Engineering CTCF DNA-Binding Specificity to Alter Gene Expression and Genome Topology in Human Cells

Engineering CTCF DNA-binding specificity to alter gene expression and genome topology in human cells

A dissertation presented

by

Rebecca Tayler Cottman

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biological and Biomedical Sciences

Harvard University

Cambridge, Massachusetts

May 2020

Dissertation Advisor: J. Keith Joung Rebecca Tayler Cottman

Engineering CTCF DNA-binding specificity to alter gene expression and genome topology in human cells

Abstract

The 3D organization of the eukaryotic genome is an integral part of cell homeostasis and differentiation. Genome organization is a multi-tiered system regulated in large part by a host of transcription factors and chromatin interacting proteins, such as CCCTC-binding factor (CTCF). CTCF is a ubiquitous DNA-binding protein involved in higher-order topological organization of the genome via establishment of topologically associated domains (TADs). CTCF is recruited to the genome by 40 bp CTCF binding sites (CBSs) that contain a highly conserved 15bp motif. CBSs are frequently mutated in cancer and developmental diseases leading to loss of CTCF binding and subsequent gene misregulation, but studying the mechanistic consequences of CTCF function is complicated by its broad functionality within eukaryotic cells. In addition, CTCF has been implicated in gene regulation through the formation of promoter-enhancer loops, but it is not clear if cohesin or RNA is the cofactor in this process. To address this limitation, I sought to define and alter the DNA-binding determinants of CTCF so as to facilitate structure/function

studies of this important regulator. Using a bacterial-two-hybrid (B2H) reporter system, I first identified the nucleotides in the CBS that are essential for CTCF binding. I used this knowledge to generate a series of variant CBSs (vCBS) that are no longer bound efficiently by wild-type CTCF. Leveraging the B2H as a selection system, I evolved CTCF variants with altered binding specificities for these vCBSs. Utilizing the CTCF-regulated proto-oncogene *MYC* as an endogenous human gene reporter, I demonstrated that these engineered CTCF variants could reproduce the normal biological role of CTCF *in cellula* and could be used to define the functional consequences of mutating CTCF domains on expression of this gene, providing evidence that RNA, not cohesin, is the facilitating cofactor of establishing the CTCF-mediated promoter-enhancer loop. I have developed a system to study the mechanistic requirements for CTCF-mediated gene expression without the confounding pleiotropic effects, allowing for site specific analysis of CTCF mediated gene expression. This work could be applied to creating a toolbox of variant transcription factors with novel DNA-recognition profiles for application in epigenetic engineering.

Table of Contents

Front Matter

Acknowledgments

Thank you Caleb Lareau for analysis of 4C data; Sowmya Iyer for analysis of ChIP-seq and RNA-seq data; Jay Jun for help with benchwork; Esther Tak and Benjamin Waldman for insightful discussion.

I would like to thank the following people for their contributions towards my growth as a scientist:

Esther (Yu Gyoung) Tak- For incredible insight and patience. Thank you for teaching me everything I know about the epigenome and how to study it. I will always bend the knee to the Queen of Epigenetics.

Julian Grünewald & Karl Petri- The German scholars of infinite wisdom. Without you two, I would have lost my sanity long ago. Thanks for keeping me humble.

Keith Joung- For keeping the lights on! Also, for challenging me, believing in me, and always giving me every opportunity to succeed. Thank you for mentoring me to become what I am today. I hope to do you proud in the real world.

Benjamin Waldman- I look forward to many more scientific discussions over morning coffee.

I dedicate this thesis to those in my life that have fundamentally shaped me: My family (Cottmans, Stubbs', Waldmans, and Jirovskys) and Benjamin Waldman.

“I can’t believe it worked!”

-Keith Joung, September 4th, 2018

Introduction

The eukaryotic genome is organized by multiple layers of protein-protein and protein-DNA interactions that coordinate gene expression, nuclear localization, and cell differentiation. Biological mechanisms of the eukaryotic cell such as differentiation, mitosis, and DNA repair, are the result of the interplay between epigenetic modifications, chromatin associated proteins, transcription factors, and the segregation of the genome in a 3D space[1–3]. Genome organization can be arranged in layers of complexity that follow the hierarchy of three main tiers of organization: chromosome territories, A/B compartmentalization, and the formation of topologically associating domains (TADs). Phase separation is an additional layer of eukaryotic genome organization that forms structured regions of interchromosomal and intrachromosomal interactions through formation of liquid-phase condensates[4,5]. Phase separation is the accumulation of alike molecules at a region of the genome to form droplets that act as membrane-less organelles. The condensates recruit molecules (or proteins) of similar function while excluding others and in this way create a reaction hub for inter or intrachromosomal regions[6–9]. The organization of the eukaryotic genome is key to spatial management of gene regulatory elements and plays a critical role in the regulation of gene expression.

Hi-C experiments determined the eukaryotic genome is organized into two main compartments (A and B) within the nucleus. Compartment A consists of euchromatic regions of the genome that are localized to the nuclear interior, and contain predominantly actively transcribed gene bodies. While Compartment B contains mostly heterochromatic genome

segments that are localized to the nuclear lamin and nucleolus[11–14]. Swapping between compartments appears to be dynamic and mediated by chromatin interacting proteins such as transcription factors[12,15,16]. Transcription factors CTCF and YY1 have demonstrated the ability to drag genomic regions from compartment A to compartment B, effectively silencing genes of that region, however cell-wide depletion of CTCF did not impact the establishment or maintenance of A/B compartmentalization, suggesting CTCF does not act in this tier of chromatin organization[17,18,19].

Within each A/B compartment, the genome is further segregated into insulated neighborhoods by the formation of topologically associating domains (TADs). TADs establish regions of the genome, spanning several MBs, that can interact with elements within the TAD boundary, but not with nearby elements outside the boundaries. 92% of mammalian TAD boundaries are established by CTCF-Cohesin DNA-protein complexes[19–21]. CTCF-independent TADs are demarcated by RNA polymerases, build-up of nascent RNA transcripts, and/or transcriptional activators[19,22–25]. A subset of CTCF-independent TADs are still occupied by cohesin and have a broader looping structure more similar to A/B compartmentalization[2,21,26]. Transcription factors KLF4 and OCT4 have been shown to interact with cohesin and may act to create these TADs[27,28]. Single cell analysis and single molecule imaging studies of genome organization found that A/B compartmentalization remains largely stagnant with any one gene remaining in their original compartment, while TADs are comparatively variable and genes can fluctuate between CTCF-cohesin loops suggesting a level of flexibility at this tier of genome organization[29–33].

TAD construction is explained by the loop extrusion model. In this model, cohesin threads along the genome until it collides with a CTCF, bound to the DNA via a CTCF binding site (CBS). The genome is then looped through the ring-like cohesin complex until it binds with a convergently oriented DNA-bound CTCF[34,35,36]. The orientation of the CTCF bound to the genome dictates the boundaries of the cohesin-extruded loop[36]. A TAD will only form between two CTCFs bound in opposite orientation of each other with the N-terminal domains arranged towards the inside of the loop. TADs can contain nested, smaller TADs further compartmentalizing the genome into smaller scale subTADs that facilitate fine-tuned gene insulation or gene activation, via promoter enhancer looping, within gene clusters[19,21,22,37–39]. Studies on CTCF localization through mitosis reveal CTCF-dependent TADs assemble on the genome in a bottom-up order, where CTCF binds across the genome first followed by a delayed and gradual accumulation of cohesin[40,41]. The smaller subTADs are built first followed by larger TADs. While TADs are highly conserved across cells and species, sub-TADs are not conserved across species or even cell types and function as flexible, topological mechanisms of gene regulation within the established TAD **(Figure 1.1a)**[42].

CTCF is a key regulator of genome organization. Global depletion of CTCF in embryonic stem cells, as well as dividing and non-dividing terminally differentiated cells, resulted in loss of 82% of TAD formation and impacted transcriptional activity for ~⅛ of all protein coding genes[19]. Although a global disruption of TAD structures only leads to gene expression changes in a fifth of total mammalian protein-coding genes, depletion of CTCF had the largest impact on genes regulated by clusters of enhancers (superenhancers), suggesting a role in gene regulation through enhancer recruitment and restraint[43,44]. In addition, CTCF has been found to regulate mRNA

splicing by influencing the rate of transcription and implicated in promoting homologous recombination repair at double-strand breaks[45,46.47]. The many roles of CTCF in the eukaryotic cell can be attributed to the many domains of CTCF and their function. CTCF can be split into three main regions; 1) the amino-terminal domain spanning amino acids 1-248, 2) the central DNA binding domain containing an 11-finger Cys_2His_2 zinc finger array, and 3) the carboxy-terminal domain spanning amino acids 580-727 **(Figure 1.1b)**. Cys_2His_2 zinc finger architecture is the most common among transcription factors in eukaryotic cells, and not unique to CTCF. Zinc finger arrays following the Cys_2His_2 architecture can often function as recognition domains to facilitate protein-DNA, protein-RNA, and/or protein-protein interactions[48]. The zinc finger array of CTCF is no exception as CTCF has been found to bind to DNA, RNA, and form protein complexes[49–52]. These interactions will be further discussed in chapters 1, 2, and 3 respectively.

CTCF binds across the genome via a 40 bp binding site (CBS) and clusters at TAD boundaries as well as at promoter and enhancer domains within TADs, due to its duplicative role in promoter-enhancer looping[12,53]. Gene activation by promoter-enhancer looping subTADs, established by CTCF, reduces the distance between the promoter region of genes and enhancer-bound transcription factors[44,54,55]. CTCF-formed TADs can have an insulatory effect on gene transcription if the TAD boundaries exist between genes and enhancer regions, but CTCF can also repress gene expression by binding on top of the transcriptional start site of genes or by RNA Polymerase II stalling[19,45]. Sequence mutations of CBSs as well as methylation can result in loss of CTCF binding, disruption of TAD and subTAD formation, and subsequent gene misregulation and disease[39,56–59]. The accumulation of indels and substitutions in CBSs has been

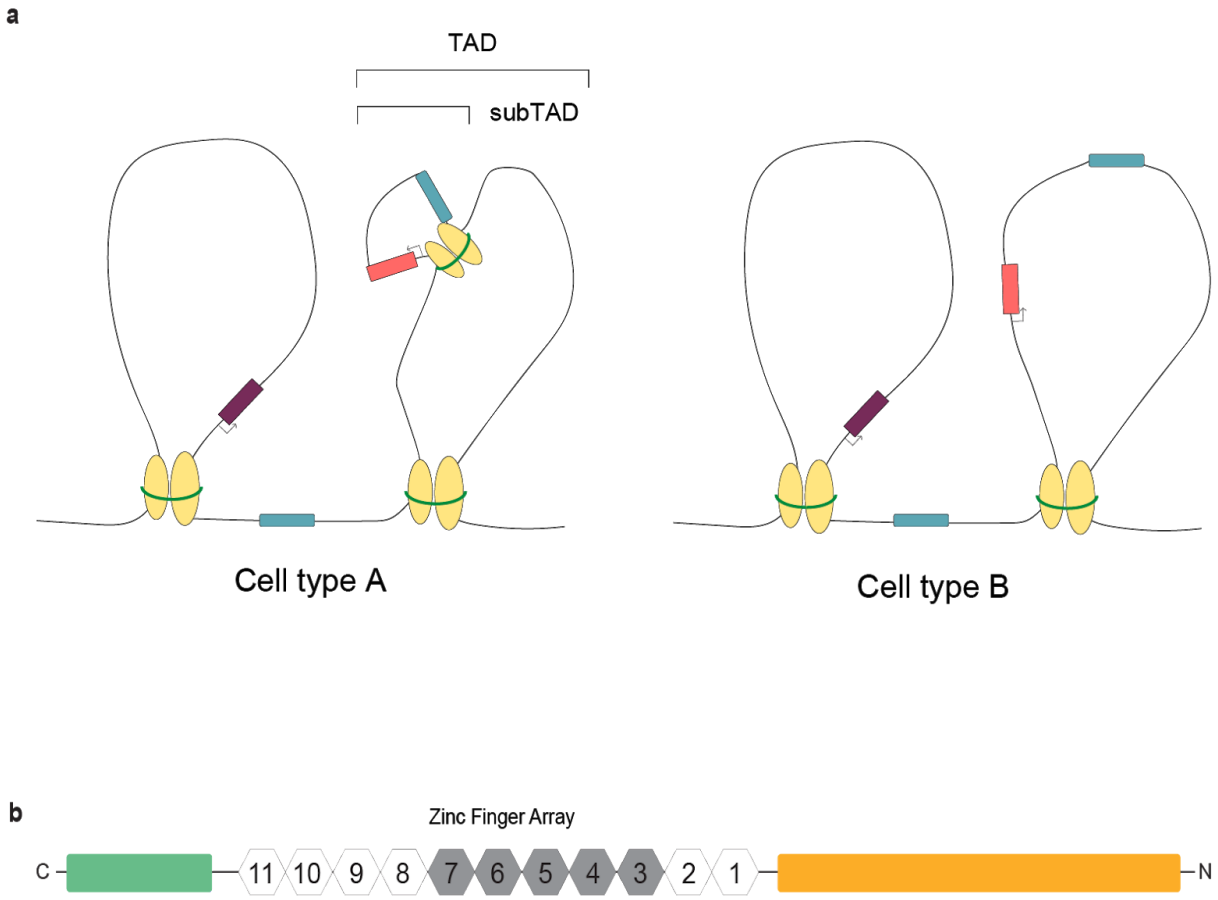

Figure 1.1: CTCF dependent topologically associated domains are established by CTCF binding to the genome via DNA binding domain. a, topologically associated domains are established by CTCF (yellow) occupancy of CTCF binding sites and subsequent looping directed by Cohesin (green ring). Genes and enhancers are depicted as boxes and ovals respectively. Genes and enhancers on either end of TAD boundaries have little interaction with each other. Genes and enhancers within a TAD will have more freedom to interact. Smaller subTADs exist within the larger insulated neighborhoods and act to direct promoter-enhancer loops. TADs are conserved across species and cell types while subTADs vary significantly. **b,** CTCF consists of three domains: N-terminal domain (orange), central DNA binding domain consisting of an 11-finger Cys_2His_2 zinc finger array, and C-terminal domain (green).

linked to chromosomal instability and tumorigenesis in gastrointestinal cancer and melanoma[60,61]. Cancer-specific CTCF binding patterns were identified in six cancer types in an analysis of over 700 CTCF ChIP-seq profiles from human tissue and cancers, a portion of which was related to CBS mutations resulting in loss of CTCF binding and the misregulation of tumor suppressor and oncogenes[10]. A trans-TAD genomic duplication in the *SOX9* gene region results in the destruction of the existing TAD and the formation of a new TAD which results in the misexpression of a previously excluded gene *KENJ2*[62]. Disruption of a TAD boundary at the gene-dense limb development loci, by genomic inversion or duplication, results in loss of association of enhancers to the limb-specific *EPHA4* gene. The same disruption leads to increased association of the *EPHA4*-specific enhancers to *WNT6-IHH* or *PAX3* resulting in concurrent loss of *EPHA4* expression and ectopic activation of *WNT6, IHH* or *PAX3*. The destruction of one TAD and formation of another, and subsequent gene misregulation, results in digit malformations in humans[39]. In all these cases, the mutations to the 40 bp CBS that result in loss of CTCF occupancy are not consistent and include deletions, insertions, or multiple substitutions. Therefore it is not clear which portion of the CBS is responsible for maintaining CTCF occupancy and TAD formation.

Due to CTCF's ubiquitous nature in the cell, it is difficult to investigate structure-function studies at CTCF regulated gene loci. Which mutations within the CBS result in loss of CTCF binding and gene misregulation is not well defined. The disruption and formation of TADs in disease cases suggest TAD level of chromosomal organization is flexible. Single-molecule imaging studies provide evidence for gene fluctuation between TADs, suggesting continuous formation and dissolution of CTCF-cohesin loops. Based on this, it may

be possible to manipulate the higher-order organization of the eukaryotic genome for targeted gene regulation. We set out to determine if we could define critical bases within the CBS that lead to loss of CTCF binding, evolve CTCF variants with new sequence specificity to target mutated CBSs and use this system to determine cofactors of CTCF promoter-enhancer looping at an endogenous site, and finally use engineered CTCF variants for alteration of gene expression by introducing variant-specific CTCF-mediated genome reorganization.

Chapter 1:

Utilizing bacterial selection for engineering CTCF proteins with novel sequence recognition

Rebecca T. Cottman[1,3], J. Keith Joung[1,2]

[1]Molecular Pathology Unit, Center for Cancer Research, and Computational and Integrative Biology, Massachusetts General Hospital, Charlestown, MA, [2]Department of Pathology, Harvard Medical School, Boston, MA, [3]Department of Biological and Biomedical Sciences, Harvard Medical School, Boston, MA

All authors listed contributed to the work described in this chapter. I designed and performed wet-lab experiments. Keith Joung oversaw research and provided direction.

Introduction

CTCF is a transcription factor conserved in most eukaryotes with the exception of yeast, C. elegans, and plants[1]. CTCF was first identified as a regulator of *c-MYC* (*MYC*) expression, it has since then been identified as a key regulator of higher order genome organization[35,49,63]. CTCF binds throughout the genome via a highly conserved 11-finger zinc finger (ZF) array with a Cys_2His_2 architecture that recognizes a 40bp CBS, of which a core region of 15 bp forms the motif defined by JASPAR motif analysis of ChIP-seq datasets **(Figure 1.2a)**[48,64–66]. Co-crystal structure of the 11-finger ZF array bound to its DNA substrate suggests that only ZFs 3-7 of the 11-finger ZF array appear to make protein-DNA contacts[67]. The site of protein-DNA contacts overlaps with the highly conserved 15bp core sequence (**Figure 1.2a**), confirming the source of conservation of particular base pairs within the larger CTCF binding site. In contrast, ZFs 8-11 and ZFs 1-2 do not appear to mediate sequence-specific contacts, which may be why the conserved motif does not extend beyond the 15 bp core sequence within the targeting range of ZF3-7 (**Figure 1.2a**)[67]. In fact, structures of CTCF zinc finger array bound to its DNA substrate were obtained with high resolution for only zinc finger 2-9 with the other fingers not visible in

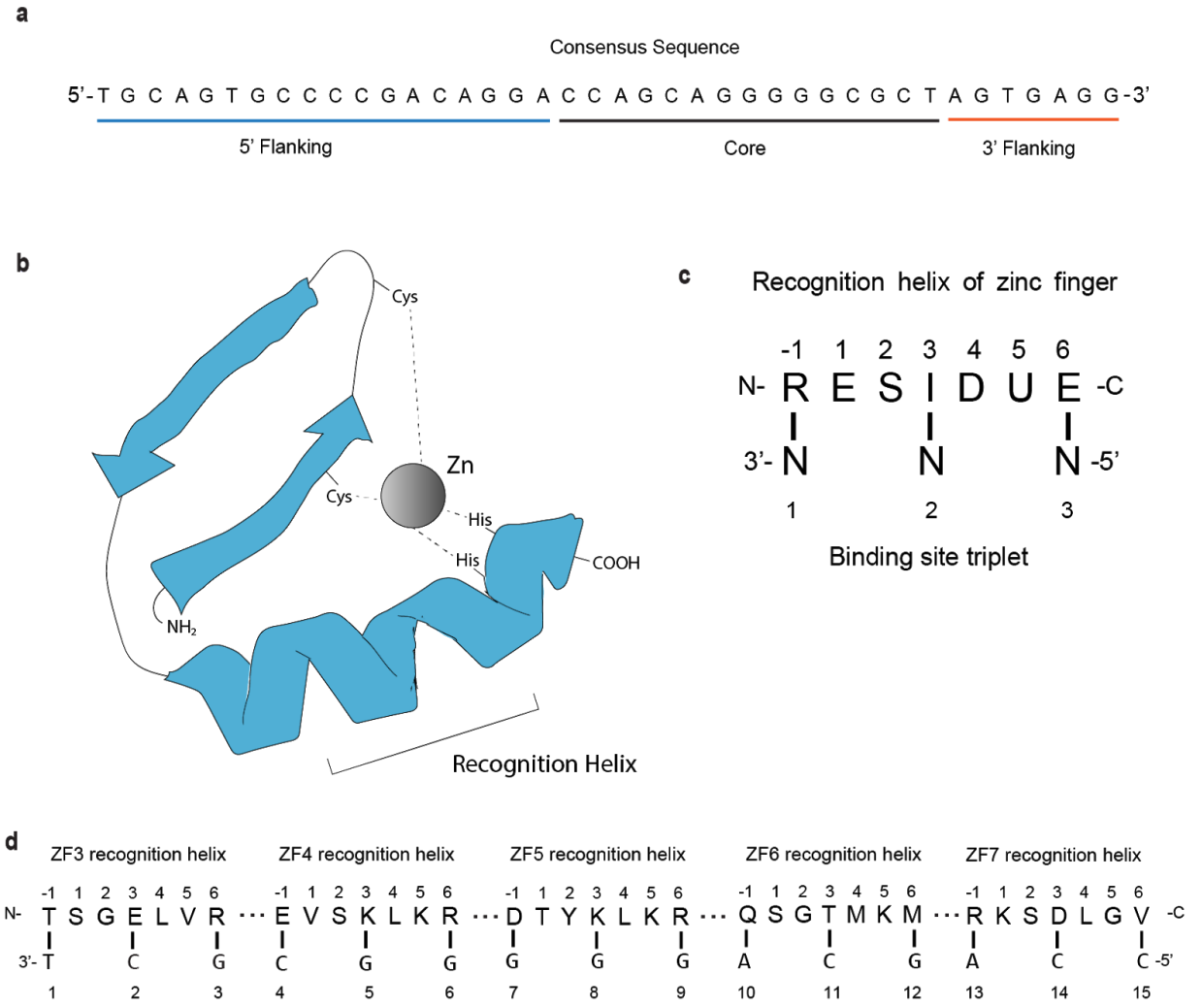

Figure 1.2: Zinc finger array of the CTCF DNA binding domain makes Protein-DNA contacts to the partially conserved CTCF binding site. a, Representative sequence of CTCF binding site. Core sequence is conserved and forms the motif derived from ChIP experiments (black underline). 5' and 3' flanking regions (blue and orange underline, respectively) are not detected by ChIP experiments, but are required for CTCF binding. **b,** Structure of a single Cys_2His_2 zinc finger in complex with a Zn+ molecule. The alpha helix contains the residues that coordinate protein-DNA contacts with the sequence of the binding site. **c,** Simplified diagram of the protein-DNA contacts typical of Cys_2His_2 zinc finger architecture. Residues of the recognition helix are numbered -1 through 6, with a subset of residues forming direct protein-DNA contacts (black bars) with a triplet of DNA sequence. **d,** Residues of each zinc finger recognition helix and the interacting DNA sequence. Zinc fingers 3-7 of the zinc finger array of CTCF make contacts with the 15 bp core motif of the CTCF binding site. Residues are numbered in relation to the recognition helix of each zinc finger, the core region of the CBS is numbered 3'-5' direction.

the structure, providing evidence they are not involved in targeting the flanking sequences of the CBS[67]. However, ChIP-exo studies of lenti-virally integrated CTCF constructs implicate the importance of ZFs 8-11 and ZFs 1-2 in maintaining CTCF occupancy genome-wide[68]. Single peptide changes in the first or second zinc-coordinating histidine within any one of the 11 ZFs resulted in a reduction in occupancy of CBSs genome-wide with the most severe reduction observed in constructs with mutations to ZFs 3-7. This, combined with DNaseI footprinting suggests that although ZFs 3-7 appear to be critical for protein-DNA interactions, the remaining fingers in the array are required for optimal binding to the target CBS in a sequence-independent manner[68,69]. It remains unclear why all 11 fingers in the array impact CTCF occupancy at CTCF binding sites. As is the case with other DNA binding proteins utilizing Cys_2His_2 zinc finger arrays such as TFIIIA, WT1, or GATA-1, ZFs 3-7 could be responsible for protein-DNA contacts while the other fingers in the array could be coordinating protein-RNA or protein-protein interactions that may stabilize CTCF on the genome[70–72].

