Brain and cancer associated binding domain mutations provide insight into CTCF's relationship with chromatin and its ability to act as a chromatin organizer

Catherine Do[1,2,*], Guimei Jiang[1,2,*] , Giulia Cova[1,2,*] , Christos C. Katsifis[3,4,5,φ], Domenic N. Narducci[3,4,5,φ], Jie Yang[6,φ], Theodore Sakellaropoulos[1,2], Raphael Vidal[1,2], Priscillia Lhoumaud[1,2], Faye Fara Regis[1,2], Nata Kakabadze[1,2], Elphege P Nora[7], Marcus Noyes[8], Xiaodong Cheng[6†], Anders S. Hansen[3,4,5†], Jane A Skok[1,2,*]

1. Department of Pathology, NYU Grossman School of Medicine, New York, NY, USA
2. Perlmutter Cancer Center, NYU Langone Health, New York, NY, USA
3. MIT Department of Biological Engineering
4. Gene Regulation Observatory, Broad Institute of MIT and Harvard, Cambridge, MA 02139, USA
5. Koch Institute for Integrative Cancer Research, Cambridge, MA, 02139, USA
6. Department of Epigenetics and Molecular Carcinogenesis, The University of Texas MD Anderson Cancer Center, Houston, TX 77030
7. Cardiovascular Research Institute, and Department of Biochemistry and Biophysics, University of California San Francisco, CA, USA
8. Institute for Systems Genetics, Department of Biochemistry and Molecular Pharmacology, NYU Langone Health, New York, NY, USA

[*, †, φ,] These authors contributed equally.
Corresponding author:
Jane.Skok@nyulangone.org

Emails:
Catherine Do: Catherine.Do2@nyulangone.org
Guimei Jiang: Guimei.Jiang@nyulangone.org
Christos Katsifis: katsifis@mit.edu
Domenic N Narducci: domenicn@mit.edu
Jie Yang: jieyang301@gmail.com
Giulia Cova: Giulia.Cova@nyulangone.org
Theodore Sakellaropoulos: Theodoros.Sakellaropoulos@nyulangone.org
Raphael Vidal: vidalraphaels@icloud.com
Priscillia Lhoumaud: priscillia.lhoumaud@ijm.fr
Faye Fara Regis: farafayedregis@yahoo.com
Nata Kakabadze: Nata.Kakabadze@nyulangone.org
Elphege Nora: elphege.nora@ucsf.edu
Marcus Noyes: marcus.noyes@nyulangone.org
Xiaodong Cheng: xcheng5@mdanderson.org
Anders S. Hansen: ashansen@mit.edu
Jane Skok: Jane.Skok@nyulangone.org

Abstract 150 words.

Although only a fraction of CTCF motifs are bound in any cell type, and few occupied sites overlap cohesin, the mechanisms underlying cell-type specific attachment and ability to function as a chromatin organizer remain unknown. To investigate the relationship between CTCF and chromatin we applied a combination of imaging, structural and molecular approaches, using a series of brain and cancer associated CTCF mutations that act as CTCF perturbations. We demonstrate that binding and the functional impact of WT and mutant CTCF depend not only on the unique binding properties of each protein, but also on the genomic context of bound sites and enrichment of motifs for expressed TFs abutting these sites. Our studies also highlight the reciprocal relationship between CTCF and chromatin, demonstrating that the unique binding properties of WT and mutant proteins have a distinct impact on accessibility, TF binding, cohesin overlap, chromatin interactivity and gene expression programs, providing insight into their cancer and brain related effects.

Introduction

The CCCTC-binding factor (CTCF) is an eleven zinc finger DNA-binding protein that plays a key role in chromatin organization and gene regulation. Key insight into the mechanisms underlying CTCF's function were revealed by molecular studies showing that cohesin binding overlaps CTCF sites on chromatin in a CTCF dependent manner (Parelho, Hadjur et al. 2008, Wendt, Yoshida et al. 2008). A more coherent picture of the co-operative function of these two factors emerged from subsequent analyses showing that acute depletion of CTCF in cell lines leads to loss of highly self-interacting topologically associated domain (TAD) structures and redistribution of cohesin on chromatin (Nora, Goloborodko et al. 2017). It is now well established that CTCF and cohesin play a key role in organizing chromatin into TAD structures by promoting the formation of loops and boundaries that are important for gene regulation. This involves a loop-extrusion mechanism in which cohesin complexes create loops by actively extruding DNA until movement of the complex is blocked by two CTCF binding sites in convergent orientation (Fudenberg, Imakaev et al. 2016, Davidson, Bauer et al. 2019). The requirement for orientation specific CTCF binding can be explained by the underlying structure of the interaction between the two proteins, which occurs between the SA2-SCC1 component of cohesin and the N terminal region of CTCF (Li, Haarhuis et al. 2020, Nishana, Ha et al. 2020, Nora, Caccianini et al. 2020, Pugacheva, Kubo et al. 2020). CTCF, in conjunction with cohesin, is enriched at TAD boundaries that function as insulators, contributing to gene regulation by restricting the interaction of regulatory elements to promoters of target genes located within the same TAD.

While the above studies provide critical insight into CTCF's global contribution to chromatin folding and gene regulation, the role of CTCF in regulating individual loci in a context specific manner is unclear. Since CTCF's binding profiles are known to be cell-type specific chromatin accessibility is presumed to be important for CTCF attachment. However, the mechanisms underlying CTCF's ability to bind and act as a chromatin organizer in a cell-type specific manner are incompletely elucidated. CTCF has been degraded by auxin in numerous cell types, and although these studies provided important insight into its role in gene regulation, these models are insufficient for (i) analyzing the contribution of genomic context to site-specific CTCF binding and function, or (ii) distinguishing direct from indirect effects that can be a confounding issue for interpreting site specific impact. Instead, information related to CTCF's direct effects on individual loci has come from the genetic ablation of CTCF binding sites at specific regions in the genome in a given cell type. Indeed, disruption of TAD boundaries by deletion of one or more proximal binding sites can alter gene regulation as a result of aberrant enhancer promoter contacts, and this can have

dramatic consequences on developmental processes and cancer initiation (Guo, Yoon et al. 2011, Xiang, Zhou et al. 2011, Guo, Xu et al. 2015, Narendra, Rocha et al. 2015, Flavahan, Drier et al. 2016, Hnisz, Weintraub et al. 2016). However, deletion of CTCF boundary elements can be contextual, impacting gene expression in certain cell types but not in others. Furthermore, interrogating the general mechanisms underlying the crosstalk between CTCF function and the cell- and locus-specific context of its binding by genetic manipulation of individual CTCF binding sites is laborious and has the limitation of providing insight into the control of only a limited number of loci in the neighboring region.

The *CTCF* locus, located on chromosome 16q band 22, corresponds to one of the smallest regions of overlap for common deletions in breast and prostate cancers (Filippova, Lindblom et al. 1998, Xiang, Zhou et al. 2011). Moreover, point mutations and deletions have been identified in many other tumors. Together these findings indicate that CTCF acts as a tumor suppressor (Debaugny and Skok 2020). Consistent with this, deletion of one *Ctcf* allele predisposes mice to spontaneous B-cell lymphomas as well as radiation- and chemically-induced cancer in a broad range of tissues (Kemp, Moore et al. 2014). Additionally, genetic alterations in *Ctcf* and changes in dosage are associated to varying degrees with intellectual disability and microcephaly, and thus CTCF is thought to play an important role in brain development and neurological disorders (Gregor, Oti et al. 2013, Konrad, Nardini et al. 2019, Valverde de Morales, Wang et al. 2023).

CTCF binds to chromatin through a subset of its eleven zinc fingers (ZFs) (Chen, Tian et al. 2012, Maurano, Wang et al. 2015). Specifically, ZF4-ZF7 make sequence specific contacts with a 12-15 base-pair (bp) core consensus binding site and it is thought that ZF2, ZF8 and ZF9 contribute to stability (Hashimoto, Wang et al. 2017). ZF1 and ZF10, contain RNA binding domains (RBDs) that contribute to binding of CTCF and CTCF-mediated looping at a subset of sites through an unknown mechanism (Saldana-Meyer, Rodriguez-Hernaez et al. 2019). While the crystal structure of CTCF in complex with a known DNA binding domain has revealed new insights into the contribution of each ZF to sequence-specific binding in an *in vitro* setting (Hashimoto, Wang et al. 2017), there has been no complementary, comprehensive analysis of cancer associated CTCF mutants and it is thus not known how mutations within each ZF impact binding stability and DNA sequence specificity, and the consequences this has on binding profile, cohesin overlap, accessibility, chromosome architecture and gene regulation in cell. Given this lack of information, we have no context in which to determine the functional relevance of mutations.

Using an inducible mouse ESC complementation system, we combined imaging, structural, molecular and bioinformatic analyses to examine the impact of nine, high frequency cancer and brain associated CTCF mutations in eight amino acid residues of ZFs that make contact with the core consensus binding motif. A subset of these are also associated with neurological disorders. Collectively, the nine mutations have a graded impact and can act as CTCF perturbations, providing an elegant system with which to investigate, (i) the relationship between WT and mutant CTCF and chromatin, and (ii) site specific mutant effects on accessibility, TF binding, cohesin overlap, chromatin interactivity and gene expression, both of which offer insight into CTCF's biology and its cancer and brain related effects. Compared to a complete knockout of CTCF or deletion of a CTCF binding site, the mutants offer a more subtle and in-depth approach for teasing out the contribution of CTCF to chromatin organization and gene regulation, since each mutation produces its own unique effect.

Here we demonstrate that the functional impact of WT and mutant CTCF depends not only on the specific binding properties of each protein, but also on the genomic context of the binding sites. Analyses of the relationship between CTCF, cohesin and accessibility revealed that although CTCF can bind both accessible and inaccessible sites, it has a weaker signal and is less competent at performing its function of blocking cohesin at inaccessible sites. Thus, changes in accessibility are linked to strength of WT and mutant CTCF binding and their function. But accessibility *per se*, does not predict CTCF binding because the majority of accessible motif containing sites are not bound by CTCF. Bioinformatic and RNA-seq analyses revealed that CTCF binding is linked to enrichment of motifs of expressed TFs abutting these sites. Moreover, motifs are also enriched within 20bps of CTCF bound inaccessible sites but in contrast to bound accessible sites, the cofactor TFs are expressed at a lower level, suggesting that a change in transcriptional program can flip an inaccessible bound site to an accessible bound site in which newly bound cofactor TFs can lead to a stronger CTCF signal capable of blocking cohesin. These findings indicate that cell-type specific transcriptional programs shape the genomic context of CTCF's binding profiles, determining which sites are functional. Our studies also highlight the reciprocal relationship between CTCF and chromatin, demonstrating that the unique binding properties of WT and mutant CTCF have a distinct impact on accessibility, TF binding, cohesin overlap, chromatin interactivity and gene expression programs. Collectively, the mutants offer a rich resource to investigate site specific CTCF-mediated effects on chromatin folding and gene regulation, while distinguishing cause from consequence and direct from indirect outcomes.

Finally, graded mutant perturbations provide mechanistic insight into CTCF biology and mutant specific cancer and brain related effects.

RESULTS

CTCF complementation system

Although it is possible to examine CTCF mutations in cancer cells to determine their functional outcome, this setting is not ideal because the cancer cells have many other genetic and epigenetic alterations that would confound a clean analysis of their impact. To circumvent these issues, we made use of a small-molecule auxin inducible degron (AID) mouse ESC (mESC) system (Nora, Goloborodko et al. 2017, Nishana, Ha et al. 2020). In these cells, both endogenous CTCF alleles are tagged with AID as well as eGFP (CTCF-AID-eGFP) and the auxin-activated ubiquitin ligase TIR1 (from *Oryza sativa*) is constitutively expressed from the *Rosa26* locus. Addition of indole acetic acid (IAA, an analogue of auxin) leads to rapid poly-ubiquitination and proteasomal degradation of the proteins tagged with the AID domain.

To study the impact of CTCF mutations, we established a rescue system wherein the mESC degron cell line was modified to express either a stable doxycycline-inducible control wild-type *Ctcf* or a mutant *Ctcf* (*mCtcf*) transgene in the absence of endogenous CTCF (Nora, Caccianini et al. 2020) (**Figure 1A**). Transgenes have a 3 x FLAG tag, which allows us to distinguish mutant or wild-type transgenic CTCF from endogenous CTCF using a FLAG antibody in Western blot and ChIP-seq analysis. An mRuby fluorescent tag additionally enables us to accurately determine expression levels of mutant or wild-type transgenic CTCF by FACS. The three conditions used for our analysis are shown in **Figure 1B**, and an example of the FACS profile for cells expressing a WT CTCF transgene in each condition is shown in **Figure 1C**. This system was previously used by us to study the interplay between CTCF and CTCFL, the paralog of CTCF (Nishana, Ha et al. 2020).

To study the impact of cancer associated CTCF mutations, we selected a subset of misssense mutations that occur at high frequency in cancer patients (**Figure S1**), focusing on those found in ZFs that make contact with the 12-15bp consensus CTCF binding motif (**Figure 1D** and highlighted with an asterisk in **Figure S1**). We analyzed mutations in **(i)** amino acids that make base-specifying contacts with the alpha helix of DNA (at locations -2, 2, 3 and 6 on the ZF), **(ii)** residues that coordinate the zinc ion which are essential for providing stability to ensure the proper folding of the ZF domain, **(iii)** two other classes of amino acid residues in the ZFs:

Figure 1

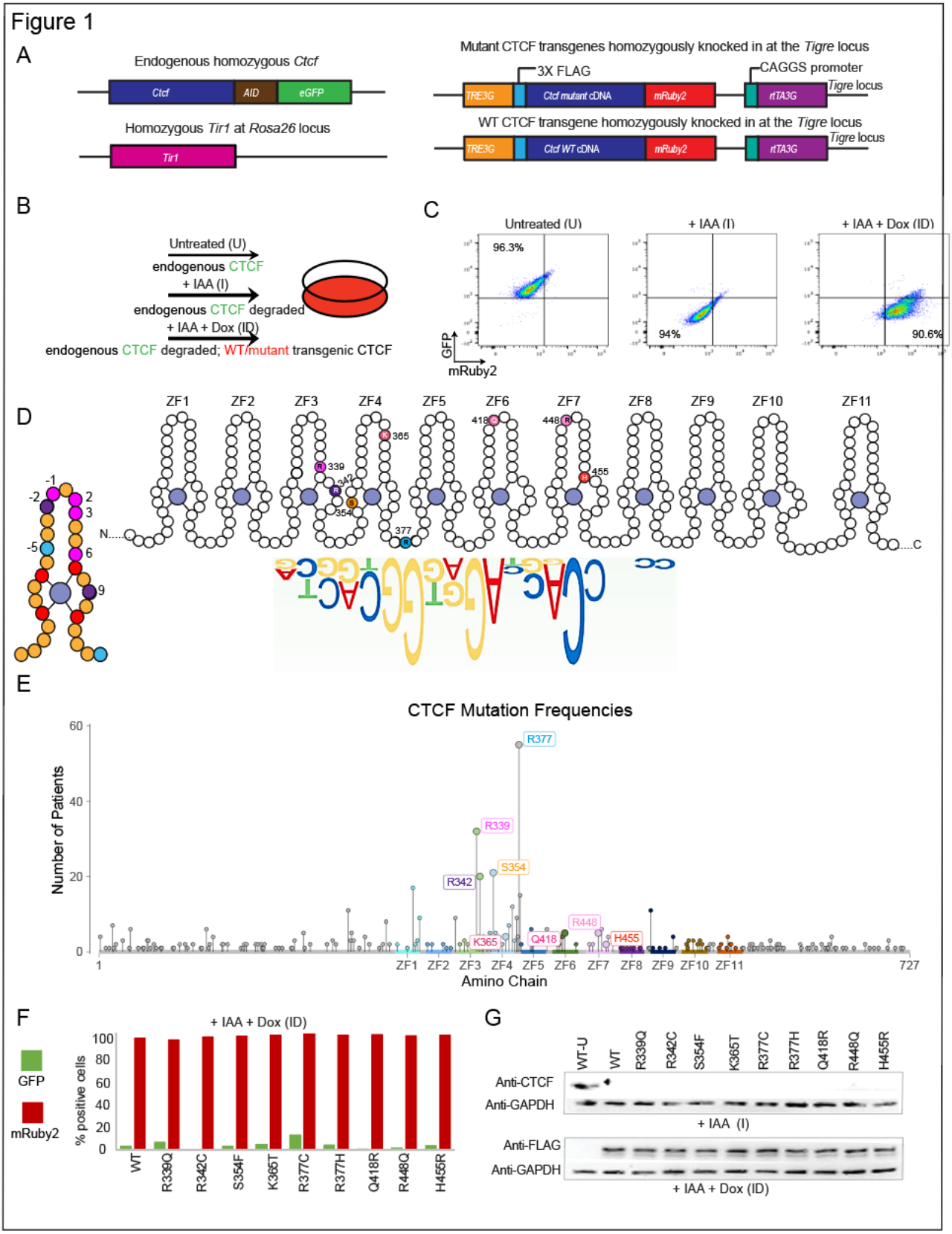
A
Endogenous homozygous Ctcf
Ctcf
AID
eGFP
Homozygous Tir1 at Rosa26 locus
Tir1
Mutant CTCF transgenes homozygously knocked in at the Tigre locus
3X FLAG
CAGGS promoter
TRE3G
Ctcf mutant cDNA
mRuby2
rtTA3G
Tigre locus
WT CTCF transgene homozygously knocked in at the Tigre locus
Ctcf WT cDNA
B
Untreated (U)
endogenous CTCF
+ IAA (I)
endogenous CTCF degraded
+ IAA + Dox (ID)
endogenous CTCF degraded; WT/mutant transgenic CTCF
C
Untreated (U)
+ IAA (I)
+ IAA + Dox (ID)
96.3%
94%
90.6%
GFP
mRuby2
D
ZF1
ZF2
ZF3
ZF4
ZF5
ZF6
ZF7
ZF8
ZF9
ZF10
ZF11
N
C
E
CTCF Mutation Frequencies
Number of Patients
Amino Chain
R377
R339
S354
R342
K365
Q418
R448
H455
F
+ IAA + Dox (ID)
GFP
mRuby2
% positive cells
WT
R339Q
R342C
S354F
K365T
R377C
R377H
Q418R
R448Q
H455R
G
WT-U
Anti-CTCF
Anti-GAPDH
+ IAA (I)
Anti-FLAG
+ IAA + Dox (ID)

