A., Resch, W., Li, X., Lee, B., Kovalchuk, A.L., Tang, Z., Nelson, S., Di Pierro, M., Cheng, R.R., Machol, I., St Hilaire, B.G., Durand, N.C., Shamim, M.S., Stamenova, E.K., Onuchic, J.N., Ruan, Y., Nussenzweig, A., Levens, D., Aiden, E.L. & Casellas, R. The Energetics and Physiological Impact of Cohesin Extrusion. *Cell* **173**, 1165-1178 e1120 (2018).

38. Wutz, G., Ladurner, R., St Hilaire, B.G., Stocsits, R.R., Nagasaka, K., Pignard, B., Sanborn, A., Tang, W., Varnai, C., Ivanov, M.P., Schoenfelder, S., van der Lelij, P., Huang, X., Durnberger, G., Roitinger, E., Mechtler, K., Davidson, I.F., Fraser, P., Lieberman-Aiden, E. & Peters, J.M. ESCO1 and CTCF enable formation of long chromatin loops by protecting cohesin(STAG1) from WAPL. *Elife* **9** (2020).

39. Brackley, C.A., Johnson, J., Michieletto, D., Morozov, A.N., Nicodemi, M., Cook, P.R. & Marenduzzo, D. Nonequilibrium Chromosome Looping via Molecular Slip Links. *Phys Rev Lett* **119**, 138101 (2017).

40. Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.M., Dostie, J., Pombo, A. & Nicodemi, M. Complexity of chromatin folding is captured by the strings and binders switch model. *Proc Natl Acad Sci U S A* **109**, 16173-16178 (2012).

41. Bianco, S., Lupianez, D.G., Chiariello, A.M., Annunziatella, C., Kraft, K., Schopflin, R., Wittler, L., Andrey, G., Vingron, M., Pombo, A., Mundlos, S. & Nicodemi, M. Polymer physics predicts the effects of structural variants on chromatin architecture. *Nat Genet* **50**, 662-667 (2018).

42. Brackley, C.A., Taylor, S., Papantonis, A., Cook, P.R. & Marenduzzo, D. Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization. *Proc Natl Acad Sci U S A* **110**, E3605-3611 (2013).

43. Buckle, A., Brackley, C.A., Boyle, S., Marenduzzo, D. & Gilbert, N. Polymer Simulations of Heteromorphic Chromatin Predict the 3D Folding of Complex Genomic Loci. *Mol Cell* **72**, 786-797 e711 (2018).

44. Conte, M., Fiorillo, L., Bianco, S., Chiariello, A.M., Esposito, A. & Nicodemi, M. Polymer physics indicates chromatin folding variability across single-cells results from state degeneracy in phase separation. *Nat Commun* **11**, 3289 (2020).

45. Di Pierro, M., Zhang, B., Aiden, E.L., Wolynes, P.G. & Onuchic, J.N. Transferable model for chromosome architecture. *Proc Natl Acad Sci U S A* **113**, 12168-12173 (2016).

46. Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., Haering, C.H., Mirny, L. & Spitz, F. Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51-56 (2017).

47. Despang, A., Schopflin, R., Franke, M., Ali, S., Jerkovic, I., Paliou, C., Chan, W.L., Timmermann, B., Wittler, L., Vingron, M., Mundlos, S. & Ibrahim, D.M. Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat Genet* **51**, 1263-1271 (2019).

48. Gribnau, J., Hochedlinger, K., Hata, K., Li, E. & Jaenisch, R. Asynchronous replication timing of imprinted loci is independent of DNA methylation, but consistent with differential subnuclear localization. *Genes Dev* **17**, 759-773 (2003).

49. Li, Y., Rivera, C.M., Ishii, H., Jin, F., Selvaraj, S., Lee, A.Y., Dixon, J.R. & Ren, B. CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS One* **9**, e114485 (2014).

50. Zhou, H.Y., Katsman, Y., Dhaliwal, N.K., Davidson, S., Macpherson, N.N., Sakthidevi, M., Collura, F. & Mitchell, J.A. A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes Dev* **28**, 2699-2711 (2014).

51. Kentepozidou, E., Aitken, S.J., Feig, C., Stefflova, K., Ibarra-Soria, X., Odom, D.T., Roller, M. & Flicek, P. Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biol* **21**, 5 (2020).

52. Frith, M.C., Saunders, N.F., Kobe, B. & Bailey, T.L. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput Biol* **4**, e1000071 (2008).

53. Nakahashi, H., Kieffer Kwon, K.R., Resch, W., Vian, L., Dose, M., Stavreva, D., Hakim, O., Pruett, N., Nelson, S., Yamane, A., Qian, J., Dubois, W., Welsh, S., Phair, R.D., Pugh, B.F., Lobanenkov, V., Hager, G.L. & Casellas, R. A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep* **3**, 1678-1689 (2013).

54. Xu, D., Ma, R., Zhang, J., Liu, Z., Wu, B., Peng, J., Zhai, Y., Gong, Q., Shi, Y., Wu, J., Wu, Q., Zhang, Z. & Ruan, K. Dynamic Nature of CTCF Tandem 11 Zinc Fingers in Multivalent Recognition of DNA As Revealed by NMR Spectroscopy. *J Phys Chem Lett* **9**, 4020-4028 (2018).

55. Yin, M., Wang, J., Wang, M., Li, X., Zhang, M., Wu, Q. & Wang, Y. Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites. *Cell Res* **27**, 1365-1377 (2017).

56. Yan, J., Chen, S.A., Local, A., Liu, T., Qiu, Y., Dorighi, K.M., Preissl, S., Rivera, C.M., Wang, C., Ye, Z., Ge, K., Hu, M., Wysocka, J. & Ren, B. Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. *Cell Res* **28**, 387 (2018).

57. Fang, R., Yu, M., Li, G., Chee, S., Liu, T., Schmitt, A.D. & Ren, B. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res* **26**, 1345-1348 (2016).

58. Mumbach, M.R., Rubin, A.J., Flynn, R.A., Dai, C., Khavari, P.A., Greenleaf, W.J. & Chang, H.Y. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13**, 919-922 (2016).

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

60. Bintu, B., Mateo, L.J., Su, J.H., Sinnott-Armstrong, N.A., Parker, M., Kinrot, S., Yamaya, K., Boettiger, A.N. & Zhuang, X. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362** (2018).

61. Mateo, L.J., Murphy, S.E., Hafner, A., Cinquini, I.S., Walker, C.A. & Boettiger, A.N. Visualizing DNA folding and RNA in embryos at single-cell resolution. *Nature* **568**, 49-54 (2019).

62. Wang, S., Su, J.H., Beliveau, B.J., Bintu, B., Moffitt, J.R., Wu, C.T. & Zhuang, X. Spatial organization of chromatin domains and compartments in single chromosomes. *Science* **353**, 598-602 (2016).

63. Su, J.H., Zheng, P., Kinrot, S.S., Bintu, B. & Zhuang, X. Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. *Cell* **182**, 1641-1659 e1626 (2020).

64. Alexander, J.M., Guan, J., Li, B., Maliskova, L., Song, M., Shen, Y., Huang, B., Lomvardas, S. & Weiner, O.D. Live-cell imaging reveals enhancer-dependent Sox2 transcription in the absence of enhancer proximity. *Elife* **8** (2019).

65. Jia, Z., Li, J., Ge, X., Wu, Y., Guo, Y. & Wu, Q. Tandem CTCF sites function as insulators to balance spatial chromatin contacts and topological enhancer-promoter selection. *Genome Biol* **21**, 75 (2020).

66. Cai, H.N. & Shen, P. Effects of cis arrangement of chromatin insulators on enhancer-blocking activity. *Science* **291**, 493-495 (2001).

67. Muravyova, E., Golovnin, A., Gracheva, E., Parshikov, A., Belenkaya, T., Pirrotta, V. & Georgiev, P. Loss of insulator activity by paired Su(Hw) chromatin insulators. *Science* **291**, 495-498 (2001).

68. Rhee, H.S. & Pugh, B.F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408-1419 (2011).

69. Benabdallah, N.S., Williamson, I., Illingworth, R.S., Kane, L., Boyle, S., Sengupta, D., Grimes, G.R., Therizols, P. & Bickmore, W.A. Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. *Mol Cell* **76**, 473-484 e477 (2019).

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

71. Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J.B. & Gregor, T. Dynamic interplay between enhancer-promoter topology and gene activity. *Nat Genet* **50**, 1296-1303 (2018).

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

73. van de Geijn, B., McVicker, G., Gilad, Y. & Pritchard, J.K. WASP: allele-specific software for robust molecular quantitative trait locus discovery. *Nat Methods* **12**, 1061-1063 (2015).

74. Juric, I., Yu, M., Abnousi, A., Raviram, R., Fang, R., Zhao, Y., Zhang, Y., Qiu, Y., Yang, Y., Li, Y., Ren, B. & Hu, M. MAPS: Model-based analysis of long-range chromatin interactions from PLAC-seq and HiChIP experiments. *PLoS Comput Biol* **15**, e1006982 (2019).