The Cys_2His_2 class of zinc finger arrays is the most common protein architecture found in the DNA binding domains of eukaryotic transcription factors[48]. The first observation of the Cys_2His_2 architecture was made in a study of Transcription Factor IIIA (TFIIIA). Much of the architecture was later elucidated from further studies with EGR1, also known as Zif268[48]. Zinc fingers of this architecture consist of two anti parallel beta sheets followed by an alpha helix that contains the residues forming the protein-DNA contacts. The two highly conserved Cysteines and Histines of each zinc finger coordinate a Zn+ ion that maintains the finger-like fold of the protein **(Figure 1.2b)**. This allows for the orientation of the alpha helix to recognize and bind to the major groove of the DNA, with only a slight widening of the major groove to accommodate

the array. Recognition of a target sequence is coordinated by the recognition helix, defined as residues -1 through 6, of each zinc finger in tandem with residues -1, 2, 3, and 6 making Van der Waals contacts between the protein and DNA **(Figure 1.2c)**. ZF3-7 of CTCF makes protein-DNA contacts with the 15 bp core sequence of the 40 bp CBS in this fashion **(Figure 1.2d)**. Linkers connecting the fingers of the array typically conform to the consensus amino acid sequence 'TGEKP'. Each finger in the array will recognize a triplet of base pairs in the recognition sequence; in the case of Zif268, three zinc fingers target a 9 bp sequence of DNA. CTCF has 11-fingers of which only 5 make protein-DNA contacts to recognize the 15 bp core region of the 40 bp CBS[67,68]. Attempts to expand the target sequence of Zif268 beyond 9 bps were initially made by adding more zinc fingers to the array with a simple 'TGEKP' linker, but any addition beyond three fingers had moderate improvements in binding affinity to the expanded target sequence[73].

Accumulation of substitutions and indels in CBSs has been implicated in tumorigenesis of gastrointestinal cancers and melanoma[60,61]. Inversions disrupting CBSs during embryogenesis results in developmental disorders, such as fragile-x syndrome, due to disruption in TAD boundaries[62,74,75]. CTCF binding sites have been reported to be mutational hotspots in cancer with cancer-associated mutations localized to the core sequence of the CTCF binding site in primary samples from gastrointestinal cancer patients and with accompanying atypical gene expression profiles of oncogenic and tumor suppressor genes[61]. Small deletions of CTCF binding sites have also been shown to lead to loss of expression of genes such as *MYC* and *PTGS2*, which both play a role in cancer development[76,77]. It would therefore be therapeutically relevant to engineer CTCF to bind to pathogenic CBS mutations and restore gene regulation. However, the

heterogeneity of alterations in the CBS sequence that results in loss of CTCF occupancy make it difficult to know which mutations are causal. EMSA-based *in vitro* studies determined base changes within the 15 bp core region resulted in reduced CTCF occupancy, however this study investigated a handful of 'triplet' mutations, where a triplet of DNA was altered at a time, and did not perform an exhaustive analysis of all possible single base alterations across the CBS[78].

The DNA binding domain of CTCF is composed of an 11-finger zinc finger array that follows the Cys_2His_2 architecture common in eukaryotic transcription factors and found in Zif268. Attempts to engineer Cys_2His_2 zinc finger arrays to target novel sequences have mainly focused on three-finger Zif268 due to its modularity and previously well characterized structure and protein-DNA interactions. It is therefore feasible to apply the previously developed selection methods used to evolve Zif268 sequence specificity to CTCF.

There are two competing selection systems for creating Cys_2His_2 zinc finger arrays that recognize novel sequences: Phage display and the bacterial-two-hybrid method. Phage display was first described as a method for clonal expansion of desired gene products or a method for raising antibodies against a displayed peptide[79]. The first application of phage display as a selection system was in selecting a library of mutant human growth hormone proteins (hGH) fused to the C-terminus of the filamentous gene III and for binding to the wild type receptor[80]. This was the first demonstration of the use of phage display to study and select for novel protein-protein interactions. Finally, it was applied to evolving proteins with novel protein-DNA interactions by fusing the short Zif268 coding sequence to the C-terminal end of the filamentous phage III gene and selecting libraries of Zif268 with degenerative coding sequence to bind to, at first, 3 base changes in the 9 base pair recognition sequence[81–84].

The bacterial-two-hybrid (B2H) selection method was developed later as an alternative, and more efficient, selection method for Zif268 variants with novel protein-DNA and protein-protein interactions[85,86]. Originally, the B2H method was applied as a screen for protein-protein interactions, where protein interacting domain (A) was fused to a DNA binding domain with known binding capacity for a DNA sequence. Another protein interacting domain (B) was fused to the alpha-subunit of *E. coli* RNA polymerase. Interaction of protein A and B would result in the assembly of transcriptional machinery at the TSS of a weak promoter upstream of a *lacZ* gene. Protein-protein interaction could then be screened with a colorimetric assay detecting beta-galactosidase, the product of *lacZ* expression **(Figure 1.3a)**[87,88]. The B2H selection system was adapted from this work by replacing *lacZ* with *HIS3-aadA* construct and replacing protein A and B with modified portions of yeast derived proteins Gal11P and Gal4[85]. Gal4 is fused to the alpha subunit of *E. coli* RNA polymerase while Gal11P is fused to the protein of interest that will have varying ability to recognize and bind a target sequence upstream of a minimal promoter directing expression of the selective yeast gene *HIS3* that can recover histidine biosynthesis is *E.coli* strains with *ΔHisB* **(Figure 1.3b)**[85]. A phagemid library with variation in residues -1 through 6 of the recognition helix of the most C-terminal zinc finger of the three-finger Zif268 zinc finger array was fused to Gal11P and through selection generated novel Zif268 variants capable of binding to three unique 3 bp subsite of the native 9 bp target sequence. The B2H selection system could negatively screen through a larger library pool of variants than the B2H reporter system and could produce proteins with altered sequence

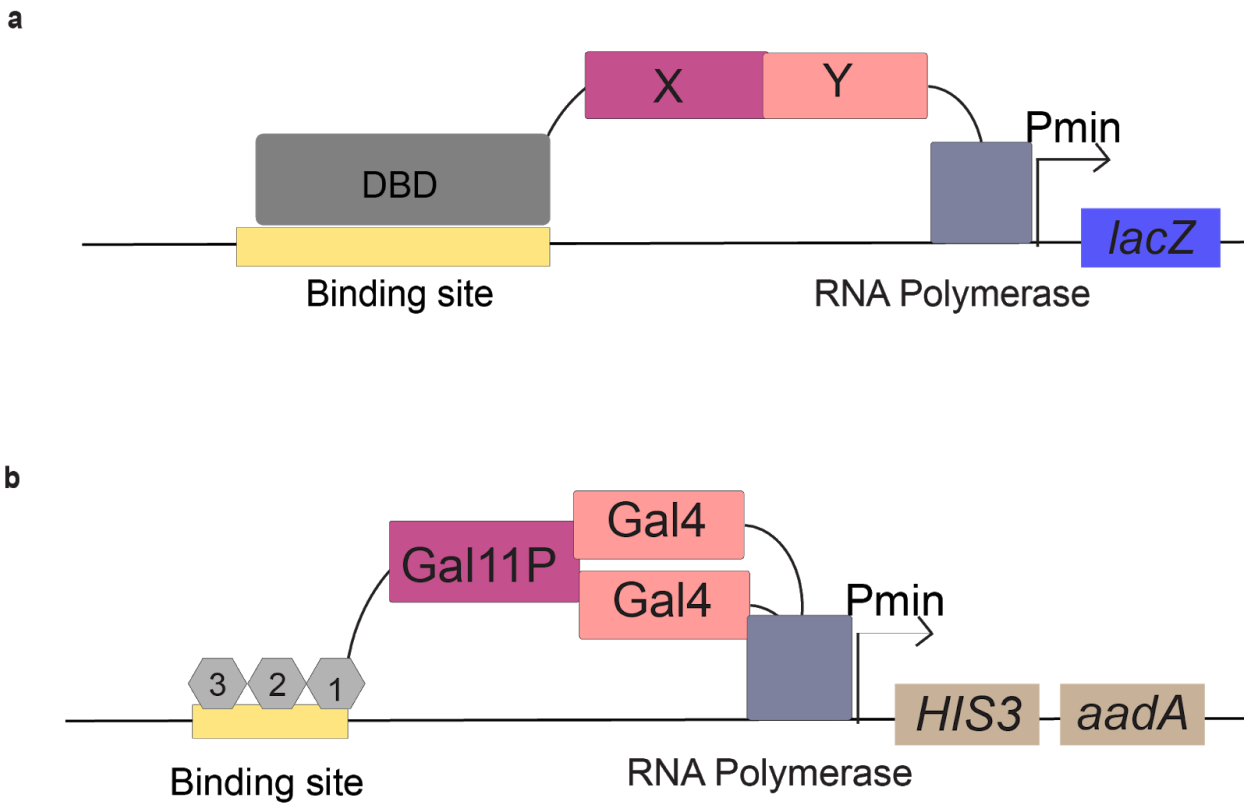

Figure 1.3: Previously described bacterial-two-hybrid (B2H) systems for screening of protein-protein and selection of protein-DNA contacts. a, B2H reporter assay of Protein-Protein interaction between protein X and Y developed by Hochschild et al.(87,89). Protein interactions were assayed by fusing protein X to a DNA binding domain with known affinity for a binding site (yellow box) upstream of a reporter gene (*lacZ*). Expression of the reporter gene is dependent on interaction of Protein X with protein Y, which is fused to an alpha subunit of RNA polymerase and is subsequently recruited to the minimal promoter. **b,** Diagram of B2H selection system developed by Joung et al., (85) for Protein-DNA interactions. Modified versions of yeast derived proteins (Gal4 and GallP) are used to replace the protein-protein interaction of X and Y. The *lacZ* reporter gene is replaced with *HIS3* and *aadA*. 3 finger array of Zif268 is fused to Gal11p and survival of the bacteria is dependent on successful binding of the zinc finger array to the sequence of the binding site.

specificity in fewer rounds of selection than the phage display method[85]. This method has also been applied to selection of novel protein-protein interactions, with minimal breakthroughs of background variants [86].Both selection methods take into consideration the nature of Cys_2His_2 zinc finger arrays and the components of the protein structure that influences protein-DNA and protein-protein interactions. Here we adapted the B2H system to create a reporter assay and selection for identifying critical bases within the CBS binding site for maintaining CTCF occupancy and then CTCF variants that can bind to those sequences. In this chapter, the protein-DNA interactions of ZFs 3-7 of the CTCF zinc finger array will be explored further and a bacterial-two-hybrid selection system will be applied to generate CTCF variants with novel sequence affinity.

Results

CTCF relies on critical residues within the CBS motif to maintain DNA contact.

CTCF is composed of a central DNA recognition domain composed of an eleven-finger zinc finger array (**Figure 1.1b**). Crystal structure of the CTCF zinc finger array bound to double-stranded DNA indicates that zinc fingers 3 through 7 make protein-DNA contacts with the 15 bp core motif of CBS and direct the sequence specificity of CTCF across the genome **(Figure 1.2d)**[67]. Indels or duplications disrupting the core motif of CBSs, which may have resulted in loss of CTCF binding, have been linked to developmental defects, instability of the genome, as well as associated with tumorigenesis[60,61,74]. Because of the heterogeneity of mutations that result in loss of CTCF binding and onset of disease, it is difficult to determine which bases within the CTCF binding site are critical for CTCF occupancy. In order to engineer CTCF variants with orthogonal CBSs that wild type CTCF could not bind to, we first needed to determine which bases were critical to maintaining CTCF occupancy.

To do this, we developed a bacterial-two-hybrid (B2H) reporter system with a *lacZ* reporter regulated by CTCF binding **(Figure 1.4a)**. First we determined the components of the CTCF DNA binding domain and CTCF binding site required for optimal binding in the B2H reporter system. We screened ga11p fusions of different subsets of the zinc finger array on a selection of known CBSs from previously published studies **(Figure 1.4b)**. We determined a full length 40 bp consensus CBS, composed of 5' flanking, core, and 3'flanking regions, paired with

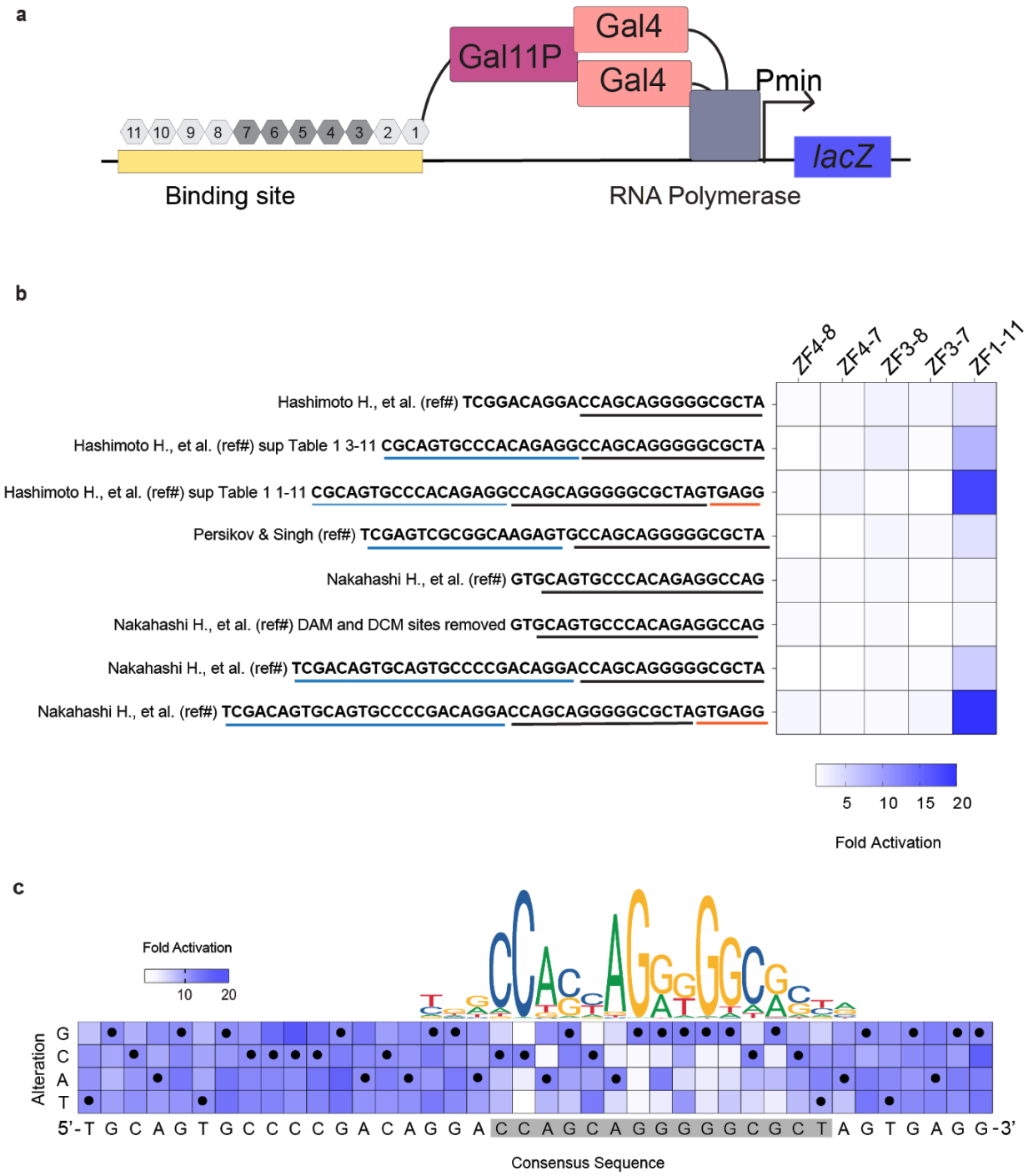

Figure 1.4: Optimization of CTCF zinc finger array and CTCF binding site sequence in an adapted bacterial-two-hybrid reporter system. a, Diagram of CTCF zinc finger array and binding site in the B2H reporter system. Full length or subsets of the 11-finger array of CTCF were fused to gal11P, zinc fingers involved in Protein-DNA contacts (dark gray) and those with

no evidence of making contact to the DNA (light gray) are indicated. Expression of the *lacZ* reporter gene is dependent on binding of the CTCF zinc finger array to the binding site. **b,** Optimization of the components of the CTCF zinc finger array and CTCF binding site that is required for optimal activation of the reporter gene. Heat map reflect fold activation of the reporter gene over background expression. Columns are labeled with the portion of the zinc finger array fused to gal11p; the rows are marked by binding site sequences and their source. Binding sites tested contained different components of the full 40 bp CTCF binding site as indicated by black underline (core), blue underline (5' flanking), or orange underline (3' flanking). **c,** Consensus CTCF binding site (CBS) with the 15 bp core region highlighted in gray. Heat map reflects fold activation of *lacZ* reporter gene above background for a single bp change at the indicated site in the consensus sequence. Each row of the heat map indicates the labeled base at that position with black dots indicating consensus nucleotide sequence at that position. Each cell represents mean fold activation of triplicate experiments.

the full 11-finger zinc finger array, led to the strongest CTCF binding. To identify the critical bases in the CBS necessary for CTCF binding, we introduced single substitutions of the remaining possible bases at each position of the binding site and assayed the ability of CTCF ZF array to bind using the B2H reporter assay. We discovered that certain single base-pair substitutions within the 15 bp core motif region (highlighted in gray) resulted in complete loss of CTCF binding, and substitutions outside of this core region of the binding site had little impact **(Figure 1.4c)**.

Engineering CTCF variants to bind to variant binding sites.

Next, utilizing a B2H selection system we engineered CTCF variant zinc finger arrays that could bind to the variant CBSs (vCBS) with single alterations in the bases critical for CTCF binding **(Figure 1.5a)**. The B2H selection is similar to the reporter system except instead of binding driving the expression of *lacZ*, successful binding results in survival of the clone via the expression of *blaC*, an antibiotic resistance gene. In this way a library of zinc finger array variants can be selected to bind to a new target sequence. The CTCF zinc finger array follows the Cys_2His_2 zinc finger architecture and as such, the recognition helix of each ZFs 3-7 recognizes a triplet of bases in the CBS **(Figure 1.5b)**. Amino acids at positions -1, 2, 3, and 6 of the recognition helix, relative to the first residue in the alpha helix of each zinc finger, establish the protein-DNA contacts with the CBS, however zinc finger arrays make context-dependent interactions and so all residues of the recognition helix must be considered when designing selections targeting an altered binding site[48]. Therefore, libraries of zinc fingers were constructed with degenerative residues limited to within the DNA-recognition helix of zinc finger 3-7, and

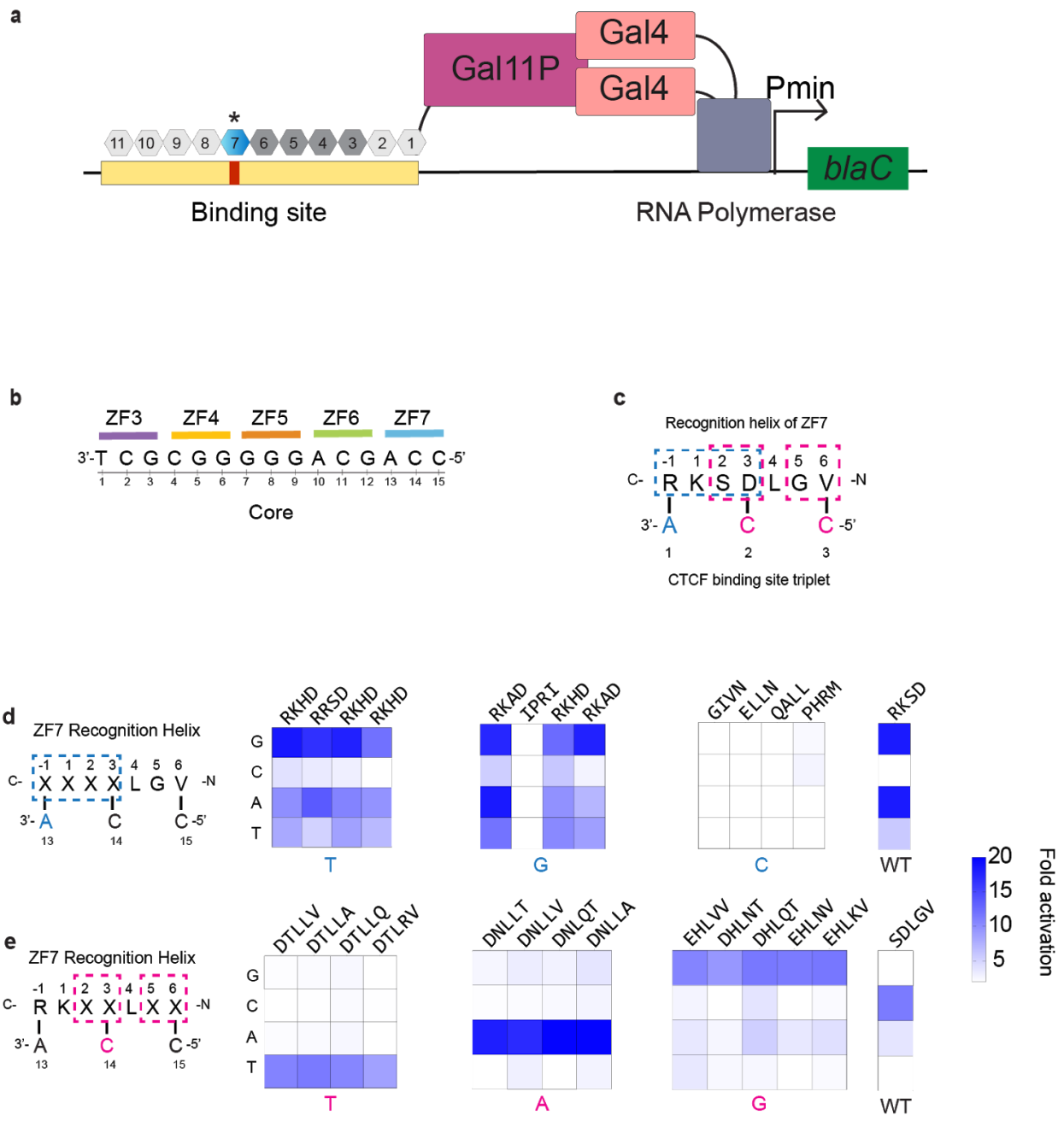

Figure 1.5: B2H selection system for generating CTCF variants with novel sequence recognition. a, A diagram of the B2H selection system used to select CTCF variants within one finger of the zinc finger array for a single base change in the CTCF binding site. Zinc finger containing VNS codons for residues in the recognition helix is marked with an asterisk and multicolored to represent a complex library of zinc finger variants within the array. Survival of

colonies is dependent on successful binding of CTCF variant to the modified sequence, the consensus CBS (yellow) with single base change (red rectangle), and subsequent expression of *blaC*, a penicillin resistance gene. **b-c,** Detail of core sequence within the consensus CBS and the zinc fingers within the array that coordinate protein-DNA contacts with each triplet. The positions within the core are numbered from 3'-5' for future reference. **d,** ZF7 recognition helix, with the protein-DNA contacts detailed (black lines), recognizes bases 13-15 of the core sequence within the CBS. The 'A' base at position 13 was altered to a 'T', 'G', or 'C' and the library of CTCF zinc finger array variants were constructed with 'VNS' codons encoding the residues at the indicated positions (X) within the ZF7 recognition helix. CTCF variants were tested for sequence specificity in the B2H reporter assay. Heat maps of mean fold activation above background in triplicate. Wild-type zinc finger array (RKSD) base specificity. Sequence preference of surviving variants shown for each base change in the CBS. Surviving variants were sequenced and their residues at the (X) positions are detailed above the heat map of binding ability as assayed by the previously described B2H reporter system. The blue base under each heat map indicates the base the variants were selected to bind to, base sequences labeling the rows indicate the base present at the same position in the B2H reporter. **e,** Schematic of library construction for ZF7 recognition helix selected to bind to alterations in the 2nd position of the triplet DNA sequence. (X) Indicate residues encoded by 'VNS' codons, 'C' base altered to 'T', 'A', or 'G'. Performance of surviving variants (sequences above heat maps) selected to bind to a single base change (listed below heat map) in the CBS were assayed for ability to bind to all possible sequence changes at that position.

only to the subset of residues that would have an impact on binding to the corresponding single base change in the CBS. The library of CTCF zinc finger arrays with degenerative residues was prepared by introducing 'VNS' at codons corresponding to the residues -1,1,2,3 or 2,3,5,6 of the recognition helix, depending on the location of the alteration in the CBS **(Figure 1.5b-c)**. We did not alter residue 4 of the recognition helix because it faces the internal core of the ZF domain and is not expected to make contacts to the DNA, but instead serve to stabilize the structure of the alpha-helix[48,90]. CTCF zinc finger array libraries were selected to bind to single base alterations at each critical residue of the CBS using the B2H selection system (**Figure 1.5a**). A sub-set of surviving clones were assayed for their ability to bind to their target vCBS using the previously described B2H reporter system and found that we had three categories of variants; specific, relaxed, weakened **(Figure 1.5d-e, Figure 1.6)**.

