Brain and cancer associated binding domain mutations provide insight into CTCF's relationship with chromatin and its contribution to gene regulation

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], Theodore Sakellaropoulos[1,2], Raphael Vidal[1,2], Priscillia Lhoumaud[1,2], Aristotelis Tsirigos [1,2,6], Faye Fara D. Regis[1,2], Nata Kakabadze[1,2], Elphege P Nora[7], Marcus Noyes[8], 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. Applied Bioinformatics Laboratories, Office of Science & Research, NYU Grossman School of Medicine, New York, NY, USA
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

Abstract 150 words.

Here we used a series of CTCF mutations to explore CTCF's relationship with chromatin and its contribution to gene regulation. CTCF's impact depends on the genomic context of bound sites and the unique binding properties of WT and mutant CTCF proteins. Specifically, CTCF's signal strength is linked to changes in accessibility, and the ability to block cohesin is linked to its binding stability. Multivariate modelling reveals that both CTCF and accessibility contribute independently to cohesin binding and insulation, however CTCF signal strength has a stronger effect. CTCF and chromatin have a bidirectional relationship such that at CTCF sites, accessibility is reduced in a cohesin-dependent, mutant specific fashion. In addition, each mutant alters TF binding and accessibility in an indirect manner, changes which impart the most influence on rewiring transcriptional networks and the cell's ability to be reprogrammed. Collectively, the mutant perturbations provide a rich resource for determining CTCF's site-specific 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 [1,2]. 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 [3]. 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 chromatin loops by actively extruding DNA until movement of the complex is blocked by two CTCF binding sites in convergent orientation [5,6]. 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 [7-10]. 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[11,12].

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 are incompletely elucidated. CTCF has been degraded by auxin in numerous cell types, and although these studies provide 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 because of aberrant enhancer promoter contacts, and this can have dramatic consequences on developmental processes and cancer initiation [13-18]. 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 [18,19]. Moreover, point mutations and deletions have been identified in many other tumors. Together these findings indicate that CTCF acts as a tumor suppressor [20]. 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 [21]. 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 [22-24].

CTCF binds to chromatin through a subset of its eleven zinc fingers (ZFs) [25,26]. Specifically, ZFs3-7 make sequence specific contacts with a 12-15 base-pair (bp) core consensus binding site. 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 [27]. While the crystal structure of CTCF in complex with a known DNA binding site has revealed new insights into the contribution of each ZF to sequence-specific binding in an *in vitro* setting, 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. 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 genetic, imaging, molecular and bioinformatic analyses to examine the impact of nine, high frequency cancer associated CTCF mutations in eight amino acid residues of ZFs that contact the core consensus binding motif. A subset of these is 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) direct or indirect site-specific mutant effects on accessibility, TF binding, cohesin overlap, chromatin interactivity and gene expression. Compared to a complete knockout of CTCF or deletion of a CTCF binding site, the mutants offer a more subtle and in-depth approach