75. Durand, N.C., Robinson, J.T., Shamim, M.S., Machol, I., Mesirov, J.P., Lander, E.S. & Aiden, E.L. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99-101 (2016).

CHAPTER 3. Identification of H3K4me1-associated proteins at mammalian enhancers

3.1 Abstract

Enhancers act to regulate cell type specific gene expression by facilitating the transcription of target genes. In mammalian cells active or primed enhancers are commonly marked by monomethylation of Histone H3 at lysine 4 (H3K4me1) in a cell-type specific manner. Whether and how this histone modification regulates enhancer-dependent transcription programs in mammals is unclear. In this study, we conducted SILAC Mass-spec experiments with mono-nucleosomes and identified multiple H3K4me1 associated proteins, including many involved in chromatin remodeling. We demonstrate that H3K4me1 augments the association of the chromatin remodeling complex BAF to enhancers *in vivo* and that *in vitro*, H3K4me1 nucleosomes are more efficiently remodeled by the BAF complex. Crystal structures of BAF component BAF45c reveal that monomethylation, but not trimethylation, is accommodated by BAF45c's H3K4 binding site. Our results suggest that H3K4me1 plays an active role at enhancers by facilitating the binding of the BAF complex and possibly other chromatin regulators.

3.2 Introduction

In cells, cis-regulatory elements such as enhancers and promoters can be defined not only by DNA sequence motifs but also by common and predictive patterns of epigenetic modifications[1]. Active promoters are enriched for H3K4me3, H3/H4 acetylation along with binding of multiple chromatin regulatory complexes [2]. Primed enhancers are marked by H3K4me1 (coupled with a depletion of H3K4me3) whereas active enhancers are enriched for H3K4me1, H3K27ac and sometimes H4K16ac and H3K122ac [2-8]. Such epigenetic signatures are commonly used to predict de novo regulatory elements in novel cell types. Numerous studies have demonstrated that H3K4me1 is highly dynamic and correlates well with cell-type specific gene expression profiles, whereas promoter-associated H3K4me3 is more invariant across cell types[9].

It has been postulated that specific histone modifications function as binding elements for effector proteins that serve to regulate transcription through manipulation of the chromatin environment or assembly of transcription machinery [10-13]. For example, promoter-associated H3K4me3 can lead to recruitment of TFIID (through direct interaction with TAF3) to positively regulate transcription [14]. On the other hand, the function of H3K4me1 at enhancers has not been well understood. In Drosophila, knockout of the Trithorax-related (Trr) histone methyltransferase results in a global loss of H3K4me1 [15] and a concomitant loss of enhancer function [15, 16]. Similarly, loss of KMT2C/D, the human homologs of Trr, abolishes H3K4me1 and reduces H3K27ac levels as well as binding of Mediator and RNA polymerase II at enhancers[16, 17]. KMT2C/D knockout cells exhibited defects in enhancer activation, cell type specific

gene expression and differentiation capacities [15, 17]. These studies, while supporting a role for H3K4me1 in enhancer function, did not reveal the mechanism of action by this histone mark. It is very likely that H3K4me1 may act by recruiting specific effector proteins.

A recent study of the H3K4 demethylase KDM5C revealed that while H3K4me3 positively regulates transcription at promoters, increased H3K4me3 serves to decrease enhancer function [18]. The correct balance of H3K4me1 and me3 at promoters is equally important for transcriptional regulation. At promoters, a decrease of H3K4me3 and repression of transcription is coupled with an increase of H3K4me1 in many cell types [19]. Additionally, H3K4me1 is known to block binding of H3K4me3-associated factors such as ING1. In fact, H3K4me1 also demarcates the boundaries of active promoters, thus limiting the recruitment of factors and specifying the promoter region [19]. Clearly these closely related modifications play very distinct roles in gene regulatory networks in cells, depending on localization and differential association with regulatory complexes. This fact underscores the need to identify factors that can specifically bind to H3K4me1, and perhaps distinguish between H3K4me1 and me3, in order to fully understand the role of this histone modification in gene regulation.

Peptide or nucleosome pulldown coupled with SILAC mass spec analysis has been utilized to identify factors associating specifically with histone tail modifications[14, 20, 21]. Such studies have successfully identified proteins associated with H3K9me, H3K4me3, and H3K27me3. However, in all previous studies, binding of complexes to methylated versus unmethylated histone states was compared. In the current study, we designed a screen to identify candidate H3K4me1 binders while simultaneously

comparing association of factors with mononucleosomes bearing the H3K4me1 versus H3K4me3 modification. Our approach identified multiple components of the transcriptional regulatory machinery including the BAF complex as enriched for H3K4me1 association. ChIP-seq analysis confirmed that these factors' binding to putative enhancers correlates with H3K4me1 genome-wide in mESCs. Importantly, binding of these H3K4me1-associating proteins was drastically reduced upon depletion of KMT2C/D and loss of H3K4me1 at enhancers. In addition, loss of H3K4me1 in a mutant mouse ES cell line bearing catalytic site mutations in KMT2C and KMT2D correlates with reduced binding of BAF components SMARCA4 (BRG1) and DPF2 (BAF45d). We characterized the subunit in the BAF complex involved in preferential recognition of H3K4me1 over H3K4me3 by X-ray crystallographic analysis. We further demonstrated that *in vitro* BAF more efficiently remodels H3K4me1 nucleosomes. Taken together, our results provide mechanistic insights by which H3K4me1 acts to regulate the function of enhancers.

3.3 Results

3.3.1 Identification of potential H3K4me1 binding partners

We assembled nucleosomes with chemically modified histone H3 and naïve H4, H2A, and H2B (Figure 3.1A)[22-25]. The H3K4me1 and H3K4me3 nucleosomes were used as baits in pulldowns from nuclear extract (NE) prepared from HeLa cells grown in media containing either light or heavy isotope-labeled amino acids as shown in Figure

3.1b[20]. Any factor specifically associating with H3K4me1 over H3K4me3 in the forward reaction would be detected by mass spec as enriched in light Lys labeled peptides, and in heavy Lys labeled peptides in the reverse reaction (Figure 3.1b). Multiple replicates were performed with similar results. For final analysis, 2 replicates were combined and ratios of light peptides to heavy peptides were averaged across replicates (Figure 3.1c and Supplementary Table 3.1). As we are only assessing H3K4me1 vs me3 affinities we cannot rule out the possibility that factors identified as H3K4me1 binders may also associate with H3K4me2 or me0. Nevertheless, our approach yielded a plethora of putative H3K4me1 associated proteins including many known chromatin regulators and chromatin associated factors (Supplementary Table 3.2). Multiple subunits of the BAF (SWI/SNF) complex, such as SMARCA4 (BRG1) and SMARCC1/2 (BAF155/170), were isolated in the precipitates. Also identified were components of other chromatin remodeling complexes such as BAZ1B from WINAC and WICH, and BAZ1A from ACF. Many factors isolated contain histone-binding domains (Supplementary Table 3.2) and, in addition, several of these factors have been found associated with H3K4me1 regions of the genome in cells by ChIP mass spectrometry [26]. Interestingly, two Cohesin subunits were found to be associated with H3K4me1-nucleosomes. Cohesin is known to associate with enhancers and facilitate enhancer-promoter looping[27]. The results implicate H3K4me1 in many facets of enhancer function from chromatin remodeling to looping of enhancers and promoters. In addition to the H3K4me1 associated factors we identified several novel H3K4me3 associated proteins such as the FACT components SSRP1 and SUPT16H.

Our mono-nucleosome pulldowns differed from previous experiments that largely employed methylated histone tail peptides as bait. For the purpose of comparison, the assay was repeated comparing H3K4me1 and H3K4me3 peptides instead of mono-nucleosomes and in this case we observed enrichment of TAF and ING family proteins as observed by other labs[14]. Notably, there was less enrichment of factors for H3K4me1 in the peptide pulldowns, compared to the use of mono-nucleosome templates. This difference could be due to histone tails adopting a distinct conformation, necessary for substrates to bind, only in the presence of intact nucleosomes [28]. Alternatively, it could be due to additional interactions that exist only in intact nucleosome substrates.

To validate association and identity of a subset of the chromatin regulators (CRs) identified in our screen we incubated methylated nucleosomes with HeLa NE and performed western blotting to identify associated factors (Figure 3.1d). Target validation was limited by availability of specific antibodies so unfortunately we were unable to conduct further analysis on several interesting candidates. However, we confirmed preferential binding of H3K4me1 over H3K4me3 by a number of known enhancer-associated factors. It should also be noted that some proteins bind to multiple methylation states, such as Sap18 to H3K4me1/me2 and SMARCC2 to H3K4me0/1 (Figure 3.1d). While some factors have domains known to bind methylated Lysine residues, such as PHD domains found in PHRF1, and BAF components, other factors identified in the screen do not have any know histone binding domains. It is clear that complex binding patterns of multiple protein complexes is involved.