Figure 1: **CTCF complementation system**. (**A)** Scheme of genetic modifications in the *Ctcf* locus and the doxycycline inducible knocked-in WT and mutant transgenic *Ctcf* at the *Tigre* locus. The endogenous *Ctcf* contains the auxin inducible degron (AID) and the eGFP tag on both alleles. All transgenes harbor an N terminal 3 X FLAG tag and *TetO-3G* element as well as a C terminal mRUBY2 and *rtTA3G* for doxycycline induced expression. (**B**) Experimental strategy for expression of dox-inducible WT and mutant transgenic CTCF in the absence of endogenous CTCF using the auxin inducible degron system. Addition of indole-3-acetic acid (IAA) a chemical analog of auxin leads to transient and reversible degradation of endogenous CTCF, while addition of doxycycline (Dox) leads to induction of transgene expression. The three conditions used in our analysis are: U, untreated cells; I, IAA treated for CTCF depletion; ID, IAA plus Dox treated for depletion of endogenous CTCF and induction of WT or mutant transgenic CTCF expression. (**C**) Flow cytometry showing the level of GFP (endogenous CTCF) and (transgenic WT CTCF) under the different conditions shown in (**B**). (**D**) Scheme showing the locations of the different types of CTCF mutations within a ZF (left). Amino acids that make base-specifying contacts with the alpha helix of DNA (at locations -2, 2, 3 and 6 on the ZF are shown in different shades of pink, residues that coordinate the zinc ion are shown in red, boundary residues are shown in purple, and residues that contact the sugar phosphate backbone of DNA are shown in blue. Schematic representation of CTCF showing the locations of each mutation under investigation. (**E**) Graph showing the incidence and location of CTCF mutations in cancer patients. The color codes match those in **D**. (**F**) Bar graph showing the expression levels of transgenic, mRuby2 labelled WT and mutant CTCF after treatment with IAA and Dox (ID). The presence of the IAA leads to degradation of GFP labelled endogenous CTCF. (**G**) Western blot showing degradation of endogenous CTCF as detected with an antibody to CTCF, and the induction of transgenic CTCF as detected using an antibody to FLAG.

boundary residues (that are important for interactions between ZFs), and residues that contact the sugar phosphate backbone of DNA, and **(iv)** other highly mutated residues in the region. Each class of mutation is colored coded as shown in **Figure 1D** (left) **and Figure S1**, and the color code is maintained throughout the figures. Mutations were identified from cBioPortal (Cerami, Gao et al. 2012, Gao, Aksoy et al. 2013, de Bruijn, Kundra et al. 2023) and COSMIC (Tate, Bamford et al. 2019) (http://www.sanger.ac.uk/genetics/CGP/cosmic/)

In total, we analyzed mutations in eight amino acid residues and for R377, the most highly mutated residue in cancer, we included two distinct mutations (R377H and R377C) that are equally represented in patients (**Figure 1E and Figure S1**). It should be noted, that a subset of the selected mutations (R339Q, R342C, R448Q, R377H and R377C) have also been shown to be implicated in neurodevelopmental disorders (**Figure S1**) (Price, Fedida et al. 2023, Valverde de Morales, Wang et al. 2023). To analyze the impact of each CTCF mutation, individual clones with comparable levels of transgene expression (mutant and WT) were selected based on FACS and Western blot analysis (**Figure 1F, G and Figure S2**).

Mutations reduce the chromatin residence time and bound fraction of CTCF

To elucidate the impact of the cancer-associated mutations on CTCF's binding dynamics, we performed fluorescence recovery after photobleaching (FRAP) on the 9 mutants and wild-type CTCF (Sprague, Pego et al. 2004, Hansen, Pustova et al. 2017). Briefly, we degraded endogenous CTCF-mAID-GFP by IAA addition, induced WT or mutant transgene expression with Dox for 48 hours, and then performed FRAP by bleaching a circular 1 μm diameter circle, and

monitored recovery for over 10 minutes, recording 30 movies per condition over three replicates. While the recovery of WT-CTCF largely matched prior CTCF FRAP results (Hansen, Pustova et al. 2017), all mutants exhibit faster recovery consistent with reduced and/or less stable DNA-binding (**Figure 2A**). To quantitatively extract dynamic parameters, we fit the FRAP curves to a reaction-dominant kinetic model (Sprague, Pego et al. 2004, Hansen, Pustova et al. 2017) and estimated the residence time - duration of CTCF binding to DNA – and bound fraction – proportion of total CTCF bound to DNA – of each mutant with subsequent comparison to the WT (**Figure 2B-C**). As shown in **Figure 2B-C**, there is a general reduction in the residence time and/or bound fraction across each mutant, with WT CTCF exhibiting the largest specific bound fraction and residence time and the mutants showing a lower or similar residence time. Each mutant displayed unique parameters, regardless of the mutant category. The R377H and R377C mutants – both mutations in the phosphate contacts – have similar bound fractions, but distinct residence times. The mutations in the amino acids making direct contact with DNA – R339Q of ZF3, K365T of ZF4, Q418R of ZF6, and R448Q of ZF7 – have varied residence times and bound fractions, despite belonging to the same group. The R448Q mutant of ZF7 showed the second lowest bound fraction, despite exhibiting a residence time comparable to WT. The Q418R mutant of ZF6 had the lowest residence time of all mutants examined with an intermediate bound fraction. Conversely, the K365T mutant of ZF4 exhibited a bound fraction and residence time most closely comparable to WT. The H455R mutation of ZF7, which assists in coordinating the zinc, had the lowest bound fraction and the second lowest residence time of all examined mutants, accurately reflecting the most deleterious mutation to canonical CTCF function. The R342C mutation of ZF3, which occurs in a boundary residue between the two zinc-ligand histidine residues of ZF3, showed a residence time similar to wild type, but had a lower bound fraction, possibly suggesting a reduced role for boundary residues in binding efficiency. Lastly, the S354F mutant of ZF4, located between the two zinc-ligand cysteine residues of ZF4, is substituted by a bulky phenylalanine in place of a serine in ZF4, showed both a moderate decrease in bound fraction as well as residence time.

In summary, our FRAP results demonstrate that all mutants show reduced DNA binding either through a reduced residence time or bound fraction. Surprisingly, some mutations strongly affect either the residence time or the bound fraction, with minimal effects on the other. This suggests

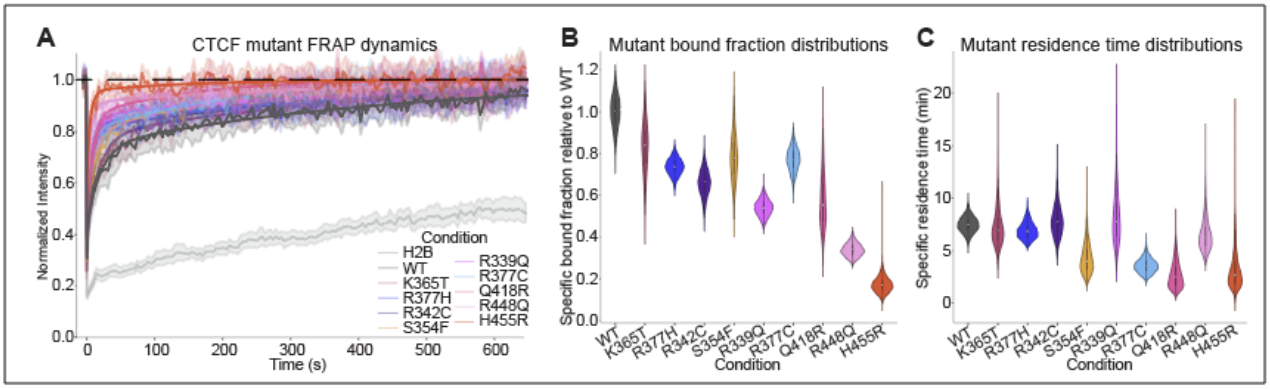

Figure 2: **CTCF mutations reduce chromatin bound fraction and residence time**. (**A)** Plots of FRAP dynamics for WT and mutant mRuby-CTCF overlaid with a two-state binding model. The scatterplots show the average recovery across all movies analyzed, and the outlines give 95% CI. The bold line shows the fitted model. (**B**) Violin plots of specific bound fraction distributions. (**C**) Violin plots of specific residence time (min) distributions. Specific bound fraction and specific residence time distributions were determined by bootstrapping (n=2500).

that the CTCF target search mechanism (ON-rate) can be somewhat decoupled from binding stability (OFF rate) (Hansen, Amitai et al. 2020), such that some mutants affect the efficiency of CTCF search for binding sites, without affecting the residence time of CTCF once bound.

CTCF mutations have unique chromatin binding profiles

To examine the binding profiles of mutant versus WT CTCF at the molecular level, we performed a FLAG ChIP-seq using cells expressing transgenic WT and mutant CTCF in the absence of endogenous CTCF (ID) using the same auxin and Dox conditions as described for the FRAP experiments. ChIP analysis identified three groups of binding sites for each mutation: WT only (sites bound by WT CTCF that mutants can no longer bind), common sites (sites that are bound by both WT and mutant proteins) and mutant only sites (*de novo* sites that only mutant proteins bind) (**Figure 3A, B** and **Figure S3A, B).** The bar graphs and profiles show the percentage of each group of binding sites and the profiles of binding, which are unique to each mutant and different from WT CTCF. It is of note that even at common sites, the strength of CTCF binding is altered in a mutant specific manner, suggesting that a 'binding dosage effect' could contribute to their functional impact. As expected, the H455R mutation in the zinc coordinating residue of ZF 7 loses the most binding sites and has a high percentage of WT only sites, mirroring the FRAP results (**Figure 2**). Together with the FRAP data, this analysis suggests that the effect of each mutation can be decoupled into three metrics, number of binding sites, occupancy/bound fraction (ON-rate) and residence time (OFF-rate), which are likely to be the result of a mutant-specific combination of altered sequence-dependent binding and stability.

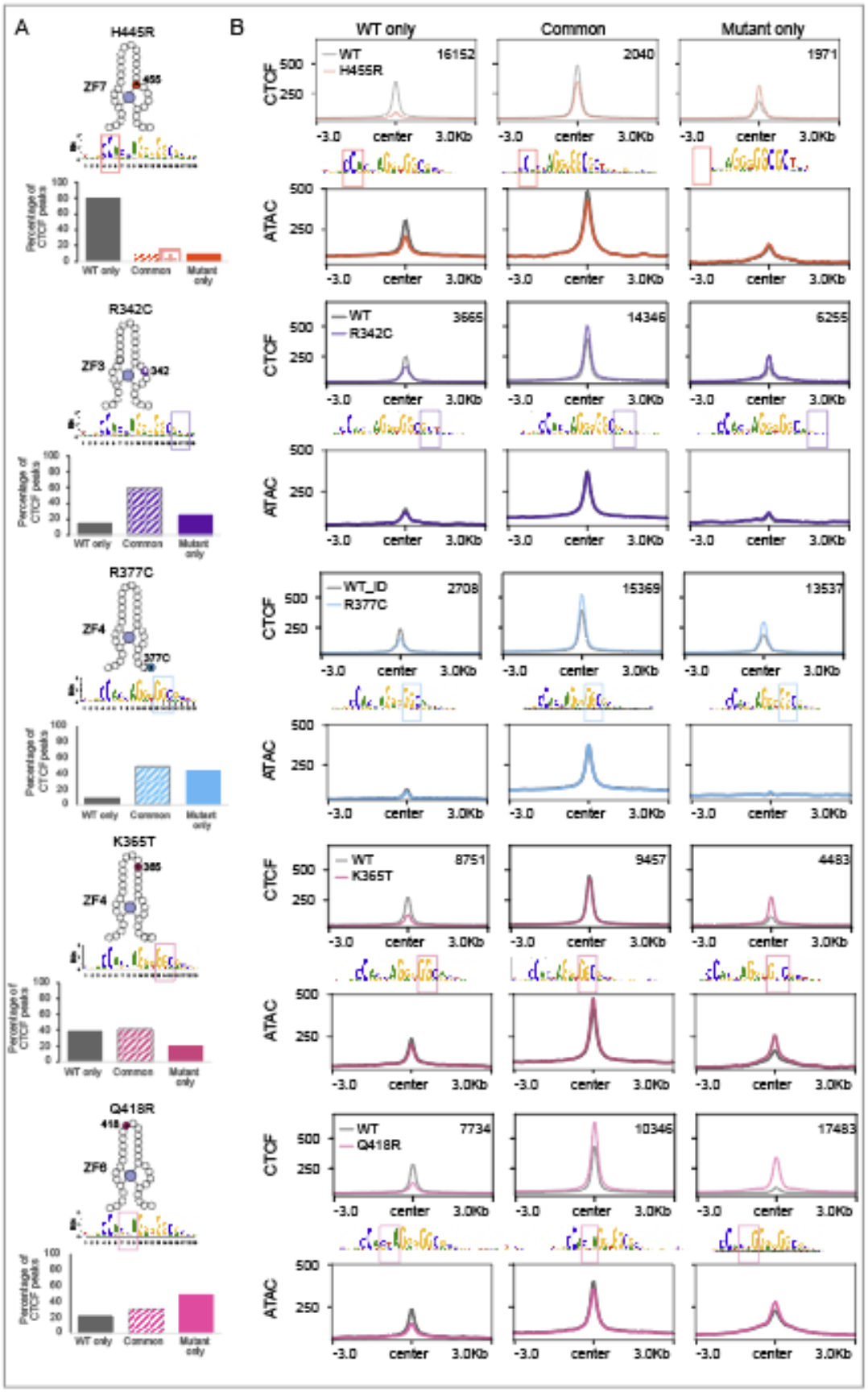
A
B
H445R
ZF7
455
R342C
ZF3
342
R377C
ZF4
377C
K365T
ZF4
365
Q418R
ZF8
418
Percentage of CTCF peaks
WT only
Common
Mutant only
WT only
Common
Mutant only
CTCF
ATAC
WT
H455R
16152
2040
1971
R342C
3665
14346
6255
WT_ID
R377C
2708
15369
13537
K365T
8751
9457
4483
Q418R
7734
10346
17483
-3.0
center
3.0Kb

Figure 3: **CTCF mutations have unique chromatin binding profiles**. (**A)** Scheme showing the locations of the different types of CTCF mutations within a ZF (top) and the percentage of WT and mutant CTCF binding at WT only, common and mutant only sites (bottom). (**B**) CTCF binding and ATAC profiles of WT (black) and mutant CTCF (colored line) for WT only, common and mutant only sites after treatment of cells with IAA + Dox (ID). The dominant CTCF motif is shown for each mutation and group below each CTCF binding profile. The remaining mutant are shown in **Figure S3**.

The strongest peaks for both WT and mutant CTCF are found at common binding sites compared to their respective binding at WT only and mutant only sites (**Figure 3B and Figure S3B**). To determine whether differences in peak intensity in the different groups could be explained by differences in the sequence of the underlying DNA, we ran the *de novo* motif discovery pipeline, Cis Diversity to identify the binding motifs of each group of sites (Biswas and Narlikar 2021). This analysis identified the consensus CTCF motif as dominant in both WT only and common sites, suggesting that the motif constraint is not sufficient to explain the weak versus strong WT CTCF binding at these sites and other factors must play a role (**Figure 3A, Figure S3B and Figure S4**). Indeed, the omniATAC (Corces, Trevino et al. 2017) profiles for the different binding groups indicate that CTCF binding strength is a function of chromatin accessibility. In contrast to the WT only and common sites, the motifs of mutant only sites largely differed from that of the consensus sequence as demonstrated by either (i) a diminished requirement for DNA bases that make contact with the zinc finger that is mutated (marked by a square box in the motif **Figure 3B**), or (ii) an alteration to the consequence of the consensus motif, which could be explained by the impact of mutations on DNA-base interacting zinc finger residues crucial for DNA sequence recognition. Of note, the R377H and R377C mutations, located in the linker region between ZF4 and ZF5, display their own unique binding profiles, although the dominant motif in each group of binding sites (WT only, common and mutant sites) is the consensus motif. Given that R377, which makes a contact with DNA backbone phosphate group and does not make direct contact with a DNA base, it would not be expected to bind a motif with an altered DNA sequence, but rather to bind with reduced stability, consistent with the FRAP results (**Figure 2C**).

In sum, ChIP-seq analysis reveals that CTCF mutants lose, retain and gain a subset of binding sites throughout the genome with mutant specific profiles. Loss of CTCF binding occurs predominantly at weak CTCF binding sites, while binding is retained at the stronger sites although with variable strength depending on the mutant. Differences in binding strength cannot be explained by differences in the binding motif as the consensus CTCF motif is dominant at both WT only and common sites, indicating that other factors such as chromatin accessibility contribute, as shown by the ATAC-seq profiles of these sites.