We were interested to see if the sub-set of evolved zinc finger arrays we characterized from the selection was representative of the population of surviving variants. We scraped the colonies growing on the highest selection stringency, isolated the zinc finger array encoding plasmid and prepared it for NGS. The resulting sequences were translated and aligned to reflect the amino acid in position -1 through 6 of the recognition helix for each selection of ZF7 libraries on a C to A, G, or T base change in the CBS **(Figure 1.7)**. The initial clones expanded from the selection appeared to be representative of the dominant variants in the population (**Figure 1.5d-e**).

We next wanted to make a vCBS that would be orthogonal to CBSs in the human genome. Multiple changes were introduced to the 40 bp CBS, all within the 15 bp core region, to generate five different vCBSs (vCBS1-5) that were then used in the B2H selection system. Five

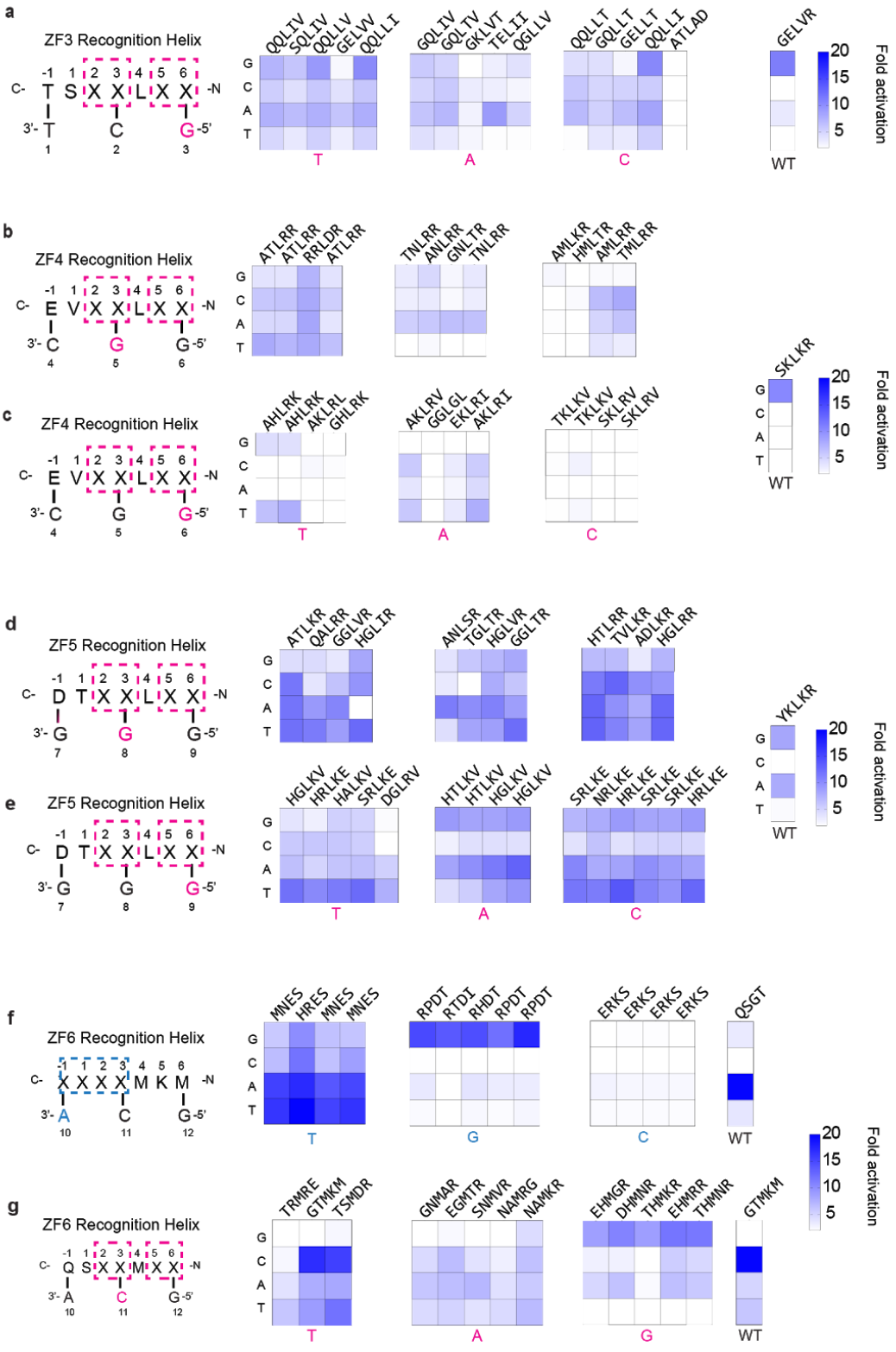
a
ZF3 Recognition Helix
b
ZF4 Recognition Helix
c
ZF4 Recognition Helix
d
ZF5 Recognition Helix
e
ZF5 Recognition Helix
f
ZF6 Recognition Helix
g
ZF6 Recognition Helix
Fold activation
WT

Figure 1.6: Zinc finger library construction and selection for alterations to critical bases in the CBS. a-g, Library construction of recognition helix and target triplet of DNA in the CBS. Residues altered varied with 'VNS' codon encoding residues within the recognition helix (X) and the target base that was altered (colored pink or blue). Each variant pulled out of selection were assayed for binding to all possible changes in the CBS at that position in the B2H reporter assay.

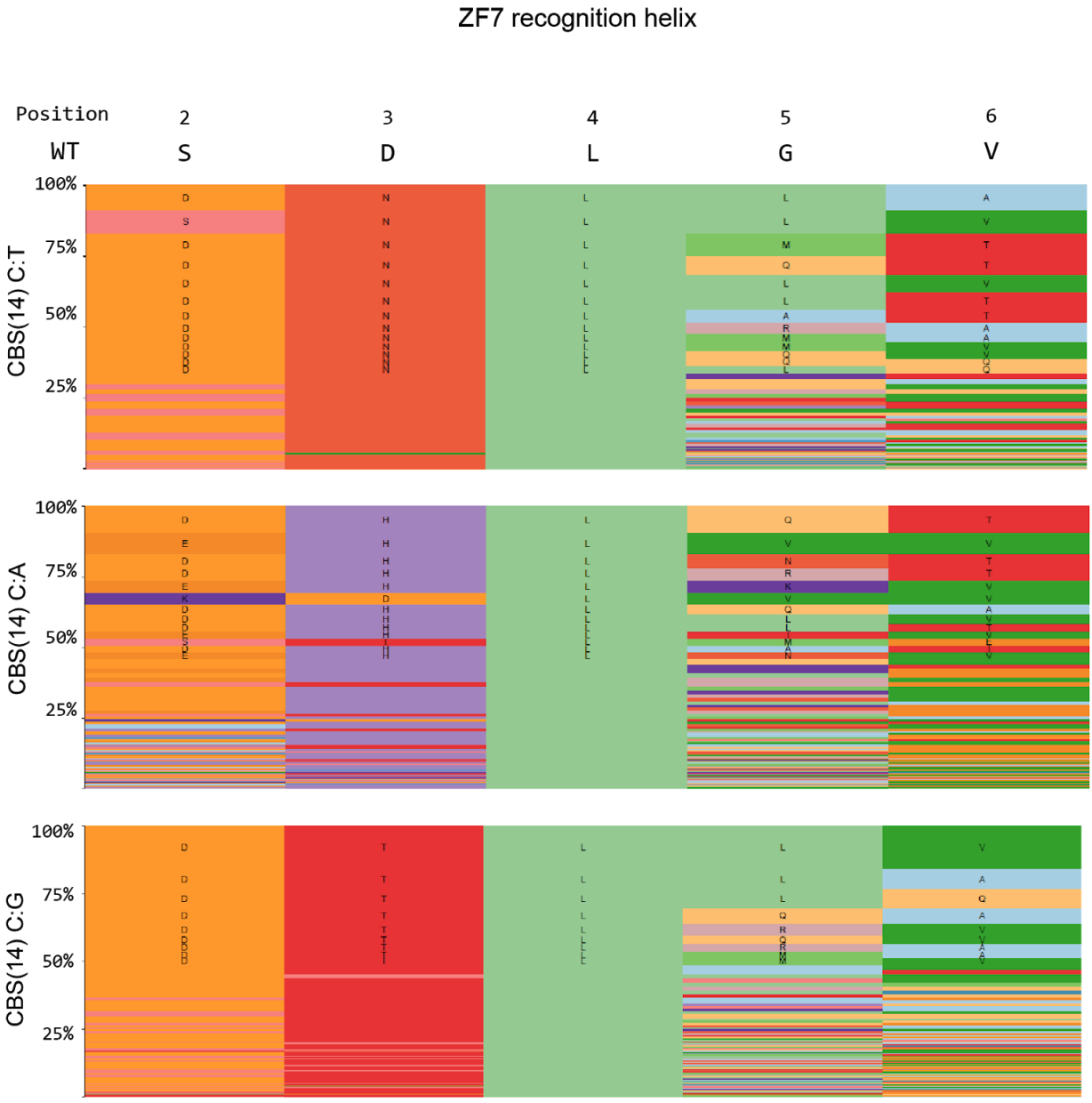

Figure 1.7: Population sequencing of selection pools demonstrates preferred residues within the recognition helix for base specific sequence recognition. Residues at varied positions in ZF7 recognition helix post selection with the library described in Figure 5e. Quilts reflect the amino acid sequence of the population of surviving variants. Wild-type residues of ZF7 recognition helix recognizing a C at position 14 of the core CBS are at the top of the quilts with the C:N change along the left side of each population.

groups of CTCF zinc finger array libraries were constructed using the pools of variants generated from the previous selections to single base changes in the CBS **(Figure 1.8a-e)**. 5 groups of libraries were selected on 5 different vCBSs using the B2H selection system. 12 colonies from each vCBS selection were assayed using the B2H *lacZ* reporter system for their ability to bind to their vCBS and the consensus CBS (**Figure 1.9a-e)**. We found that in selection groups 3, 4, and 5, many of the observed CTCF variants were represented more than once, highlighting the strength of our selection system to enrich for the small number of strong binders from a pool with an initial estimated complexity of $1.0x10^8$. In all but one vCBS, we were able to select for CTCF variants that recognized multiple alterations to the CBS that did not retain wild-type CTCF occupancy.

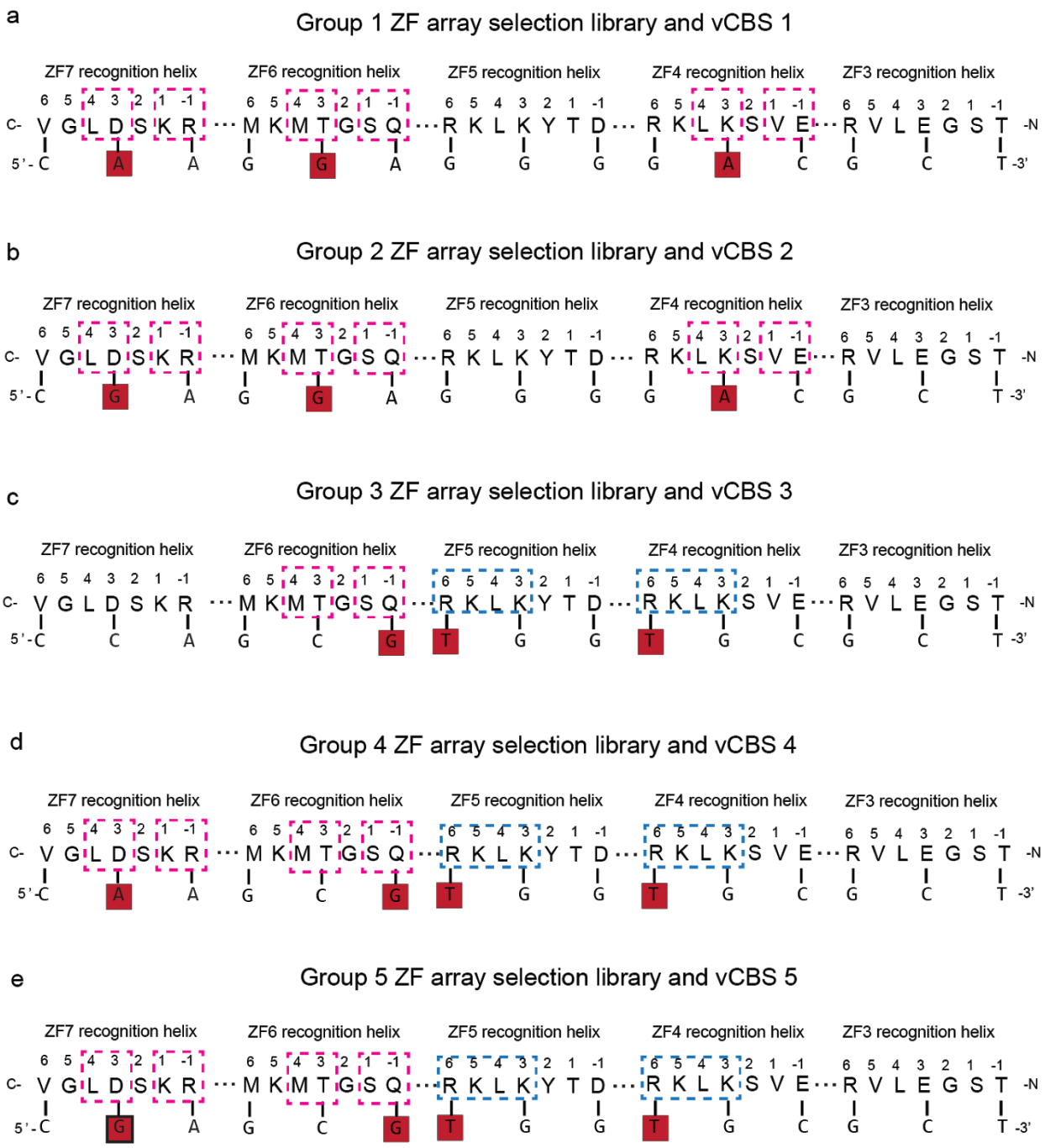

Figure 1.8: Library construction of CTCF zinc finger arrays targeting multiple simultaneous alterations to the CBS sequence. a-e, alterations to the consensus CBS are shown and highlighted with red boxes. Zinc fingers and their residues within the recognition helix requiring alteration are indicated with colored brackets. Selected variant pools from the previous selections in Figure 5 and 6 for the alteration in the binding sequence indicated were assembled into the zinc finger array at the zinc fingers indicated.

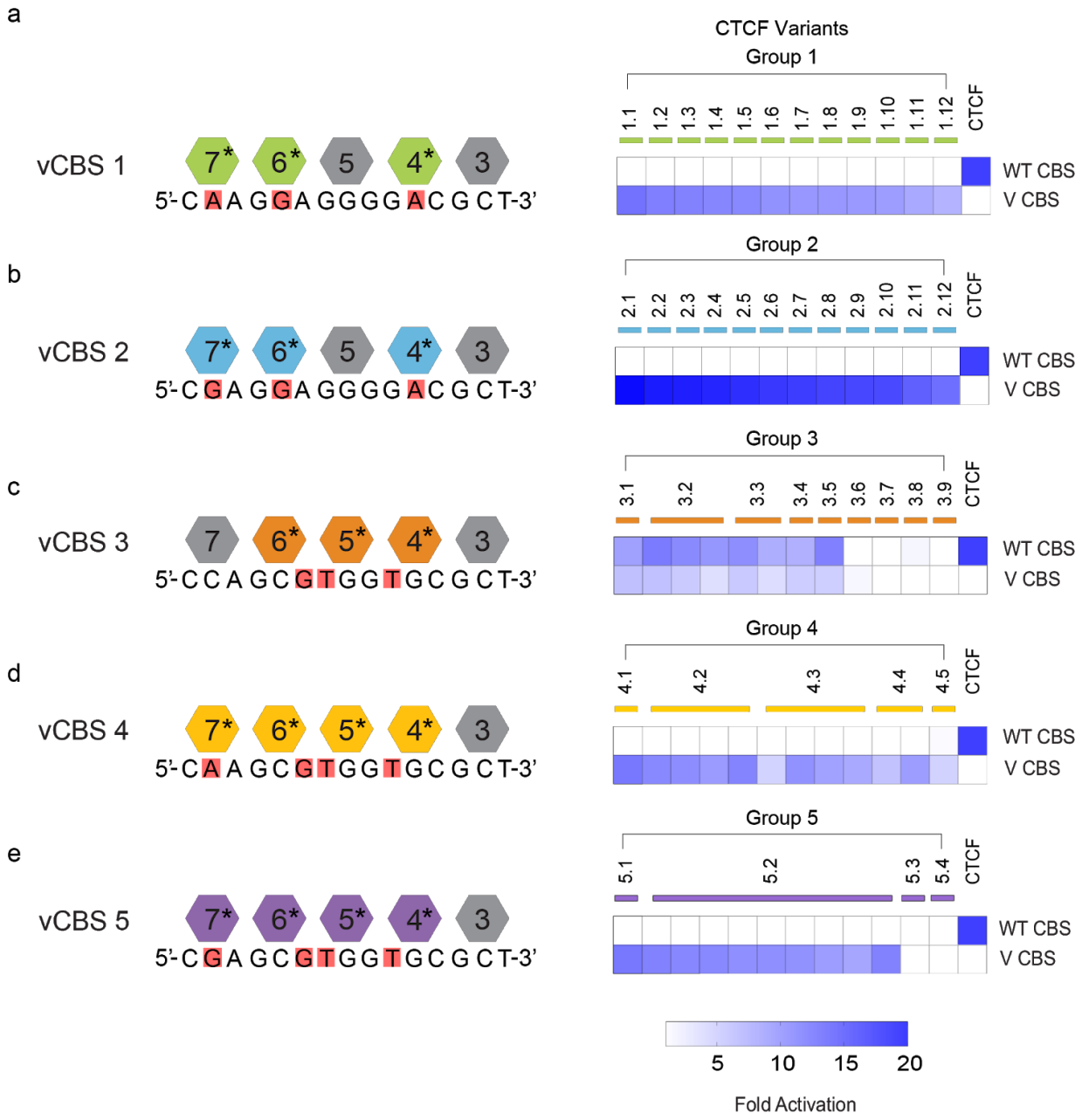

Figure 1.9: CTCF variants bind specifically to variant CBS with multiple, simultaneous sequence alterations. a-e, Variant zinc finger (colored,*) of the 11 finger array and target vCBS used in the selection. Heat map of binding affinity of surviving variants to the target vCBS sequence and consensus WT CBS. Original wild-type CTCF zinc finger array (CTCF) also assayed.

Discussion

Previous attempts to characterize the sequence requirements of CTCF binding sites have been made, with alterations made to the sequence in sets of 3[78]. Here we describe the first characterization of single base alterations made sequentially across the entire CTCF binding site. These results determine the sequence requirements within the entire 40 bp CBS for CTCF occupancy is restricted to the 15 bp core region **(Figure 1.4)**. In addition, our results corroborate the ChIP-seq derived motif for CTCF binding as well as indicate causal mutations of those reported to be enriched within CBSs in cancers[60,61,64,65]. The full 40 bp CTCF binding site was required for optimal occupancy of wild-type CTCF zinc finger array in the B2H reporter assay. EMSA-based *in vitro* studies of sequence requirements for CTCF binding do not always require the full 40 bp binding site. In one study, a 28 bp DNA substrate containing the 15 bp core was enough to recruit CTCF[78]. However, the presence of the full length sequence is required for optimal occupancy in eukaryotic cells, similar to what is observed in the B2H reporter assay[68]. What is surprising is the flexibility in the sequence at the regions flanking the core sequence. It is likely that the DNA-protein contacts of the zinc fingers 3-7 maintain the conservation of bases within the core sequence and the lack of conservation in the flanking regions suggests the remaining zinc finger arrays do not make constructive contacts with the flanking regions of the CBS. In eukaryotic cells, the flanking sequences may serve to recruit co-factors to CTCF binding sites, but these sequences are not required in the *E. coli* based B2H system where there is no homologue of CTCF, making it unlikely that any co-factors are at play in the prokaryotic system. Therefore, it remains unclear what the role of the flanking sequences are in the CBS, but our

results further confirm their presence is required for CTCF recruitment. In addition, we provide evidence that whatever function the flanking sequences of the CBS play in CTCF binding, it is sequence independent. The necessity for the full zinc finger array in the B2H assay for optimal occupancy of the binding site reflects the findings *in cellula*, where mutations in any one of the zinc fingers of the 11-finger array had an impact on occupancy, with the largest impact being in the DNA-binding fingers 3-7[68].

There were more variants selected with relaxed specificity rather than altered specificity, this could be a result of sampling as only a subset of surviving variants were assayed for binding **(Figure 1.5, 1.6)**. It may also be a reflection of the nature of the bacterial-two-hybrid selection protocol as the system allows for anything that binds to the sequence and does not have a selective pressure for specificity. In previous attempts of Zif268 evolution with phage display, after a couple rounds of selection of the library pool on the target sequence, the enriched pools would be subject to a negative selection with a wild-type, or off-target, sequence as bait[83]. This would allow for a depletion from the selection pool of variants that had non-specific sequence preference. This approach could be used to improve the specificity of our selection further by having a degenerative sequence that included off-target sites and non-inclusive of the on-target site, upstream of a toxic gene, *ccdB* for example, in the B2H system. Finally, current CTCF library design excludes Aromatics and Cysteine, as well as stop codons. However, aromatics are often required to coordinate binding of a C in the first or second position of a sequence triplet. We can potentially improve CTCF variant recognition of C bases within the CBS if we allow for these residues to exist in the selection library.

One caveat of the B2H system is CTCF and other Cys_2His_2 zinc fingers are not native to prokaryotes. The *E. coli* genome may function as a non-specific competitor to the CTCF zinc finger array, sequestering potential variant CTCFs that could have bound to the binding site. We also observed long-term expression of CTCF zinc finger arrays in *E. coli* resulting in an unusually high rate of missense and nonsense mutations in CTCF coding sequence suggesting the zinc-finger array may have been interacting with the bacterial genome or interfering with a molecular pathway that impacted viability. This may result in the loss of variants that bind to the target sequence if they also bind strongly to the *E.coli* genome and result in a fitness cost.

Materials and Methods

All antibiotics and chemicals were obtained from Sigma-Aldrich or ThermoFisher.