3.3.2 CRs are localized to H3K4me1 rich regions of the genome

Next we performed ChIP-seq for 16 CRs and 4 histone modification marks in mouse embryonic stem cells (mESCs) to determine the localization of the candidate H3K4me1-binding chromatin regulators (CRs). Clustering analysis of the ChIP-seq profiles of these factors along with three histone H3 lysine 4 methylation states (me1, me2 and me3) showed that nearly all of the CRs tested cluster together with H3K4me1 in a branch separate from H3K4me2 and H3K4me3 (Figure 3.2a). We further assayed the binding of the CRs to a subset of previously validated enhancers [29] and negative control regions by ChIP-qPCR, and found CRs to be enriched at all enhancers tested (Figure 3.2b-c and Figure S3.1a-d). Enrichment of H3K4me1-associated CRs was observed at a previously validated Sox2 enhancer[30], and several factors are also enriched at the Sox2 promoter overlapping with the promoter-flanking H3K4me1 domains. Interestingly we observed consistently higher CR enrichment at regions with both H3K4me1 and H3K27ac (Figure S3.2c-d). Next we investigated CR association with poised (n=28,008) and active (n=13,811) enhancer regions, defined as H3K4me1-positive regions with or without concomitant H3K27ac signals. For this specific analysis "active" enhancers were defined based on H3K27ac signals and not H3K16ac or H3K122ac. We discovered that active enhancer regions tend to be occupied by multiple CRs while poised enhancer regions show individual CR binding patterns (Figure 3.2d and Figure S3. 2d left vs right panel). The majority of CRs tested bound a high fraction of H3K27ac containing enhancer regions (Fig. S3.1e). That acetylation of H3K27 at enhancers coincides with binding by multiple co-activators implies that binding of multiple CRs might be necessary for full activation of the enhancers.

3.3.3 H3K4me1 dependent association of CRs with enhancers

The above results confirmed the association of CR complexes to H3K4me1 *in vitro* and *in vivo*. To determine if chromatin association of CRs is dependent upon H3K4me1, we carried out ChIP-seq analyses of these protein complexes in mouse ESC deleted of KMT2C/D [31]. Previous studies have demonstrated that KMT2C/D are responsible for H3K4me1 deposition at enhancers in multiple species[15-17]. Consistent with previous data from mouse pre-adipocytes and human colon cancer cells, knockout of both of these enzymes in mouse ESCs results in a general decrease in H3K4me1 but has little effect on the global level of H3K4me3 [31]. We performed H3K4me1, H3K4me2, and H3K4me3 ChIP-seq in mESCs deleted of both KMT2C and KMT2D genes (DKO) and compared the results with the data from WT mESCs. We observed that the majority of H3K4me3 distribution remains unaltered between WT and DKO (Figure 3.3c) whereas H3K4me2 levels are mildly effected (Figure 3.3b and Fig. S3.2a-b). Consistent with the previous studies we observed a dramatic reduction in H3K4me1 signal throughout the genome (Figure 3.3a and Fig. S3.2a-b): 47% of H3K4me1 peaks detected in WT mESCs were lost in DKO mESC (Figure 3.3d, Fig. S3.2c). The KMT2C/D-dependent H3K4me1 peaks are enriched at enhancers (Figure 3.3d), consistent with previously suggested function of KMT2C/D at these sites [15-17]. KMT2C/D-independent H3K4me1 peaks, on the other hand, overlap not only with enhancers and but also promoters (Figure 3.3d). We also detected both KMT2C/D dependent and independent H3K4me2 peaks (Fig. S3.2d). However, in contrast to H3K4me1 peaks, the KMT2C/D-dependent-H3K4me2 is found at both enhancers and

promoters at equal proportions. Additionally, as seen in pre-adipocytes, KMT2C/D dependent loss of H3K4me1 also coincides with a moderate decrease in H3K27ac at the same regions (Figure 3.3e and Fig. S3.2a-b).

Both KMT2C/D dependent and independent peaks are bound by CRs but the fraction of associated peaks is highly variable (Figure 3.3f). CRs should be reduced at KMT2C/D dependent sites in DKO cells if H3K4me1 acts to facilitate or stabilize their binding. To test this hypothesis, we performed ChIP-seq for a subset of the H3K4me1-associated CRs and demonstrate and overlap with H3K4me1 occupancy in the wild-type cells. All CRs tested were reduced at KMT2C/D dependent H3K4me1 sites compared to KMT2C/D independent sites in the DKO mESCs (Figure 3.3f and S3.2). We obtained similar results assessing CR association with known mESC enhancers using ChIP-qPCR (Fig. S3.2e).

A recent study by Dorighi and colleagues highlights a role for KMT2C/D in transcription regulation independent of H3K4me1 deposition[32]. Our data suggests that H3K4me1 is important for CR binding, however this new study raised the possibility that loss of KMT2C/D could directly affect binding of CRs independently of H3K4me1 loss. We therefore utilized the KMT2C/D catalytically inactive cell line (dCD) to distinguish between the role of H3K4me1 and KMT2C/D in binding of CRs. We performed ChIP-seq for H3K4me marks, H3K27ac, and BAF complex components SMARCA4 (BRG1) and DPF2 (BAF45d) (Figure 3.4a). In the dCD cells 38% of the distal H3K4me1 sites had reduced levels of H3K4me1. Interestingly, a small fraction of H3K4me1 sites also gained H3K4me1 signal, which is consistent with the previous data[32], and these sites are located closer to promoters than the H3K4me1 depleted regions. As in DKO cells,

H3K4me2 and me3 levels were less affected than H3K4me1 (Figure 3.4b-c, S3.3b). At regions where we observed specific loss of H3K4me1 signal we likewise observed a decrease in binding of both SMARCA4 and DPF2 (Figure 3.4e-f, S3.3c-d). Reduced BAF complex binding is specific for sites where H3K4me1 is depleted (Figure 3.4f) and was not seen at sites where H3K4me1 is unchanged, confirming the role of H3K4me1 in facilitating BAF binding to these regions. Taken together, our data from KMT2C/D KO and catalytically inactive cells supports the hypothesis that H3K4me1 plays an important role in binding of multiple CR complexes to enhancers.

3.3.4 BAF complex preferentially binds to and remodels H3K4me1 nucleosomes

The BAF complex is known to co-localize with H3K4me1 in the genome [6]. Our data suggests that H3K4me1 may play a direct role in stabilizing BAF complex binding to chromatin. To confirm that H3K4me1 can indeed serve to facilitate binding of BAF complexes in the absence of other co-factors or transcription factors, we repeated the mono-nucleosome pulldown assays with BAF complex purified from HeLa cells (Fig. S3.4a). We demonstrate that purified BAF complex binds to H3K4me1 with higher affinity than H3K4me3 on mono-nucleosomes (Figure 3.5a) and, to a lesser extent, H3 tail peptides (Fig. S3.4b). These data demonstrate that protein complexes can recognize and distinguish between closely related H3K4 methylation states, and this could be important for their recruitment to enhancers. The BAF complex regulates transcription by remodeling nucleosomes at sites of H3K4me1, suggesting a link between histone methylation and BAF activity. Utilizing *in vitro* nucleosome remodeling

assays [33] we find that the BAF complex more efficiently remodels H3K4me1 mono-nucleosomes, than H3K4me0, H3K4me2, and H3K4me3 mono-nucleosomes (Figure 3.5b-c, S3.4c). This data suggests a functional link between enhancer-specific histone modifications and the activity of recruited chromatin regulatory complexes.

3.3.5 Crystal structure of DPF3 binding preferentially H3K4me1

Based on peptide binding and NMR/X-ray structures, the PHD1 domain of BAF component DPF3 (BAF45c) recognizes H3K14ac, while the PHD2 domain in these proteins binds to H3K4me0 [34]. BAF subunits DPF1, DPF2, DPF3, and PHF10 (BAF45B, C, D, and A isoforms respectively) have cell type specific expression patterns[35]. Our data demonstrates that mESC specific DPF2 associates with H3K4me1. To determine if the DPF3 (BAF45c) PHD2 domain could contribute to H3K4me1 recognition as well, we purified the PHD1/2 region of DPF3 of this family of proteins, and used isothermal titration calorimetry to measure its affinity for H3 tail peptides containing H3K14ac plus H3K4me0, H3K4me1, or H3K4me3. Consistent with our biochemical studies, we found that the isolated BAF45c PHD1/2 region strongly preferred H3K4me1 (K_d of 20 μM for H3K4me1/K14ac) over H3K4me3 binding (K_d of 115 μM for H3K4me3/K14ac) (Fig. S3.5a-c). However, in contrast to our findings with the intact BAF complex and mono-nucleosomes the DPF3 PHD1/2 region bound to the H3K4me0 peptide with slightly higher affinity (K_d of 7.8 μM for H3K4me0/K14ac) than the H3K4me1 peptide. These data suggest that additional factors in the BAF complex and/or nucleosomes may influence H3K4me1 specificity.