Mutations uniquely impact CTCF's interaction with DNA

To complement the molecular profiling of the CTCF mutants, computational modeling of the CTCF protein structure was generated for the eight mutated residues in ZF3-ZF7 under investigation. The potential effects of these missense mutants were examined in the context of available CTCF protein structures, including the core DNA binding domain containing fragments of ZF1-ZF7 (PDB 8SSS), K365T (PDB 8SST), ZF2-ZF7 (PDB 5TOU), ZF3-ZF7 (PDB 5KKQ and 5T00), ZF4-ZF7

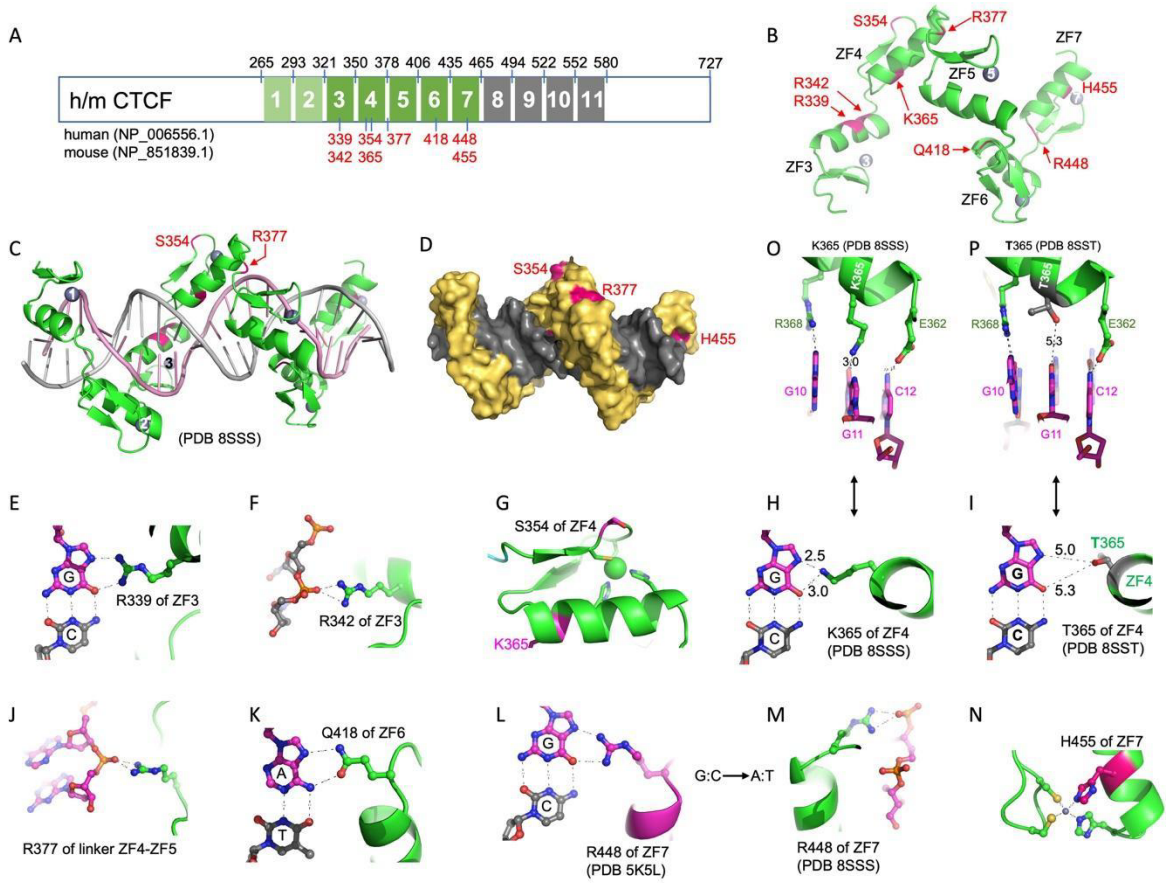

Figure 4: **Each mutation uniquely impacts CTCF's interaction with DNA.** (**A**) The schematic 11 tandem zinc fingers (ZF) of human (or mouse) CTCF proteins. The amino acid numbers in black indicate the start and end residue of each ZF. The numbers in magenta indicate the mutants observed in ZF3-ZF7 examined in this study. (**B**) The 8 mutants are mapped onto ZF3-ZF7. For clarity, the bound DNA molecule and ZFs outside of ZF3-ZF7 are removed. **(C)** Ribbon model of ZF1-ZF7 follows the right-handed twist of the DNA, with each finger occupying the DNA major groove (PDB 8SSS). (**D**) Space filling model of ZF1-ZF7 bound with DNA. Three mutated residues (Ser354, Arg377 and His455) are visible from the surface, whereas the others are buried in the protein-DNA interface. (**E**) Arg339 of ZF3 recognizes a G:C base pair. (**F**) Arg342 interacts with a DNA backbone phosphate group. (**G**) Ser354 of ZF4 is located between two zinc-coordination cysteine residues and is exposed on the surface away from DNA binding. (**H**) Lys365 of ZF4 interacts with a G:C base pair by forming two hydrogen bonds with the O6 and N7 atoms of guanine. (**I**) The substitution of Lys365 with threonine (K365T) results in loss of direct interactions with the corresponding base pair (PDB 8SST). (**J**) Arg377, located in the linker between ZF4 and ZF5 interacts with a DNA backbone phosphate group. (**K**) Gln418 of ZF6 recognizes an A:T base pair. (**L**) Arg448 of ZF7 recognizes a G:C base pair (PDB 5K5L), and (**M**) when the G:C base pair becomes A:T, Arg448 undergoes a conformational change to interact with a DNA backbone phosphate group (PDB 8SSS). (**N**) His455 of ZF7 is one of the four Zn-coordination residues. (**O**) Lys365-containing ZF4 recognizes a triplet of three base pairs (GGC). (**P**) Thr365-containing ZF4 interacts with the same GGC triplet without a direct contact with the central guanine.

(PDB 5K5H), ZF5-ZF8 (PDB 5K5I and 5K5J), ZF6-8 (PDB 5K5L) and ZF4-ZF9 (PDB 5UND) in complex with cognate DNA (Hashimoto et al., 2017).

The location of the eight mutated residues within the CTCF protein is shown in **Figure 4A**. These are mapped onto ZF3-ZF7, with the DNA molecule and ZFs outside of ZF3-ZF7 removed for clarity (**Figure 4B**). A ribbon model of ZF1-ZF7 following the right-handed twist of the DNA, showing each finger occupying the DNA major groove (PDB 8SSS) is depicted in **Figure 3C**. The latter, and a space filling model of ZF1-ZF7 bound to DNA, shows the three mutated residues, Ser354, Arg377 and His455, which are visible from the surface, while the others are buried in the protein-DNA interface (**Figure 4C, D**). **Figures 4E-P** outline the interactions between each of the mutated residues and the DNA.

Four of the eight missense mutants in ZF3-ZF7s are located at DNA-base interacting residues crucial for DNA sequence recognition. These include Arg339 of ZF3, Lys365 of ZF4, Gln418 of ZF6, and Arg448 of ZF7, which are involved in interactions with guanine (via arginine or lysine) and adenine (via glutamine), and thus the substitutions lead to loss of specific DNA interactions and gain of additional binding sites by accommodating other possible base pairs, which explains alterations in the consensus motif found at mutant *de novo* binding sites identified in the ChIP-seq data (**Figure 3**). Lys365-containing ZF4 recognizes a triplet of three base pairs (GGC) and when this is mutated to Thr365, ZF4 interacts with the same GGC triplet without a direct contact with the central guanine, making a less stable interaction. Indeed, the requirement for the middle G in the GGC triplet is lost at *de novo* sites. Arg342 of ZF3 and Arg377, located between ZF4 and ZF5, are involved in interactions between the two zinc fingers and DNA backbone phosphate interactions. Thus, changes to these arginine residues might be disruptive for DNA binding stability, but the imaging analyses demonstrate that each mutant has a distinct effect.

Ser354 of ZF4 lies between two zinc coordination cysteine residues, and is positioned on the surface of CTCF, pointing away from DNA binding. S354F/Y introduces a bulky aromatic residue that creates a surface hydrophobic residue that might affect other (non-DNA binding) interactions. Alternatively, S354F/Y could destroy the integrity of ZF4 by affecting zinc binding. However, if the latter were the case one would expect a bigger effect on binding and residence time than that observed (**Figure 2B and C**). His455 of ZF7 is one of the (C2H2) residues crucial for coordination of Zn binding to ZF7. The H455R substitution results in loss of zinc ion binding, which would be expected to have a dramatic effect, consistent with the observed loss of binding

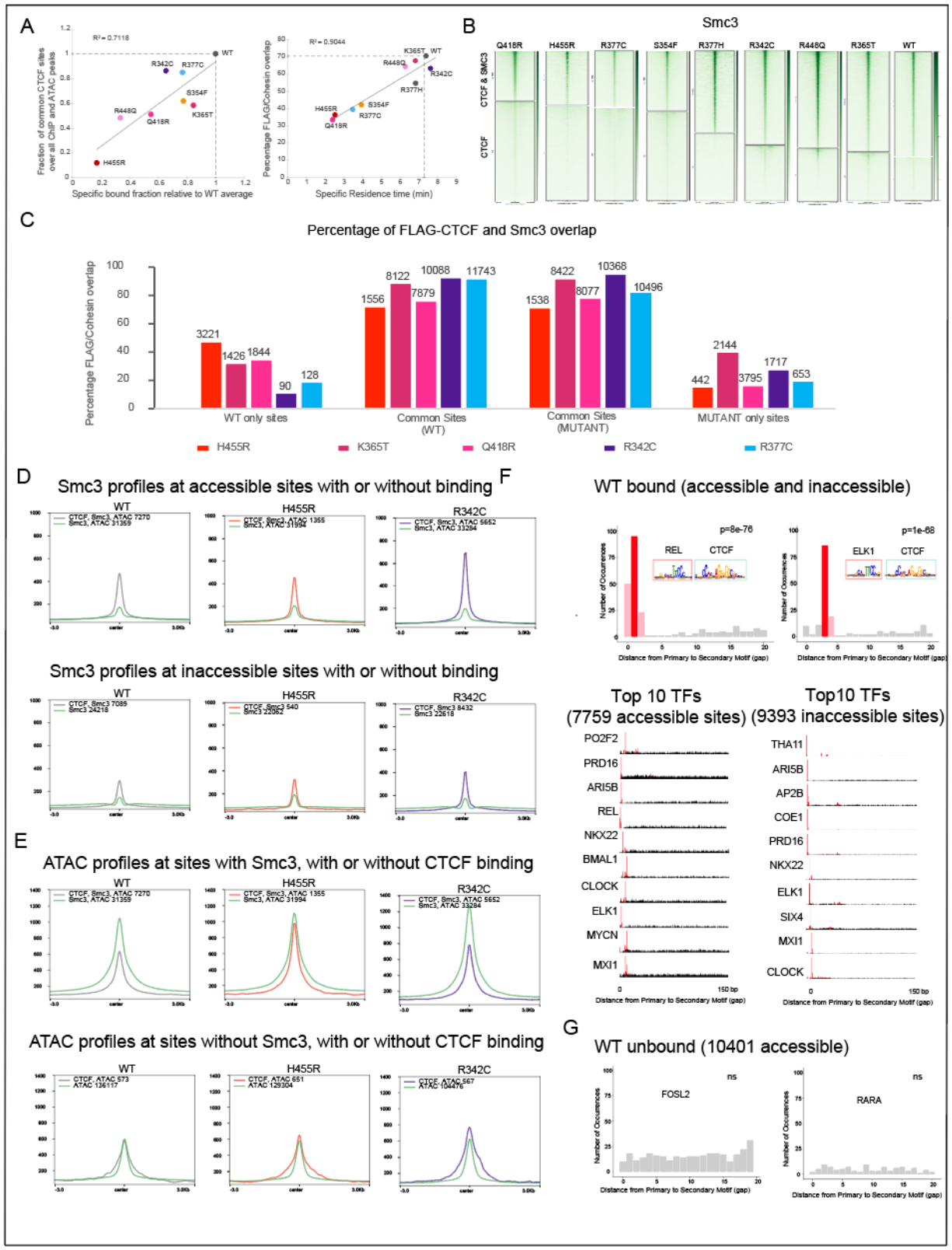
A
R² = 0.7118
Fraction of common CTCF sites over all ChIP and ATAC peaks
Specific bound fraction relative to WT average
WT
R342C
R377C
S354F
K365T
R448Q
Q418R
H455R
R² = 0.9044
Percentage FLAG/Cohesin overlap
Specific Residence time (min)
B
Smc3
Q418R
H455R
R377C
S354F
R377H
R342C
R448Q
R365T
WT
CTCF & SMC3
CTCF
C
Percentage of FLAG-CTCF and Smc3 overlap
Percentage FLAG/Cohesin overlap
3221
1426
1844
90
128
1556
8122
7879
10088
11743
1538
8422
8077
10368
10496
442
2144
3795
1717
653
WT only sites
Common Sites (WT)
Common Sites (MUTANT)
MUTANT only sites
H455R
K365T
Q418R
R342C
R377C
D
Smc3 profiles at accessible sites with or without binding
WT
H455R
R342C
Smc3 profiles at inaccessible sites with or without binding
E
ATAC profiles at sites with Smc3, with or without CTCF binding
ATAC profiles at sites without Smc3, with or without CTCF binding
F
WT bound (accessible and inaccessible)
p=8e-76
REL
CTCF
p=1e-68
ELK1
CTCF
Number of Occurrences
Distance from Primary to Secondary Motif (gap)
Top 10 TFs (7759 accessible sites)
Top10 TFs (9393 inaccessible sites)
PO2F2
PRD16
ARI5B
REL
NKX22
BMAL1
CLOCK
ELK1
MYCN
MXI1
THA11
ARI5B
AP2B
COE1
PRD16
NKX22
ELK1
SIX4
MXI1
CLOCK
150 bp
G
WT unbound (10401 accessible)
ns
FOSL2
RARA

Figure 5: **The relationship between CTCF binding, cohesin overlap and accessibility** (**A)** Correlation (R^2=0.713) between the chromatin bound mutant versus WT fraction detected by FRAP and the fraction of mutant binding at common CTCF sites relative to all FLAG ChIP-seq and ATAC-seq peaks. (**B**) Correlation (R^2=0.8967) between residence time and the percentage of CTCF-cohesin overlap for each mutant (right). Heatmaps show the proportion of CTCF-cohesin versus CTCF only binding (right). All CTCF peaks correspond to the peaks found in each mutant or WT. (**C)** Bar graph showing the overlap between CTCF and cohesin (Smc3) binding at WT only, common and mutant only sites, in cells expressing either WT or mutant CTCF. IAA condition (degradation of WT CTCF, no transgene expression) is shown for comparison to a dramatic loss of CTCF binding. (**D)** Profiles of Smc3 binding in WT and mutant CTCF expressing cells is shown for accessible (top) and inaccessible (bottom) sites with (color-coded for each mutant) or without (green) CTCF binding. The remaining mutants are shown in Figuire S6A. (**E**) Profiles of ATAC-seq in WT and mutant CTCF expressing cells is shown for accessible Smc3 sites (top) and accessible sites without Smc3, with (color-coded for each mutant) or without (green) CTCF binding. The remaining mutants are shown in Figure S6B. CTCF bound accessible sites have a narrower ATAC-seq peak compared to unbound sites at Smc3 sites. Each CTCF mutant has a differential ability to decrease the breadth of ATAC-seq peaks. (F) Enrichment of TF motifs within tight intervals ~10 bp at bound CTCF accessible and inaccessible sites. Two examples are shown in the top panel and the top 10 enriched TFs are shown below. The bars under the motifs represent the frequency and position of highlighted TFs in the region. The red bar represent the position with significant enrichment. (G) "Spaced motif" analysis at unbound accessible CTCF sites showing no significant enrichment of TF motifs close to unbound accessible CTCF sites.

and residence time shown by our imaging and molecular analyses. Together the structural changes inherent to each mutation provide insight into the mutant specific binding properties observed by FRAP and ChIP-seq.

CTCF mutations have a graded impact on binding and function

To further understand the relationship between FRAP and ChIP-seq data, we compared the specific bound fraction estimated by FRAP, to the bound fraction estimated by ChIP-seq using FLAG ChIP-seq and ATAC-seq peaks to define all potential binding sites. While we did not find a good correlation when assessing the overall CTCF ChIP-seq bound fraction, we found a strong and significant correlation (R^2=0.713, p=0.008) between the chromatin bound fraction of the mutants detected by FRAP and the fraction of mutant binding at common CTCF sites (**Figure 5A**). These data suggest that the FRAP bound fraction mostly captures the effect of the mutations on the number of strong accessible binding sites.

We next asked whether the residence time detected by FRAP is linked to CTCF's binding stability and ability to block cohesin. For this, we performed ChIP-seq for cohesin (using an Ab to the cohesin component, Smc3) and found that the residence time was strongly associated (R^2=0.8967, p=0.001) with the percentage of CTCF-cohesin overlap for each mutant (**Figure 5A**) which reflects the impact of binding stability on CTCF's function in loop extrusion. As shown in the heatmaps, the proportion of global CTCF-cohesin versus CTCF only binding is graded across mutants, to some extent mirroring a gene dosage effect that occurs as a consequence of altered residence time (**Figure 5B**).