Reporter (ONPG β-galactosidase) Assay

CTCF binding to the variant CBS (vCBS) was determined using a modified bacterial-two-hybrid reporter system. The system relies on co-expression of two low-copy plasmids, zinc finger array plasmid and the binding site plasmid, in chemically competent delta lambda (Δλ) *Escherichia coli* cells, a gift from the lab of Dr. George Church (Harvard University, Boston, MA). The zinc finger array plasmid contained a Kanamycin resistance gene (Kan^R) and the Gal11p-ZF array fusion, the zinc finger array of wild-type CTCF or a CTCF variant, under the control of a lac promoter. This same plasmid contained the Gal4 and alpha subunit of RNA polymerase fusion under control of the lac operon. The zinc finger array plasmid was maintained at low copy in *E. coli* via pBR322 derived origin containing the *rop* gene. The binding site plasmid, the second component of the reporter system, conferred Chloramphenicol resistance (Cam^R) and contained the CBS upstream of a minimal *lacZ* gene under the control of a minimal promoter (Pmin). The binding site plasmid was maintained at low copy number via origin *oriV*. Δλ cells co-transformed with both plasmids plated on LB agar plates containing 50 μg/mL kanamycin and 25 μg/mL chloramphenicol. Colonies from the plates were grown overnight in 1 mL cultures of LB medium supplemented with 50 μg/mL kanamycin, 25 μg/mL chloramphenicol, 1M IPTG

and 10 mM $ZnCl_2$. 25 ml of overnight culture was transferred to fresh 1 mL medium and grown for 2 hours, towards the end, growth was monitored by up to three OD_{595} readings using a microplate reader (Bio-Rad Model 680 168100XTU). At an OD value range between 0.157 and 0.260, 100 ml of culture was lysed with a mixture of Popculture (Novagen #71092-75mL) and rLysozyme solution (Novagen #71110-3) for 15 minutes. 135 ml of Z buffer (60 mM NaH_2PO_4, 10 mM KCl, 1mM $MgSO_4$, pH adjusted to 7.0) containing β-ME (2.7 uL/1 mL of Z buffer) and 30 ml of ONPG (4 μg/mL) were added to each lysed sample and rate of hydrolysis of ONPG by the gene product of *lacZ* (β –galactosidase) was monitored by a microplate reader (OD_{410}). Fold activation of the *lacZ* reporter gene was calculated as a ratio of the reaction rate of samples with the ZF-array fusion and the target binding site over samples with an empty fusion and the same target binding site, adjusted for their respective culture density (OD_{595}).

Assembly of the CTCF zinc finger array selection library for single and multiple alterations in the CBS

The residues within the recognition helix of each finger considered for variation in the library depended on the position of the altered sequence in the CBS. In the case of single base alterations to the binding site, a library of variants was generated with 'VNS' codon replacing the coding sequence of either residues [-1, 1, 2, and 3] or residues [2, 3, 5, and 6] of the recognition

helix within the finger of the array that maintained direct protein-DNA contacts with the altered base. The residues [-1,1,2,3] were varied when the altered base was in the 1 position of the triplet of sequence, the zinc finger recognized (3'-5' direction). Residues [2,3,5,6] were varied when the altered base was in position 2 or 3 of the triplet of sequence the zinc finger recognized. Oligos with degenerative 'VNS' sequences within the CTCF array were ordered via IDT, with complementary sequences on either end to act as cloning 'handles' for annealing and ligating into the target zinc finger within the array of a double-digested backbone. Fill in of the un-complimented 'VNS' portion of the insert was achieved by *E.coli* plasmid repair mechanisms. Each library was assembled into the digested Kan^R, zinc finger-gal11P fusion plasmid by annealing the oligos and pooling 16x T4 ligation reactions (NEB #M0202S) following the manufacturer's protocol. The pooled ligation reactions were purified by miniElute PCR purification columns and eluted in 20 uL of ddH_2O following manufacturer's instructions (Qiagen #28004). 5 uL of the reaction was electroporated into 70 uL of XL1 Blue *E. coli* cells made electro competent following standard vacuum-based protocols. Each sample of electroporated cells was recovered for 1 hour in 1 mL of pre-warmed SOC at 37°C, shaking at 900rpm, and pooled in triplicate into 100 mL of LB media with 50 μg/mL kanamycin for expansion at 10 hours until OD_{600} 0.400 was reached. Plasmid was then extracted from cell culture by MIDIprep kit following manufacturer's protocol and eluting with 100 uL of provided EB (Qiagen #12943).

Selection pools of CTCF variants against multiple alterations of the CBS were constructed from pools of variants that were enriched from selection on the single base alterations to the CBS. Construction of CTCF variant pools against multiple CBS alterations was performed by targeted PCR amplification and Gibson assembly of required zinc finger arrays into a zinc finger array-gal11p fusion plasmid with BsaI restriction site landing pads replacing the target zinc finger array. Gibson reactions and clonal expansion were carried out as described for the initial selections.

Selection of engineered CTCF that binds vCBS

Selections for CTCF variants that bind to their target vCBS were performed in two stages, using a B2H system with a β-lactamase resistance gene (*blaC*) as selective pressure. Individual libraries with variation introduced into one of the zinc fingers 3 through 7 in the eleven zinc finger array was transformed into electrocompetent Δλ *E. coli* cells already containing the target vCBS and recovered in 1 mL of pre-warmed SOC in a 96w block for an hour at 37°C, shaking at 900rpm. Subsequently, the culture of transformants were transferred to 3 mL of LB supplemented with 50 μg/mL kanamycin, 12.5 μg/mL chloramphenicol and 10 mM $ZnCl_2$ and induced with 200 mM IPTG. After 3 hours of induction, 1 mL of the culture was plated on low stringency rectangle agar plates (50 μg/mL kanamycin, 12.5 μg/mL chloramphenicol, 200 mM

IPTG, 100 μg/mL carbenicillin, 10 mM $ZnCl_2$, and a low concentration (0.45 μg/mL) of clavulanic acid, a β-lactamase inhibitor. After 24 hours of incubation, the colonies on the plate were harvested in 1 mL of LB of which a 50 ml aliquot was used to inoculate 1 mL of Terrific Broth (TB) medium. The inoculated TB cultures were grown overnight (12-16 hours) in a 96w block with agitation (900rpm) at 37°C. Plasmids were isolated using miniprep kits (Qiagen 27106) to yield an enriched library of CTCF variants which was then subjected to a second, higher stringency round of selection. For the high stringency selection, 600 ng of the enriched CTCF variant library from low stringency selections were transformed into chemically competent Δλ *E. coli* cells already containing the vCBS selection plasmid and subsequent recovery and induction were performed as described above. Following induction, 1 mL of cultures were plated on rectangle gradient selection agar plates, a gradient of low to high selection stringency. Gradient stringency selection plates were constructed by plating 20 mL of molten LB agar containing (4 μg/mL of clavulanic acid, 50 μg/mL kanamycin, 12.5 μg/mL chloramphenicol, 200 mM IPTG, 100 μg/mL carbenicillin, 10 mM $ZnCl_2$) in a rectangle plate (ThermoFisher #264728) and solidified at ~30° angle (elevated by resting one edge on the fat end of a p200 tip). Once the bottom layer was solidified a second layer of 20 mLs of molten LB agar containing (50 μg/mL kanamycin, 12.5 μg/mL chloramphenicol, 200 mM IPTG, 100 μg/mL carbenicillin, 10 mM $ZnCl_2$) was spread evenly on top. Two wedges were created, one with clavulanic acid and one without, for the gradient effect across the plate. Colonies were picked

from the highest point of growth on the gradient plate and grown overnight in 1 mL of LB and 50 µg/mL of kanamycin in 96w blocks, 37C, 900rpm. CTCF variant containing plasmids were isolated from overnight cultures by Qiagen miniprep (Qiagen #27106) and retransformed into chemically competent XL1 Blues for expansion and subsequent sequencing to identify the amino acids in the CTCF variant.

Chapter 2:

CTCF variants can replicate the biological function of wild-type CTCF in human cells

Rebecca T. Cottman[1,3], Jay W. Jun[1], Caleb A. Lareau[1,4,5], Martin J. Aryee[1,4,5], J. Keith Joung[1,2]

[1]Molecular Pathology Unit, Center for Cancer Research, and Computational and Integrative Biology, Massachusetts General Hospital, Charlestown, MA, [2]Department of Pathology, Harvard Medical School, Boston, MA, [3]Department of Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, [4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, [5]Broad Institute of MIT and Harvard, Cambridge, MA.

All authors listed contributed to the work described in this chapter. I designed and performed all wet-lab experiments. Jay Jun helped perform all wet-lab experiments except for NGS library prep and 4C experiments. Caleb Lareau performed informatic analysis of 4C-seq datasets. Martin Aryee provided direction in 4C data analysis. Keith Joung oversaw research and provided direction.

Introduction

CTCF establishes TADs across the genome to create neighborhoods of structure that facilitate the interaction of genomic elements within, while insulating them from elements outside the TAD boundaries[91–96,97]. CTCF acts to regulate gene expression with its function mostly dictated by the location of CTCF binding in relation to the transcriptional start site[19]. Selective gene activation of gene clusters within TAD neighborhoods by CTCF can occur by prevention of nucleosome occlusion of the TSS or by formation of promoter-enhancer loops via recruitment of enhancer regions to the gene promoters. The exact mechanism of bringing specific enhancer and promoter elements in physical proximity is not completely defined, but DNA binding proteins such as the transcription factor YY1 and CTCF, play a role[98]. Although the majority of CTCF binding sites cluster at TAD boundaries, CTCF binding can occur within the enhancer and promoter regions of genes, and has been demonstrated to contribute to increased association between enhancer-promoter regions by looping, or subTADs[54,19,21,99,100]. While subTAD formation was initially believed to be exclusively CTCF-cohesin mediated, recent work has provided evidence for the complexes of CTCF-cohesin-RNA or potentially CTCF-RNA assembling and directing sub-TAD formation, especially those involved in promoter-enhancer looping events at CTCF-regulated gene loci[20,76,101,102]. The *MYC* locus is an example of a gene under the control of a CTCF-directed promoter-enhancer loop, but it is unclear whether cohesin and/or RNA is the co-mediator in establishing the topological change.

Originally CTCF was thought to be a negative regulator of the proto-oncogene *c-MYC* (*MYC*)[49]. Subsequent studies now indicate that CTCF is a positive regulator of *MYC*, with several binding sites within the intronic regions as well as upstream of the transcriptional start site (TSS)[63,76]. The *MYC* gene resides within a 2.8 Mb TAD with three CTCF binding sites at the *MYC* locus; 2kb upstream of the TSS, one overlapping the TSS, and the final one binding ~1 kb downstream of the TSS within the first intron[63]. These binding sites and the larger TAD structure are conserved across cell-types and species. In contrast, the promoter-enhancer looping structures within the 2.8 Mb TAD vary dramatically in number and size across cell types and species[64,76]. In K562s, a human erythroid lymphoblastoid cell line, *MYC* expression has been shown to be dependent on CTCF-mediated promoter-enhancer looping between the enhancer docking site, a CTCF binding site 2 kb upstream of the *MYC* TSS, and two enhancer regions ~2 Mb downstream **(Figure 2.1)**[20,76]. Deletion or methylation of this CBS results in a loss of CTCF binding concurrent with a reduction of *MYC* expression and loss of enhancer-promoter looping[20,76]. We used this locus to determine if CTCF variants pulled out of the bacterial selection can reproduce the biological function of wild-type CTCF in creating promoter-enhancer loops in K562s.

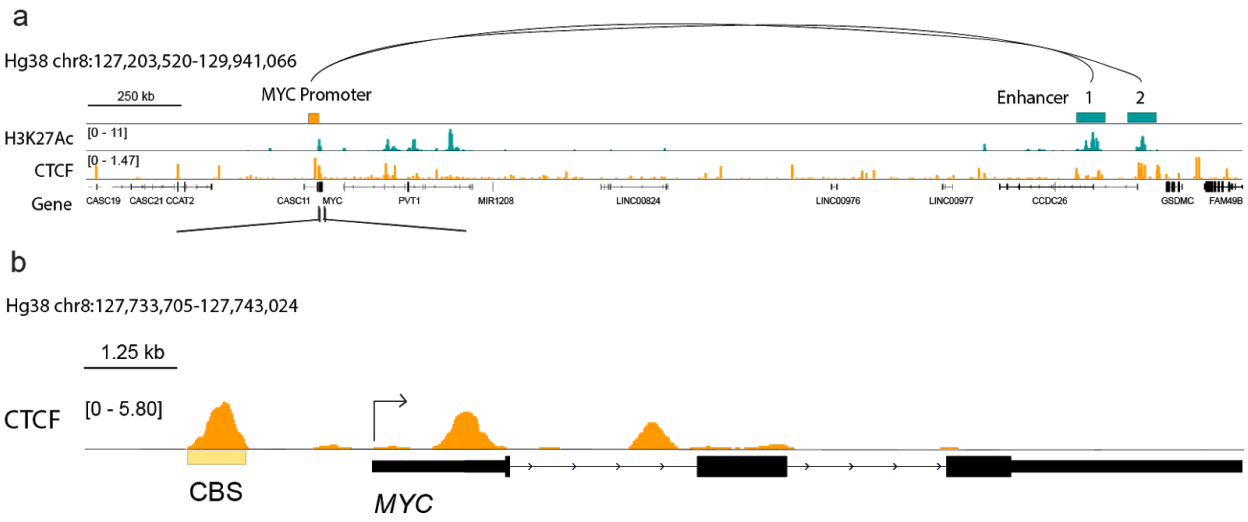

Figure 2.1: *MYC* expression is driven by a CTCF-mediated promoter-enhancer loop. a, Chromosome region within one large TAD containing the *MYC* locus. ChIP-seq profiles of CTCF and H3K27Ac at the Hg38-chr8:127,203,520-129,941,066 locus of K562s. Promoter-enhancer looping of the *MYC* locus promoter region (yellow box) and two distal enhancer regions (#'ed teal boxes) is represented by black arcing line. **b,** ChIP-seq peaks of CTCF at the *MYC* locus in K562 cell line. The CBS critical for maintenance of *MYC* expression (yellow box) is ~2kb upstream of *MYC* TSS. Two additional CTCF binding events occur within the first exon and intron of *MYC*.

Results

Engineered CTCFs can fulfill the biological role of native CTCF at the *MYC* locus.

We investigated the ability of the CTCF variants to bind to their target sequence *in cellula* and replicate the function of native CTCF. To demonstrate this, we used the ubiquitously expressed *MYC* locus which is enclosed in a 2.8 Mb TAD recognized as an insulated neighborhood containing no other annotated protein-encoding genes[103,104]. *MYC* expression is regulated in part by CTCF-mediated looping of distal enhancer regions to the promoter region of *MYC*. It has been previously shown that introduction of indels at the CBS ~2kb upstream of the *MYC* TSS prevents CTCF binding which results in a loss of association to distal enhancer regions ~2 Mb downstream of the *MYC* TSS and a subsequent reduction of *MYC* expression[20,76]. As a first step in testing the CTCF variants *in cellula*, we engineered clonal cell lines that replaced the critical *MYC*-proximal CBS with each one of the five vCBSs used in the bacterial selection **(Figure 2.2a)**. A 'Deletion' cell line, with a 14 bp deletion of the 15 bp core sequence of the CBS, and a 'Concensus' cell line, with the native CBS replaced with the wt CBS used in the B2H system, were also created as clonal negative and positive controls, respectively. Chromatin Immunoprecipitation (ChIP) of CTCF followed by qPCR with primers specific to the modified CBS locus was performed to detect endogenous CTCF occupancy of the region in each of the vCBS, 'Deletion', and 'Consensus' cell lines. CTCF could not bind to the vCBSs, with the exception of vCBS 3, and *MYC* expression was reduced to a similar level as cell lines with a deletion of the target CBS (**Figure 2.2b, c**). Although CTCF was able to bind to vCBS 3, it did

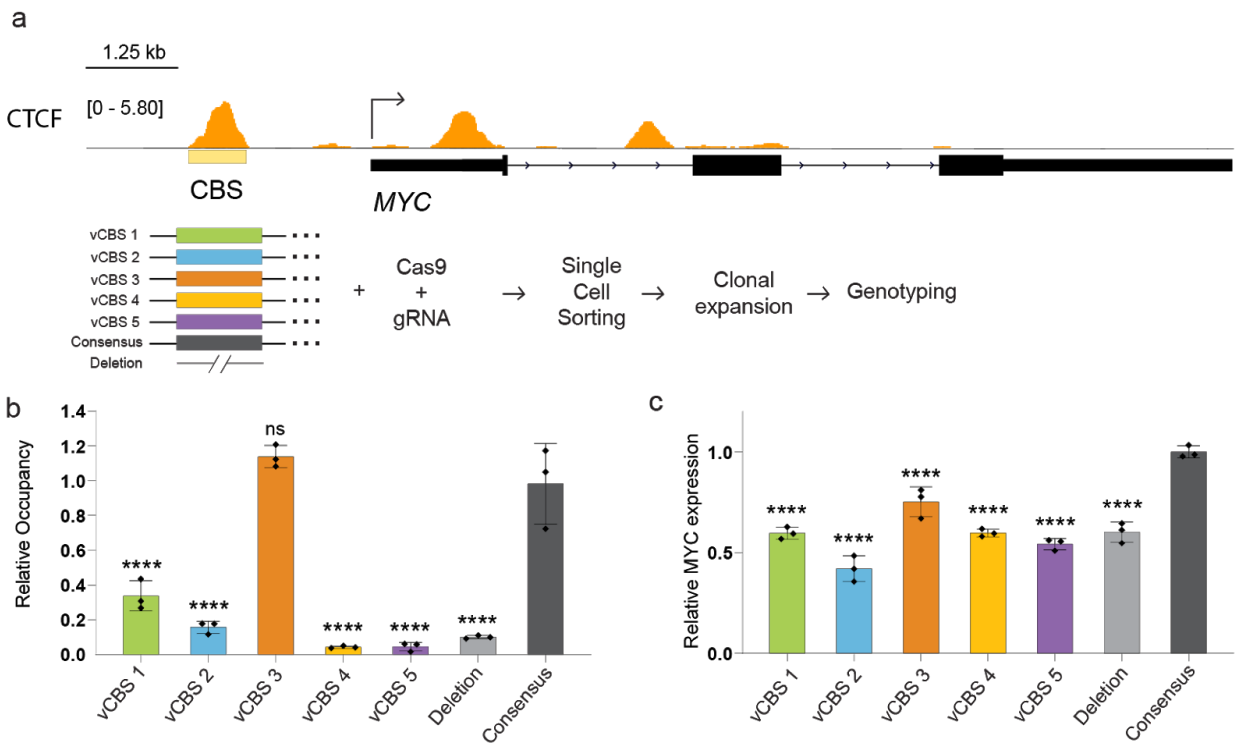

Figure 2.2: CTCF variants restore MYC expression in engineered cell lines by recreating CTCF-mediated topological loop. a, *MYC* locus on Chr8 with ChIP-seq peaks of CTCF in K562s. Workflow of generating clonal cell lines with variant CBSs (vCBS)s, consensus CBS (consensus), or indels (deletion) in place of the critical CBS through CRISPR-based genome editing. **b-c**, ChIP-qPCR quantification of endogenous CTCF occupancy of clonal cell lines with vCBS, deletion, or consensus CBS and concurrent effect on *MYC* expression. Occupancy of endogenous CTCF and *MYC* expression levels are relative to levels in the 'Consensus' cell line. *MYC* expression levels were normalized across samples to *HPRT*, a house-keeping gene. Values reflect triplicate replicates with p values (ns ≥ .05, **** < .0001) determined by ANOVA compared to the control (Consensus cell line).

not have a full recovery of *MYC* expression. This may be due to vCBS 3 sequence specific effects on CTCF cofactors required for *MYC* expression.

We next used these cell lines to assay the binding affinity of each group of CTCF variants generated from the B2H selections to their respective vCBS cell line **(Figure 2.3a-e)**. Variants from vCBS groups 2, 4, and 5 had the most promising occupancy of their target vCBS, while variants from group 1 proved to have weak affinity for their binding sequence and were not further investigated. Next, we expressed the top two CTCF variants from each group in their cognate vCBS cell lines and observed *MYC* expression was restored concurrently with binding of the CTCF variants to the vCBSs (**Figure 2.4a-f)**. Finally, we confirmed that the recovery of *MYC* expression in the engineered cell lines was due to the CTCF variant restoring the looping of distal enhancers to the promoter of *MYC*, using circularized chromosome conformation capture (4C) analysis **(Figure 2.4g-i)**. Cells expressing the CTCF variant showed increased association between the *MYC* promoter region and the two distal enhancer regions, while transfection of wild-type CTCF or GFP plasmid in the vCBS cell lines did not restore the looping. CTCF variants expressed in a cell line with a deletion of the critical CBS did not recover *MYC* expression, indicating binding at the engineered vCBS site is cause for restoration of *MYC* expression **(Figure 2.5)**.

We took the opportunity of the lack of CTCF occupancy at the vCBS sites in the engineered cell lines to determine if wild-type CTCF fused to dCas9 could be directed to the vCBS region and used to restore *MYC* expression. Not only could *MYC* expression not be recovered in either N-terminal or C-terminal conformation of the dCas9-CTCF fusion, but dCas9 fusion to a CTCF variant, which is able to bind to the same vCBS site when not fused to dCas9,

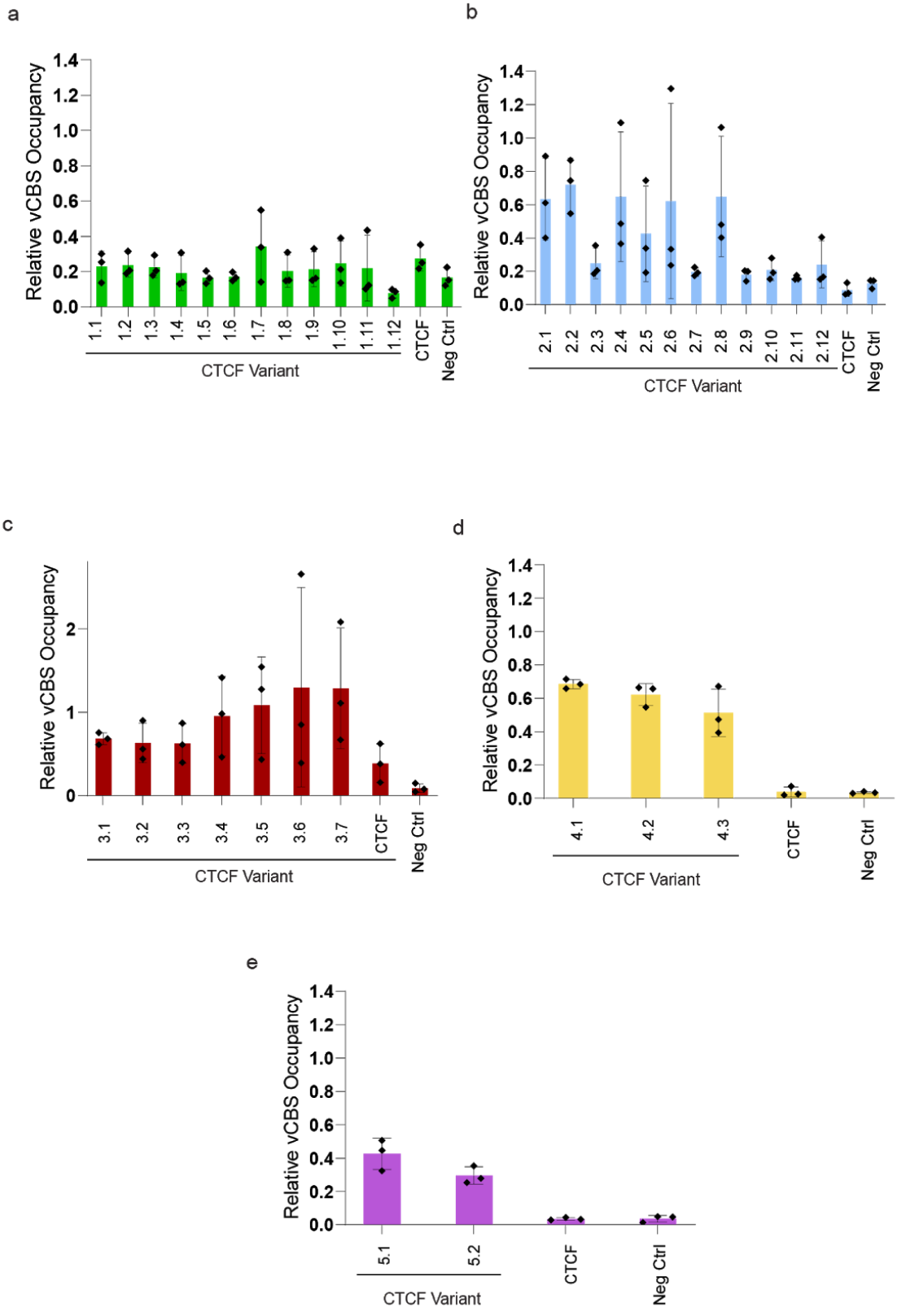
a
Relative vCBS Occupancy
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
1.10
1.11
1.12
CTCF
Neg Ctrl
CTCF Variant
b
Relative vCBS Occupancy
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
2.9
2.10
2.11
2.12
CTCF
Neg Ctrl
CTCF Variant
c
Relative vCBS Occupancy
2
1
0
3.1
3.2
3.3
3.4
3.5
3.6
3.7
CTCF
Neg Ctrl
CTCF Variant
d
Relative vCBS Occupancy
4.1
4.2
4.3
CTCF
Neg Ctrl
CTCF Variant
e
Relative vCBS Occupancy
5.1
5.2
CTCF
Neg Ctrl
CTCF Variant

Figure 2.3: Binding capacity of CTCF variants to their target vCBS in the engineered cell lines. a-e, ChIP-qPCR assay of occupancy of vCBS in vCBS cell lines transfected with CTCF variants (#.#), wild-type CTCF (CTCF), and pMaxGFP as transfection control (Neg Ctrl). ChIP performed with the HA antibody for plasmid delivered construct specific detection of occupancy at the vCBS. Occupancy is relative to endogenous CTCF occupancy of consensus CBS in the consensus cell line.