To reveal the atomic basis of the preferential recognition of DPF3 PHD1/2 for H3K4me1 over H3K4me3, we next determined two high-resolution (1.2 Å) crystal structures of the DPF3 PHD1-2 region bound to H3 tail peptides (residues 1-18) containing H3K14ac and either H3K4me0 or H3K4me1 (Supplementary Table 3.3). The two structures show a nearly identical overall structure of DPF3 (<0.04 Å overall Cα r.m.s.d.), and largely agree with prior structures of this protein, with a 1.5 Å overall Cα r.m.s.d. to a prior NMR structure (PDB ID 2KWJ) and 0.8 Å overall Cα r.m.s.d. to a prior X-ray crystal structure (PDB ID 5I3L) [31, 34]. In our two structures, the two PHD domains are intimately associated with one another, with a binding pocket in PHD1 that recognizes H3K14ac, and a pocket in PHD2 that recognizes H3K4 (Figure 6a-c) leading to virtually identical bound conformations of the H3K4me0 and H3K4me1 peptides. In both complexes, H3K4 is nestled tightly in a surface cavity made up of the hydrophobic side chains of I314, L331, and F333. In addition, the main chain carbonyl groups of residues 314, 315, and 317 are all close enough to the H3K4 amino group to form hydrogen-bonding interactions. These interactions likely contribute to the preferential binding of unmethylated or monomethylated H3K4, the amino groups of which can form two (K4me1) or three (K4me0) hydrogen bonds, over di-or trimethylated H3K4. In addition, the H3K4 mono-methyl group packs in a preformed cavity that is just large enough for a single methyl group. Hence, these carbonyls may sterically disfavor di- or trimethylated H3K4 binding.

In contrast to earlier NMR structures of the DPF3-H3 tail complex [34], but in agreement with a recent crystal structure [31], our structures show that H3 residues 4-10 adopt an α-helical conformation. Additionally, we find that H3R8 forms a “lid” over the

binding site, extending directly over H3K4 and forming a hydrogen-bonding network with DPF3 residues E315 and D328 on opposite sides of the H3K4 binding pocket (Figure 3.6b-c and S3.5d-e); this residue's position was not well-resolved in the previous crystal structure [31]. Both the α-helical conformation of the H3 tail and the H3R8 "lid" most closely mirror earlier observations in crystal structures of the MYST family acetyltransferase KAT6A (MOZ), which possesses a double-PHD finger domain at its N-terminus that recognizes unmodified H3K4 and acetylated H3K14 [28] or propionylated/ butyrylated/crotonylated H3K14[36]. This H3 tail-binding mode may also be shared in other double-PHD finger protein families; for instance, an unpublished NMR structure of KMT2C (PDB code 2YSM) shows that this protein possesses a pair of acidic residues bracketing the H3K4 binding site that could participate in H3R8 binding. This mode of H3K4 recognition may also have functional relevance as it leaves the H3K4me1 group solvent exposed in the complex, creating the possibility that additional factors in BAF or in the nucleosome itself could associate with the composite DPF3-H3K4me1 surface and provide additional specificity for H3K4me1 over H3K4me0.

3.4 Discussion

In summary, we carried out SILAC mass spectrometry analysis to systematically identify nuclear proteins that bind H3K4me1. Our experiments uncovered components of multiple chromatin regulatory complexes, including the BAF chromatin remodeling complex, as H3K4me1-associating proteins. We further validated the binding of a

subset of these complexes to H3K4me1 mononucleosomes *in vitro* and to genomic regions bearing the histone mark in embryonic stem cells. We showed that deletion of H3K4 methyltransferaseses KMT2C/D leads to a loss of occupancy by these complexes at KMT2C/D-dependent H3K4me1 regions. Importantly, we confirmed that loss of H3K4me1 in both KMT2C/D knock out and catalytically null mutant cells correlated with a decrease in binding of CRs to enhancers, supporting our hypothesis that H3K4me1 plays an important role in binding of key chromatin regulatory factors. We chose to focus on the BAF complex, and obtained strong evidence suggesting that H3K4me1 is directly involved in the association of this complex to chromatin. The BAF complex belongs to the SWI/SNF family of ATP-dependent chromatin remodeling complexes [35]. Containing between 10 to 12 components, BAF complexes are necessary for early embryogenesis, activation of lineage specific genes during cellular differentiation, and maintenance of pluripotency of embryonic stem cells. Genome-wide profiling studies have shown that BAF complexes generally localize to distal enhancers where they are required for histone acetylation during differentiation of ES cells. A recent study involving *in situ* capture of specific genomic regions also identified BAF as an enhancer bound complex[37]. However, exactly how BAF complex is recruited to the enhancers has not been fully understood [35]. Here, we provided multiple lines of evidence that H3K4me1 may play a role in the recruitment of BAF complex to enhancers. BAF complexes fail to localize to promoter-distal enhancers in KMT2C/D double-KO, and in KMT2C/D catalytically inactive mutant cells. Using protein-pull down assays, we showed that the BAF complex interacts directly with H3K4me1 mononucleosome *in vitro* via the PHD2 domain in DPF3 (BAF45c). X-ray crystallography experiments further

revealed a surface cavity in the PHD2 domain of DPF3 that readily accommodates monomethylated lysine 4 of histone H3, but not tri-methylation. Finally, nucleosome remodeling assays demonstrated that H3K4me1 facilitates the BAF complex' nucleosome remodeling activity above all other H3K4me states. These results, taken together, support a model in which the histone modification H3K4me1 directly helps to recruit BAF complex to enhancers, and therefore plays an active role in enhancer function.

While this work was under revision, Dorighi and colleagues[32] reported that KMT2C/D promotes RNA synthesis at enhancers and nearby promoters independently of the H3K4 monomethylation activities. While this observation suggests that H3K4me1 may not be necessary for loading of RNA polymerase II at enhancers and subsequent activation of target promoters, it does not rule out other functions of H3K4me1 at enhancers. Another recent study demonstrated that Drosophila bearing catalytically inactive Trr (H3K4me1 histone methyltransferase) survive to adulthood with only subtle gene expression changes. However, if subjected to temperature stress conditions developmental abnormalities were observed[38]. In addition, this and other studies have found that loss of KMT2C/D in mESCs does not affect self-renewal[17, 38]. This can be partially explained by the fact that at poised enhancers in mESCs H3K4me1 is KMT2C/D independent[32], suggesting a role for other methyltransferases in H3K4me1 deposition and enhancer function in higher organisms. This is in agreement with our current study demonstrating that ~50% of H3K4me1 peaks in mESCs are KMT2C/D independent. Therefore, additional experiments are needed to better define the role of H3K4me1 in enhancer function during cellular differentiation and animal development.