To examine the effect of mutations on CTCF's function in loop extrusion, we assessed the overlap of WT and mutant CTCF with Smc3 at WT only, common and mutant only sites (**Figure 5C and Figure S5**). For our analysis we first removed any Smc3 sites that were not specific to WT binding at WT only CTCF sites, or mutant CTCF binding at mutant only sites since those Smc3 sites are likely CTCF-independent (**Figure S5**). The data demonstrate that Smc3 overlaps with both WT and mutant CTCF at a significantly higher percentage at common sites (60-80%) compared to WT only and mutant only binding sites (below 40%), consistent with stronger and more accessible WT and mutant CTCF peaks at these locations (**Figure 5C**). While we showed that the percentage of common sites varies between mutants (**Figure 3** and **Figure S3**), the mutants retain their function to block cohesin at these accessible locations.

A comparison of Smc3 profiles at CTCF bound versus unbound sites demonstrates that TFs (which we assume are responsible for accessibility) contribute very little to Smc3 signal compared to CTCF. Smc3 profiles at CTCF bound sites confirmed that CTCF is most functional at blocking cohesin at accessible sites and that each mutant CTCF performs this function with variable effectiveness, (**Figure 5D and Figure S6A**). This trend is also observed at inaccessible Smc3 sites but with a lower overall Smc3 intensity in both WT and mutants, suggesting that CTCF bound at inaccessible sites is less likely to be associated with loop extrusion. It is of note that the majority of both WT and mutant CTCF were found at inaccessible sites although CTCF binding is weaker at these sites (**Figure S6B**), suggesting that for a given cell type, only the subset of strong and accessible CTCF binding is involved in loop extrusion. However, mutant binding at *de novo* sites can be functional and block cohesin at low levels depending on the mutant (**Figure 5C** and **Figure S5**).

CTCF-dependent loop extrusion reduces ATAC-seq signal

To further investigate the link between accessibility and CTCF binding, we compared the ATAC-seq signal at CTCF bound versus unbound sites in the presence or absence of cohesin (**Figure 5E** and **Figure S6C**). Surprisingly, while functional bound CTCF sites overlapping Smc3 showed strong ATAC-seq signals, accessibility was reduced when CTCF was bound compared to accessible Smc3 sites without CTCF binding. This was observed for both WT and mutants. Each CTCF mutant has a differential ability to decrease the breadth of ATAC-seq peaks mirroring its binding stability, suggesting that CTCF is driving the change in accessibility and that the relationship between CTCF and accessibility is bidirectional. Changes in accessibility are not seen

at CTCF bound versus unbound sites without Smc3. These sites are less likely to form loops, lending support to the idea that loop formation decreases accessibility.

In sum, these results demonstrate the importance of identifying the molecular and epigenetic features that drive both CTCF site-specific and stable binding. Furthermore, they highlight the usefulness of the mutants in distinguishing cause from effect: without the mutants it would not have been possible to conclude that binding of CTCF involved in loop extrusion drives changes in accessibility.

Bound versus unbound CBSs can be distinguished by neighboring TF binding sites

Using the FIMO pipeline with a default significance threshold, $p<10^{-4}$) (Grant, Bailey and Noble 2011), we identified ~700,000 predicted CTCF motifs in the genome, the vast majority of which are low confidence, but potentially cell-type specific sites. However, only a fraction of these sites are bound by CTCF in any cell type, even though many of the unbound sites are accessible. We used our ATAC-seq data and the SPAMO pipeline (Whitington, Frith et al. 2011) to identify factors that could distinguish bound from unbound sites: Our analyses revealed that accessible bound sites, can be distinguished from unbound sites by the enrichment of TF motifs within tight intervals of the bound CTCF site (**Figure 5F, 5G and Table S1**), suggesting that cofactor binding is important for CTCF function. Enriched TF motifs were located within 10bp away at 59% of bound sites and within 20 bp at 63% of bound sites. Of note, while SPAMO could detect TF motifs enriched for accessible unbound sites, the aligned CTCF and cofactor motifs showed the presence of highly similar regions, suggested by the high information content of the flanking nucleotides and the repetitive sequences, indicating that the spacing is due to duplication rather than to a functional biological relationship at CTCF unbound sites.

TF motifs were also enriched within 10-20 bps of CTCF bound inaccessible sites. Only a subset of enriched TFs were common in both accessible and inaccessible sites (**Figure 5F** and **Table S1**). We, therefore, asked whether the TFs with motifs found at accessible CTCF bound sites reflect TFs expressed in mESC. Using the RNA-seq from the WT CTCF expressing mESCs, we found that 65% of these TFs were expressed in mESCs (TPM>2) which represents a modest but significant enrichment in expressed genes (p=0.04, OR=1.6). Their expression level was also slightly but not significantly higher than the level of other TFs (median TPM of 9 versus 4). No mESC enrichment was observed for TFs with motifs found at inaccessible CTCF bound sites, which were expressed at lower levels (median TPM of 6 versus 11). This result indicates that a

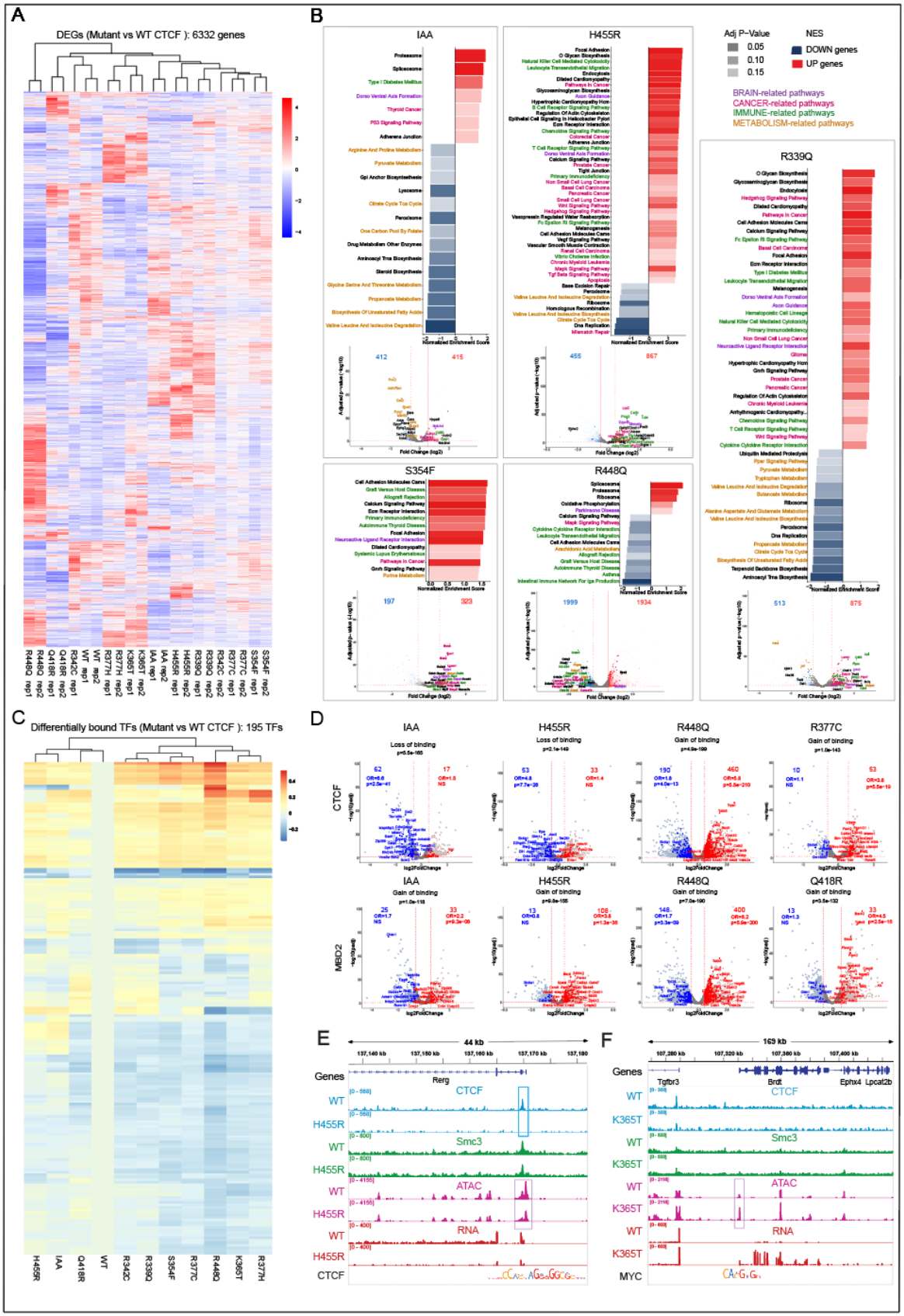
A
DEGs (Mutant vs WT CTCF): 6332 genes
B
IAA
H455R
R339Q
S354F
R448Q
Adj P-Value
0.05
0.10
0.15
NES
DOWN genes
UP genes
BRAIN-related pathways
CANCER-related pathways
IMMUNE-related pathways
METABOLISM-related pathways
C
Differentially bound TFs (Mutant vs WT CTCF): 195 TFs
D
IAA
H455R
R448Q
R377C
Q418R
CTCF
MBD2
E
44 kb
Genes
Rerg
WT
H455R
CTCF
Smc3
ATAC
RNA
F
169 kb
Genes
Tgfbr3
Brdt
Ephx4
Lpcat2b
WT
K365T
CTCF
Smc3
ATAC
RNA
MYC

Figure 6: **CTCF mutations alter gene expression and TF binding.** (**A)** Heatmap showing unsupervised clustering of standardized and normalized expression levels of differentially expressed genes (DEGs) in cells that have endogenous CTCF degraded in the absence or presence of WT transgene or mutant CTCF transgenes. Genes are considered differentially expressed in comparison to cells expressing the WT CTCF transgene in the absence of endogenous CTCF. (**B**) Gene set enrichment analysis of DEGs. Examples shown are cells with endogenous CTCF degraded in the absence (IAA), and presence of the mutant CTCF transgenes, H455R, S354T, R448Q and R339Q. All other mutants are shown in **Figure S7A**. For each condition, bar graphs show significantly enriched KEGG pathways. The volcano plots below highlight the DEGs belonging to these enriched pathways with genes outside of these pathways shown in black. Brain related pathways are shown in purple, cancer in pink, immune in green and metabolism in mustard. (**C**) Heatmap showing differentially bound TFs (predicted from footprinting analysis of ATAC-seq data using the TOBIAS pipeline) in cells which have endogenous CTCF degraded in the absence (IAA) or presence of WT or mutant CTCF transgenes (ID condition). For this analysis replicates were merged and compared with cells expressing the WT CTCF transgene in the absence of endogenous CTCF. (**D**) Volcano plots show examples of two differentially bound TFS (CTCF, MYC) overlapping the promoters of DEGs in CTCF degraded (IAA) and CTCF mutant expressing cells (ID). The remaining mutant comparisons are shown in **Figure S7B**. The enrichment of the TF target genes among the up- and down-regulated genes are reported on top on the volcanos (Odds Ratios (ORs), and p-values). (**E,F**) Examples of altered CTCF binding at the promoter of differentially expressed *Rerg*, and predicted differentially

change in transcriptional program might be able to flip an inaccessible CTCF bound site to an accessible bound site in which newly bound TFs lead to a stronger CTCF signal capable of blocking cohesin.

Each CTCF mutation alters gene expression and TF binding in a unique manner.

To determine how CTCF mutations impact gene expression we performed RNA-seq and did an unsupervised clustering analysis, comparing gene expression in cells in which endogenous CTCF was degraded in the absence (IAA) or presence of the WT or mutant CTCF transgenes (ID). Alterations in transcriptional output were determined by comparison with cells expressing WT transgenic CTCF (ID) (**Figure 6A**). Overall, 6332 genes had altered gene expression across all mutants and including the IAA condition, a selection of which are labelled in the heatmap. As expected, H455R, which loses the most binding sites has the closest profile to the IAA condition. Also of note is the difference in gene expression changes between R377C and R377H, two distinct mutants of the same residue which give rise to different sets of differentially expressed genes (DEGs).

Gene set enrichment analysis using the KEGG database (Kanehisa and Goto, 2000) showed a strong enrichment in functional pathways related to cancer, brain, immune and metabolic processes (color coded in **Figure 6B**, with genes outside these pathways represented in black**)** among DEGs across mutants. This finding is compatible with the human diseases associated with CTCF mutations. The top differentially expressed genes of those pathways are shown in volcano plots below with the same color-code. Examples in **Figure 6B** show changes in pathway and gene expression in the absence of CTCF (IAA condition) and in cells expressing the mutant CTCF

transgenes, H455R of ZF7, S354F of ZF4, R448Q of ZF7 and R339Q of ZF3 (ID condition). All other mutants are shown in **Figure S7**.

The data in **Figures 6A, B** demonstrate that each mutant displays a distinct profile of gene expression changes linked to alterations in a unique set of functional pathways. Interestingly, cells in which endogenous CTCF is degraded (IAA condition) in general show fewer gene expression changes compared to mutant expressing cells, with the exception of the R377C mutation (**Figure S6**). Some of the mutants have alterations in many pathways (R399Q of ZF3), while others highlight changes in only a few (R418Q of ZF6 and R377C) as shown in **Figure 6B and Figure S7A**, suggesting that for those mutants, the alteration in gene expression does not accumulate in specific pathways.

To investigate whether mutant specific transcriptional changes could be explained by altered TF binding, we performed a footprinting analysis of ATAC-seq data using the TOBIAS pipeline (Bentsen et al., 2020) in cells which have endogenous CTCF degraded in the absence (IAA) or presence of WT or mutant transgenes (ID). For this analysis replicates were merged and comparisons made with WT ID cells. Overall, 195 TFs had predicted differential binding across all mutants and including the IAA condition, a selection of which are labelled in the heatmap. As a proof of principle, we detected predicted differential binding for CTCF, which we know from our analyses has altered binding in the mutants, and consistent with the profiles in **Figures 2, 3 and 5** we see the most depletion in the H455R mutant and the IAA condition and most enriched in the R448Q mutant of ZF7. Additionally, the two R377 mutants give rise to distinct profiles of predicted differential CTCF binding with their distinct binding profiles. Numerous other predicted differentially bound TFs (n = 194) were identified as shown in **Figure 5C**.

To analyze the impact of predicted differential binding of TFs on gene expression, we overlapped differentially bound TFs with a region of 2kb around the promoters of DEGs in cells expressing different CTCF mutants**.** Although TF binding at gene promoters regulates gene expression, TFs can also bind distal and proximal enhancers to exert transcriptional control. However, since we cannot connect enhancers to their target genes with any certainty, these regulatory elements are excluded from our analysis and we focus only on DEGs with predicted differentially bound TFs at their promoters. We observed a strong enrichment of predicted differentially bound TF sites among the DEGs suggesting that part of the transcriptional changes associated with CTCF mutations could be explained by the disruption of specific TF pathways (**Table S2**). **Figure 6D**

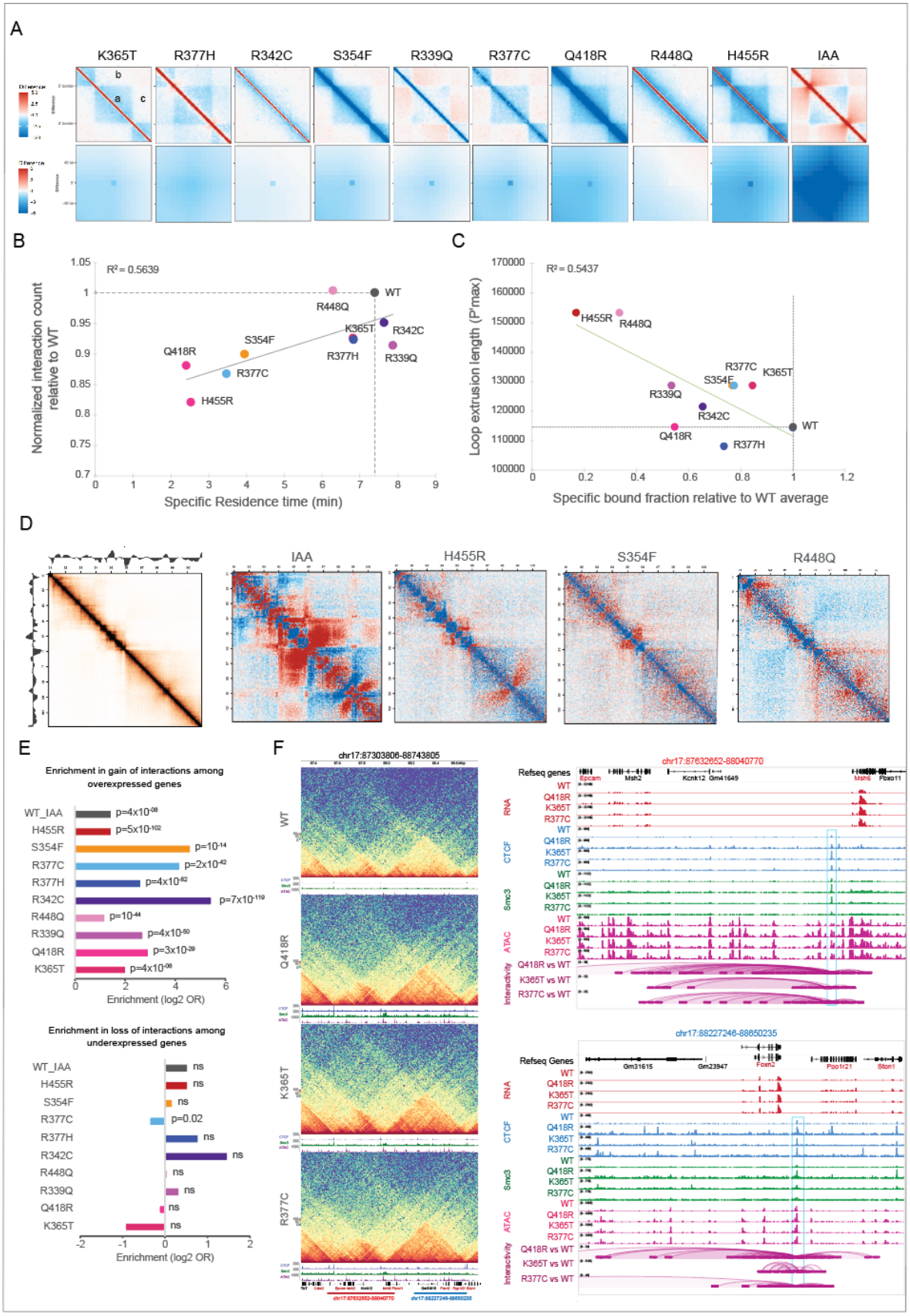
A
K365T
R377H
R342C
S354F
R339Q
R377C
Q418R
R448Q
H455R
IAA
B
R² = 0.5639
Normalized interaction count relative to WT
Specific Residence time (min)
WT
R448Q
K365T
R342C
S354F
R377H
R339Q
Q418R
R377C
H455R
C
R² = 0.5437
Loop extrusion length (P'max)
Specific bound fraction relative to WT average
H455R
R448Q
R377C
K365T
S354F
R339Q
R342C
WT
Q418R
R377H
D
IAA
H455R
S354F
R448Q
E
Enrichment in gain of interactions among overexpressed genes
WT_IAA p=4x10-08
H455R p=5x10-102
S354F p=10-14
R377C p=2x10-42
R377H p=4x10-52
R342C p=7x10-119
R448Q p=10-44
R339Q p=4x10-50
Q418R p=3x10-29
K365T p=4x10-08
Enrichment (log2 OR)
Enrichment in loss of interactions among underexpressed genes
WT_IAA ns
H455R ns
S354F ns
R377C p=0.02
R377H ns
R342C ns
R448Q ns
R339Q ns
Q418R ns
K365T ns
Enrichment (log2 OR)
F
chr17:87303806-88743805
WT
Q418R
K365T
R377C
chr17:87632652-88040770
Refseq genes
Epcam
Msh2
Kcnk12
Gm41649
Msh6
Fbxo11
RNA
CTCF
Smc3
ATAC
Interactivity
Q418R vs WT
K365T vs WT
R377C vs WT
chr17:88227246-88650235
Refseq Genes
Gm31615
Gm23947
Foxn2
Ppp1r21
Ston1