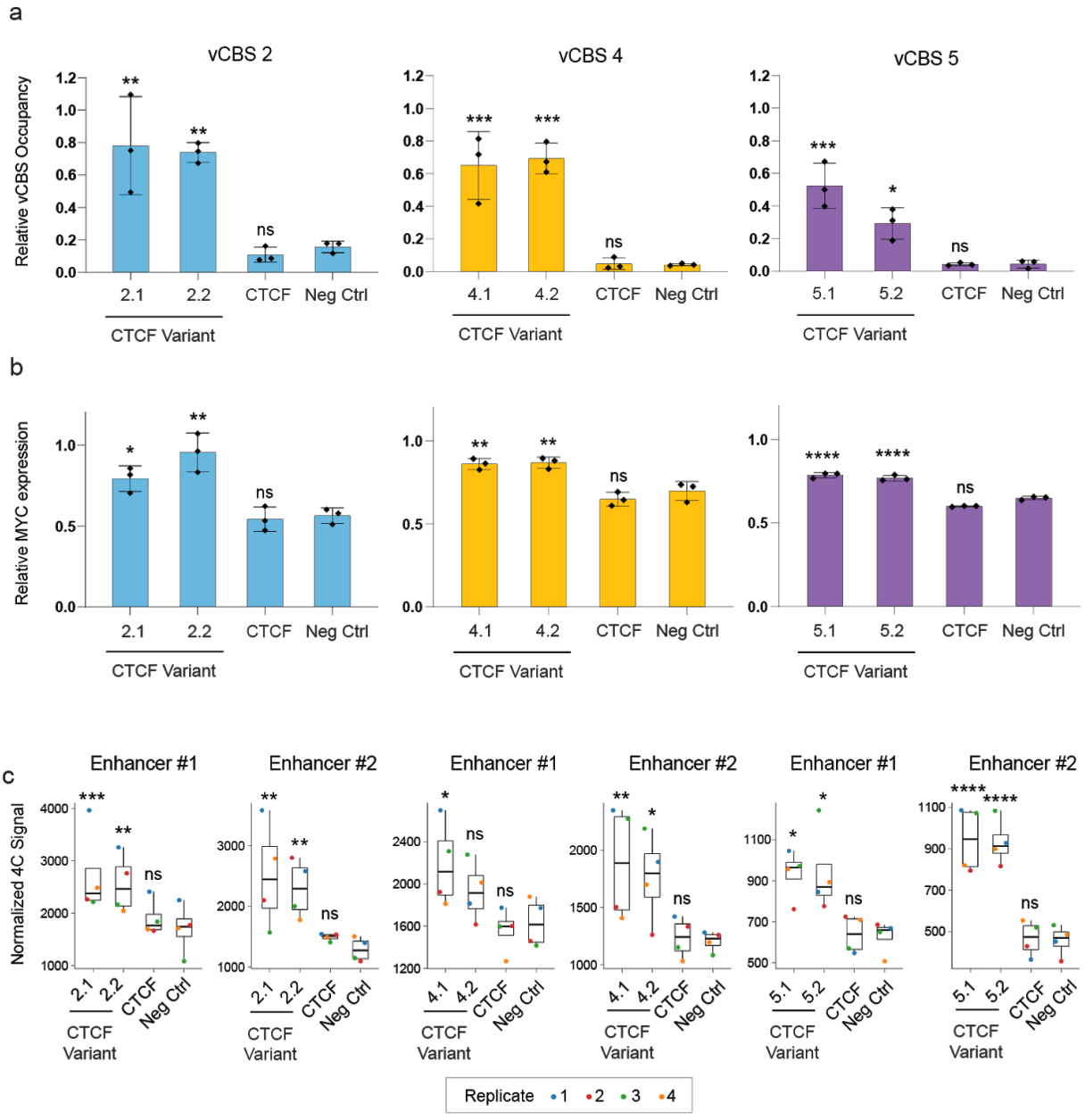

Figure 2.4: Expression of CTCF variants in respective engineered cell lines bind to the variant CBS and recover *MYC* expression. a, Occupancy of vCBS was measured by ChIP-qPCR, IP performed with CTCF antibody. Graph of the mean occupancy of vCBS of top two variants from each selection group, wild-type CTCF expressed from plasmid (CTCF) and pMaxGFP (Neg Ctrl) as a transfection control, relative to endogenous CTCF occupancy of consensus CBS in the consensus cell line. Experiment performed in triplicate with ANOVA relative to Neg Ctrl derived p-values above each sample (ns ≥ .05, * < .05; ** < .01, *** < .001). **b,** Endogenous *MYC* expression levels of corresponding samples normalized to a housekeeping

gene (*HPRT*). Data reflects triplicates of normalized endogenous *MYC* expression levels relative to *MYC* expression in the 'Consensus' cell line as quantified by RT-qPCR. Significance was determined in the same way as described in (b) **c,** 4C analysis of conditions in (a-b) to observe association of two distal enhancer regions (Enh #1; Enh #2) to the viewpoint (*MYC* promoter region). 4C signal across samples was normalized to reads accumulated at the viewpoint. Box plots are result of quadruplicate replicates with ANOVA p-values reflecting significance compared to the transfection control (Neg Ctrl) indicated by asterisks above each box plot (ns ≥ .05, * < .05; ** < .01, *** < .001, **** < .0001).

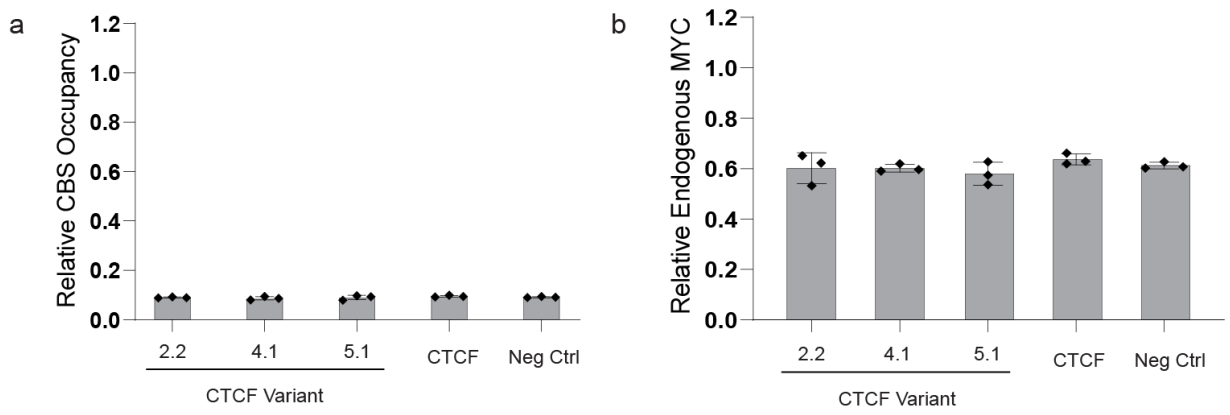

Figure 2.5: Recovery of *MYC* expression is specific to CTCF variants binding to the vCBS ~2kb upstream of the *MYC* TSS. a, ChIP-qPCR, with CTCF antibody, of the CBS region region in the 'deletion' CBS cell line transfected with CTCF variants (2.2, 4.1, or 5.1), wild-type CTCF (CTCF), or pMaxGFP (Neg Ctrl). in triplicate. Occupancy of the region with the top variant CTCF of each group relative to occupancy of consensus CBS by endogenous CTCF. **b,** Endogenous *MYC* expression relative to expression in the consensus cell line. Bar graphs reflect the mean value of triplicate replicates quantified by RT-qPCR.

appears to hinder its ability to occupy the vCBS site, suggesting the fusion in and of itself interferes with normal CTCF function **(Figure 2.6)**.

CTCF directed looping at the *MYC* locus is RNA mediated and Cohesin-independent

Cohesin and RNA have been shown to be involved in the formation of CTCF-mediated subTADs[101,102]. Currently it is not known if either cohesin or RNA plays a role in CTCF-directed looping at the *MYC* locus. RNA-CTCF interactions are coordinated by a series of RNA binding regions (RBRs) scattered throughout CTCF 11-finger zinc finger array while cohesin interaction is coordinated by the interaction between a subunit of cohesin, SA2, with three peptides within the C-terminal and N-terminal domain of CTCF**(Figure 2.7a)**[105–107]. The co-mediators of CTCF looping at the *MYC* locus can not be easily determined with endogenous CTCF. Our CTCF variants and cell lines can be applied as a method for locus-specific identification of cofactors in CTCF-mediated gene activation.

To investigate the RNA-CTCF interactions at the *MYC* locus, we deleted the RBRs within zinc finger 1 (ZF1) and zinc finger 10 (ZF10) of CTCF variant 2.2, 4.1, and 5.1, zinc fingers that have been shown to be non-essential for DNA binding, but essential for maintaining CTCF-RNA interaction[68,101]. The double RBR deletion (ΔZF1+ΔZF10 RBRs) in the CTCF variants resulted in reduced *MYC* expression despite maintaining the ability to bind to the target vCBS (**Figure 2.7b,c).** When we deleted the C-terminal domain of the CTCF variants, a region previously thought to mediate cohesin-CTCF looping, we did not observe any effect on binding to vCBS or *MYC* expression. It has been shown that removal of the N-terminal region of CTCF

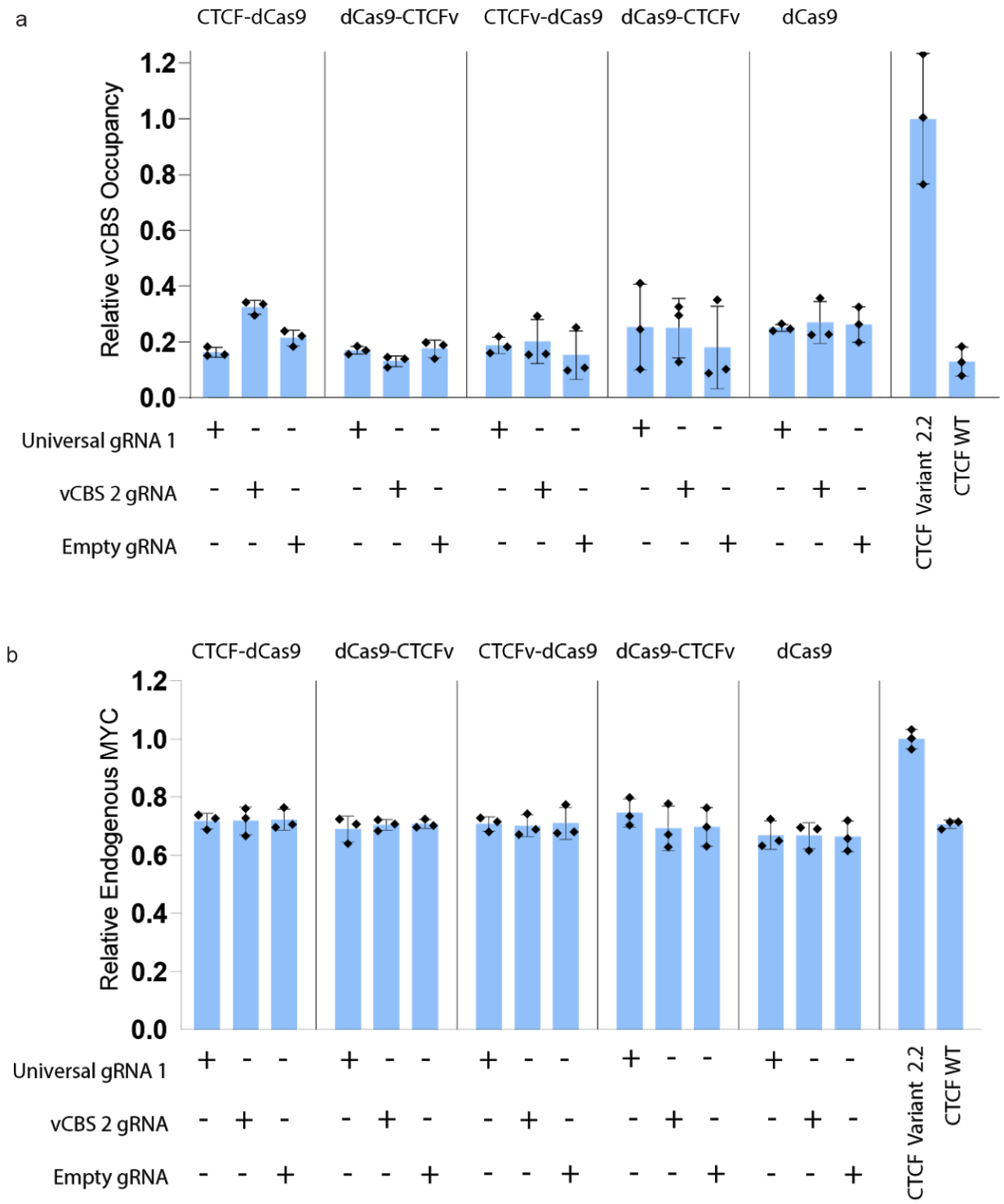

Figure 2.6: Permutations of dCas9 fusion to CTCF or a CTCF variant do not recover *MYC* expression. a, Occupancy of the vCBS in vCBS 2 cell line transfected with dCas9 and dCas9 CTCF variant 2.2 fusion constructs was quantified by ChIP-qPCR of the vCBS region,

immunoprecipitation performed with HA antibody for detection of dCas9 and dCas9 fusion constructs. Data reflects triplicate samples with or without the specified gRNAs localizing the dCas9 fusion to the specified target sites. Universal gRNA 1 and 2 target sites 20 bp up and downstream, respectively, of the vCBS site. vCBS2 gRNA overlaps the vCBS site and mimics distance CTCF variant 2.2 binding would be to the TSS of *MYC*. Occupancy is relevant to the vCTCF 2.2 positive control with plasmid-based expression of wild-type CTCF (CTCF) as a negative control. **b,** Endogenous *MYC* expression levels of samples in (**a**) relative to endogenous *MYC* expression levels in the positive control (CTCF variant 2.2). Data reflects the mean of triplicate experiments quantified by RT-qPCR with endogenous *MYC* specific primers.

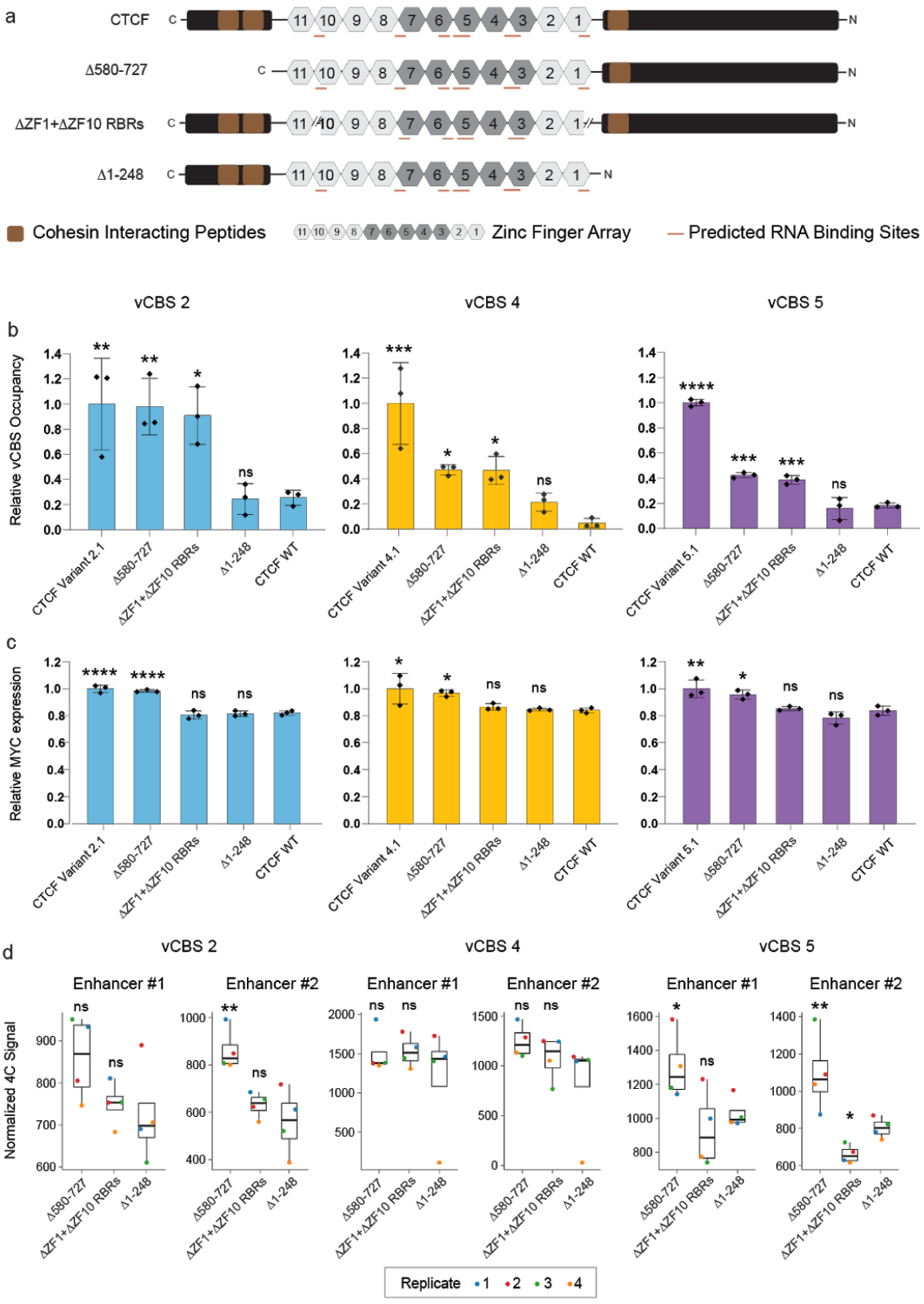
a
CTCF
Δ580-727
ΔZF1+ΔZF10 RBRs
Δ1-248
C
N
11 10 9 8 7 6 5 4 3 2 1
Cohesin Interacting Peptides
Zinc Finger Array
Predicted RNA Binding Sites
b
vCBS 2
vCBS 4
vCBS 5
Relative vCBS Occupancy
CTCF Variant 2.1
CTCF Variant 4.1
CTCF Variant 5.1
Δ580-727
ΔZF1+ΔZF10 RBRs
Δ1-248
CTCF WT
c
Relative MYC expression
d
Enhancer #1
Enhancer #2
Normalized 4C Signal
Replicate 1 2 3 4

Figure 2.7: RNA coordinates CTCF-mediated promoter-enhancer looping at the *MYC* locus. a, Diagram of full length CTCF and various modifications. Important Cohesin interacting and RNA interacting regions of CTCF are marked and labeled. (Δ 580-727) removes 2 out of 3 peptides that have been identified to mediate cohesin recruitment. (Δ ZF1+ΔZF10 RBRs) modification removes two out of 6 predicted RNA binding regions and has been shown to result in loss of RNA-dependent TADs genome-wide. (Δ 1-248) is known to have lost the ability to bind to DNA and serves as a negative control. **b-c**, Occupancy of indicated vCBS by specified modified CTCF variants from groups 2,4, and 5 and the concurrent *MYC* expression levels. Occupancy assayed by ChIP-qPCR, IP performed with HA antibody, in triplicate, ANOVA p-values indicated significance compared to negative control (Δ 1-248) (ns ≥ .05, * < .05; ** < .01, *** < .001, **** < .0001). **d,** 4C analysis of association of enhancer region #1 and #2 to the *MYC* promoter region for the same conditions in (b-c). Viewpoint same as previously described for the *MYC* locus. 4C signal was normalized across samples to reads mapped to the viewpoint. ANOVA p-values of quadruplicates calculated in relation to negative control (Δ 1-248) (ns ≥ .05, * < .05; ** < .01, *** < .001, **** < .0001).

results in loss of binding of CTCF that is independent of the DNA-binding domain[105]. We used this modification in our CTCF variants as a negative control and saw the expected loss of occupancy, reduction of *MYC* expression, and lack of association of distal enhancer regions to the promoter region of *MYC*. Removal of the two C-terminal cohesin interacting peptides did not result in a loss of the promoter-enhancer loop. In contrast, deletion of the two RBRs in ZF1 and ZF10 of each of the CTCF variants resulted in a loss of association of the distal enhancer regions to the *MYC* promoter as quantified by 4C (**Figure 2.7d**). The reduction in *MYC* expression due to the deletion of RBSs coincides with a loss of enhancer association with the *MYC* TSS, similar to levels in the negative control. These data suggest that the CTCF directed promoter-enhancer loop at the *MYC* locus relies on RNA as a co-mediator.

Discussion

We have demonstrated that not only do the CTCF variants bind to their selected for sequence in K562s, but the use of these variants as a method for site-specific investigation of cofactors in CTCF-directed subTADs. In most cases, occupancy of the vCBS tracked with recovery of *MYC* expression **(Figure 2.2).** However, the ability of endogenous CTCF to bind to vCBS 3 sequence did not result in a full recovery of *MYC* expression. Although CTCF is able to bind to vCBS 3, it could be that the modification in the sequence do not allow for ideal wrapping around the DNA that is required for comediation with a cofactor. Alterations in this sequence are targeted by ZF 4 and 5, which also have RNA binding regions. It is possible that ZF 7,6, and 3 of

CTCF could be enough to sustain the protein-DNA interaction, but 4 and 5 are not participating fully, which in turn may impact their ability to interact with their RNA substrate required for MYC expression. In addition, linkers between ZFs in an array are unstructured in the unbound-state and become rigid and structured, making contacts to the phosphate backbone, when the array is bound to DNA[90,108,109]. ZF4 and ZF5 of the array may not be perfectly bound to the vCBS 3 sequence and subsequently have unstructured linkers that may interfere with the proper interaction with cofactors. This may explain why occupancy of a vCBS does not always correlate with the level of recovery of *MYC* expression. However, binding of the ZF to DNA, and stabilization of the linker between zinc fingers in the array, does not seem to impact RNA recognition in other transcription factors with the Cys_2His_2 architecture[70]. Still, the influence of CTCF-DNA interaction and subsequent interaction with substrates may be an explanation for the discrepancy.