3.5 Figures

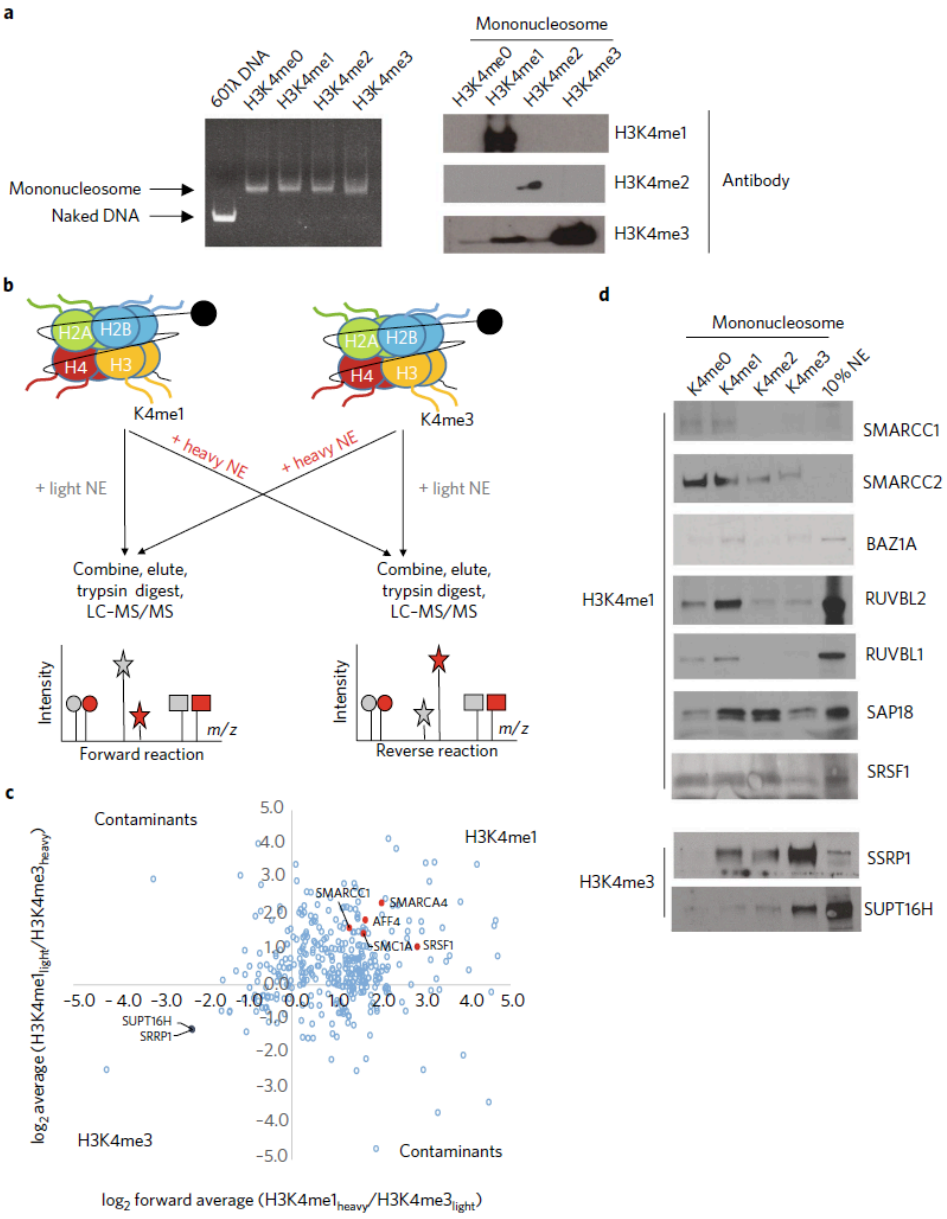


Figure 3.1: **Identification of H3K4me1 binding proteins using SILAC and Mass-spec analysis**. **a)** Left – Mononucleosomes assembled from biotin tagged 601λ positioning sequence and methylated octamers. Right – Chemically modified nucleosomes are recognized by specific antibodies against various H3K4 methylation. 3 independent chemical modifications were tested yielding similar results. **b)** Schematic of SILAC mass spec screen. **c)** Average Log2 L/H of forward reactions on X-axis and log2 H/L of reverse reactions on y-axis (from 4 independent biological replicates). Top right quadrant is H3K4me1 associated factors and bottom left quadrant contains H3K4me3 associated factors. **d)** Biotin-tagged methylated nucleosomes used as bait for pulldowns from HeLa NE. The bound proteins detected by western blotting with specific antibodies are listed, experiments were repeated at least twice with similar results.

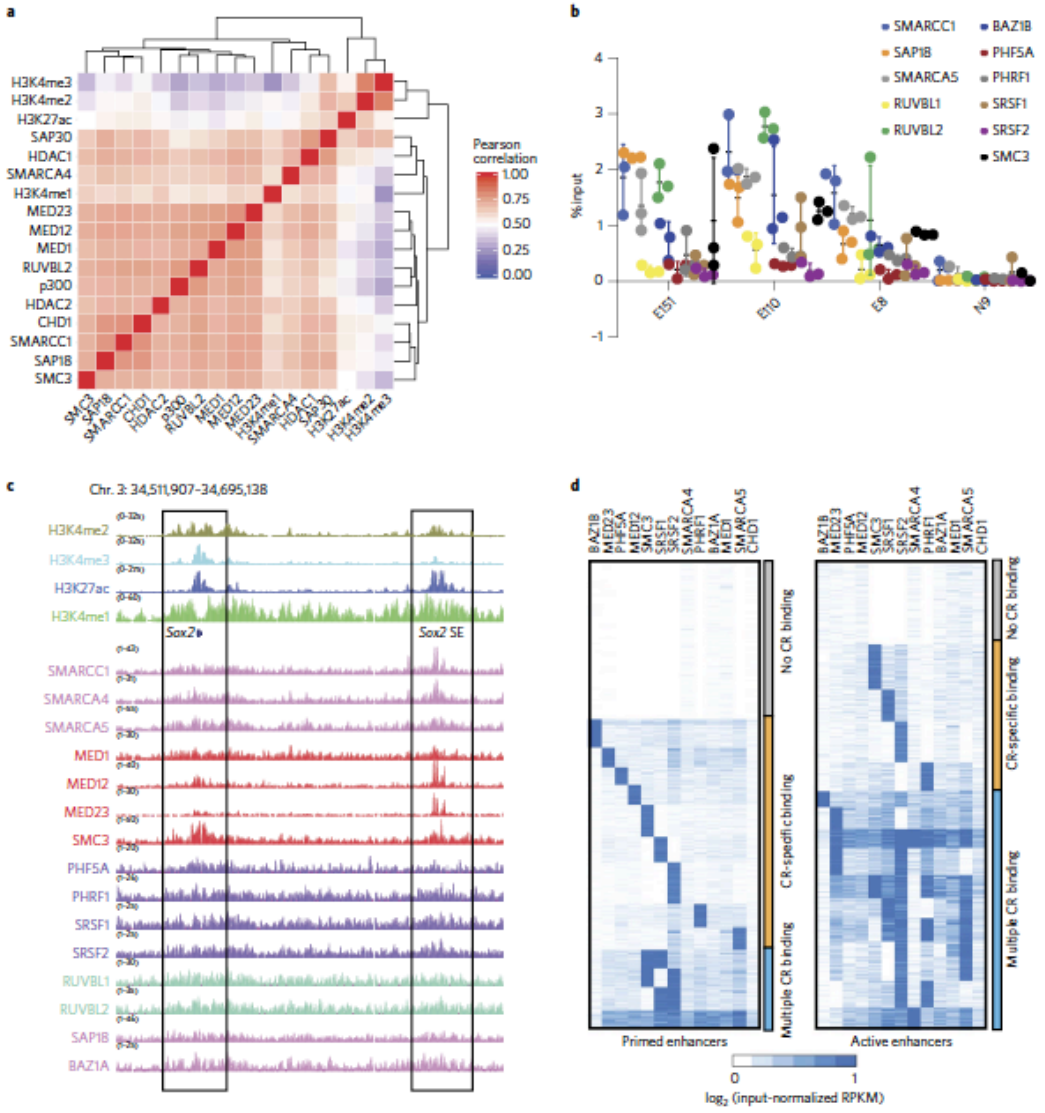


Figure 3.2: **Binding of CRs at H3K4me1 regions and enhancers**. **a)** Hierarchical clustering of genome-wide ChIP-seq signals (RPKM) for H3K4me1, H3K4me2, H3K4me3 and chromatin binding proteins with 1kb-binning, n=2,435,743. The heatmap shows pair-wise Pearson correlation coefficient between different ChIP-seq datasets. **b)** ChIP-qPCR in mESC with antibodies listed, primers designed for validated enhancers E110, E151, E8 and negative control region N9. Error bars, mean ±SD for n=3 biological replicates. **d)** Browser shot of candidate H3K4me1 readers at the Sox2 enhancer. Active enhancer with high H3K27ac boxed left, poised enhancer with low H3K27ac boxed right. **d)** Heat maps for K-means clustering results of input normalized CR signals according to poised enhancers versus active enhancers. Each cluster was manually classified as 'Multiple CR bind', 'CR-specific bind', and 'No CR bind' according to CR binding patterns. Experiments were repeated at least twice with each antibody.