Figure 7: CTCF mutations alter chromatin interactivity. (A) Aggregated analyses. The top panels show the aggregated differential TAD analysis (ATA) for WT and each CTCF mutant. The middle square represents the aggregated TAD flanked by the upstream and downstream TAD**. (a)** Reflects the intra-TAD interaction, while **(b)** and **(c)** highlight the inter-TAD interactions**.** The lower panels show the aggregated differential peak analysis (APA). For the ATA, we performed a pairwise comparison using the union of the boundaries detected in WT and the corresponding mutant. For the APA, we performed a pairwise comparison using the union of loops found in WT and the corresponding mutant. The gain of interactions is color-coded in red and the loss of interactions in blue. These analyses were performed at 10 kb resolution. (**B**) Correlation between the normalized interaction count (at the aggregated loops as described in **Figure 7A**) relative to WT estimated by Hi-C and the specific residence time detected by FRAP. These analyses were performed at 10 kb resolution. (**C**) Correlation between the loop extrusion length estimated by the P'max values derived from Hi-C and the specific bound fraction detected by FRAP. These analyses were performed at 10 kb resolution. (**D**) Example of differential interactions between CTCF mutant and WT. The left matrix shows the interaction in WT within a 10 Mb region at 40 kb resolution. The graphs on the side of the matrix represent the insulation score. The differential matrices are shown for IAA, H455R, S354F and R448Q as examples. Gained interactions are color-coded in red and lost interactions in blue. (**E**) Bar graphs showing the enrichment in over-expressed (top) or under-expressed (bottom) genes among gained (top) or lost (bottom) loops with one anchor overlapping the promoter of the DEGs in IAA and CTCF mutants. (**F**) Example of 2 loci (blue and red rectangles) with a direct effect of gain in CTCF binding and chromatin interactivity. The left panel shows the Hi-C interaction matrices in WT, Q418R, K365T and R377C with both gain of intra- and inter-TAD interactions in the mutant compared to WT. The left panels show the zoom-in tracks of CTCF (blue) Smc3 (green), ATAC (red) and significant differential chromatin loops (purple) with one anchor overlapping the differential CTCF binding sites at both loci. Overexpressed genes are highlighted in red. Of note, only gained interactions were detected at these loci. This analysis was performed at 10 kb resolution.

and Figure S7B highlight two predicted differentially bound TFs (CTCF and MYC) at promoters of DEGs. These are shown in volcano plots for cells in which endogenous CTCF is degraded (IAA) and cells expressing different CTCF mutations (ID). Examples of (predicted and validated) altered CTCF binding at the promoter of differentially expressed *Rerg*, and predicted differentially bound MYC at the promoter of differentially expressed *Brdt*, are shown in **Figure 6E, F.** Loss of CTCF binding at the promoter of the *Rerg* gene leads to a decrease in its expression, while expression of *Brdt* increases in line with increased predicted binding of MYC at its promoter. The latter is an example of an indirect effect because there is no CTCF bound at this gene.

Consistent with our finding that each CTCF mutation alters TF binding and gene expression in a unique manner, we show that for each CTCF mutant, predicted differentially bound TFs target the promoters of a distinct set of up (red) and down (blue) regulated genes. Upregulated genes could reflect enriched binding of a TF at a promoter if the factor acts as an activator, but we also cannot exclude the possibility that upregulation of expression could occur through depletion of a TF that acts as a silencer. Similarly, downregulated genes could reflect reduced or enriched binding of a TF at a promoter where a factor is respectively acting as an activator or repressor. In general, however, we find that predicted enrichment and depletion of binding of TFs (**Figure 6A**) is respectively associated overall with up and downregulation of gene expression (**Figure 6D and Figure S7B and Table S2**).

In sum, combined analysis of RNA-seq and ATAC-seq highlight the variable effects of each CTCF mutation, connecting gene expression changes with 195 predicted differentially bound TFs, to provide a deeper understanding of the factors that underlie the unique transcriptomic profile of each CTCF mutation**.**

CTCF mutant chromatin interactivity is linked to each mutant's binding properties

To determine how CTCF mutants impact chromatin interactions we performed Hi-C. We found that each mutant has a distinct interaction profile and exhibits (i) variable loss of intra TAD interactions and boundary strength as well as (ii) aggregated loop strength. The IAA condition, in which CTCF is degraded shows a loss of intra TAD interactions and a gain of inter TAD interactions as well as the most reduction in aggregated loop strength (**Figure 7A**), while mutants display their own unique effects on TAD interactions and loop strength.

Consistent with our finding that cohesin overlap is correlated with CTCF mutant residence time as determined by FRAP analysis (**Figure 5B**), we found that residence time is also associated with chromatin interactivity (**Figure 7B**). Gassler et al., showed that the first derivative of the relative contact probability curve after log-log transformation (P'max) provides an estimate of the average loop size and cohesin density (Gassler, Brandao et al. 2017). We found that the loop extrusion length (P'max) for each mutant is associated with the fraction that is bound to chromatin, although with a modest correlation (**Figure 7C**), and the longest loops are associated with mutants that have the lowest bound fraction indicating a lower density of CTCF-cohesin anchors in these mutants (**Figure 5B**). These studies demonstrate that imaging and molecular analyses can be functionally integrated to provide new insight into mutant specific effects. The variability in the interaction profile of each mutant can be clearly seen across a 10Mb region of chromosome (**Figure 7D**). Finally, we found that the promoters of overexpressed genes were enriched in gain of chromatin loops (**Figure 7D**). This enrichment was not observed among the under-expressed genes, suggesting that under-expression observed in mESC expressing mutant CTCF might be independent of CTCF's function in chromatin organization (**Figure 7E**), while overexpression might reflect a direct CTCF-mediated effect on loop extrusion as depicted in the examples showing the gain of interactions from CTCF binding sites toward the promoters of overexpressed genes, including *Msh6*, *Epcam*, *Foxn2* and *Cox7a2l* (**Figure 7F** and **S8)**. Interestingly, tall hese genes are implicated in tumorigenesis. Indeed, MSH6 is a mismatch repair factor and its mutation is associated with Lynch syndrome and cancer susceptibility (Edelmann, Yang et al. 1997). Overexpression of MSH6 might also exert an oncogenic function in glioblastoma by promoting

cell proliferation (Chen, Liu et al. 2019). *Epcam* encodes for a membrane glycoprotein involved in epithelial cell adhesion and has been described as both an oncogene and tumor suppressor depending on the microenvironment (van der Gun, Melchers et al. 2010). FOXN2 acts as a tumor suppressor in multiple cancers, including breast, lung and liver (Ma, Lu et al. 2018, Ye and Duan 2019, Liu, Liu et al. 2021). Overexpression of COX7A2L might promote hypoxia tolerance in breast and endometrial cancer (Ikeda, Horie-Inoue et al. 2019).

DISCUSSION

Using a combination of imaging, structural and molecular approaches we have examined the impact of nine, high frequency cancer associated CTCF mutations in ZFs that make contact with the core consensus binding motif. Collectively the graded mutant perturbations offer a platform for investigating the molecular and epigenetic features that govern CTCF's choice of binding sites, and the degree of stability with which CTCF binds at individual locations and is able to function as a chromatin organizer. Since the mutants allow us to distinguish direct from indirect CTCF-mediated effects, and cause from consequence, they provide a deeper understanding of CTCF's downstream impact on chromatin and gene expression.

Our studies revealed that the specificity of interaction between the ZFs in CTCFs binding domain and its target DNA sequence is a key determinant of attachment. However, if the same sequence is found in both accessible and inaccessible chromatin, CTCF will preferentially bind the accessible sites as highlighted by the fact that mutants bind common accessible sites in preference to less accessible WT only sites. Thus, sites in open chromatin dominate the search space. This is underscored by the strong correlation between the fraction of chromatin bound mutants detected by FRAP, and mutant binding at common, accessible CTCF sites.

We observed a good correlation between each mutant's residence time and CTCF-cohesin overlap, suggesting a model in which binding stability is important for blocking cohesin movement on chromatin. A similar correlation was found for residence time and chromatin interactivity, linking cohesin overlap with loop formation. Interestingly, CTCF can bind both inaccessible and accessible motifs and at both these sites the presence of a TF binding site in close proximity (within 10bps) distinguishes bound from non-bound locations. Motifs identified at accessible sites were modestly enriched for TFs expressed in mESCs, as opposed to motifs at inaccessible sites that were not associated with mESC expressing TFs. CTCF bound accessible sites are more likely to overlap with cohesin suggesting that the presence of a TF bound in close proximity to

CTCF can influence its binding stability. We speculate that an inaccessible CTCF bound site could potentially switch to an accessible site depending on the transcriptional program of the cell and the TFs that are expressed that can bind these sites. Theoretically this should convert a weak ChIP-seq signal to a stronger signal with cohesin overlapping at these sites. As previously shown by et Narendra et al. (Narendra, Rocha et al. 2015, Narendra, Bulajic et al. 2016) we identified a CTCF binding site at the *Hoxa5/6* locus in our WT expressing mESC. This site acts as a boundary and regulates the activation of the *HoxA* cluster in motoneurons but not in mESCs where the whole locus is decorated in H3K27me3, consistent with the low accessibility detected in our ATAC-seq data. Our SPAMO analysis identified an E2F2 motif abutting the *Hoxa5/6* CTCF binding site. E2F2 regulates cell cycle exit and has been associated with the neuronal transcriptome and maintenance of the postmitotic state of differentiated neurons (Persengiev, Li et al. 2001, Fleck, Jansen et al. 2023). Thus, CTCF binding at inaccessible loci might reflect 'poised' binding sites that will become functional upon activation of a cell-type specific transcriptional program.

In principle, all these features could have been uncovered by analyzing WT CTCF in the context of cohesin overlap and accessibility. However, collectively the graded effects of the mutants act as perturbations that lend support to our models, underscoring how different aspects of CTCFs binding properties contribute. Moreover, they reveal that each mutant's search for binding sites can be decoupled from its chromatin bound residence time, which is a reflection of its binding stability. For example, the Q418R mutant binds an intermediate number of sites but at these sites it is has a very low residence time and is less competent at blocking cohesin. In contrast, R342C binds fewer sites with a similar residence time compared to WT CTCF, and it is slightly better at blocking cohesin than WT protein. Furthermore, H455R and R448Q, which have very different chromatin residence times, have similarly low chromatin bound fractions, correlated to their P'-max. The increased loop length of both mutants can be explained by the unobstructed passage of cohesin past what would normally be a bound CTCF site.

The crosstalk between accessibility and CTCF binding goes in both directions, such that accessibility affects binding and binding affects accessibility. Although binding at inaccessible sites has no impact on ATAC-seq signal, binding at accessible sites, narrows and reduces the ATAC-seq peak. Evidence for this being a CTCF driven effect comes from analysis of the CTCF mutants. Those with low binding stability (H455R and Q418R) are unable to function in this capacity, while the R342C mutant has a stronger effect compared to WT CTCF. We speculate

that these differences reflect decreased accessibility at the loop bases of CTCF-Smc3-mediated and TF-Smc3-mediated loops, consistent with our finding that CTCF is much better at blocking cohesin than TFs.

While loss or gain of CTCF binding at promoters can account for direct mutant specific effects on accessibility and gene expression, we could only demonstrate a direct effect of gained CTCF binding associated with gained loops the involving promoters of over-expressed genes. Under-expression could, therefore, more often reflect indirect effects of CTCF binding alterations. Indeed, all mutants have their own unique indirect impact in globally changing accessibility at sites where CTCF is not bound. This effect, links changes in TF footprinting with altered transcriptional output, explaining some of the mutant specific effects that we observed.

Aside from the R377C mutant, we observed fewer changes in gene expression in IAA treated cells compared to cells expressing CTCF mutants. Although the mutants all uniquely affect gene expression pathways, we detected enrichment in pathways affecting the brain, immune system, cancer and metabolism, which is consistent with the clinical setting in which the mutations are found, namely cancer and brain disorders. The added effect on immune and metabolic pathways provides insight into other changes that could occur in CTCF mutant-mediated diseases.

Taken together the binding domain mutants we have analyzed here, provide a new appreciation of CTCF's bidirectional relationship with chromatin, its ability to bind and function in different contexts as well as the potential impact of each mutation in clinical settings. Furthermore, our analyses provide a better understanding of how any genetic or epigenetic disorder that alters the landscape of chromatin can in turn impact CTCF binding and function.

Methods

Cell lines

Mouse embryonic stem cells E14Tg2a (karyotype 19, XY; 129/Ola isogenic background) and all clones derived from these were cultured under feeder-free conditions in 0.1% gelatin (Sigma ES-006-B) coated dishes (Falcon, 353003) at 37°C and 5% CO_2 in a humidified incubator. The cells were grown in DMEM (Thermo Fisher, 11965-118) supplemented with 15% fetal bovine serum (Thermo Fisher, SH30071.03), 100 U/ml penicillin - 100 µg/ml streptomycin (Sigma, P4458), 1 X GlutaMax supplement (Thermo Fisher, 35050-061), 1 mM sodium pyruvate (Thermo Fisher, 11360-070), 1 X MEM non-essential amino-acids (Thermo Fisher, 11140-50), 50 µM b-

mercaptoethanol (Sigma, 38171), 10^4 U/ml leukemia inhibitory factor (Millipore, ESG1107), 3 μM CHIR99021 (Sigma, SML1046) and 1 μM MEK inhibitor PD0325901 (Sigma, PZ0162). The cells were passaged every alternate day by dissociation with TrypLE (Thermo Fisher, 12563011).

DNA constructs

Construction of vector for cloning transgenic, doxycycline-inducible expression of WT and mutant mouse *Ctcf* cDNA were obtained from GenScript in pUC19 vectors. The cDNA was amplified such that it harbors an AflII sequence at the 3' end of the gene and fused to a FLAG tag (that harbors NotI sequence) at the 5' end with the help of a fusion PCR. The resultant fragment was digested with NotI and AflII. The *Ctcf* gene was removed from pEN366 (Nora, Goloborodko et al. 2017) by digesting with the same enzymes. This backbone was used for insertion of each *Ctcf* mutant.

In brief, the cDNA region corresponding to each of the C and N terminals and zinc fingers were PCR amplified in such a way that it included a short stretch of the 5′ and/or 3′ region of the neighboring fragment to be connected. The desired PCR products were then annealed, amplified by PCR and cloned into the NotI and AflII sites of the pEN366 backbone (Addgene #156432). All of the constructs were verified by DNA sequence analysis. For all transgenes, the final vector harbors an N terminal 3 X FLAG tag and a C terminal *mRuby* as in-frame fusion to WT and mutant *Ctcf*. The vector also harbors a *TetO-3G* element and *rtTA3G* for doxycycline induced expression of the transgene, and homology arms surrounding the sgRNA target site of the *Tigre* locus for locus-specific insertion. The selection of stable integrants was achieved by virtue of *FRT-PGK-puro-FRT* cassette. Further details of the vector are described elsewhere (Nora, Goloborodko et al. 2017). The vector pX330-EN1201 (Nora, Goloborodko et al. 2017) harboring spCas9 nuclease and *Tigre*-targeting sgRNA was used for targeting of *Tigre* locus (Addgene #92144).