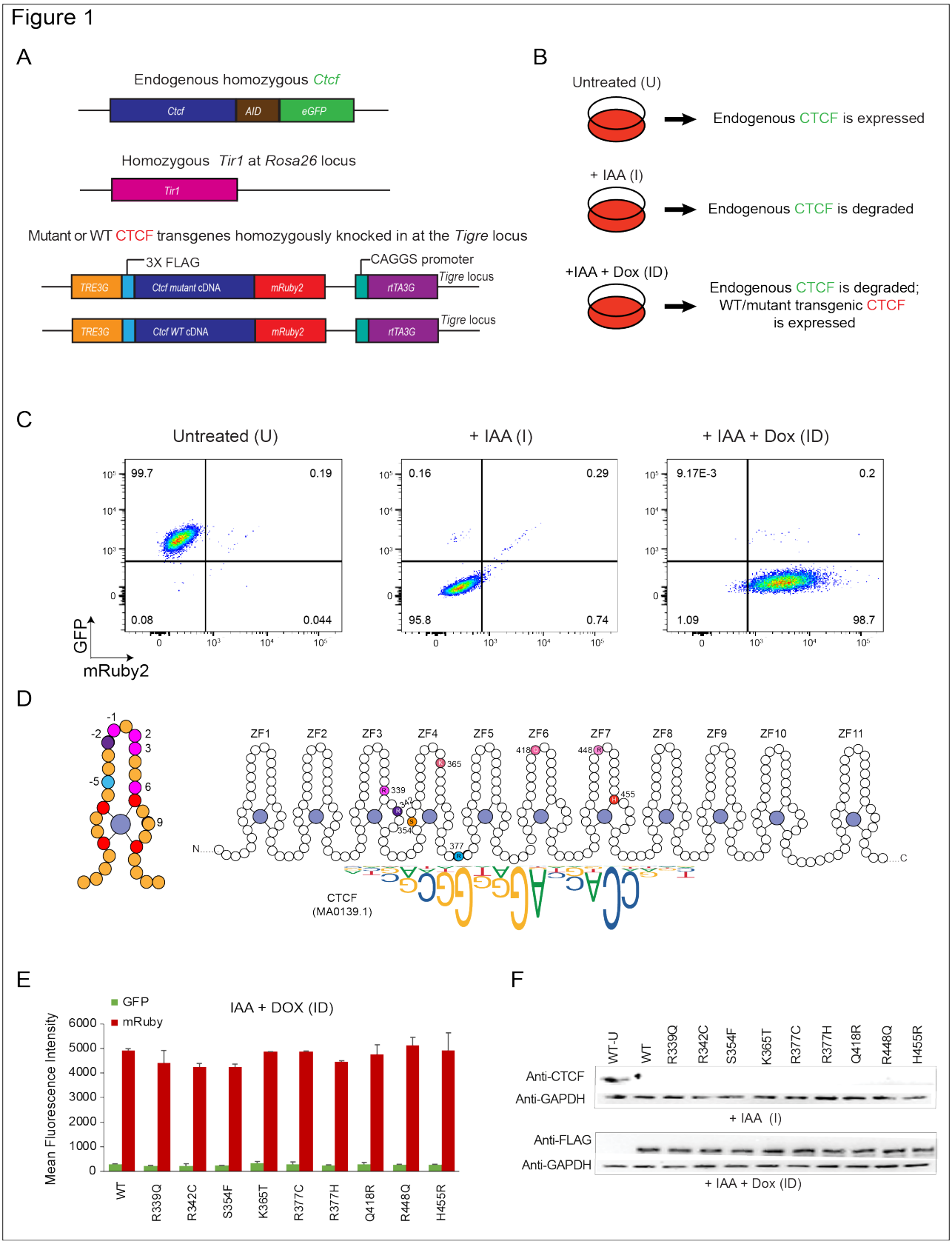
for teasing out the contribution of CTCF to chromatin organization and gene regulation, since each mutation produces its own unique effect.

Here we demonstrate that the functional impact of WT and mutant CTCF depends both on the genomic context of bound sites and the unique binding properties of each WT and mutant CTCF protein. Analyses of the relationship between CTCF, cohesin and accessibility revealed that although CTCF can bind both accessible and inaccessible sites, inaccessible bound sites predominantly have weaker signals and are less competent at blocking cohesin. Multivariate modelling reveals that both CTCF and accessibility contribute independently to cohesin binding, however CTCF signal strength has a stronger effect, linked to a stronger insulation score. Indeed, CTCFs ability to block cohesin is linked to the residence time or 'OFF' rate of each mutant on chromatin as determined by FRAP experiments, while occupancy or 'ON' rate is correlated with their binding at common accessible sites, indicating that accessibility dominates the search space. Together, the mutants highlight the fact that occupancy and residence time can be uncoupled such that although some mutants have fewer binding sites than WT CTCF, they have a stronger residence time at these locations which increases their ability to block cohesin and contribute to chromatin folding. The reverse situation also occurs such that mutants which occupy more sites than WT, do so in an unstable manner with low residence times that impair their ability to engage cohesin and form loops. Our analyses also highlight that in contrast to CTCF-bound sites, cohesin binding at CTCF-free sites is weak and does not track with accessibility. This is also true at highly accessible CTCF-free sites, such as enhancers and promoters which are involved in dynamic looping events, and where cohesin signals are smaller than those found at inaccessible CTCF-bound sites, perhaps because TFs that block cohesin have considerably shorter residence times compared to CTCF (lasting seconds compared to minutes).