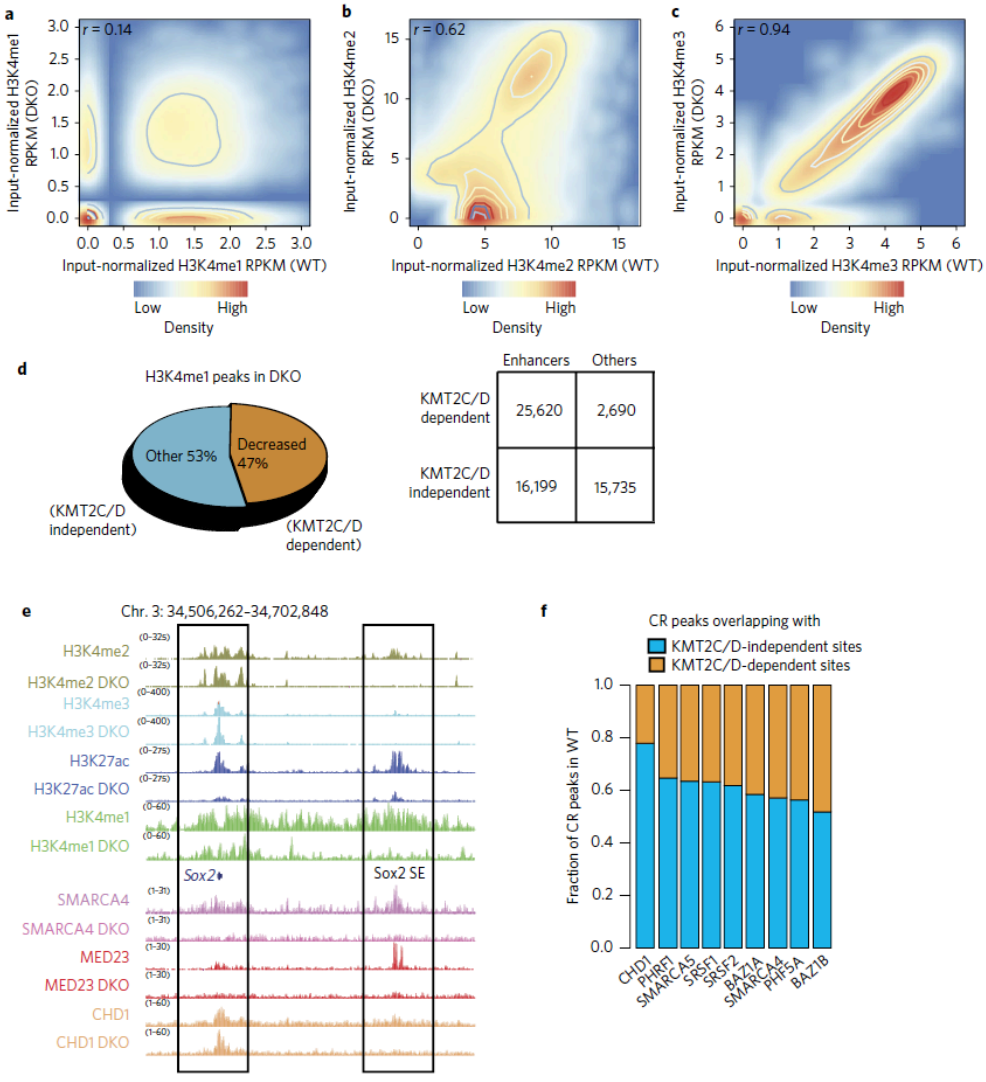


Figure 3.3: Concomitant loss of H3K4me1 and CR binding at enhancers in KMT2C/D DKO mouse ES cells. **a)** A scatter density plot of input normalized H3K4me1 RPKMs between wild-type and KMT2C/D DKO cell lines at H3K4me1 peaked regions, n=43,918. **b)** A scatter density plot of input normalized H3K4me2 RPKMs between WT and KMT2C/D DKO cell lines at H3K4me2 peaked regions, n=33,197. **c)** A scatter density plot of input normalized H3K4me3 RPKMs between wild-type and KMT2C/D DKO cell lines at H3K4me3 peaked regions, n=22,157. **d)** Upper panel - A pie chart for the fraction of H3K4me1 peaks in DKO KMT2C/D mESCs according to KMT2C/D dependent or KMT2C/D independent patterns. Lower panel - 2 by 2 table of the relationship with enhancer regions according to KMT2C/D dependent and independent H3K4me1 peaked regions. **e)** Browser shot of H3K4me1, H3K4me2, H3K4me3, H3K27ac, and CR levels in WT vs DKO KMT2C/D mESCs at the Sox2 enhancer. For each factor top track in form WT and bottom track is DKO. **f)** Bar plots are shown for the fraction of CR peaks in wild-type (y-axis) according to overlap with KMT2C/D independent (blue) and dependent sites (orange). Total number of CR peaks identified are : CHD1 (n=14,846), PHRF1 (n=21,924), SMARCA5 (n=13,891), SRSF1 (n=23,221), SRSF2 (n=31,200), BAZ1A (n=13,806), SMARCA4 n=10,897), PHRF5a (n=11,926), BAZ1B (n=5,405). Experiments were repeated at least twice in each cell type.

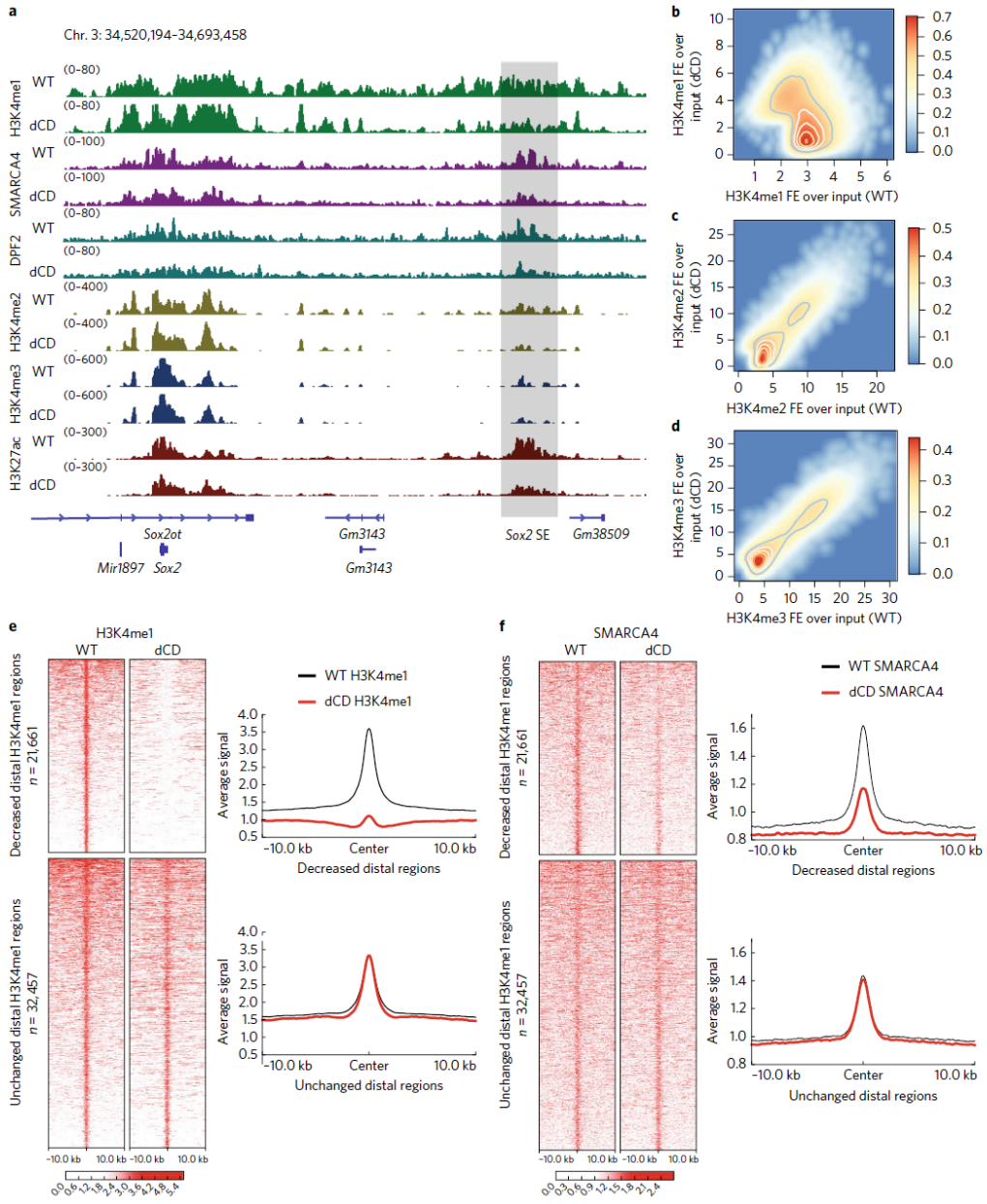


Figure 3.4: Reduced BAF complex binding is associated with depletion of H3K4me1 in KMT2C/D catalytically null (dCD) cells. a) Browser shot of ChIP-seq signal (RPKM) for SMARCA4 and DPF2 at the Sox2 locus. The Sox2 super-enhancer is shaded on right. Experiments were repeated independently twice with similar results. **b,c,d)** Scatter density plots of input normalized fold enrichment between WT and dCD at H3K4me1(n=82,053), H3K4me2(n=53,501) and H3K4me3(n=34,553) peaked regions . **e)** Left - Heatmap of input normalized H3K4me1 ChIP signal in WT and dCD over 21,661 distal H3K4me1 regions with decreased signals in dCD and 32,475 distal H3K4me1 regions with invariable signals, with regions sorted by strength of H3K4me1 signal. Right - aggregate plot showing the average signal in WT and dCD. **f)** Left - Heatmap of input normalized SMARCA4 ChIP-seq signal in WT and DCD over the same regions in (**e**). Right - aggregate plot showing the average signal in WT and dCD.