Gene targeting

Mouse embryonic stem cell E14Tg2a harboring *Ctcf-AID-eGFP* on both alleles and a knock-in of pEN114 - *pCAGGS-Tir1-V5-BpA-Frt-PGK-EM7-PuroR-bpA-Frt-Rosa26* at *Rosa26* locus was used as the parental cell line for making all the transgenes (Nora, Goloborodko et al. 2017). pEN366 derived vectors harboring the rescue transgenes (WT and mutant *Ctcf*) were used for targeting transgenes to the *Tigre* locus (clone ID# EN156.3.5) (Nishana, Ha et al. 2020). For nucleofections, 15 μg each of plasmids harboring the transgenes and 2.5 μg of those with sgRNA targeting the *Tigre* locus was used. Nucleofection were performed using Amaxa P3 Primary Cell kit (Lonza, V4XP-3024) and 4D- transfector. 2 million cells were transfected with program CG-

104 in each case. The cells were recovered for 48 h with no antibiotic followed by selection in puromycin (1 µg/mL) (Thermo Fisher, A1113803). Single colonies were manually picked and expanded in 96 well plates. Clones were genotyped by PCR and FACS was performed to confirm that the level of expression of transgenes were comparable. All the clones that were used for the analyses were homozygous for the integration of the transgenes and their levels of expression were comparable.

Induction of auxin inducible degradation of CTCF and doxycycline induced expression

For degradation of endogenous CTCF, the auxin-inducible degron was induced by adding 500 µM indole-3-acetic acid (IAA, chemical analog of auxin) (Sigma, I5148) to the media. Expression of transgenes was achieved by the addition of doxycycline (Dox, 1 µg/ml) (Sigma, D9891) to the media. The cells were treated with IAA and/or Dox for 2 days.

Western Blotting

mESCs were dissociated using TrypLE, washed in PBS, pelleted and used for western blotting. Approximately 2 million cells were used to prepare cell extract. Cell pellets were resuspended in RIPA lysis buffer (Thermo Fisher, 89900) with 1X HALT protease inhibitors (Thermo Fisher, 78430), incubated on ice for 30 min, spun at $4^{O}C$ at 13,000 rpm for 10 min and supernatant was collected. For the western blot of CTCF, low salt lysis buffer (0.1 M NaCl, 25 mM HEPES, 1 mM $MgCl_2$, 0.2 mM EDTA and 0.5% NP40) was used supplemented with 125 U/ml of benzonase (Sigma E1014). Protein concentration was measured using the Pierce BCA assay kit (Thermo Fisher, 23225). 20 µg of protein were mixed with Laemmli buffer (Biorad, 1610737) and β-mercaptoethanol, heated at $95^{O}C$ for 10 min and run on a Mini-protean TGX 4%-20% polyacrylamide gel (Biorad, 456-1095). The proteins were transferred onto PVDF membranes using the Mini Trans-Blot Electrophoretic Transfer Cell (Bio-Rad, 170-3930) at 80 V, 120 mA for 90 min. PVDF membranes were blocked with 5% BSA in 1 X TBST prior to the addition of antibody. The membranes were probed with appropriate antibodies overnight at $4^{O}C$ (anti-rabbit histone H3 (abcam, ab1791; 1: 2,500 dilution), anti-mouse FLAG antibody (Sigma, F1804; 1: 1,000 dilution), anti CTCF (active motif, 61311), anti Rad21 (ab992). Membranes were washed five times in PBST (1 × PBS and 0.1% Tween 20) for 5 min each and incubated with respective secondary antibodies in 5% BSA at room temperature for 1 h. The blots were rinsed in PBST and developed using enhanced chemiluminescence (ECL) and imaged by Odyssey LiCor Imager (Kindle Bioscien ces).

Flow cytometric analysis

Cells were dissociated with TrypLE, washed and resuspended in MACS buffer for flow cytometric analysis on LSRII UV (BD Biosciences). Analysis was performed using the FlowJo software.

Cell culture – FRAP

Control cells expressing H2B-mRuby were engineered using the same background as the CTCF transgene expressing cells, E14Tg2a mESCs. Briefly, pEN114 - *pCAGGS-Tir1-V5-BpA-Frt-PGK-EM7-PuroR-bpA-Frt-Rosa26* was used as the parental cell line to be nucleofected by the pASH40-mRuby-neo with the PiggyBac transposase vector. For the FRAP experiment, CTCF and H2B expressing cells were cultured for two days on 35 mm no 1.5H glass-bottom imaging dishes (MatTek, Ashland, MA, P35G-1.5-14C) coated with Geltrex (Gibco, A1413201) according to manufacturer's instructions. 48 hours prior to the experiment, the culture medium was changed to medium supplemented with 500 μM 3-Indoleacetic acid (IAA; Sigma-Aldrich, I2886-5G) and 1 μg/mL doxycycline (Sigma-Aldrich, D3072-1ML). Just prior to imaging, the medium was changed to imaging medium: phenol red free DMEM with all other aspects of the medium the same.

Fluorescence recovery after photobleaching (FRAP)

FRAP experiments were performed on an LSM900 confocal microscope (Zeiss, Germany) equipped with a full, humidified incubation chamber that was maintained at 37°C with 5.5% CO_2. The time series were acquired in confocal mode using two-stage acquisition with the following excitation parameters: 561nm excitation laser at 0.8% power, 53μm pinhole size corresponding to 1AU, scan speed 7, and 3x crop factor corresponding to a pixel dwell time of 3.06μs. During the first phase, we imaged 25 frames with 0.5 seconds between frames, and during the second phase we imaged 160 frames with 4 seconds between frames, resulting in a total movie length of 657 seconds. For bleaching, a 1μm diameter ROI was selected. Bleaching was performed after frame 8 using both the 488nm and 561nm lasers at 100% power, 53μm pinhole size, and scan speed 7 corresponding to a pixel dwell time of 3.06μs. For each condition, 30 movies were acquired and between 3 and 11 were excluded due to drift.

Fluorescence recovery after photobleaching (FRAP) analysis

FRAP analysis was performed using custom Python scripts. Briefly, movies were loaded into a custom Python GUI (https://github.com/ahansenlab/FRAP-drift-correction-GUI). The bleached nucleus was segmented using manually selected segmentation parameters, and was used to estimate $I_{nonbleach}(t)$ – the average nuclear intensity – and $I_{background}(t)$ – the average background intensity. To correct for cell drift, the ROI positions were manually updated at a

number of frames and the ROI position was linearly interpolated between manual updates. Cells with excessive drift were manually excluded and the mean fluorescence intensity of the bleach – $I_{bleach}(t)$ – was taken as the average intensity within the ROI.

To correct for photobleaching, $I_{nonbleach}(t)$ was smoothed using an averaging filter, and a series of correction factors ($C(t)$) were computed as the ratio of these values to their average prebleach value as follows:

$$C(t) = \frac{I_{nonbleach,smooth}(t) - I_{background}(t)}{\langle I_{nonbleach}(t) - I_{background}(t)\rangle_{prebleach}}$$

Subsequently, the FRAP curves were normalized according to the following equation:

$$\underline{I_{bleach}(t)} = \frac{C(t)(I_{bleach}(t) - I_{background}(t))}{\langle C(t)(I_{bleach}(t) - I_{background}(t))\rangle_{prebleach}}$$

where $\underline{I_{bleach}(t)}$ is the normalized, background-subtracted, and photobleaching corrected FRAP signal.

Finally, since previous measurements of CTCF's free diffusion coefficient have indicated a reaction-dominant model is most appropriate, we fit the following, previously demonstrated model(Sprague, Pego et al. 2004, Hansen, Pustova et al. 2017).

$$FRAP(t) = 1 - C_{eq,fast}e^{-k_{off,fast}t} - C_{eq,slow}e^{-k_{off,slow}t}$$

where $k_{off,slow} < k_{off,fast}$. $C_{eq,slow}$ was taken as the specific bound fraction and the specific residence time was taken as $\frac{1}{k_{off,slow}}$. To estimate parameter distributions, we implemented a bootstrapping routine in which a set of 16 movies were randomly resampled from the full dataset 2500 times with replacement. For each sample, the above equation was fit and the parameter values were taken as one data point. The 2500 resampled points were then used as distributions for plotting the specific bound fraction and residence time. For ease of comparison, the specific bound fractions were normalized to the average WT value when plotting.

ChIPmentation

mESCs were dissociated using TrypLE (Thermofisher #12605010) and washed once in 1X PBS. After counting, cells were divided in 10 million aliquots and resuspended in fresh 1X PBS (1million cells/1ml). For double cross linking 25mM EGS (ethylene glycol bis(succinimidyl succinate); Thermofisher #21565) were added and cells were put in rotation for 30 min at room temperature,

followed by addition of 1% formaldehyde (Tousimis #1008A) for 10 min also in rotation at room temperature. Quenching was performed by adding glycine to a final concentration of 0.125M followed by incubations of 5 min at room temperature in rotation. Fixed cells were washed twice with 5ml of 1X PBS containing 0.5% BSA and centrifuged at 3000rpm for 5 min at 4°C. Pellets were finally resuspended in 500ul 1X PBS containing 0.5% BSA, transferred to 1.5ml Eppendorf and centrifuged at 3000rpm for 3 min at 4°C. Supernatant was completely removed, pellets were snap-frozen in liquid nitrogen and stored at -80°C. Fixed cells (10 million) were thawed on ice, resuspended in 350 µl ice cold lysis buffer (10 mM Tris-HCl (pH 8.0), 100 mM NaCl, 1 mM EDTA (pH 8.0), 0.5 mM EGTA (pH 8.0), 0.1% sodium deoxycholate, 0.5% N-lauroysarcosine and 1X protease inhibitors) and lysed for 10 min by rotating at 4° C. Chromatin was sheared using a bioruptor (Diagenode) for 15 minutes (30 sec on, 30 sec off, high output level). 100ul of cold lysis buffer and 50ul of 10% Triton X-100 (f inal concentration of 1) were then added and the samples were centrifuged for 5 min at full speed at 4°C. Supernatant was collected, transferred to a new tube (Protein Low Binding tube) and shearing was continued for another 10 min, then the chromatin was quantified. FLAG M2 Magnetic Beads (Sigma, M8823) were used for FLAG immunoprecipitation. In other cases (CTCF, Cohesin, IgG) antibodies were bound to protein A magnetic beads by incubation on a rotator for one hour at room temperature. 10 µl each of antibody was bound to 50 µl of protein-A magnetic beads (Dynabeads) and added to the sonicated chromatin for immunoprecipitation at 4°C overnight. Next day, samples were washed and tagmentation were performed as per the original ChIPmentation protocol (Schmidl et al., 2015). In short, the beads were washed successively twice in 500 µl cold low-salt wash buffer (20 mM Tris-HCl (pH 7.5), 150 mM NaCl, 2 mM EDTA (pH 8.0), 0.1% SDS, 1% tritonX-100), twice in 500 µl cold LiCl-containing wash buffer (10 mM Tris-HCl (pH 8.0), 250 mM LiCl, 1 mM EDTA (pH 8.0), 1% triton X-100, 0.7% sodium deoxycholate) and twice in 500 µl cold 10 mM cold Tris-Cl (pH 8.0) to remove detergent, salts and EDTA. Subsequently, the beads were resuspended in 25 µl of the freshly prepared tagmentation reaction buffer (10 mM Tris-HCl (pH 8.0), 5 mM MgCl2, 10% dimethylformamide) and 1 µl Tagment DNA Enzyme from the Nextera DNA Sample Prep Kit (Illumina #20034198) and incubated at 37°C for 10 min in a thermomixer. Following tagmentation, the beads were washed successively twice in 500 µl cold low-salt wash buffer (20 mM Tris-HCl (pH 7.5), 150 mM NaCl, 2 mM EDTA (pH8.0), 0.1% SDS, 1% triton X-100) and twice in 500 µl cold Tris-EDTA-Tween buffer (0.2% tween, 10 mM Tris-HCl (pH 8.0), 1 mM EDTA (pH 8.0)). Chromatin was eluted and de-crosslinked by adding 70 µl of freshly prepared elution buffer (0.5% SDS, 300 mM NaCl, 5 mM EDTA (pH 8.0), 10 mM Tris-HCl (pH 8.0) and 10 ug/ml proteinase K for 1 hour at 55°C 850rpm and overnight at 65°C 850rpm. Next day, the supernatant was

collected, transferred to new DNA Low Binding tubes and supplemented with an additional 30 µl of elution buffer. DNA was purified using MinElute Reaction Cleanup Kit (Qiagen #28204) and eluted in 20 ul. Purified DNA (20 µl) was amplified as per the ChIPmentation protocol (Schmidl, Rendeiro et al. 2015) using indexed and non-indexed primers and NEBNext High-Fidelity 2X PCR Master Mix (NEB M0541) in a thermomixer with the following program: 72°C for 5 m; 98°C for 30 s; 14 cycles of 98°C for 10 s, 63°C for 30 s, 72°C for 30 s and a final elongation at 72°C for 1 m. DNA was purified using Agencourt AMPure XP beads (Beckman, A63881) to remove fragments larger than 700 bp as well as the primer dimers. Library quality and quantity were estimatred using Tapestation (Agilent High Sensitivity D1000 ScreenTape #5067-5584 and High Sensitivity D1000 reagents #5067-5585) and quantified by Qubit (Life Technologies Qubit™ 1X dsDNA High Sensitivity (HS) #Q33230). Libraries were then sequenced with the Novaseq6000 Illumina technology according to the standard protocols and with around 200 million 150bp paired-end total per sample.

Structural analyses of CTCF mutants

To complement the molecular analysis of CTCF mutants, the computational protein structure modeling of the CTCF protein was generated for the 8 mutated residues in ZF3-ZF7 under investigation. The potential effects of the 8 missense mutants were examined in the context of available CTCF protein structures, including the core DNA binding domain containing fragments of ZF1-ZF7 (PDB 8SSS), K365T (PDB 8SST), ZF2-ZF7 (PDB 5TOU), ZF3-ZF7 (PDB 5KKQ and 5T00), ZF4-ZF7 (PDB 5K5H), ZF5-ZF8 (PDB 5K5I and 5K5J), ZF6-8 (PDB 5K5L) and ZF4-ZF9 (PDB 5UND) in complex with cognate DNA (Hashimoto, Wang et al. 2017). Substitutions and side chain adjustments were made in PyMOL version 2.5.2 (Schrödinger, LLC), which was further used for the production of molecular graphics (Moore, Rabaia et al. 2012, Rosignoli and Paiardini 2022).

Total RNA-seq

mESCs were dissociated using TrypLE, washed in 1X PBS, pelletted and 2.5 million cells were used for extracting RNA with RNeasy plus kit (Qiagen #74134) in each case. RNA quality was checked in Tapestation (Agilent High Sensitivity RNA ScreenTape #5067-5579, High Sensitivity RNA Sample Buffer #5067-5580 and High Sensitivity RNA Ladder #5067-5581) and quantified by Nanodrop. 500ng were used for Total RNA libraries preparation using Illumina kit (Illumina Stranded Total RNA Prep, Ligation with Ribo-Zero Plus #20040529). Library concentrations were estimated using Tapestation (Agilent High Sensitivity D1000 ScreenTape #5067-5584 and High

Sensitivity D1000 reagents #5067-5585) and quantified by Qubit (Life Technologies Qubit™ 1X dsDNA High Sensitivity (HS) #Q33230). Libraries were then sequenced with the Novaseq6000 Illumina technology according to the standard protocols and with around 100 million 150bp paired-end total per sample.

Hi-C

Hi-C was performed in duplicates using around 1 million cells each. mESCs were dissociated using TrypLE (Thermofishe #12605010), washed once in 1X PBS and resuspended in fresh 1X PBS (1million cells/1ml). For double cross linking 25mM EGS (ethylene glycol bis(succinimidyl succinate); Thermofisher #21565) were added and cells were put in rotation for 30 min at room temperature, followed by addition of 1% formaldehyde (Tousimis #1008A) for 10 min also in rotation at room temperature. Quenching was performed by adding glycine to a final concentration of 0.125M followed by incubations of 5 min at room temperature in rotation. Fixed cells were washed twice with 5ml of 1X PBS containing 0.5% BSA and centrifuged at 3000rpm for 5 min at 4°C. Pellets were finally resuspended in 500ul 1X PBS containing 0.5% BSA, transferred to 1.5ml Eppendorf and centrifuged at 3000rpm for 3 min at 4°C. Supernatant was completely removed, pellets were snap-frozen in liquid nitrogen and stored at -80°C. Samples were subsequently processed using the Arima Hi-C kit as per the manufacturer's protocol and sequenced with the Novaseq6000 Illumina technology according to the standard protocols and with around 600 million 150bp paired-end reads per sample.