These data suggest that CTCF variants are able to replicate the function of CTCF at endogenous sites. The ability of CTCF variants to recruit the distal enhancer region indicates the variants are able to interact with cofactors of endogenous CTCF and with endogenous CTCF itself to re-form topological alterations in the genome. There is also the application of the CTCF variants as a tool for studying the molecular biology of CTCF *in cellula*, at endogenous sites and genome-wide.

It is unclear if CTCF has distinct subgroups of zinc fingers within the array that coordinate protein-DNA interaction and protein-RNA interaction, or if they have dual roles like other transcription factors with the Cys_2His_2 zinc finger array architecture. Wilm's Tumor suppressor 1 (WT1) is an example of a transcription factor where the fingers in the Cys_2His_2 zinc

finger array have distinct DNA or RNA binding roles[71,110]. While Transcription Factor IIIA (TFIIIA) is an example of zinc fingers in the array having a dual role where DNA recognition is established through zinc fingers in the 9-finger zinc finger array while a subset of the array, zinc fingers 4-6, has the capacity to bind DNA and RNA substrates[111,112]. Crystal structure of TFIIIA bound to its RNA substrate reveals zinc finger 4 and 6 bind to the RNA substrate by residues -1,1,2 of the recognition helix while finger 5 recognition helix contacts the phosphate backbone of RNA. Zinc fingers 4-6 have an inverse behavior when bound to a DNA substrate, where finger 5 makes protein-DNA contacts by residues -1, 2, 3, and 6 of the recognition helix and finger 4 and 6 act as non-binding linkers[111,112].

The structure of CTCF bound to a DNA substrate indicates that only zinc fingers 3-7 make stable protein-DNA contacts; our studies confirmed the reported role of ZF1 and ZF10 of the array in maintaining protein-RNA contacts[101]. It would be interesting to determine the residues within ZF1 and ZF10 that coordinate the protein-RNA contact with the RNA substrate and if they recognize RNA in the same way as the fingers in the array of TFIIIA. As of now, the peptide deletions of ZF1 and ZF10 likely lead to restructuring of the entire finger rather than identifying certain residues, perhaps introducing alanine substitutions at residues -1,1, and 2 of the recognition helix for ZF1 and ZF10 may provide some indication of function.

There is a president for CTCF-RNA mediated gene regulation, with some cases not involving cohesin[50]. There is a significant portion of subTADs that form on the genome after mitosis before Cohesin is detected which suggests that not all CTCF mediated sub-TADs rely on cohesin[41]. CTCF interacts with RNA for locus specific targeted inactivation on the X-chromosome[50]. CTCF binds to Wrap53, the antisense RNA transcript of p53 to regulate p53

expression[113]. The ability of the Cys_2His_2 zinc finger array of CTCF to interact with RNA transcripts is not surprising as many Cys_2His_2 zinc fingers have been demonstrated to make protein-RNA interactions **(Figure2.7)**. However, we can not definitively say that cohesin is not involved in the promoter-enhancer looping at the *MYC* locus as work published after these experiments were performed identified two residues in the N-terminal region of CTCF that are required for recruitment of cohesin, and the peptides in the C-terminal region are not required[106,107]. This work does not conflict with the conclusion that RNA is essential for formation of the CTCF mediated loop, but allows for the possibility that it is a complex of CTCF, Cohesin, and RNA or a complex of CTCF and RNA that build the subTAD. We could distinguish between these two possibilities by introducing two mutations in the N-terminal region of the CTCF variants, Y226A and F228A, that were identified to be essential for Cohesin interaction[107]. Then test these mutants for ability to recover *MYC* expression and create the promoter-enhancer loop. In this way we can determine if the CTCF-mediated subTAD at the *MYC* locus requires Cohesin in addition to RNA.

Another outstanding question from this study is what RNA substrate interacts with CTCF to establish the loop. Two likely candidates are the nascent transcript from the *MYC* locus or lncRNA produced from the distal enhancer regions. Performing RNA-based gel mobility shift assays would be a simple *in vitro* method of determining a multitude of RNA that is able to interact with CTCF, but would not identify the causal RNA substrate for coordinating the CTCF-directed promoter-enhancer loop. Crosslinking of CTCF variants with its RNA substrate in one of our vCBS cell lines followed by tag-based enrichment of the CTCF-RNA crosslinked complexes would allow for extraction and analysis of RNA substrates specifically involved with

looping at the *MYC* locus. However, it is possible that the sequence of the RNA does not matter and just having a large concentration of negatively charged sequence through CTCF recruitment is what establishes the environment for transcriptional activation. Therefore, scrambled RNA sequences orthognanal to the transcriptome could be included in the previously described experiments. Any sequence preference can be determined by a modified cut and run protocol where enrichment of RNA fragments bound to a purified CTCF variant or CTCF protein can be isolated and sequenced. Enrichment of any sequence from the original library can then be determined and a motif of sequence preference, if any, can be derived through motif analysis.

The lack of functionality of the dCas9 fusion constructs targeted to the vCBS sites was a surprise, but may lend insight into the function of CTCF **(Figure 2.6)**. In the case of other proteins with Cys_2His_2 architecture, the linkers between zinc fingers in an array are flexible and unstructured in an unbound state, but become rigid and form water-mediated contacts with the phosphate backbone of DNA when the fingers in the array are bound to their target[90,108]. This may be an explanation for why the CTCF-dCas9 nor the dCas9-CTCF fusions did not recover *MYC* expression as the CTCF may undergo slight conformational changes when bound to its target that enables functional interaction with RNA and other cofactors. The dCas9 fusion itself may inhibit these interactions as demonstrated with the performance of dCas9-CTCF variant (dCas9-CTCFv) or CTCF variant-dCas9 (CTCFv-dCas9) fusion at the engineered vCBS. Both constructs were unable to bind to the vCBS and recover *MYC* expression even when there was no competing gRNA-mediated recruitment to an adjacent site. The gRNAs used in these experiments were previously tested for quality by a cleavage assay of these targets using the gRNAs to recruit Cas9. These data suggest that CTCF must be able to bind to its substrate and

interact freely with co-mediators in order to alter topology of the genome and the dCas9 fusion itself had a destabilizing effect on the CTCF variant that prevented it from binding to its target site. Expression analysis should be performed to confirm that there is no difference in expression of the fusion constructs.

Materials and Methods

Mammalian cell culture

K562 cells were cultured in a 37° C incubator at 5% carbon dioxide, in RPMI 1640 medium + L-glutamine (ThermoFisher 11875119 with 10% FBS, 1% Penicillin and Streptomycin. Polyclonal, virally transduced exogenous *MYC* K562 cell line (exoMYC.K562, pCMVmurine-TdTomato) was a gift from Dr. Richard Young (Massachusetts Institute of Technology, Cambridge, MA). exoMYC.K562 cells were cultured in the medium described above for the K562 cell line, supplemented with puromycin (2 μg/mL) for maintenance of the exogenous *MYC* gene. vCBS cell lines derived from a monoclonal expansion of exoMYC.K562 #3.6 (exoMYC.K562-vCBS) were cultured the same as exoMYC.K562s with puromycin (2 μg/mL).

Generating vCBS Cell lines

The polyclonal, TdTomato positive exoMYC.K562 cell line was subjected to single cell sorting to generate a monoclonal exoMYC.K562 cell line. The number and location of integration events of the exoMYC cassette in the clonal lines were determined via the modified ATAC-Seq protocol. Clones that had integration events of the exoMYC cassette that may interfere with

downstream analysis as well as expression levels of endogenous and exogenous *MYC* were excluded. A clonal cell line that had minimal integration events while maintaining approximately similar levels of exogenous and endogenous *MYC* expression was labeled as exoMYC.K562 #3.6 and used to derive all the vCBS cell lines using CRISPR-Cas9-based editing. To generate the vCBS cell lines, pCMV-SpCas9-3xFlag-P2A-EGFP plasmid (RC5175), was co-transfected with both the gRNA targeting the CBS upstream of *MYC,* and the ssODN carrying the desired vCBS flanked by 55 nt of sequence homologous to the target site. 72 hours post transfection, cells that were positive for both GFP and TdTomato were single cell sorted into a 96w dish, expanded, and screened for the presence of vCBS at the targeted location. Clones containing the vCBS at the target site with no other indels were used for further experiments. A clonal line with a 14bp deletion at the target CBS was used in this study as a negative control.

Transfections

K562, exoMYC.K562, or exoMYC.K562-vCBS cell lines were seeded at 3 x 10^5 cells/mL in culture plates 24 hours prior to transfection following described culture conditions. On the day of transfection, 1x10^6 cells/reaction were collected by centrifugation at room temperature for 10 min at 90g, resuspended in 100 mL room temperature Nucleofector Solution (Lonza VCA-1003), and transfected with 5 µg of CTCF variant plasmid (pCMV-HA-CTCFvariant) or wildtype

CTCF plasmid (pCMV-HA-CTCF) DNA using Amaxa Biosystems Nucleofector II, program T-016. Following nucleofection, the cells were transferred to 25 mL of pre-warmed RPMI 1640 medium + L-glutamine (Thermo 11875119) with 10% FBS, 1% penicillin, 1% streptomycin and puromycin (2 μg/mL). Transfections were done in triplicate per CTCF variant or wildtype CTCF condition and pooled post-nucleofection when transferred to pre-warmed RPMI media. Pooling of transfected biological replicates was necessary to achieve the minimum number of cells (1 x 10^7) 3 days post-nucleofection needed for ChIP. 72 hours after transfection, cells were either crosslinked for ChIP processing or harvested for RNA isolation.

Determination of CTCF binding by ChIP-qPCR

ChIP-qPCR was used to determine the binding of CTCF variants and wild-type CTCF to the vCBS at the *MYC* locus. Approximately 1 x 10^7 exoMYC.K562 cells transfected with either CTCF variant or wild-type CTCF plasmid DNA were crosslinked with 1% formaldehyde (Sigma F8775) in the culture medium for 10 min at 37°C followed by quenching with 1.2 mL of 2.5 M glycine for 5 minutes. Cells were washed twice with ice cold PBS, collected by centrifugation, and frozen at -80°C. Nuclei were isolated by resuspending the cell pellet in lysis buffer (50 mM Tris-HCl, pH 7.4, 1% SDS, 0.25% DOC, protease inhibitors) followed by dilution in ChIP buffer (50 mM Tris-HCl, pH 7.4, 0.1% SDS, 150 mM NaCl, 1.84% Triton-X, protease inhibitors).

Samples were sonicated on ice in a Branson 250 Sonicator for a total run time of 5.5 minutes (0.7 s on and 1.3 s off in each cycle) and the supernatant of the lysate collected after centrifugation (20,000g). The CTCF-DNA complex in the supernatant was immunoprecipitated by overnight incubation with HA (2 μg) or CTCF (5 μg) antibody at 4°C. HA antibody was used to enrich for crosslinked protein-DNA fragments specific to the exogenous CTCF variant while CTCF antibody allowed detection of binding from both exogenous CTCF variant and endogenous wild-type CTCF. The immunoprecipitated complex was collected by incubating with magnetic Protein G Dynabeads (Thermo 10003D) for 2 hours at 4°C. The beads were then washed in wash solutions made in the following specifications: 1 mL of ice-cold Wash Buffer 1 (0.1% SDS, 0.1% DOC 1% Triton X-100, 1 mM EDTA, 10 mM Tris-HCl pH 8, 150 mM NaCl), 1 mL of Wash buffer 2 (0.1% SDS, 0.1% DOC 1% Triton X-100, 1 mM EDTA, 10 mM Tris-HCl pH 8, 500 mM NaCl), 1 mL of Wash Buffer 3 (10 mM Tris-HCl, pH 8, 250 mM LiCl, 0.5% Triton X-100, 0.5% DOC), and 1 mL of Wash Buffer 4 (10 mM Tris-HCl, pH 8.5) three times each sequentially. Beads were then resuspended in elution buffer with 5 mM DTT (1x Tris-HCl, pH 8, 0.1% SDS, 150 mM NaCl) and incubated at 65°C for an hour. The protein-DNA complex was eluted from the beads (in 1x Tris-HCl, pH 8, 0.1% SDS, 150 mM NaCl and 5 mM DTT), subjected to RNAse (Roche 11119915001) treatment at 37°C for 30 minutes, reverse crosslinked by Proteinase K (Lifetech 100005393) overnight incubation at 65°C and the DNA purified using SPRI beads (Beckman Coulter Agencourt AMPXpure A63882). Purified ChIP DNA was

quantified by qPCR using a set of on-target primers (oRC2731, oRC2732) overlapping the vCBS ~2KB upstream of MYC TSS along with set of negative site primers (oRC2739, oRC2740) targeting a region with no CTCF enrichment by publicly available ChIP-seq data, to serve as background binding values. qPCR experiments were run on a LightCycler 480 System (Roche 05015243001) under established cycling conditions (1 denaturation cycle at 95°C for 20 s, 45 amplification cycles at 95°C for 3 s followed by 60°C for 30 s). All ChIP experiments shown were performed in biological triplicates and each individual biological sample was qPCR amplified in technical triplicate.

Quantitation of *MYC* expression by qRT-PCR

RNA was isolated from ~1 x 10^6 cells from the same samples that were used for ChIP-qPCR, using the NucleoSpin RNA Plus kit (TakaraBio 740984.250) and reverse transcribed with the High Capacity RNA-to-cDNA Kit (Thermo 4387406) following the manufacturer's instructions. Quantitative real-time PCR was performed on the LightCycler 480 System (Roche 05015243001) using primers specific to the endogenous copy of *MYC* and SYBR green PCR master mix (Thermo 4309155). qPCR was performed with the same cycling conditions as described above for ChIP-qPCR. Ct values over 35 were marked as 35 because Ct values have been shown to considerably fluctuate for very low expressed transcripts. All experiments shown

were performed in biological triplicates and each individual biological sample was assayed in technical triplicates. Expression of *HPRT*, endogenous *MYC*, and exogenous *MYC* were quantified by the following primers: HPRT: oRC3150, 3151 (5'-CATTATGCTGAGGATTTGGAAAGG -3'; 5'- CTTGAGCACACAGAGGGCTACA -3'); endogenous *MYC*: oRC3045, oRC3046 (5'- AACCTCACAACCTTGGCTGA -3'; 5'-TTCTTTTATGCCCAAAGTCCAA -3'); oRC3047, oRC3048 (5'-TGATCCTAGCAGAAGCACAGG -3'; 5'- TGGACGAGCTGTACAAGAGC -3'). Expression was normalized across samples to *HPRT*, and relative endogenous *MYC* expression of experimental conditions were calculated based on endogenous *MYC* expression levels in the consensus cell line (wild-type control).

Chapter 3:

CTCF variants alter gene expression via de novo binding events and topological alteration of the genome

Rebecca T. Cottman[1,3], Jay W. Jun[1], Sowmya Iyer[1], Caleb A. Lareau[1,4,5], Martin J. Aryee[1,4,5], J. Keith Joung[1,2]

[1]Molecular Pathology Unit, Center for Cancer Research, and Computational and Integrative Biology, Massachusetts General Hospital, Charlestown, MA, [2]Department of Pathology, Harvard Medical School, Boston, MA, [3]Department of Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, [4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, [5]Broad Institute of MIT and Harvard, Cambridge, MA.

All authors listed contributed to the work described in this chapter. I designed and performed all wet-lab experiments. Jay Jun helped with ChIP experiments. Sowmya Iyer performed informatic analysis of ChIP-seq and RNA-seq datasets. Caleb Lareau performed informatic analysis of 4C-seq datasets. Martin Aryee provided direction in 4C data analysis. Keith Joung oversaw research and provided direction.

Introduction

The mechanisms of CTCF-mediated gene regulation vary from physical barrier, protein-protein interactions or recruitment of transcriptional machinery. General trends have been observed for CTCF-mediated positive regulation of genes. 90% of genes under the positive regulation of CTCF have a CTCF binding in an orientation that points the N-terminal region in the direction of gene transcription[19]. The N-terminal region of CTCF has been identified to recruit and interact with the largest subunit of RNA polymerase II and colocalize to overlapping regions of the genome at intergenic regions suggesting a mechanism for transcriptional control through polymerase recruitment[114]. CTCF interaction with RNA polymerase II does not alway have a positive effect on gene expression. Stalling of the RNA polymerase II by CTCF has been implicated in regulation of splicing events at the *CD5* locus by binding on top of exon 5 and resulting in polymerase pausing[45]. CTCF mediated expression of a gene can also be the result of physical occlusion of the spread of silencing epigenetic markers across the gene TSS. Binding of a CTCF upstream or within the intronic regions of a gene has been demonstrated to prevent the spread of nucleosomes and displace nucleosomes from its binding site as well as prevent the spread of repressive methylation marks[63,115]. Downregulation of genes by CTCF is much less complex and often the result of TAD boundaries insulating a gene body away from active enhancer regions[2,19,37]. Genes downregulated by CTCF are enriched for binding events further away from the gene promoter region while genes that are under the positive regulation of CTCF have a higher frequency of binding events proximal to the gene TSS[19,21]. CTCF mediated

protein-protein interactions have been implicated in regulation of gene transcription with corecruitment of CTCF with SIN3A or GATA-1 for locus specific regulation[51,72]. CTCF binding genome wide and subsequent gene regulation can be a result of large or small-scale topological changes. We endeavored to investigate whether CTCF variants are capable of binding across the genome and if so, if they alter gene expression suggesting they are capable of reproducing the mechanisms of CTCF-driven gene regulation.

Results

Engineered CTCF variants bind to their variant CBS genome-wide and alter gene expression.

CTCF has a role in gene regulation through the formation of TADs and subTADs. Using ChIP-seq analyses we determined the genome-wide occupied sites of the top variant from group 2, 4, and 5 selections (CTCF variant 2.2, 4.1, and 5.1). We performed ChIP with HA antibody to detect variant specific binding events and not conflate them with endogenous CTCF binding. CTCF variant 2.2 had 27,749 new binding sites across the genome, while CTCF variants 4.1 and 5.1 had fewer *de novo* binding events (54 and 24,666 respectively). Homer Motif analysis of the ChIP-seq dataset was used to identify sequence motifs of the CTCF variants in the context of the eukaryotic genome (**Figure 3.1a)**. Next, we explored changes in transcriptome that may implicate the CTCF variant binding events in directing topological changes in the genome. RNA-seq analysis was performed on K562 cells transfected with CTCF variants 2.2, 4.1, or 5.1. 141 genes and 1,141 genes in K562 cells transfected with CTCF variant 2.2 and 5.1 respectively, had altered transcript levels when compared to the cells overexpressing wild-type CTCF, while cells expressing CTCF variant 4.1 had no changes in transcription **(Figure 3.1b-d)**. Volcano plots of all protein-encoding genes were segregated based on the proximity of a CTCF variant binding event to within 2 kb of a gene TSS (green), overlapping an H3K27Ac region within 1 Mb of the gene TSS (red), binding within 2 kb of the gene TSS and overlapping an H3K27Ac region (blue), or no detected CTCF variant binding following any of the criteria **(Figure 3.1c-d)**.

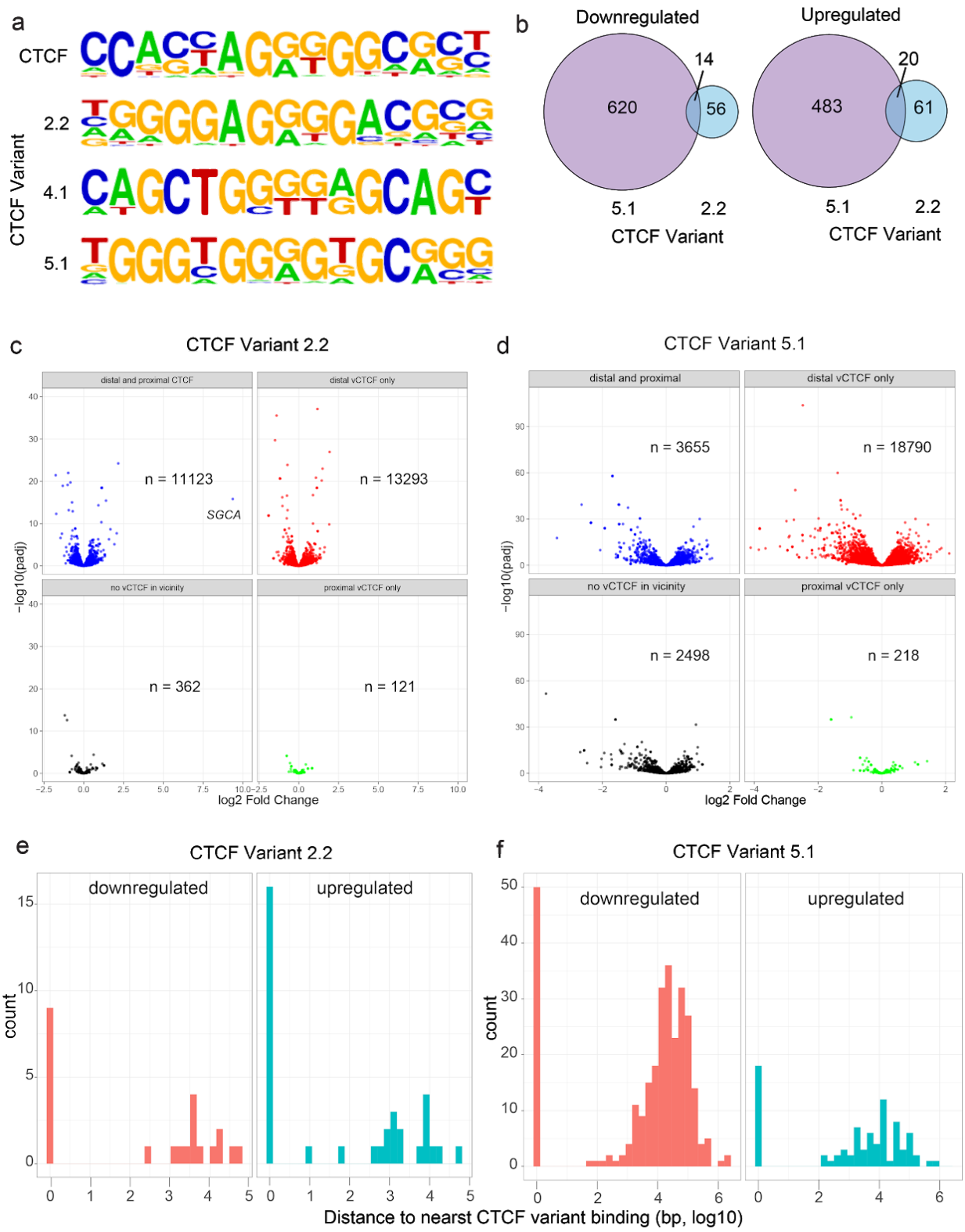
a
CTCF
CTCF Variant
2.2
4.1
5.1
b
Downregulated
620
14
56
5.1
2.2
CTCF Variant
Upregulated
483
20
61
5.1
2.2
CTCF Variant
c
CTCF Variant 2.2
distal and proximal CTCF
distal vCTCF only
n = 11123
SGCA
n = 13293
no vCTCF in vicinity
proximal vCTCF only
n = 362
n = 121
−log10(padj)
log2 Fold Change
d
CTCF Variant 5.1
distal and proximal
distal vCTCF only
n = 3655
n = 18790
no vCTCF in vicinity
proximal vCTCF only
n = 2498
n = 218
−log10(padj)
log2 Fold Change
e
CTCF Variant 2.2
downregulated
upregulated
count
f
CTCF Variant 5.1
downregulated
upregulated
count
Distance to nearst CTCF variant binding (bp, log10)

Figure 3.1: CTCF variants bind to unique sites across the genome and alter gene expression. a, CTCF and CTCF variant 15 bp motifs. Motifs generated from analysis of the top 10k ChIP-seq peaks with a p value cutoff of 0.0001 for each. **b,** Venn diagrams of the total number of downregulated and upregulated genes in K562 cells transfected with CTCF variant 2.2 or 5.1 CTCF variant 4.1 did not result in any significant changes in gene expression. Genes with altered gene expression in both group 2 and 5 are in the overlapping region of the diagram. Genes included as altered had a significance threshold of p value 0.05. **c-d,** Volcano plot of gene expression data of all protein coding genes in samples transfected with CTCF variants 2.2 or 5.1. Plots are quartered into groups based on distance from the gene TSS of a CTCF variant binding. CTCF variant overlapping an H3K27Ac site within 1 Mb (red). within 2 kb of TSS (green), both within 2kb of TSS and within the H3K27Ac region (blue), and finally none detected (black). Plots reflect fold changes compared to expression levels of K562 cells transfected with CTCF plasmid off of the same promoter. Values are the result of quadruplicate replicates. **e-f,** Histograms of CTCF variant binding frequency across distance from the TSS of genes with altered expression.