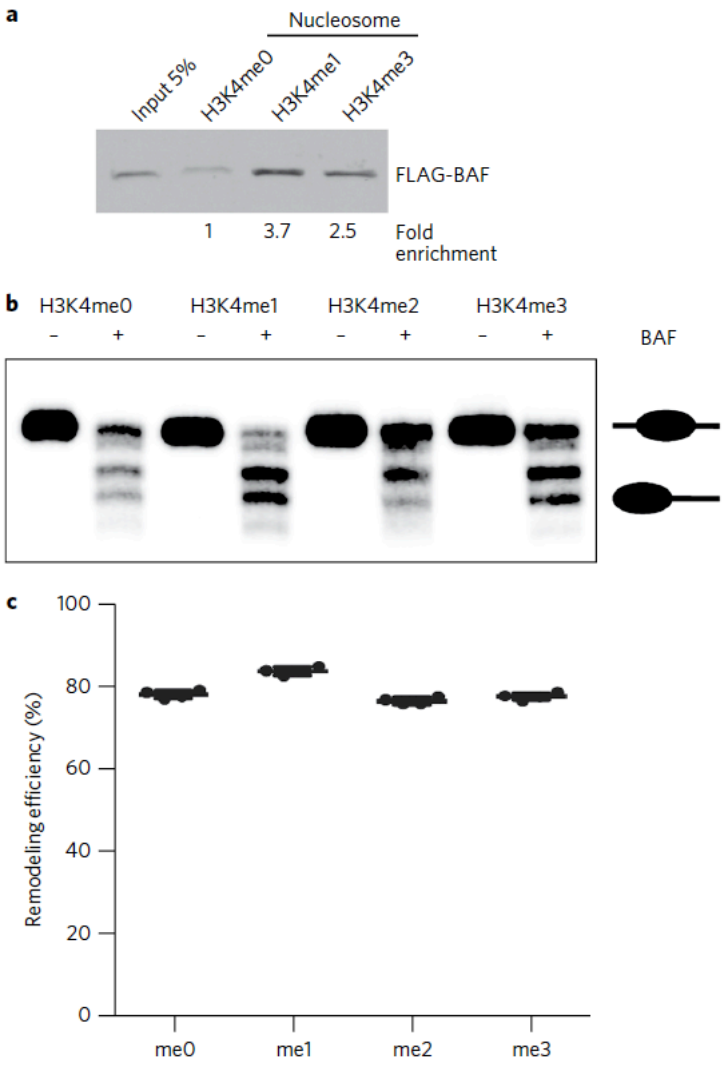


Figure 3.5: BAF complex preferentially binds and remodels H3K4me1 modified nucleosomes. **a)** Purified Flag-BAF complex binding to H3K4 methylated-nucleosomes, western blotted with anti-FLAG antibody (M2). Pulldown repeated 3 times yielding the same result. **b)** Polyacrylamide gel showing representative (n=4) *in vitro* remodeling assay. After incubation with BAF complex, nucleosomes are slid to the end of the 216-bp DNA fragment resulting in a change in mobility in the gel. Top band is un-remodeled nucleosome, and lower four bands are slid nucleosomes with different positions away from 146-bp Widom601 binding sites in the middle. **c)** Quantification of nucleosome remodeling assays. Error bars, mean ±SD n=4 biological replicates, see Figure S4C. The reduced percentage of the top band is defined as remodeling efficiency.

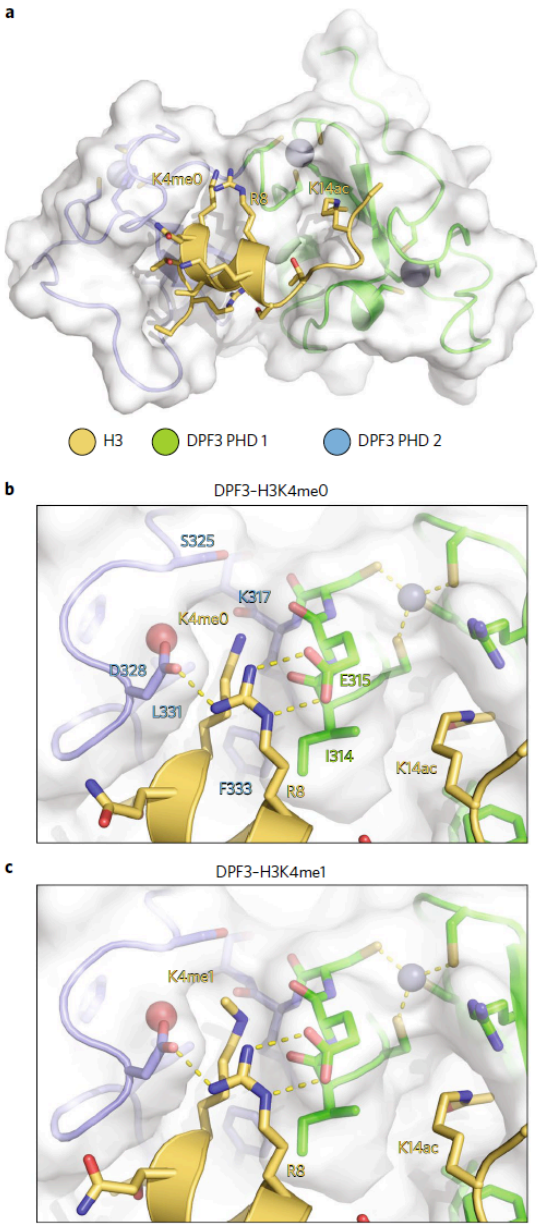


Figure 3.6: **Structural basis for H3K4 recognition by DPF3**. **a)** Overall structure of DPF3:H3K4me0 complex. DPF3 PHD1 domain is shown in green, PHD2 in blue, and histone H3 tail peptide shown in yellow. **b)** Close-up view of the DPF3 PHD1-2 region (light blue, white surface) with H3 residues 1-18 with H3K4me0 and H3K14ac (yellow). PHD1 binds H3K14ac as previously observed, while PHD2 binds H3K4 and H3R8. **c)** Close-up view of DPF3 binding H3 1-18 with H3K4me1 and H4K14ac. The mono-methyl group is accommodated in a pre-formed surface pocket on DPF3. For views of the overall structure and electron density maps, see Figure S3.5.

3.6 Author Contributions

A.L. and B.R. conceived the study and prepared the manuscript. A.L. designed and carried out the SILAC experiments, nucleosome pull-down experiments, and ChIP-seq experiments, and prepared the manuscript. H.H. performed H3K4me2 ChIP-seq analysis and all experiments with the dCD cell lines. A.Y.L. prepared sequencing libraries. C.A. ran the mass spec samples in the laboratory of H.Z. and provided expertise in mass spec analysis. H.H. performed ChIP-seq data analysis. C.W. and K.G. provided KMT2C/D DKO mESCs and shared expertise and data. W.W. and D.W. designed and executed the remodeling assays. A.K.S. designed/supplied the H3 tail peptides and, along with J.E.H., provided advice on their use in biochemical studies. N.S. purified Baf45c, performed H3 tail peptide binding measurements, and determined crystal structures under the direction of K.D.C..

3.7 Competing Interests

The authors declare no competing financial interests.

3.8 Funding Resources

Funding to this work is provide by the Ludwig Institute for Cancer Research, National Institutes of Health (5R01GM115961). AL was supported by an NIH training grant 5 T32 AI007469.

3.9 Acknowledgements

The authors thank Samantha Kuan and Bin Li for processing of ChIP-seq samples, Jason Liang and Gary Hon for help and advice in SILAC mass spec analysis, Drs. Joanna Wysocka and Kristel Dorighi for sharing the MLL3/4 dCD mESC line, and Inkyung Jung for advice on ChIP-seq data analysis. We also thank Timothy Gahman for arranging for peptide synthesis and Andrey Bobkov for assistance with isothermal titration calorimetry.

Chapter 3, in full, is a reprint of the paper published in *Nature Genetics* 2018. Andrea Local[*], Hui Huang[*], Claudio P. Albuquerque, Namit Singh, Ah Young Lee, Wei Wang, Chaochen Wang, Judy E. Hsia, Andrew K. Shiau, Kai Ge, Kevin D. Corbett, Dong Wang, Huilin Zhou and Bing Ren. "Identification of H3K4me1-associated proteins at mammalian enhancers." *Nature Genetics* 50.1 (2018): 73-82. [*]authors contributed equally to this work. The dissertation author was one of the primary investigators and authors of this paper.

3.10 References

1. Hardison, R.C. & Taylor, J. Genomic approaches towards finding cis-regulatory modules in animals. *Nat Rev Genet* **13**, 469-483 (2012).

2. Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., Wang, W., Weng, Z., Green, R.D., Crawford, G.E. & Ren, B. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39**, 311-318 (2007).

3. Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., Boyer, L.A., Young, R.A. & Jaenisch, R. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-21936 (2010).

4. Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., Ching, K.A., Antosiewicz-Bourget, J.E., Liu, H., Zhang, X., Green, R.D., Lobanenkov, V.V., Stewart, R., Thomson, J.A., Crawford, G.E., Kellis, M. & Ren, B. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-112 (2009).

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

6. Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A. & Wysocka, J. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283 (2011).

7. Taylor, G.C., Eskeland, R., Hekimoglu-Balkan, B., Pradeepa, M.M. & Bickmore, W.A. H4K16 acetylation marks active genes and enhancers of embryonic stem cells, but does not alter chromatin compaction. *Genome Res* **23**, 2053-2065 (2013).

8. Pradeepa, M.M., Grimes, G.R., Kumar, Y., Olley, G., Taylor, G.C., Schneider, R. & Bickmore, W.A. Histone H3 globular domain acetylation identifies a new class of enhancers. *Nat Genet* **48**, 681-686 (2016).

9. Calo, E. & Wysocka, J. Modification of enhancer chromatin: what, how, and why? *Mol Cell* **49**, 825-837 (2013).

10. Musselman, C.A., Lalonde, M.E., Cote, J. & Kutateladze, T.G. Perceiving the epigenetic landscape through histone readers. *Nat Struct Mol Biol* **19**, 1218-1227 (2012).

11. Seet, B.T., Dikic, I., Zhou, M.M. & Pawson, T. Reading protein modifications with interaction domains. *Nat Rev Mol Cell Biol* **7**, 473-483 (2006).

12. Smith, E. & Shilatifard, A. The chromatin signaling pathway: diverse mechanisms of recruitment of histone-modifying enzymes and varied biological outcomes. *Mol Cell* **40**, 689-701 (2010).

13. Strahl, B.D. & Allis, C.D. The language of covalent histone modifications. *Nature* **403**, 41-45 (2000).

14. Vermeulen, M., Mulder, K.W., Denissov, S., Pijnappel, W.W., van Schaik, F.M., Varier, R.A., Baltissen, M.P., Stunnenberg, H.G., Mann, M. & Timmers, H.T. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell* **131**, 58-69 (2007).