ATAC-seq

We used an improved ATAC-seq protocol (Omni-ATAC) adopted from Corces et al., (Corces, Trevino et al. 2017) with small adaptations. Briefly, cells in culture were treated with 200 U/ml DNase (Worthington # LS002007) for 30 min at 37°C to remove free-floating DNA and any DNA from dead cells. Cells were then harvested via trypsinization and resuspended in regular medium. After counting, 500,000 cells were collected, resuspended in 1X cold PBS and spin down at 500g for 5 min at 4°C in a fixed-angle centrifuge. After centrifugation, the supernatant was removed and the pellet resuspended in 500µl of cold ATAC-seq resuspension buffer 1 (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP-40 and 0.1% Tween-20 in water). 50ul (50.000 cells) were transferred to a new 1.5ml Eppendorf tube and 0.5ul 1% Digitonin (Promega #G9441) was added by pipetting well few times. The cell lysis reaction was incubated on ice for 3 min. After lysis, 1 ml of cold ATAC-seq resuspension buffer 2 (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$ and 0.1% Tween-20 in water) was added. Tube was inverted three times to mix and nuclei

were pelleted by centrifugation for 10 min at 500g at 4°C. Supernatant was very carefully removed (pellet could be almost invisible at this step) and nuclei resuspended in 45µl of transposition mix (25µl 2X TD buffer (Illumina #20034198), 16.5µl 1X PBS, 0.5µl 1% digitonin, 0.5µl 10% Tween-20 and 2.5µl water) by pipetting up and down few times. 5ul of transposase enzyme (Illumina #20034198) was then added and the transposition reaction was incubated at 37°C for 30 min in a thermomixer with shaking at 1000 rpm. Reaction was cleaned up with Zymo DNA Clean and Concentrator-5 kit (Zymo#11-302C). Transposed DNA fragments were then amplified as described previously in Buenrostro et al., (Buenrostro, Wu et al. 2015). Final libraries were purified with with Zymo DNA Clean and Concentrator-5 kit (Zymo#11-302C; Note: better to use two separate kits for pre- and post-amplification clean up), checked by Tapestation (Agilent High Sensitivity D1000 ScreenTape #5067-5584 and High Sensitivity D1000 reagents #5067-5585) and quantified by Qubit (Life Technologies Qubit™ 1X dsDNA High Sensitivity (HS) #Q33230). Libraries were then sequenced with the Novaseq6000 Illumina technology according to the standard protocols and with around 50 million 150bp paired-end reads per sample.

QUANTIFICATION AND STATISTICAL ANALYSIS

ChIP-seq data processing and quality control

Data were processed using Seq-N-Slide pipeline (Dolgalev 2022). Briefly, after trimming for NEXTERA adaptor sequence using trimGalore (Krueger 2021). The reads were aligned to mm10 genome with Bowtie2 (Langmead and Salzberg, 2012). Ambiguous reads were filtered to use uniquely mapped reads in the downstream analysis. PCR duplicates were removed using Sambamba (Tarasov, Vilella et al. 2015). Narrow peaks were called using MACS2 (Zhang, Liu et al. 2008) in pair-end mode and with IgG as control for Smc3 ChIP-seq or WT untreated condition (in which FLAG CTCF is not expressed) for CTCF FLAG to control for unspecific binding, Peaks overlapping ENCODE blacklisted regions were filtered out (Amemiya, Kundaje and Boyle 2019). These procedures allowed us to minimize the false positive calls, as reflected by the high percentage of called peaks containing CTCF consensus motifs as detected by MEME (Bailey, Johnson et al. 2015) with 78% for WT and ranging from 52 to 77% for the mutants. To ensure comparability between conditions, samples were down-sampled to the same number of aligned and good quality reads before peak calling using samtools (Danecek, Bonfield et al. 2021). PCA was performed after extracting the signal overlapping the union of peaks called across all samples using Deeptools (Ramirez, Ryan et al. 2016) to assess the reproducibility between replicates. Bigwigs were obtained for visualization on individual as well as merged bam files using Deeptools

(Ramirez, Ryan et al. 2016). For CTCF FLAG visualization, the differential signals were generated after subtracting non-specific signal present in the WT untreated condition. Heatmaps and average profiles were performed on merged bigwig files using Deeptools/2.3.3. Differential CTCF binding sites (FDR<0.05) were detected using DiffBind package (Stark 2011).

RNA-seq data processing and quality control

Data were processed using Seq-N-Slide pipeline (Dolgalev 2022). Briefly, reads were aligned against the mouse reference genome (mm10) using the STAR (Dobin, Davis et al. 2013) aligner and differentially expressed genes were called using DESeq2(Love, Huber and Anders 2014) with an adjusted p-value of 0.05 and a fold change cutoff of 1.5. PCA was performed using the TPM values to assess the reproducibility between replicates. Heatmap was performed using pheatmap R package (Kolde 2019) and GSEA using fgsea R package (Sergushichev 2016) and KEGG genesets (Kanehisa and Goto, 2000).

ATAC-seq data processing and quality control

Data were processed using Seq-N-Slide pipeline (Dolgalev 2022). Briefly, Reads were aligned to mm10 genome with Bowtie2 (Langmead and Salzberg 2012). Ambiguous reads were filtered to use uniquely mapped reads in the downstream analysis. PCR duplicates were removed using Sambamba (Tarasov, Vilella et al. 2015). Mitochondrial contamination was assessed using samtools (Danecek, Bonfield et al. 2021). All the samples showed less than 0.6% of mitochondrial DNA, Peaks were called using MACS2 (Zhang, Liu et al. 2008) and peaks overlapping ENCODE blacklisted regions were filtered out (Amemiya, Kundaje and Boyle 2019). PCA was performed after extracting the signal overlapping the union of peaks called across all samples using Deeptools (Ramirez, Ryan et al. 2016) to assess the reproducibility between replicates. Bigwigs were obtained for visualization on individual as well as merged bam files using Deeptools (Ramirez, Ryan et al. 2016). Heatmaps and average profiles were performed on merged bigwig files using Deeptools/2.3.3.

Hi-C Processing and Quality Control.

HiC-Pro (Servant, Varoquaux et al. 2015) was used to align and filter the Hi-C data. To generate Hi-C filtered contact matrices, the Hi-C reads were aligned against the mouse reference genome (mm10) by bowtie2 (version 2.3.1) using local mode for the second step of HiC-Pro alignment (--very-sensitive -L 20 --score-min G,20,8 --local) to ignore the molecular barcode UMI sequences and dark bases introduced by Arima library prep kit at the 5' end of the read. Singleton, ambiguous

and duplicated reads were filtered out. Ligation sites for Arima HiC kit were set up using GATCGATC,GANTGATC,GANTANTC,GATCANTC. Low mapping quality (MAPQ<30), self circle, dangling ends and re-ligation reads were filtered out through the HiC-pro pipeline. Valid pairs represented more than 80% for all the samples. Samples were downscaled to the same number of good quality uniquely aligned reads using HiCExplorer (Wolff, Bhardwaj et al. 2018). Of note, we did not use the number of valid pairs to downscale the samples to not overcorrect some conditions such as IAA and mutant H455R which were expected to show a dramatic biological decrease in valid interactions.

PCA was performed using the insulation score calculated using GENOVA (van der Weide, van den Brand et al. 2021) and the scaled interaction frequencies at 10 kb resolution to assess reproducibility between replicates. Downscaled interaction matrices were generated separately for each replicate and merged for visualization and downprocessing as mcool files using cooler (Abdennur and Mirny 2020). Samples were balanced using the ICE correction method(Imakaev, Fudenberg et al. 2012) with HiCExplorer. HiC maps were generated from the merged downscaled and corrected matrices using GENOVA and HiCExplorer. RCP (relative contact probabilities) were generated using GENOVA. After log-log transformation and Loess smoothing using R, the first derivative (P'max) was estimated as described in Gassler et al. (Gassler, Brandao et al. 2017).

DOWNSTREAM ANALYSIS:

CTCF peak motif analysis

Motifs were called using CisDiversity (Biswas and Narlikar 2021) on CTCF peaks called using MACS2 after separating the peaks into 3 groups for each pairwise comparison between WT and mutant CTCF: WT only, common and mutant only peaks. Of note common sites included both non-significant and significant differential binding sites detected by Diffbind as long as the peak was significantly detected by MACS in both conditions. Peak intersection sets were generated using the intervene package (Khan 2017) and motifs were searched within 250 bp of each peak summit.

ATAC-seq footprinting analysis

Footprinting analysis was performed on ATAC-seq signal using TOBIAS (Bentsen, Goymann et al. 2020). Briefly, ATAC-seq signals were first corrected for Tn5 cutting sequence bias using the

ATACorrect function and footprinting score calculated using FootprintScores at peaks detected by MACS. Finally differential footprinting estimated pariwise between WT and mutant CTCF using BINDetect and HOCOMOCO v11 core motif probability weight matrices (Kulakovskiy, Vorontsov et al. 2018). Peaks were annotated for gene promoter using UROPA (Kondili, Fust et al. 2017) and GENCODE version M25 (Frankish, Diekhans et al. 2021). Gene promoters were defined as regions within 2kb upstream and 1 kb downstream the gene transcription starting site.

Spaced motif analysis (SPAMO)

Motif spacing analysis was performed using SPAMO (Whitington, Frith et al. 2011). CTCF and ATAC-seq peaks containing CTCF consensus motifs were first identified using FIMO (Grant, Bailey and Noble 2011). SPAMO analysis was then performed using the sequences of the 250 bp region centered by the CTCF motif. CTCF consensus motif was used as the primary motif and HOCOMOCO core motif set as secondary motifs to test.

Significant spacing was defined with E-value (the lowest p-value of any spacing of the secondary motif times the number of secondary motifs) <0.05 after ignoring TFs where the aligned motif of the primary and secondary motifs showed the presence of highly similar regions and DNA repeats. This analysis was performed for accessible CTCF bound, accessible CTCF unbound and inaccessible CTCF bound peaks after intersecting CTCF motif containing CTCF and ATAC-seq peaks. Of note, the analysis was not performed on inaccessible unbound sites since the analysis would not be informative given the high number of predicted motifs in the genome and the lack of statistical power resulting from the dilution of functional and cell-specific CTCF binding sites among those sites.

Loop calling and annotation

Significant loops (q-value<0.05) were called on the downscaled and corrected interaction matrices using FitHIC at 10kb resolution (Ay, Bailey and Noble 2014) and annotated for CTCF peaks and gene TSS using pairToBed (Quinlan and Hall 2010).

Differential aggregated peak and TAD analyses

Aggregated peak and TAD analysis on downscaled and corrected interaction matrices was performed using GENOVA at 10kb resolution (van der Weide, van den Brand et al. 2021). For aggregated peaks analyses, the union of loops called by FitHIC in WT and mutant CTCF for each pairwise comparison was used to generate and compare aggregated interaction at loop anchors. For aggregated TAD analyses, the insulation score was first calculated to identify TAD and TAD

boundaries. The union of TADs called by GENOVA in WT and mutant CTCF was used to generate and compare aggregated intra and inter-TAD interactions.

Differential loop analysis

Differential loops (adjusted p-value<0.05) were called for each pairwise comparison between WT and mutant CTCF using MultiHiCCompare which performed a cross-sample normalization before testing for differential loops using an exact test (Stansfield, Cresswell and Dozmorov 2019). Enrichment for differentially expressed genes among differential loops was tested using logistic regression after annotating loop anchors with gene TSS.

Availability of data and materials

All raw and processed sequencing data files are deposited at NCBI's Gene Expression Omnibus (GEO) and will be available to public on publication of the manuscript.

Ethics approval and consent to participate

Not applicable

Competing Interests

The authors declare no competing interests.

Funding

This work was supported by 1R35GM122515 (J.S) and NIH P01CA229086 (J.S). N.M was supported National Cancer Center and A.T. by the American Cancer Society (RSG-15-189-01-RMC) and St. Baldrick's foundation (581357). ASH additionally acknowledges support from US National Institutes of Health grants R00GM130896, DP2GM140938, R33CA257878 and UM1HG011536, National Science Foundation grant 2036037, the Mathers Foundation and a Pew-Stewart Cancer Research Scholar grant.

Author' contributions

These studies were designed by Jane Skok and Catherine Do. The analysis was performed by Catherine Do and Theodore Sakellaropoulos. Experiments were performed by Guimei Jiang and Giulia Cova. Structural analyses was performed by Jie Yang and Xiaodong Chen. FRAP analysis

was performed by Christos C. Katsifis, Domenic N. Narducci and Anders S. Hansen. The paper was written by Jane Skok and Catherine Do.

Acknowledgements

The authors thank Skok lab members for helpful scientific discussions, New York University School of Medicine High Performance Computing Facility (HPCF) for computing technical support, Adriana Heguy and the Genome Technology Center (GTC) core for sequencing efforts, NYU Flow Cytometry and Cell Sorting Center for FACS analysis and sorting. GTC is a shared resource partially supported by the Cancer Center Support Grant P30CA016087 at the Laura and Isaac Perlmutter Cancer Center.

Funding

These studies were supported by a P01CA229086 (JS) and 2R35GM122515 (JS). GC and GJ were supported by fellowships from the NCC. The work in the Cheng laboratory is supported by R35 GM134744 (XC).

References:

Abdennur, N. and L. A. Mirny (2020). "Cooler: scalable storage for Hi-C data and other genomically labeled arrays." Bioinformatics **36**(1): 311-316.

Amemiya, H. M., A. Kundaje and A. P. Boyle (2019). "The ENCODE Blacklist: Identification of Problematic Regions of the Genome." Sci Rep **9**(1): 9354.

Ay, F., T. L. Bailey and W. S. Noble (2014). "Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts." Genome Res **24**(6): 999-1011.

Bailey, T. L., J. Johnson, C. E. Grant and W. S. Noble (2015). "The MEME Suite." Nucleic Acids Res **43**(W1): W39-49.

Bentsen, M., P. Goymann, H. Schultheis, K. Klee, A. Petrova, R. Wiegandt, A. Fust, J. Preussner, C. Kuenne, T. Braun, J. Kim and M. Looso (2020). "ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation." Nat Commun **11**(1): 4267.

Biswas, A. and L. Narlikar (2021). "A universal framework for detecting cis-regulatory diversity in DNA regions." Genome Res **31**(9): 1646-1662.

Buenrostro, J. D., B. Wu, H. Y. Chang and W. J. Greenleaf (2015). "ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide." Curr Protoc Mol Biol **109**: 21 29 21-21 29 29.

Cerami, E., J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander and N. Schultz (2012).

"The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data." Cancer Discov **2**(5): 401-404.

Chen, H., Y. Tian, W. Shu, X. Bo and S. Wang (2012). "Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome." PLoS One **7**(7): e41374.

Chen, Y., P. Liu, P. Sun, J. Jiang, Y. Zhu, T. Dong, Y. Cui, Y. Tian, T. An, J. Zhang, Z. Li and X. Yang (2019). "Oncogenic MSH6-CXCR4-TGFB1 Feedback Loop: A Novel Therapeutic Target of Photothermal Therapy in Glioblastoma Multiforme." Theranostics **9**(5): 1453-1473.

Corces, M. R., A. E. Trevino, E. G. Hamilton, P. G. Greenside, N. A. Sinnott-Armstrong, S. Vesuna, A. T. Satpathy, A. J. Rubin, K. S. Montine, B. Wu, A. Kathiria, S. W. Cho, M. R. Mumbach, A. C. Carter, M. Kasowski, L. A. Orloff, V. I. Risca, A. Kundaje, P. A. Khavari, T. J. Montine, W. J. Greenleaf and H. Y. Chang (2017). "An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues." Nat Methods **14**(10): 959-962.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies and H. Li (2021). "Twelve years of SAMtools and BCFtools." Gigascience **10**(2).

Davidson, I. F., B. Bauer, D. Goetz, W. Tang, G. Wutz and J. M. Peters (2019). "DNA loop extrusion by human cohesin." Science **366**(6471): 1338-1345.

de Bruijn, I., R. Kundra, B. Mastrogiacomo, T. N. Tran, L. Sikina, T. Mazor, X. Li, A. Ochoa, G. Zhao, B. Lai, A. Abeshouse, D. Baiceanu, E. Ciftci, U. Dogrusoz, A. Dufilie, Z. Erkoc, E. Garcia Lara, Z. Fu, B. Gross, C. Haynes, A. Heath, D. Higgins, P. Jagannathan, K. Kalletla, P. Kumari, J. Lindsay, A. Lisman, B. Leenknegt, P. Lukasse, D. Madela, R. Madupuri, P. van Nierop, O. Plantalech, J. Quach, A. C. Resnick, S. Y. A. Rodenburg, B. A. Satravada, F. Schaeffer, R. Sheridan, J. Singh, R. Sirohi, S. O. Sumer, S. van Hagen, A. Wang, M. Wilson, H. Zhang, K. Zhu, N. Rusk, S. Brown, J. A. Lavery, K. S. Panageas, J. E. Rudolph, M. L. LeNoue-Newton, J. L. Warner, X. Guo, H. Hunter-Zinck, T. V. Yu, S. Pilai, C. Nichols, S. M. Gardos, J. Philip, A. P. G. C. Aacr Project Genie Bpc Core Team, K. L. Kehl, G. J. Riely, D. Schrag, J. Lee, M. V. Fiandalo, S. M. Sweeney, T. J. Pugh, C. Sander, E. Cerami, J. Gao and N. Schultz (2023). "Analysis and Visualization of Longitudinal Genomic and Clinical Data from the AACR Project GENIE Biopharma Collaborative in cBioPortal." Cancer Res **83**(23): 3861-3867.

Debaugny, R. E. and J. A. Skok (2020). "CTCF and CTCFL in cancer." Curr Opin Genet Dev **61**: 44-52.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras (2013). "STAR: ultrafast universal RNA-seq aligner." Bioinformatics **29**(1): 15-21.

Dolgalev, I. (2022). "Seq-N-Slide [Computer software]."

Edelmann, W., K. Yang, A. Umar, J. Heyer, K. Lau, K. Fan, W. Liedtke, P. E. Cohen, M. F. Kane, J. R. Lipford, N. Yu, G. F. Crouse, J. W. Pollard, T. Kunkel, M. Lipkin, R. Kolodner and R. Kucherlapati (1997). "Mutation in the mismatch repair gene Msh6 causes cancer susceptibility." Cell **91**(4): 467-477.