We also wanted to know if there were any trends unique to upregulation or downregulation of genes based on distance from the gene TSS of the CTCF variant binding event. It has been shown that genes downregulated by CTCF have a greater frequency of CTCF binding events overlapping the TSS of the target gene[19]. Using the combined RNA-seq and ChIP-seq datasets, we binned all protein expressing genes in the human genome based on the proximity of a CTCF variant binding event **(Figure 3.1e-f)**. For cells transfected with CTCF variant 2.2 or 5.1, we looked at how many genes had a CTCF variant binding event within a distance range of 0-6 kb from the gene TSS for the downregulated and upregulated genes. For variant 2.2, a higher percentage of upregulated genes had a CTCF variant binding overlapping the gene TSS (26.7% of upregulated genes vs 16.1% of downregulated genes). CTCF variant 5.1 had a trend more similar to wild-type CTCF where downregulated genes had a greater percentage of binding events overlapping the gene TSS (3.9% of upregulated genes vs 8.1% of downregulated genes). The remainder of the binding events occurred more than 2 kb away from the TSS with a slight shift of CTCF variant binding events closer to the TSS for upregulated genes.

CTCF variant directed alteration in gene expression suggests the formation of *de novo* looping events.

Of the genes with altered expression, *SGCA*, a gene specific for striated muscle cells and not normally expressed in the erythroid lymphoblastoid K562 cell line, was upregulated ~662 fold in K562 cells transfected with CTCF variant 2.2 (**Figure 3.1c**). ChIP-seq reveals a *de novo* CTCF variant 2.2 occupied site ~300 bp upstream of the *SGCA* TSS oriented in the same

direction of gene transcription (**Figure 3.2a**). CTCF variants 4.1 and 5.1 did not alter *SGCA* expression nor were there detectable binding events within a 1 Mb region of the TSS of *SGCA*. Based on this, we hypothesized that the activation of *SGCA* is driven by the variant specific binding event of CTCF variant 2.2. To determine whether the binding of CTCF variant 2.2 was activating expression of *SGCA* via topological changes in the region, we performed 4C analysis on samples transfected with wild-type CTCF or CTCF variant 2.2 using the *de novo* binding site proximal to the TSS of *SGCA* as the viewpoint. We found that in samples transfected with the CTCF variant 2.2, there was an overall increase in association with the region compared to samples transfected with exogenous CTCF or the transfection control, however it was not statistically significant compared to the negative control (p value 0.32) **(Figure 3.2b)**. We also took a more targeted approach and looked at the association between the viewpoint and binding events of CTCF and CTCF variant 2.2 with compatible orientation for loop formation with the CTCF variant bound proximally to the TSS of *SGCA* within the same 1.5 Mb window. Of the 29 compatible binding events investigated, only one, involving a CTCF bound region, had a significant alteration in association in samples transfected with CTCF variant 2.2 **(Figure 3.2b-inset)**. The enhancer region immediately upstream of the CTCF bound region had a slight increase in association with the *SGCA* promoter region, but it was not significant (p value 0.57).

CA11 is another gene identified to be upregulated (~4 fold) in cells transfected with the CTCF variant 2.2. As with *SGCA*, there was a new CTCF variant specific peak ~600 bp upstream of the *CA11* TSS. This binding event was also oriented in the same direction of transcription with the N-terminus proximal to the TSS. We performed 4C with a viewpoint

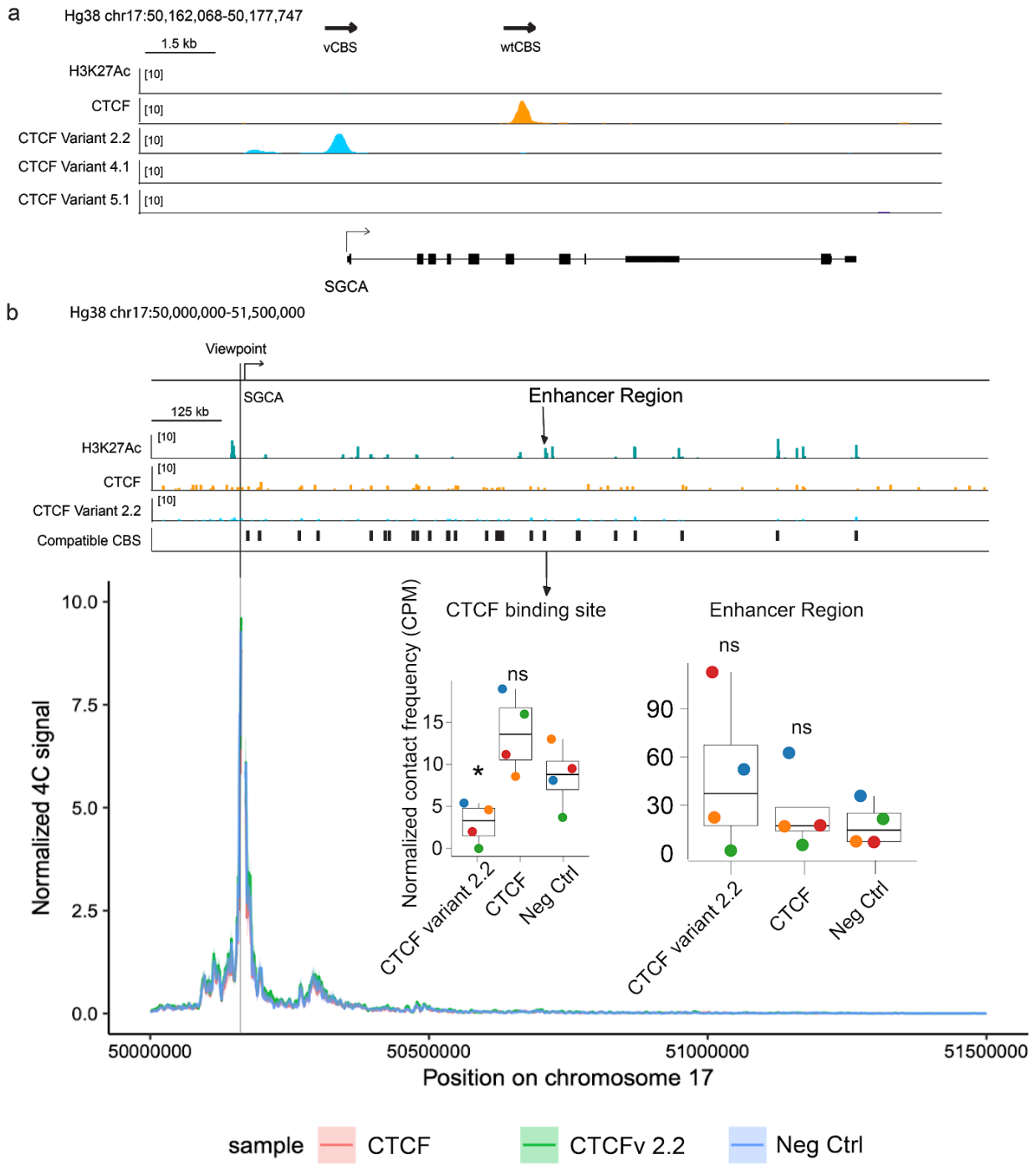

Figure 3.2: CTCF variant binding event results in *de novo* topological gene activation of a striated muscle specific gene *SGCA*. a, *SGCA* loci with ChIP-seq tracks H3K37Ac (teal) in K562s, CTCF binding (orange), and CTCF variant 2.2 binding events (blue). ChIP-seq tracks for CTCF variant 4.1 (yellow) and 5.1 (purple) are also shown. Orientation of the CTCF variant and CTCF binding is indicated by arrows (C-terminal to N-terminal direction). **b,** 4C analysis of

alterations in genome organization around the *SGCA* loci. CTCF variant binding orientation is indicated by direction of arrow, with the head of the arrow reflecting the N-terminal end of CTCF. Loop-compatible variant and wildtype CBSs are indicated by black bars. 4C analysis was performed on quadruplicate replicates. Viewpoint is ~700 bp downstream of CTCF variant specific binding event and marked with a line. 4C traces are the mean normalized 4C reads, with SEM as shaded area, across the region in cells transfected with CTCF variant 2.2 (green), wt CTCF (blue), or a pCMV-GFP plasmid as a negative control (red). Inset box plots are the frequency of association to the indicated viewpoint of the regions marked by arrows. CTCF binding site is a region with a CTCF binding, enhancer region overlaps an H3K27Ac mark immediately upstream of the CTCF binding site. Box plots are mean of quadruplicate replicates Significance indicated by p values (ns ≥ .05, * < .05)

~600bp upstream of the CA11 TSS and found that cells transfected with CTCF variant 2.2 had an increased association with the region downstream of *CA11* when compared to the CTCF only or negative control samples **(Figure 3.3a-b)**. There are three regions that appear to have a dramatic increase in association with the *CA11* promoter region in the presence of CTCF variant 2.2 (green) compared to the controls. These regions appear to coincide with loop-compatible CTCF and CTCF variant binding events. *HPF1* and *POGK* are two examples of genes downregulated (~3 and ~2 fold respectively) by CTCF variant 2.2. In both cases, CTCF variant 2.2 binds directly on top of the gene TSS **(Figure 3.4a-b)**. Both loci normally have H3K27Ac marks overlapping the region flanking the TSS suggesting this is an active promoter region with gene transcription.

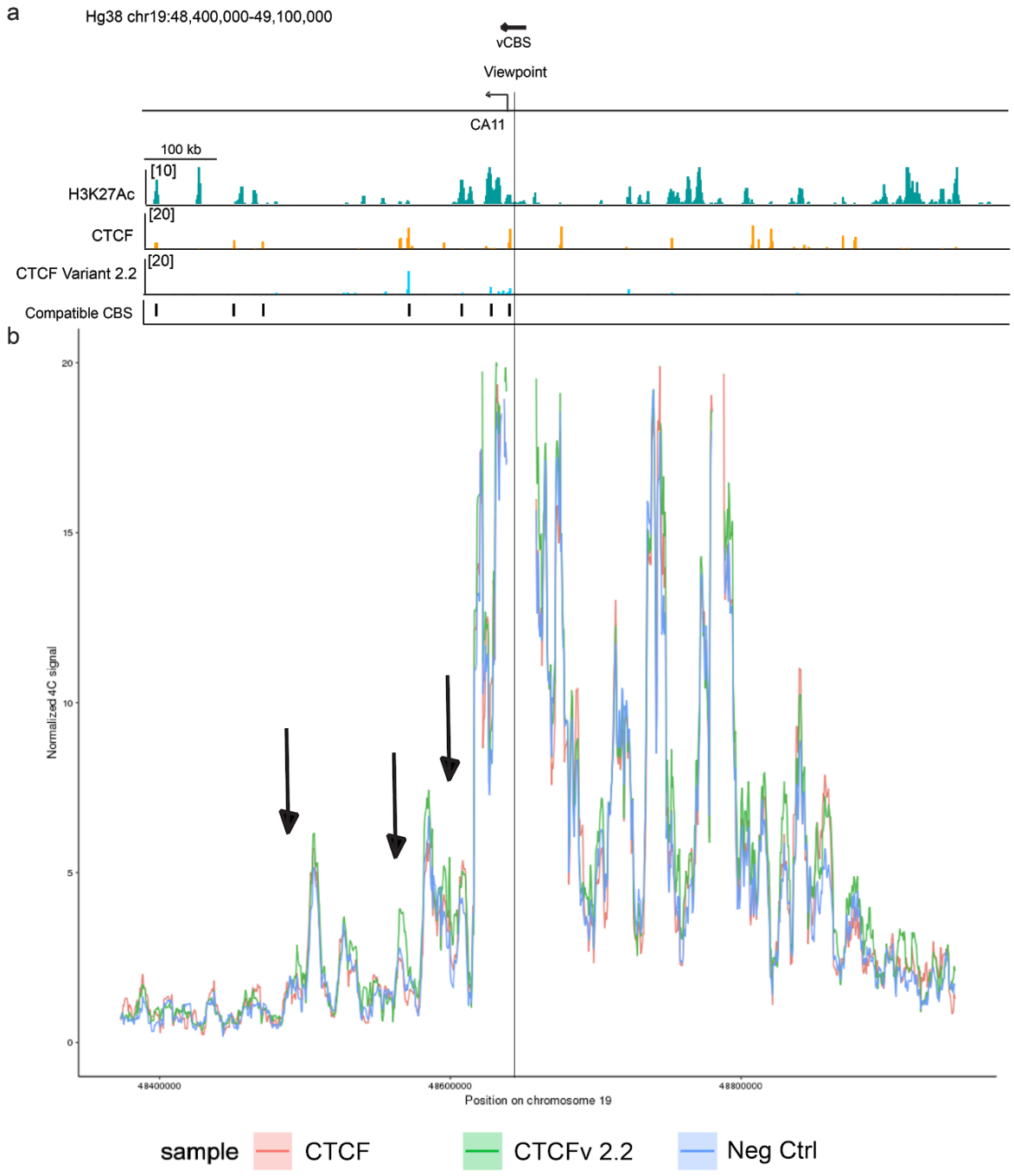

Figure 3.3: CTCF variant binding at the CA11 loci results in gene activation and topological changes. a, *CA11* loci with ChIP-seq tracks showing H3K37Ac (teal) regions in K562s, CTCF binding (orange), and CTCF variant 2.2 (blue). CTCF variant binding orientation is indicated by direction of arrow, with the head of the arrow reflecting the N-terminal end of CTCF variant. **b,** 4C analysis of alterations in genome organization around the *CA11* loci. Loop-compatible variants and wildtype CBSs are indicated by black bar. Viewpoint is ~700 bp

downstream of the CTCF variant specific binding event and marked with a line. 4C traces are the mean normalized 4C reads, with SEM as shaded area, across the region in cells transfected with CTCF variant 2.2 (green), wt CTCF (blue), or a pCMV-GFP plasmid as a negative control (red). Arrows indicate regions of increased association with the viewpoint in samples transfected with CTCF variant 2.2. 4C analysis was performed on quadruplicate replicates and data reflects the mean of normalized 4C reads across the region.

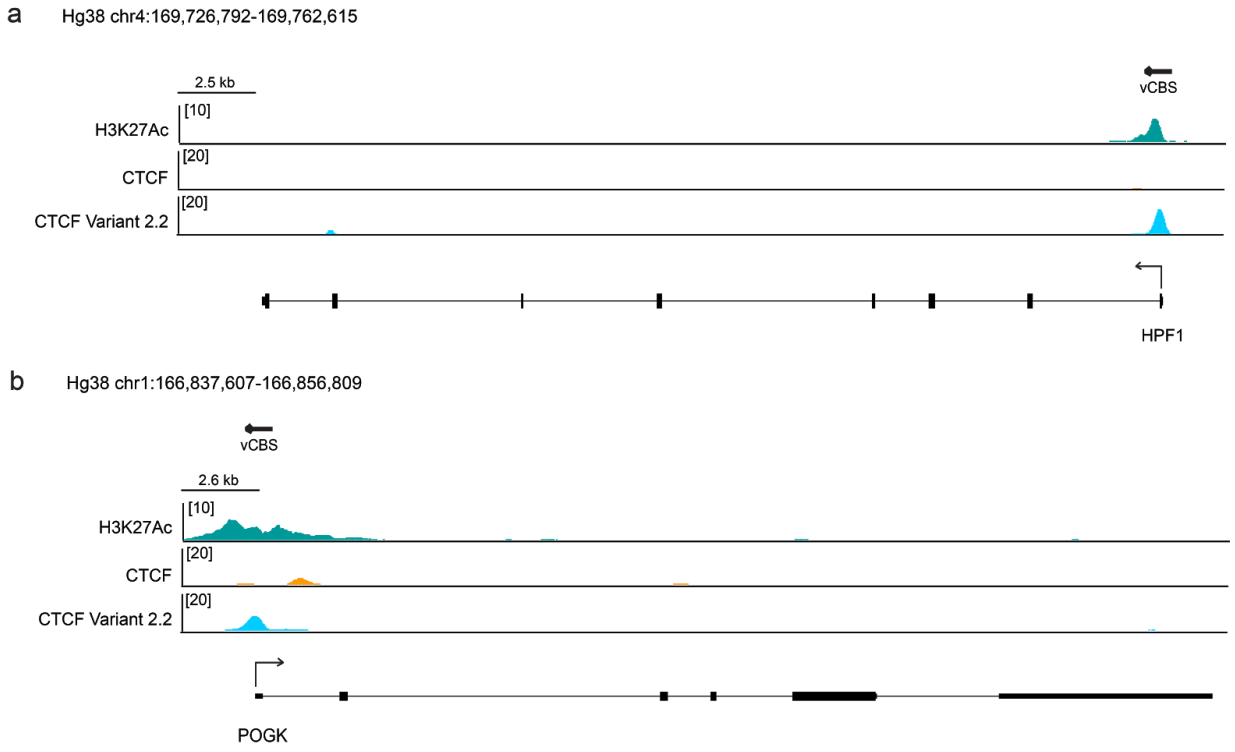

Figure 3.4: CTCF variant binding overlaps the TSS of downregulated genes HPF1 and POGK. a-b, *HPF1* and *POGK* locus with CTCF variant 2.2 ChIP-seq tracks (blue), CTCF binding events (orange), H3K27Ac regions in K562 cells (teal). Arrows indicate orientation of the CTCF variant 2.2 on the genome. *HPF1* is downregulated by ~2.4 fold and *POGK* is downregulated by ~1.7 fold in K562s transfected with CTCF variant 2.2.

Discussion

For the most part, CTCF variants had a binding motif that reflected the sequence the variant was selected to bind to (**Figure 3.1a**). However, motifs generated from CTCF variant binding events across the genome are likely impacted by two variables: the frequency of expected vCBS target sequence and the specificity of the CTCF variants. The vCBS sequence that existed most frequently across the genome with a mismatch of 1 or less was vCBS 2 with 67 sites. vCBS 4 and 5 had 13 and 10 sites of 1 mismatch or less. Therefore, it is not surprising that the vCBS 2 sequence used in the B2H selection was the most similar to the vCBS motif derived *in cellula*. Although vCBS 4 had few exact matches to the vCBS sequence *in cellula*, the motif still resembled the vCBS sequence. The conflict in motif generated for CTCF variant 4.1 may also be biased by a lack of sample size as there were only ~50 binding events across the genome. CTCF variants 2.2 and 5.1 had a similar number of binding events in K562s, but only CTCF variant 2.1 had a motif similar to its target sequence. The motif for CTCF variant 5.1 is the most dissimilar from the vCBS sequence used in the selection and is likely the result of both the vCBS sequence not existing with great frequency across the genome and the lack of specificity of this variant for its target sequence.

Gene expression changes in cells transfected with CTCF variant 2.2 and 5.1 show a distinct profile of upregulated or downregulated genes for each, with CTCF variant 5.1 resulting in a larger number of downregulated genes than upregulated **(Figure 3.1b)**. Observations have been made connecting the position of CTCF binding in relation to the gene TSS and the outcome

of gene regulation[19]. ~80% of Genes under positive regulation of CTCF have a CTCF occupied binding site within 2 kb of the gene TSS[19,21]. Genes that were upregulated by the CTCF variants did not appear to follow this pattern (**Figure 3.1b-e)**. No binding events were detected within 1-2 kb of the gene TSS for genes upregulated by CTCF variant 5.1 and there was no increased frequency of binding events in active regulatory regions for upregulated genes compared to downregulated genes. It is still possible that binding of the CTCF variant upregulates genes through prevention of nucleosome occlusion of the gene TSS rather than promoter-enhancer looping. In fact, of the genes under positive regulation by endogenous CTCF, only ~38% of those genes have an additional CTCF binding event in a regulatory region, either overlapping an SMC1a anchor site or an H3k27Ac region[19]. This suggests the remaining 62% of genes with a CTCF binding in the promoter region of genes upregulates gene expression, possibly by acting as a barrier to nucleosome occlusion[19,20,41]. In genes that are upregulated by CTCF, almost all binding events (90%) proximal to the gene TSS are in the same orientation as gene transcription, which may be a better predictor of CTCF variant activity genome-wide[19]. It may be possible to query the ChIP-seq and RNA-seq datasets for the orientation of CTCF variant binding sites proximal to the TSS of upregulated and downregulated genes to find if they follow a pattern similar to endogenous CTCF. For both CTCF variants, there were genes that had altered gene expression with no detectable CTCF variant binding event within a 1 Mb range of the TSS (**Figure 3.1d-e**). These genes may be downstream in an expression regulatory pathway controlled by another gene altered in expression by the CTCF variant. It is also possible that these genes are being mediated by a CTCF variant interaction greater than 1 Mb away from the gene TSS, which has been observed for endogenous CTCF[20,76].

We observed the presence of weak CTCF peaks overlapping CTCF variant peaks in ChIP-seq samples only transfected with variant CTCF and not in samples transfected with an endogenous source of CTCF or pMaxGFPAnother. This may support a potential confounding factor with this ChIP-seq based analysis as the CTCF variants can interact with endogenous CTCF to affect the transcriptome. Performing this analysis on only CTCF variant binding events will always render an incomplete image of the actual mechanisms at play for genome-wide effects on transcription. In addition, the previous studies were able to draw patterns of CTCF binding distance to gene TSS based on ~60,000 CTCF binding events and almost 5,000 gene transcription changes while we were working with ~25,000 binding events and 1,100 gene expression changes for this analysis. The limited power of our dataset and frequency of binding events impedes any conclusion on functional profiling of CTCF variant binding distance to gene TSS. A further topic of research would be to use the ChIP-seq and RNA-seq data sets to try and determine a pattern of fold change and location of binding. The biological mechanisms of CTCF-mediated gene activation is not entirely clear and as such it is difficult to identify the cofactors at play in the gene expression changes seen with expression of the CTCF variants.

This decline in association with the promoter region of *SGCA* at a CTCF bound site may hint at a topological alteration that results in the activation of *SGCA* **(Figure 3.2)**. Based on the evidence. However, it is also possible that the CTCF variant binding event at the *SGCA* locus prevents occlusion of the TSS by nucleosomes, a form of CTCF-mediated gene activation previously described[19,63]. The topological changes observed by 4C of the genomic region downstream of *CA11* is the closest evidence linking CTCF variant topological changes with gene activation **(Figure 3.3)**. These data suggests a *de novo* loop is formed between a CTCF variant

binding upstream of the TSS of *CA11* and a second, or possibly multiple, CTCF variant and wild-type CTCF binding event downstream. The most proximal increase in association overaps is overlapped by a H3K27Ac region, indicating the increased activation of *CA11* may be the result of this enhancer looping to the promoter region via the new CTCF variant binding binding in a loop-compatible orientation. In both the case of *SGCA* and *CA11*, the causal enhancer or binding site can be validated by doing targeted deletion of implicated enhancer region, or more elegantly, the CTCF or CTCF variant binding sites, and observing the impact on gene expression as well as association to the gene promoter.