15. Herz, H.M., Mohan, M., Garruss, A.S., Liang, K., Takahashi, Y.H., Mickey, K., Voets, O., Verrijzer, C.P. & Shilatifard, A. Enhancer-associated H3K4 monomethylation by Trithorax-related, the Drosophila homolog of mammalian Mll3/Mll4. *Genes Dev* **26**, 2604-2620 (2012).

16. Hu, D., Gao, X., Morgan, M.A., Herz, H.M., Smith, E.R. & Shilatifard, A. The MLL3/MLL4 branches of the COMPASS family function as major histone H3K4 monomethylases at enhancers. *Mol Cell Biol* **33**, 4745-4754 (2013).

17. Lee, J.E., Wang, C., Xu, S., Cho, Y.W., Wang, L., Feng, X., Baldridge, A., Sartorelli, V., Zhuang, L., Peng, W. & Ge, K. H3K4 mono- and di-methyltransferase MLL4 is required for enhancer activation during cell differentiation. *Elife* **2**, e01503 (2013).

18. Outchkourov, N.S., Muino, J.M., Kaufmann, K., van Ijcken, W.F., Groot Koerkamp, M.J., van Leenen, D., de Graaf, P., Holstege, F.C., Grosveld, F.G. & Timmers, H.T. Balancing of histone H3K4 methylation states by the

Kdm5c/SMCX histone demethylase modulates promoter and enhancer function. *Cell Rep* **3**, 1071-1079 (2013).

19. Cheng, J., Blum, R., Bowman, C., Hu, D., Shilatifard, A., Shen, S. & Dynlacht, B.D. A role for H3K4 monomethylation in gene repression and partitioning of chromatin readers. *Mol Cell* **53**, 979-992 (2014).

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

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

22. Carruthers, L.M., Tse, C., Walker, K.P., 3rd & Hansen, J.C. Assembly of defined nucleosomal and chromatin arrays from pure components. *Methods Enzymol* **304**, 19-35 (1999).

23. Luger, K., Rechsteiner, T.J. & Richmond, T.J. Expression and purification of recombinant histones and nucleosome reconstitution. *Methods Mol Biol* **119**, 1-16 (1999).

24. Luger, K., Rechsteiner, T.J. & Richmond, T.J. Preparation of nucleosome core particle from recombinant histones. *Methods Enzymol* **304**, 3-19 (1999).

25. Simon, M.D., Chu, F., Racki, L.R., de la Cruz, C.C., Burlingame, A.L., Panning, B., Narlikar, G.J. & Shokat, K.M. The site-specific installation of methyl-lysine analogs into recombinant histones. *Cell* **128**, 1003-1012 (2007).

26. Engelen, E., Brandsma, J.H., Moen, M.J., Signorile, L., Dekkers, D.H., Demmers, J., Kockx, C.E., Ozgur, Z., van, I.W.F., van den Berg, D.L. & Poot, R.A. Proteins that bind regulatory regions identified by histone modification chromatin immunoprecipitations and mass spectrometry. *Nat Commun* **6**, 7155 (2015).

27. Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., Taatjes, D.J., Dekker, J. & Young, R.A. Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430-435 (2010).

28. Dreveny, I., Deeves, S.E., Fulton, J., Yue, B., Messmer, M., Bhattacharya, A., Collins, H.M. & Heery, D.M. The double PHD finger domain of MOZ/MYST3 induces alpha-helical structure of the histone H3 tail to facilitate acetylation and methylation sampling and modification. *Nucleic Acids Res* **42**, 822-835 (2014).

29. Yue, F., Cheng, Y., Breschi, A., Vierstra, J., Wu, W., Ryba, T., Sandstrom, R., Ma, Z., Davis, C., Pope, B.D., Shen, Y., Pervouchine, D.D., Djebali, S., Thurman, R.E., Kaul, R., Rynes, E., Kirilusha, A., Marinov, G.K., Williams, B.A., Trout, D., Amrhein, H., Fisher-Aylor, K., Antoshechkin, I., DeSalvo, G., See, L.H., Fastuca, M., Drenkow, J., Zaleski, C., Dobin, A., Prieto, P., Lagarde, J., Bussotti, G., Tanzer, A., Denas, O., Li, K., Bender, M.A., Zhang, M., Byron, R., Groudine, M.T., McCleary, D., Pham, L., Ye, Z., Kuan, S., Edsall, L., Wu, Y.C., Rasmussen, M.D., Bansal, M.S., Kellis, M., Keller, C.A., Morrissey, C.S., Mishra, T., Jain, D., Dogan, N., Harris, R.S., Cayting, P., Kawli, T., Boyle, A.P., Euskirchen, G., Kundaje, A., Lin, S., Lin, Y., Jansen, C., Malladi, V.S., Cline, M.S., Erickson, D.T., Kirkup, V.M., Learned, K., Sloan, C.A., Rosenbloom, K.R., Lacerda de Sousa, B., Beal, K., Pignatelli, M., Flicek, P., Lian, J., Kahveci, T., Lee, D., Kent, W.J., Ramalho Santos, M., Herrero, J., Notredame, C., Johnson, A., Vong, S., Lee, K., Bates, D., Neri, F., Diegel, M., Canfield, T., Sabo, P.J., Wilken, M.S., Reh, T.A., Giste, E., Shafer, A., Kutyavin, T., Haugen, E., Dunn, D., Reynolds, A.P., Neph, S., Humbert, R., Hansen, R.S., De Bruijn, M., Selleri, L., Rudensky, A., Josefowicz, S., Samstein, R., Eichler, E.E., Orkin, S.H., Levasseur, D., Papayannopoulou, T., Chang, K.H., Skoultchi, A., Gosh, S., Disteche, C., Treuting, P., Wang, Y., Weiss, M.J., Blobel, G.A., Cao, X., Zhong, S., Wang, T., Good, P.J., Lowdon, R.F., Adams, L.B., Zhou, X.Q., Pazin, M.J., Feingold, E.A., Wold, B., Taylor, J., Mortazavi, A., Weissman, S.M., Stamatoyannopoulos, J.A., Snyder, M.P., Guigo, R., Gingeras, T.R., Gilbert, D.M., Hardison, R.C., Beer, M.A., Ren, B. & Mouse, E.C. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355-364 (2014).

30. Li, Y., Rivera, C.M., Ishii, H., Jin, F., Selvaraj, S., Lee, A.Y., Dixon, J.R. & Ren, B. CRISPR reveals a distal super-enhancer required for Sox2 expression in mouse embryonic stem cells. *PLoS One* **9**, e114485 (2014).

31. Wang, C., Lee, J.E., Lai, B., Macfarlan, T.S., Xu, S., Zhuang, L., Liu, C., Peng, W. & Ge, K. Enhancer priming by H3K4 methyltransferase MLL4 controls cell fate transition. *Proc Natl Acad Sci U S A* **113**, 11871-11876 (2016).

32. Dorighi, K.M., Swigut, T., Henriques, T., Bhanu, N.V., Scruggs, B.S., Nady, N., Still, C.D., 2nd, Garcia, B.A., Adelman, K. & Wysocka, J. Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Mol Cell* **66**, 568-576 e564 (2017).

33. Phelan, M.L., Sif, S., Narlikar, G.J. & Kingston, R.E. Reconstitution of a core chromatin remodeling complex from SWI/SNF subunits. *Mol Cell* **3**, 247-253 (1999).

34. Zeng, L., Zhang, Q., Li, S., Plotnikov, A.N., Walsh, M.J. & Zhou, M.M. Mechanism and regulation of acetylated histone binding by the tandem PHD finger of DPF3b. *Nature* **466**, 258-262 (2010).

35. Ho, L. & Crabtree, G.R. Chromatin remodelling during development. *Nature* **463**, 474-484 (2010).

36. Xiong, X., Panchenko, T., Yang, S., Zhao, S., Yan, P., Zhang, W., Xie, W., Li, Y., Zhao, Y., Allis, C.D. & Li, H. Selective recognition of histone crotonylation by double PHD fingers of MOZ and DPF2. *Nat Chem Biol* **12**, 1111-1118 (2016).

37. Liu, X., Zhang, Y., Chen, Y., Li, M., Zhou, F., Li, K., Cao, H., Ni, M., Liu, Y., Gu, Z., Dickerson, K.E., Xie, S., Hon, G.C., Xuan, Z., Zhang, M.Q., Shao, Z. & Xu, J. In Situ Capture of Chromatin Interactions by Biotinylated dCas9. *Cell* **170**, 1028-1043 e1019 (2017).

38. Rickels, R., Herz, H.M., Sze, C.C., Cao, K., Morgan, M.A., Collings, C.K., Gause, M., Takahashi, Y.H., Wang, L., Rendleman, E.J., Marshall, S.A., Krueger, A., Bartom, E.T., Piunti, A., Smith, E.R., Abshiru, N.A., Kelleher, N.L., Dorsett, D. & Shilatifard, A. Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nat Genet* **49**, 1647-1653 (2017).