Filippova, G. N., A. Lindblom, L. J. Meincke, E. M. Klenova, P. E. Neiman, S. J. Collins, N. A. Doggett and V. V. Lobanenkov (1998). "A widely expressed transcription factor with multiple DNA sequence specificity, CTCF, is localized at chromosome segment 16q22.1 within one of the smallest regions of overlap for common deletions in breast and prostate cancers." Genes Chromosomes Cancer **22**(1): 26-36.

Flavahan, W. A., Y. Drier, B. B. Liau, S. M. Gillespie, A. S. Venteicher, A. O. Stemmer-Rachamimov, M. L. Suva and B. E. Bernstein (2016). "Insulator dysfunction and oncogene activation in IDH mutant gliomas." Nature **529**(7584): 110-114.

Fleck, J. S., S. M. J. Jansen, D. Wollny, F. Zenk, M. Seimiya, A. Jain, R. Okamoto, M. Santel, Z. He, J. G. Camp and B. Treutlein (2023). "Inferring and perturbing cell fate regulomes in human brain organoids." Nature **621**(7978): 365-372.

Frankish, A., M. Diekhans, I. Jungreis, J. Lagarde, J. E. Loveland, J. M. Mudge, C. Sisu, J. C. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, C. Boix, S. Carbonell Sala, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. Garcia Giron, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, K. L. Howe, T. Hunt, O. G. Izuogu, R. Johnson, F. J. Martin, L. Martinez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, F. C. Riera, M. Ruffier, B. M. Schmitt, E. Stapleton, M. M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, M. Y. Wolf, J. Xu, Y. T. Yang, A. Yates, D. Zerbino, Y. Zhang, J. S. Choudhary, M. Gerstein, R. Guigo, T. J. P. Hubbard, M. Kellis, B. Paten, M. L. Tress and P. Flicek (2021). "Gencode 2021." Nucleic Acids Res **49**(D1): D916-D923.

Fudenberg, G., M. Imakaev, C. Lu, A. Goloborodko, N. Abdennur and L. A. Mirny (2016). "Formation of Chromosomal Domains by Loop Extrusion." Cell Rep **15**(9): 2038-2049.

Gao, J., B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander and N. Schultz (2013). "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal." Sci Signal **6**(269): pl1.

Gassler, J., H. B. Brandao, M. Imakaev, I. M. Flyamer, S. Ladstatter, W. A. Bickmore, J. M. Peters, L. A. Mirny and K. Tachibana (2017). "A mechanism of cohesin-dependent loop extrusion organizes zygotic genome architecture." EMBO J **36**(24): 3600-3618.

Grant, C. E., T. L. Bailey and W. S. Noble (2011). "FIMO: scanning for occurrences of a given motif." Bioinformatics **27**(7): 1017-1018.

Gregor, A., M. Oti, E. N. Kouwenhoven, J. Hoyer, H. Sticht, A. B. Ekici, S. Kjaergaard, A. Rauch, H. G. Stunnenberg, S. Uebe, G. Vasileiou, A. Reis, H. Zhou and C. Zweier (2013). "De novo mutations in the genome organizer CTCF cause intellectual disability." Am J Hum Genet **93**(1): 124-131.

Guo, C., H. S. Yoon, A. Franklin, S. Jain, A. Ebert, H. L. Cheng, E. Hansen, O. Despo, C. Bossen, C. Vettermann, J. G. Bates, N. Richards, D. Myers, H. Patel, M. Gallagher, M. S. Schlissel, C. Murre, M. Busslinger, C. C. Giallourakis and F. W. Alt (2011). "CTCF-binding elements mediate control of V(D)J recombination." Nature **477**(7365): 424-430.

Guo, Y., Q. Xu, D. Canzio, J. Shou, J. Li, D. U. Gorkin, I. Jung, H. Wu, Y. Zhai, Y. Tang, Y. Lu, Y. Wu, Z. Jia, W. Li, M. Q. Zhang, B. Ren, A. R. Krainer, T. Maniatis and Q. Wu (2015). "CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function." Cell **162**(4): 900-910.

Hansen, A. S., A. Amitai, C. Cattoglio, R. Tjian and X. Darzacq (2020). "Guided nuclear exploration increases CTCF target search efficiency." Nat Chem Biol **16**(3): 257-266.

Hansen, A. S., I. Pustova, C. Cattoglio, R. Tjian and X. Darzacq (2017). "CTCF and cohesin regulate chromatin loop stability with distinct dynamics." Elife **6**.

Hashimoto, H., D. Wang, J. R. Horton, X. Zhang, V. G. Corces and X. Cheng (2017). "Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA." Mol Cell **66**(5): 711-720 e713.

Hnisz, D., A. S. Weintraub, D. S. Day, A. L. Valton, R. O. Bak, C. H. Li, J. Goldmann, B. R. Lajoie, Z. P. Fan, A. A. Sigova, J. Reddy, D. Borges-Rivera, T. I. Lee, R. Jaenisch, M. H. Porteus, J. Dekker and R. A. Young (2016). "Activation of proto-oncogenes by disruption of chromosome neighborhoods." Science **351**(6280): 1454-1458.

Ikeda, K., K. Horie-Inoue, T. Suzuki, R. Hobo, N. Nakasato, S. Takeda and S. Inoue (2019). "Mitochondrial supercomplex assembly promotes breast and endometrial tumorigenesis by metabolic alterations and enhanced hypoxia tolerance." Nat Commun **10**(1): 4108.

Imakaev, M., G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker and L. A. Mirny (2012). "Iterative correction of Hi-C data reveals hallmarks of chromosome organization." Nat Methods **9**(10): 999-1003.

Kemp, C. J., J. M. Moore, R. Moser, B. Bernard, M. Teater, L. E. Smith, N. A. Rabaia, K. E. Gurley, J. Guinney, S. E. Busch, R. Shaknovich, V. V. Lobanenkov, D. Liggitt, I. Shmulevich, A. Melnick

and G. N. Filippova (2014). "CTCF haploinsufficiency destabilizes DNA methylation and predisposes to cancer." Cell Rep **7**(4): 1020-1029.

Khan, A. M., A (2017). "Intervene: a tool for intersection and visualization of multiple genomic region sets."

Kolde, R. (2019). "pheatmap: Pretty Heatmaps. R package version 1.0.12."

Kondili, M., A. Fust, J. Preussner, C. Kuenne, T. Braun and M. Looso (2017). "UROPA: a tool for Universal RObust Peak Annotation." Sci Rep **7**(1): 2593.

Konrad, E. D. H., N. Nardini, A. Caliebe, I. Nagel, D. Young, G. Horvath, S. L. Santoro, C. Shuss, A. Ziegler, D. Bonneau, M. Kempers, R. Pfundt, E. Legius, A. Bouman, K. E. Stuurman, K. Ounap, S. Pajusalu, M. H. Wojcik, G. Vasileiou, G. Le Guyader, H. M. Schnelle, S. Berland, E. Zonneveld-Huijssoon, S. Kersten, A. Gupta, P. R. Blackburn, M. S. Ellingson, M. J. Ferber, R. Dhamija, E. W. Klee, M. McEntagart, K. D. Lichtenbelt, A. Kenney, S. A. Vergano, R. Abou Jamra, K. Platzer, M. Ella Pierpont, D. Khattar, R. J. Hopkin, R. J. Martin, M. C. J. Jongmans, V. Y. Chang, J. A. Martinez-Agosto, O. Kuismin, M. I. Kurki, O. Pietilainen, A. Palotie, T. J. Maarup, D. S. Johnson, K. Venborg Pedersen, L. W. Laulund, S. A. Lynch, M. Blyth, K. Prescott, N. Canham, R. Ibitoye, E. H. Brilstra, M. Shinawi, E. Fassi, D. D. D. Study, H. Sticht, A. Gregor, H. Van Esch and C. Zweier (2019). "CTCF variants in 39 individuals with a variable neurodevelopmental disorder broaden the mutational and clinical spectrum." Genet Med **21**(12): 2723-2733.

Krueger, F. (2021). "TrimGalore."

Kulakovskiy, I. V., I. E. Vorontsov, I. S. Yevshin, R. N. Sharipov, A. D. Fedorova, E. I. Rumynskiy, Y. A. Medvedeva, A. Magana-Mora, V. B. Bajic, D. A. Papatsenko, F. A. Kolpakov and V. J. Makeev (2018). "HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis." Nucleic Acids Res **46**(D1): D252-D259.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods **9**(4): 357-359.

Li, Y., J. H. I. Haarhuis, A. Sedeno Cacciatore, R. Oldenkamp, M. S. van Ruiten, L. Willems, H. Teunissen, K. W. Muir, E. de Wit, B. D. Rowland and D. Panne (2020). "The structural basis for cohesin-CTCF-anchored loops." Nature **578**(7795): 472-476.

Liu, X. H., L. P. Liu, X. M. Xu, M. Hua, Q. Kang, A. Li and L. Huang (2021). "FOXN2 suppresses the proliferation and invasion of human hepatocellular carcinoma cells." Eur Rev Med Pharmacol Sci **25**(2): 731-737.

Love, M. I., W. Huber and S. Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." Genome Biol **15**(12): 550.

Ma, J., Y. Lu, S. Zhang, Y. Li, J. Huang, Z. Yin, J. Ren, K. Huang, L. Liu, K. Yang, G. Wu and S. Xu (2018). "beta-Trcp ubiquitin ligase and RSK2 kinase-mediated degradation of FOXN2 promotes tumorigenesis and radioresistance in lung cancer." Cell Death Differ **25**(8): 1473-1485.

Maurano, M. T., H. Wang, S. John, A. Shafer, T. Canfield, K. Lee and J. A. Stamatoyannopoulos (2015). "Role of DNA Methylation in Modulating Transcription Factor Occupancy." Cell Rep **12**(7): 1184-1195.

Moore, J. M., N. A. Rabaia, L. E. Smith, S. Fagerlie, K. Gurley, D. Loukinov, C. M. Disteche, S. J. Collins, C. J. Kemp, V. V. Lobanenkov and G. N. Filippova (2012). "Loss of maternal CTCF is associated with peri-implantation lethality of Ctcf null embryos." PLoS One **7**(4): e34915.

Narendra, V., M. Bulajic, J. Dekker, E. O. Mazzoni and D. Reinberg (2016). "CTCF-mediated topological boundaries during development foster appropriate gene regulation." Genes & development **30**: 2657-2662.

Narendra, V., P. P. Rocha, D. An, R. Raviram, J. A. Skok, E. O. Mazzoni and D. Reinberg (2015). "Transcription. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation." Science **347**(6225): 1017-1021.

Nishana, M., C. Ha, J. Rodriguez-Hernaez, A. Ranjbaran, E. Chio, E. P. Nora, S. B. Badri, A. Kloetgen, B. G. Bruneau, A. Tsirigos and J. A. Skok (2020). "Defining the relative and combined contribution of CTCF and CTCFL to genomic regulation." Genome Biol **21**(1): 108.

Nora, E. P., L. Caccianini, G. Fudenberg, K. So, V. Kameswaran, A. Nagle, A. Uebersohn, B. Hajj, A. L. Saux, A. Coulon, L. A. Mirny, K. S. Pollard, M. Dahan and B. G. Bruneau (2020). "Molecular basis of CTCF binding polarity in genome folding." Nat Commun **11**(1): 5612.

Nora, E. P., A. Goloborodko, A. L. Valton, J. H. Gibcus, A. Uebersohn, N. Abdennur, J. Dekker, L. A. Mirny and B. G. Bruneau (2017). "Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization." Cell **169**(5): 930-944 e922.

Parelho, V., S. Hadjur, M. Spivakov, M. Leleu, S. Sauer, H. C. Gregson, A. Jarmuz, C. Canzonetta, Z. Webster, T. Nesterova, B. S. Cobb, K. Yokomori, N. Dillon, L. Aragon, A. G. Fisher and M. Merkenschlager (2008). "Cohesins functionally associate with CTCF on mammalian chromosome arms." Cell **132**(3): 422-433.

Persengiev, S. P., J. Li, M. L. Poulin and D. L. Kilpatrick (2001). "E2F2 converts reversibly differentiated PC12 cells to an irreversible, neurotrophin-dependent state." Oncogene **20**(37): 5124-5131.

Price, E., L. M. Fedida, E. M. Pugacheva, Y. J. Ji, D. Loukinov and V. V. Lobanenkov (2023). "An updated catalog of CTCF variants associated with neurodevelopmental disorder phenotypes." Front Mol Neurosci **16**: 1185796.

Pugacheva, E. M., N. Kubo, D. Loukinov, M. Tajmul, S. Kang, A. L. Kovalchuk, A. V. Strunnikov, G. E. Zentner, B. Ren and V. V. Lobanenkov (2020). "CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention." Proc Natl Acad Sci U S A **117**(4): 2020-2031.

Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842.

Ramirez, F., D. P. Ryan, B. Gruning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dundar and T. Manke (2016). "deepTools2: a next generation web server for deep-sequencing data analysis." Nucleic Acids Res **44**(W1): W160-165.

Rosignoli, S. and A. Paiardini (2022). "Boosting the Full Potential of PyMOL with Structural Biology Plugins." Biomolecules **12**(12).

Saldana-Meyer, R., J. Rodriguez-Hernaez, T. Escobar, M. Nishana, K. Jacome-Lopez, E. P. Nora, B. G. Bruneau, A. Tsirigos, M. Furlan-Magaril, J. Skok and D. Reinberg (2019). "RNA Interactions Are Essential for CTCF-Mediated Genome Organization." Mol Cell **76**(3): 412-422 e415.

Schmidl, C., A. F. Rendeiro, N. C. Sheffield and C. Bock (2015). "ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors." Nat Methods **12**(10): 963-965.

Sergushichev, A. (2016). "An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation." bioRxiv. doi: 10.1101/060012.

Servant, N., N. Varoquaux, B. R. Lajoie, E. Viara, C. J. Chen, J. P. Vert, E. Heard, J. Dekker and E. Barillot (2015). "HiC-Pro: an optimized and flexible pipeline for Hi-C data processing." Genome Biol **16**: 259.

Sprague, B. L., R. L. Pego, D. A. Stavreva and J. G. McNally (2004). "Analysis of binding reactions by fluorescence recovery after photobleaching." Biophys J **86**(6): 3473-3495.

Stansfield, J. C., K. G. Cresswell and M. G. Dozmorov (2019). "multiHiCcompare: joint normalization and comparative analysis of complex Hi-C experiments." Bioinformatics **35**(17): 2916-2923.

Stark, R. B., G (2011). "DiffBind: differential binding analysis of ChIP-Seq peak data."

Tarasov, A., A. J. Vilella, E. Cuppen, I. J. Nijman and P. Prins (2015). "Sambamba: fast processing of NGS alignment formats." Bioinformatics **31**(12): 2032-2034.

Tate, J. G., S. Bamford, H. C. Jubb, Z. Sondka, D. M. Beare, N. Bindal, H. Boutselakis, C. G. Cole, C. Creatore, E. Dawson, P. Fish, B. Harsha, C. Hathaway, S. C. Jupe, C. Y. Kok, K. Noble, L. Ponting, C. C. Ramshaw, C. E. Rye, H. E. Speedy, R. Stefancsik, S. L. Thompson, S. Wang, S. Ward, P. J. Campbell and S. A. Forbes (2019). "COSMIC: the Catalogue Of Somatic Mutations In Cancer." Nucleic Acids Res **47**(D1): D941-D947.

Valverde de Morales, H. G., H. V. Wang, K. Garber, X. Cheng, V. G. Corces and H. Li (2023). "Expansion of the genotypic and phenotypic spectrum of CTCF-related disorder guides clinical management: 43 new subjects and a comprehensive literature review." Am J Med Genet A **191**(3): 718-729.

van der Gun, B. T., L. J. Melchers, M. H. Ruiters, L. F. de Leij, P. M. McLaughlin and M. G. Rots (2010). "EpCAM in carcinogenesis: the good, the bad or the ugly." Carcinogenesis **31**(11): 1913-1921.

van der Weide, R. H., T. van den Brand, J. H. I. Haarhuis, H. Teunissen, B. D. Rowland and E. de Wit (2021). "Hi-C analyses with GENOVA: a case study with cohesin variants." NAR Genom Bioinform **3**(2): lqab040.

Wendt, K. S., K. Yoshida, T. Itoh, M. Bando, B. Koch, E. Schirghuber, S. Tsutsumi, G. Nagae, K. Ishihara, T. Mishiro, K. Yahata, F. Imamoto, H. Aburatani, M. Nakao, N. Imamoto, K. Maeshima, K. Shirahige and J. M. Peters (2008). "Cohesin mediates transcriptional insulation by CCCTC-binding factor." Nature **451**(7180): 796-801.

Whitington, T., M. C. Frith, J. Johnson and T. L. Bailey (2011). "Inferring transcription factor complexes from ChIP-seq data." Nucleic Acids Res **39**(15): e98.

Wolff, J., V. Bhardwaj, S. Nothjunge, G. Richard, G. Renschler, R. Gilsbach, T. Manke, R. Backofen, F. Ramirez and B. A. Gruning (2018). "Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization." Nucleic Acids Res **46**(W1): W11-W16.

Xiang, Y., X. Zhou, S. L. Hewitt, J. A. Skok and W. T. Garrard (2011). "A multifunctional element in the mouse Igkappa locus that specifies repertoire and Ig loci subnuclear location." J Immunol **186**(9): 5356-5366.

Ye, H. and M. Duan (2019). "FOXN2 is downregulated in breast cancer and regulates migration, invasion, and epithelial- mesenchymal transition through regulation of SLUG." Cancer Manag Res **11**: 525-535.

Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li and X. S. Liu (2008). "Model-based analysis of ChIP-Seq (MACS)." Genome Biol **9**(9): R137.