In the cases of downregulation of *HPF1* and *POGK*, the binding of the CTCF variant directly on top of the gene TSS is the likely mechanistic explanation **(Figure 3.4)**. It is interesting to note that the orientation of the CTCF variant bound to the site is not in the opposite direction of transcription. This would suggest that the method of repression is through physical occlusion of the start codon, perhaps acting as a physical barrier to read through of the transcription elements assembled at the 5' UTR.

Materials and Methods

4C-Seq

Transfection and chromatin fixation of the exoMYC.K562-vCBS cells lines with either CTCF variant or WT CTCF plasmid DNA was done as previously described for ChIP using 1 x 10^7 cells per sample. 4C sample processing was performed following published protocol (van de Werken et al., 2012, Ref. 116) with a slight modification in the isolation of nuclei (Rao et al., 2014, Ref. 117). NlaIII (New England Biolabs) was used as the primary restriction enzyme followed by BfaI (New England Biolabs) as the secondary restriction enzyme. 4C-seq library was prepared by amplifying 3.2 µg of 4C sample across 16 individual PCR reactions with primers oRC3050 (5'- GCGCGCGTAGTTAATTCATG -3'), oRC3051 (5'- AAAGAAGGGTATTAATGGGC -3') following Roche Expand Long-template Polymerase protocol (Sigma #11681842001). The PCR reactions were purified, eluted, and pooled as described in (van de Werken et al., 2012). Each 4C sample was prepared with a unique UMI using Rubicon DNA-seq library kit, 48S (Takara #R400675) following manufacturer's protocol. The prepared libraries were pooled in equal molar amounts and sequenced to produce at least 10 million 76 cycle single-end reads per sample, on the Illumina NextSeq 500 system.

4C-Seq data processing

Reads containing specific 4C adapters for each viewpoint were identified and trimmed using cutadapt (Martin 2011) and aligned to the hg38 reference genome with bowtie2 (Langmead and Salzberg 2012). Counts overlapping regions of interest were determined using the GenomicRanges countOverlaps function (Lawrence et al., 2013) and normalized to the total number of fragments detected on the viewpoint chromosome. Statistical significance was determined from the Wald statistic p-value for the biological treatment term from the linear model for the log2 normalized counts while accounting for the replicate as a fixed effect covariate. Smoothed 4C track visualizations were computed taking the mean over replicates in 6000bp bins with a 500bp step size as previously reported (Schuijers et al., 2018).

ChIP-Seq Sample handling

Samples for ChIP-seq were processed in the same manner as described above for ChIP-qPCR, with the exception of a final elution of DNA in 13 mL of 10 mM Tris pH 7.5. The eluted DNA was used to prepare libraries for sequencing using the Rubicon DNA-seq library kit, 48S (Takara #R400675). Final libraries were sequenced in single-read mode for 76 bases on an Illumina NextSeq 500 system.

ChIP-Seq data processing

Reads were aligned to the hg38 reference genome using bwa (PMID:20080505). After removal of PCR duplicates, normalized signals were generated using bedtools (https://doi.org/10.1093/bioinformatics/btq033). MACS2(PMID:18798982) was used for calling peaks using input DNA as control with a setting of 0.0001 for q-value thresholds. Top 10000 peaks based on peak score were selected for motif identification using Homer (PMID: 20513432)

RNA-Seq

RNA-seq was performed to identify transcriptome-wide gene expression changes that were specific to engineered CTCF binding events. EGFP-tagged CTCF variants were transfected into K562 cell line and sorted for EGFP positivity following RNA extraction. K562 cells were transfected in triplicate using Lonza Kit V K562 using the manufacturer's protocol. 3 x 10^6 cells were nucleofected with a total of 15 µg of CTCF variant-P2A-EGFP plasmid and cultured for 72 hours in previously described culture conditions. A minimum of 2 x 10^5 EGFP positive cells were sorted and RNA extracted using Nucleospin RNA Plus (Takara Clonetech # 740984.250) following manufacturer's protocol. Total RNA cDNA synthesized from 50 ng of total RNA was

used for preparing NGS libraries using TruSeq Stranded Total RNA Library Prep Gold (Illumina 20020598) following the manufacturer's instructions. Replicates of each uniquely molecularly indexed CTCF variant NGS library were pooled and sequenced on a NextSeq 500 system, with the aim of achieving 20 million 76 cycle, dual indexed, paired end reads for each sample.

RNA-Seq data processing

Reads were aligned to hg38 using STAR(https://doi.org/10.1093/bioinformatics/bts635), followed by removal of PCR duplicates and removal of reads that aligned to ribosomal RNA. Gene expression was quantified using featureCounts(doi: 10.1093/bioinformatics/btt656) and differential expression analysis was performed using DESeq2(doi: 10.1186/s13059-014-0550-8).

Conclusion

Through the bacterial-2-hybrid system we showed which residues within the CBS were critical to maintaining CTCF-binding, interestingly these critical residues are also sites where mutations that lead to tumorigenesis are located, granting insight into which of the accumulated mutations are causal in driving cancer progression, which can be used as a target for drug design. A majority of eukaryotic transcription factors contain a zinc finger array in their DNA binding domain, such as YY1, TFIIIA, FOG, and GATA-1[70,98,118,119]. The bacterial selection system could be applied to these transcription factors to make a suite of gene regulating proteins that recognize a desired sequence.

We used this system that places the endogenous human *MYC* gene under heterologous CTCF regulation to study the mechanistic requirements for CTCF-mediated *MYC* expression. Previous studies have shown that CTCF can generate topological loops with either Cohesin or RNA as a cofactor. Mutations of the RNA-binding domain of the CTCF variants did not affect binding to its cognate vCBS, but resulted in reduced *MYC* expression. By contrast, removal of the Cohesin-interacting peptides within the C-terminal region of the CTCF variant did not affect binding or reduce *MYC* expression. We could not conclude however, that the loop established at the *MYC* locus was Cohesin independent as more recent publications indicate the peptide critical for Cohesin-CTCF interaction resides in the N-terminal domain of CTCF[106,107]. Looping at the *MYC* locus is therefore mediated by a CTCF-Cohesin-RNA model or a CTCF-RNA model as there are examples of CTCF relying solely on RNA in regulating gene expression[38,50].

We also demonstrate that the vCBS and the cognate CTCF variants can be potentially used to induce heterotopic gene expression by altering the topological organization of the genome. CTCF mediated gene activation could be the result of promote-enhancer looping, prevention of promoter occlusion by nucleosomes, or recruitment of transcriptional machinery to a TSS. While CTCF mediated gene repression is the result of insulation of genes from near-by enhancers and blocking transcription machinery or occlusion of the gene TSS by overlapping binding. The gene expression changes observed with expression of the CTCF variants can provide a system for independent study of what, along with CTCF, created a genomic environment for gene regulation. As CTCF has been observed to interact with RNA, protein, and DNA, the machinery assembled for gene activation or repression may vary by locus. The orthogonality of some of the CTCF variants could be used to investigate the complex networks of CTCF-mediated gene regulation at these loci. Furthermore, Hi-C experiments with the CTCF variants will provide a clear picture of the interaction between CTCF variants and endogenous CTCF. Comparing Hi-C datasets from cells expressing the CTCF variants and the wild-type condition would allow for the isolation of CTCF variant-mediated topological changes in the genome and provide evidence for TAD structures composed of CTCFv-CTCFv or CTCFv-CTCFwt complexes.

We have developed a system to study the mechanistic requirements for CTCF-mediated gene expression without the confounding pleiotropic effects caused by altering the endogenous CTCF protein. Our results demonstrate the functionality of variant CTCFs as a substitute for endogenous CTCF function at promoter-enhancer loops. Using variant CTCFs we were able to elucidate RNA as a cofactor in maintaining *MYC* expression. The observed topological changes

in cells expressing our CTCF variants highlights how the topology of the genome can be manipulated for gene regulation. Generating CTCF variants with the ability to bind to vCBSs allows for new promoter-enhancer looping events for targeted gene regulation. This system could be used for creating synthetic gene circuits or treating diseases that result from haploinsufficient gene expression. Our results support existing literature about the mechanistic function of CTCF, but also highlights new discoveries and illustrates a method for determining co-factors in CTCF-driven gene regulation. In conclusion, this novel system we have developed enables the investigation of the mechanisms behind CTCF-driven loops at a single locus. This method also allows for epigenome engineering via directed topological changes to alter gene expression.

References

1. Heger, P., Marin, B., Bartkuhn, M., Schierenberg, E. & Wiehe, T. The chromatin insulator CTCF and the emergence of metazoan diversity. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 17507–17512 (2012).

2. Stadhouders, R., Filion, G. J. & Graf, T. Transcription factors and 3D genome conformation in cell-fate decisions. *Nature* **569**, 345–354 (2019).

3. Van Bortle, K. & Corces, V. G. Nuclear organization and genome function. *Annu. Rev. Cell Dev. Biol.* **28**, 163–187 (2012).

4. Chen, H. *et al.* Dynamic interplay between enhancer–promoter topology and gene activity. *Nat. Genet.* **50**, 1296–1303 (2018).

5. Lim, B., Heist, T., Levine, M. & Fukaya, T. Visualization of Transvection in Living Drosophila Embryos. *Mol. Cell* **70**, 287–296.e6 (2018).

6. Shin, Y. *et al.* Liquid Nuclear Condensates Mechanically Sense and Restructure the Genome. *Cell* **175**, 1481–1491.e13 (2018).

7. Hyman, A. A., Weber, C. A. & Jülicher, F. Liquid-liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol.* **30**, 39–58 (2014).

8. Maeshima, K., Ide, S., Hibino, K. & Sasai, M. Liquid-like behavior of chromatin. *Curr. Opin. Genet. Dev.* **37**, 36–45 (2016).

9. 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).

10. Fang, C. *et al.* Cancer-specific CTCF binding facilitates oncogenic transcriptional dysregulation. *bioRxiv* 2020.01.17.910687 (2020) doi:10.1101/2020.01.17.910687.

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

12. van Steensel, B. & Belmont, A. S. Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression. *Cell* **169**, 780–791 (2017).

13. Vieux-Rochas, M., Fabre, P. J., Leleu, M., Duboule, D. & Noordermeer, D. Clustering of mammalian Hox genes with other H3K27me3 targets within an active nuclear domain. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 4672–4677 (2015).

14. Wijchers, P. J. *et al.* Characterization and dynamics of pericentromere-associated domains in

mice. *Genome Res.* **25**, 958–969 (2015).

15. Wijchers, P. J. *et al.* Cause and Consequence of Tethering a SubTAD to Different Nuclear Compartments. *Mol. Cell* **61**, 461–473 (2016).

16. Therizols, P. *et al.* Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. *Science* **346**, 1238–1242 (2014).

17. Harr, J. C. *et al.* Directed targeting of chromatin to the nuclear lamina is mediated by chromatin state and A-type lamins. *J. Cell Biol.* **208**, 33–52 (2015).

18. Yusufzai, T. M., Tagami, H., Nakatani, Y. & Felsenfeld, G. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol. Cell* **13**, 291–298 (2004).

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

20. Hyle, J. *et al.* Acute depletion of CTCF directly affects MYC regulation through loss of enhancer-promoter looping. *Nucleic Acids Res.* **47**, 6699–6713 (2019).

21. Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 996–1001 (2014).

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

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

24. Bailey, S. D. *et al.* ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat. Commun.* **2**, 6186 (2015).

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

26. Rada-Iglesias, A., Grosveld, F. G. & Papantonis, A. Forces driving the three-dimensional folding of eukaryotic genomes. *Mol. Syst. Biol.* **14**, e8214 (2018).

27. Wei, Z. *et al.* Klf4 organizes long-range chromosomal interactions with the oct4 locus in reprogramming and pluripotency. *Cell Stem Cell* **13**, 36–47 (2013).

28. Abboud, N. *et al.* A cohesin-OCT4 complex mediates Sox enhancers to prime an early embryonic lineage. *Nat. Commun.* **6**, 6749 (2015).

29. Bintu, B. *et al.* Super-resolution chromatin tracing reveals domains and cooperative

interactions in single cells. *Science* **362**, (2018).

30. Flyamer, I. M. *et al.* Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. *Nature* **544**, 110–114 (2017).

31. Hansen, A. S., Cattoglio, C., Darzacq, X. & Tjian, R. Recent evidence that TADs and chromatin loops are dynamic structures. *Nucleus* **9**, 20–32 (2018).

32. Stevens, T. J. *et al.* 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**, 59–64 (2017).

33. Tan, L., Xing, D., Chang, C.-H., Li, H. & Xie, X. S. Three-dimensional genome structures of single diploid human cells. *Science* **361**, 924–928 (2018).

34. Sanborn, A. L. *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E6456–65 (2015).

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

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

37. Ali, T., Renkawitz, R. & Bartkuhn, M. Insulators and domains of gene expression. *Curr. Opin. Genet. Dev.* **37**, 17–26 (2016).

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

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

40. Oomen, M. E., Hansen, A. S., Liu, Y., Darzacq, X. & Dekker, J. CTCF sites display cell cycle-dependent dynamics in factor binding and nucleosome positioning. *Genome Res.* **29**, 236–249 (2019).

41. Zhang, H. *et al.* Chromatin structure dynamics during the mitosis-to-G1 phase transition. *Nature* **576**, 158–162 (2019).

42. Rudan, M. V. *et al.* Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Rep.* **10**, 1297–1309 (2015).

43. Beagrie, R. A. *et al.* Complex multi-enhancer contacts captured by genome architecture mapping. *Nature* **543**, 519–524 (2017).

44. Andrey, G. *et al.* A switch between topological domains underlies HoxD genes collinearity

in mouse limbs. *Science* **340**, 1234167 (2013).

45. Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74–79 (2011).

46. Hilmi, K. *et al.* CTCF facilitates DNA double-strand break repair by enhancing homologous recombination repair. *Science Advances* **3**, (2017).

47. Han, D., Chen, Q., Shi, J., Zhang, F. & Yu, X. CTCF participates in DNA damage response via poly(ADP-ribosyl)ation. *Sci. Rep.* **7**, 43530 (2017).

48. Wolfe, S. A., Nekludova, L. & Pabo, C. O. DNA Recognition by Cys2His2 Zinc Finger Proteins. *Annual Review of Biophysics and Biomolecular Structure* vol. 29 183–212 (2000).

49. Lobanenkov, V. V. *et al.* 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).

50. Kung, J. T. *et al.* Locus-specific targeting to the X chromosome revealed by the RNA interactome of CTCF. *Mol. Cell* **57**, 361–375 (2015).

51. Lutz, M. *et al.* Transcriptional repression by the insulator protein CTCF involves histone deacetylases. *Nucleic Acids Res.* **28**, 1707–1713 (2000).

52. Weth, O. *et al.* CTCF induces histone variant incorporation, erases the H3K27me3 histone mark and opens chromatin. *Nucleic Acids Res.* **42**, 11941–11951 (2014).

53. Monahan, K., Horta, A. & Lomvardas, S. LHX2- and LDB1-mediated trans interactions regulate olfactory receptor choice. *Nature* **565**, 448–453 (2019).

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

55. de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499–506 (2013).

56. Pugacheva, E. M. *et al.* Familial cases of point mutations in the XIST promoter reveal a correlation between CTCF binding and pre-emptive choices of X chromosome inactivation. *Hum. Mol. Genet.* **14**, 953–965 (2005).

57. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016).

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

59. Flavahan, W. A. *et al.* Altered chromosomal topology drives oncogenic programs in

SDH-deficient GISTs. *Nature* **575**, 229–233 (2019).

60. Poulos, R. C. *et al.* Functional Mutations Form at CTCF-Cohesin Binding Sites in Melanoma Due to Uneven Nucleotide Excision Repair across the Motif. *Cell Rep.* **17**, 2865–2872 (2016).

61. Guo, Y. A. *et al.* Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nat. Commun.* **9**, 1520 (2018).

62. Franke, M. *et al.* Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature* **538**, 265–269 (2016).

63. Gombert, W. M. & Krumm, A. Targeted deletion of multiple CTCF-binding elements in the human C-MYC gene reveals a requirement for CTCF in C-MYC expression. *PLoS One* **4**, e6109 (2009).

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

65. Fornes, O. *et al.* JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).

66. Persikov, A. V. & Singh, M. De novo prediction of DNA-binding specificities for Cys2His2 zinc finger proteins. *Nucleic Acids Res.* **42**, 97–108 (2014).

67. Hashimoto, H. *et al.* Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA. *Mol. Cell* **66**, 711–720.e3 (2017).

68. Nakahashi, H. *et al.* A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep.* **3**, 1678–1689 (2013).

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

70. Ryan, R. The role of zinc finger linkers in p43 and TFIIIA binding to 5S rRNA and DNA. *Nucleic Acids Research* vol. 26 703–709 (1998).

71. Davies, R. *et al.* Multiple roles for the Wilms' tumor suppressor, WT1. *Cancer Res.* **59**, 1747s–1750s; discussion 1751s (1999).

72. Kang, Y., Kim, Y. W., Kang, J., Yun, W. J. & Kim, A. Erythroid specific activator GATA-1-dependent interactions between CTCF sites around the β-globin locus. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* vol. 1860 416–426 (2017).

73. Laity, J. H., Lee, B. M. & Wright, P. E. Zinc finger proteins: new insights into structural and functional diversity. *Curr. Opin. Struct. Biol.* **11**, 39–46 (2001).

74. Sun, J. H. *et al.* Disease-Associated Short Tandem Repeats Co-localize with Chromatin Domain Boundaries. *Cell* **175**, 224–238.e15 (2018).

75. Kraft, K. *et al.* Serial genomic inversions induce tissue-specific architectural stripes, gene misexpression and congenital malformations. *Nat. Cell Biol.* **21**, 305–310 (2019).

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

77. Kang, J. Y. *et al.* Disruption of CTCF/cohesin-mediated high-order chromatin structures by DNA methylation downregulates PTGS2 expression. *Oncogene* **34**, 5677–5684 (2015).

78. Li, W. *et al.* Identification of critical base pairs required for CTCF binding in motif M1 and M2. *Protein Cell* **8**, 544–549 (2017).

79. Smith, G. P. Filamentous Fusion Phage: Novel Expression Vectors that Display Cloned Antigens on the Virion Surface. *Science* **228**, 1315–1317 (1985).

80. Bass, S., Greene, R. & Wells, J. A. Hormone phage: an enrichment method for variant proteins with altered binding properties. *Proteins* **8**, 309–314 (1990).

81. Choo, Y. & Klug, A. Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 11168–11172 (1994).

82. Jamieson, A. C., Kim, S. H. & Wells, J. A. In vitro selection of zinc fingers with altered DNA-binding specificity. *Biochemistry* **33**, 5689–5695 (1994).

83. Rebar, E. J. & Pabo, C. O. Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science* **263**, 671–673 (1994).

84. Wu, H., Yang, W. P. & Barbas, C. F. Building zinc fingers by selection: toward a therapeutic application. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 344–348 (1995).

85. Joung, J. K., Ramm, E. I. & Pabo, C. O. A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 7382–7387 (2000).

86. Giesecke, A. V., Fang, R. & Joung, J. K. Synthetic protein-protein interaction domains created by shuffling Cys2His2 zinc-fingers. *Mol. Syst. Biol.* **2**, 2006.2011 (2006).

87. Dove, S. L., Joung, J. K. & Hochschild, A. Activation of prokaryotic transcription through arbitrary protein-protein contacts. *Nature* **386**, 627–630 (1997).

88. Dove, S. L. & Hochschild, A. Conversion of the omega subunit of Escherichia coli RNA polymerase into a transcriptional activator or an activation target. *Genes Dev.* **12**, 745–754

(1998).

89. Kornacker, M. G., Remsburg, B. & Menzel, R. Gene activation by the AraC protein can be inhibited by DNA looping between AraC and a LexA repressor that interacts with AraC: possible applications as a two-hybrid system. *Mol. Microbiol.* **30**, 615–624 (1998).

90. Elrod-Erickson, M., Rould, M. A., Nekludova, L. & Pabo, C. O. Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. *Structure* **4**, 1171–1180 (1996).

91. Bonev, B. & Cavalli, G. Organization and function of the 3D genome. *Nat. Rev. Genet.* **17**, 661–678 (2016).

92. Buecker, C. & Wysocka, J. Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet.* **28**, 276–284 (2012).

93. Bulger, M. & Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell* **144**, 327–339 (2011).

94. de Wit, E. *et al.* The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature* **501**, 227–231 (2013).

95. Spitz, F. Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles. *Semin. Cell Dev. Biol.* **57**, 57–67 (2016).

96. Müeller-Storm, H. P., Sogo, J. M. & Schaffner, W. An enhancer stimulates transcription in trans when attached to the promoter via a protein bridge. *Cell* **58**, 767–777 (1989).

97. Fraser, J. *et al.* Hierarchical folding and reorganization of chromosomes are linked to transcriptional changes in cellular differentiation. *Mol. Syst. Biol.* **11**, 852 (2015).

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

99. Splinter, E. *et al.* CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Blood Cells, Molecules, and Diseases* vol. 38 178 (2007).

100. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **18**, 285–298 (2017).

101. Hansen, A. S. *et al.* Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region in CTCF. *Mol. Cell* **76**, 395–411.e13 (2019).

102. Saldaña-Meyer, R. *et al.* RNA Interactions Are Essential for CTCF-Mediated Genome Organization. *Mol. Cell* **76**, 412–422.e5 (2019).

103. Montavon, T. & Duboule, D. Landscapes and archipelagos: spatial organization of gene regulation in vertebrates. *Trends Cell Biol.* **22**, 347–354 (2012).

104. Ovcharenko, I. *et al.* Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15**, 137–145 (2005).

105. Xiao, T., Wallace, J. & Felsenfeld, G. Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity. *Mol. Cell. Biol.* **31**, 2174–2183 (2011).

106. Pugacheva, E. M. *et al.* CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 2020–2031 (2020).

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

108. Laity, J. H., Jane, W. D. H. & Wright, P. E. DNA-induced a-Helix Capping in Conserved Linker Sequences is a Determinant of Binding Affinity in Cys-His Zinc Fingers. *Journal of Molecular Biology* **295**, 719–727 (2000).

109. Wolfe, S. A., Ramm, E. I. & Pabo, C. O. Combining structure-based design with phage display to create new Cys(2)His(2) zinc finger dimers. *Structure* **8**, 739–750 (2000).

110. Caricasole, A. *et al.* RNA binding by the Wilms tumor suppressor zinc finger proteins. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 7562–7566 (1996).

111. Nolte, R. T., Conlin, R. M., Harrison, S. C. & Brown, R. S. Differing roles for zinc fingers in DNA recognition: Structure of a six-finger transcription factor IIIA complex. *Proceedings of the National Academy of Sciences* vol. 95 2938–2943 (1998).

112. Lu, D., Alexandra Searles, M. & Klug, A. Crystal structure of a zinc-finger–RNA complex reveals two modes of molecular recognition. *Nature* vol. 426 96–100 (2003).

113. Saldaña-Meyer, R. *et al.* CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53. *Genes Dev.* **28**, 723–734 (2014).

114. Chernukhin, I. *et al.* CTCF interacts with and recruits the largest subunit of RNA polymerase II to CTCF target sites genome-wide. *Mol. Cell. Biol.* **27**, 1631–1648 (2007).

115. Owens, N. *et al.* CTCF confers local nucleosome resiliency after DNA replication and during mitosis. *Elife* **8**, (2019).

116. van de Werken, H. J. G. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat. Methods* **9**, 969–972 (2012).

117. Rao, S. *et al.* A dual role for autophagy in a murine model of lung cancer. *Nat. Commun.* **5**,

3056 (2014).

118. Tsang, A. P. *et al.* FOG, a Multitype Zinc Finger Protein, Acts as a Cofactor for Transcription Factor GATA-1 in Erythroid and Megakaryocytic Differentiation. *Cell* vol. 90 109–119 (1997).

119. Merika, M. & Orkin, S. H. Functional synergy and physical interactions of the erythroid transcription factor GATA-1 with the Krüppel family proteins Sp1 and EKLF. *Molecular and Cellular Biology* vol. 15 2437–2447 (1995).