Interestingly, although accessibility can impact CTCF binding, CTCF can in turn reduce accessibility in a cohesin-dependent manner, suggesting a role for loop extrusion in this process. The mutations enable us to determine that CTCF binding drives the reduction in accessibility, rather than the other way around, because each mutant has a distinct impact. Overall, reduced accessibility at loop bases correlates with the fraction bound, consistent with CTCF functioning in a dose-dependent manner. The unique properties of WT and mutant CTCF also perturb TF binding and cancer, brain, immune and metabolic gene expression pathways. Mutants have a more pronounced effect on transcriptional networks than auxin inducible degradation, because of indirect effects that can be explained by altered TF binding profiles. Changes in each mutant's

150 transcriptional program in turn have an impact on the cell's ability to exit from pluripotency and
151 undergo reprogramming. Thus, while we do not investigate the pathogenesis of human disorders,
152 our investigations offer insight into cancer and brain-related effects. Collectively, the mutants offer
153 a rich resource to investigate site specific CTCF-mediated effects on chromatin folding and gene
154 regulation, while distinguishing cause from consequence and direct from indirect outcomes.
155
156 **RESULTS**
157 **CTCF complementation system**
158 To explore CTCF's relationship with chromatin and its contribution to gene regulation, we made
159 use of a small-molecule auxin inducible degron (AID) mouse ESC (mESC) system [3,8]. In these
160 cells, both endogenous CTCF alleles are tagged with AID as well as eGFP (CTCF-AID-eGFP)
161 and the auxin-activated ubiquitin ligase TIR1 (from *Oryza sativa*) is constitutively expressed from
162 the *Rosa26* locus (**Fig. 1A)**. Addition of indole acetic acid (IAA, an analogue of auxin) leads to
163 rapid poly-ubiquitination and proteasomal degradation of the proteins tagged with the AID domain.
164
165 To study the impact of CTCF mutations, we established a rescue system wherein the mESC
166 degron cell line was modified to express either a stable doxycycline-inducible control wild-type
167 *Ctcf* or a mutant *Ctcf* (*mCtcf*) transgene in the absence of endogenous CTCF[9]. Transgenes,
168 knocked into the *Tigre* locus, have a 3 x FLAG tag, which allows us to distinguish mutant or wild-
169 type transgenic CTCF from endogenous CTCF using a FLAG antibody in Western blot and ChIP-
170 seq analysis. An mRuby fluorescent tag enables us to accurately determine expression levels of
171 mutant or wild-type transgenic CTCF by flow cytometry (**Fig. 1A**). The three conditions used for
172 our analysis are shown in **Fig. 1B**, and an example of the FACS profile for cells expressing a WT
173 CTCF transgene in each condition is shown in **Fig. 1C**. This system was previously used by us
174 to study the interplay between CTCF and CTCFL, the paralog of CTCF [8].
175
176 We selected a subset of missense mutations that occur at high frequency in cancer patients
177 (**Extended Data Fig. 1A**), focusing on those found in ZFs that contact the 12-15bp consensus
178 CTCF binding motif (**Fig. 1D** and highlighted with an asterisk in **Extended Data Fig. 1B**). We
179 analyzed mutations in **(i)** amino acids that make base-specifying contacts with the alpha helix of
180 DNA (at locations -2, 2, 3 and 6 on the ZF), **(ii)** residues that coordinate the zinc ion which are
181 essential for providing stability to ensure the proper folding of the ZF domain, **(iii)** two other
182 classes of amino acid residues in the ZFs: boundary residues (that are important for interactions
183 between ZFs), and residues that contact the sugar phosphate backbone of DNA, and **(iv)** other

Figure 1


A
Endogenous homozygous Ctcf
Ctcf
AID
eGFP
Homozygous Tir1 at Rosa26 locus
Tir1
Mutant or WT CTCF transgenes homozygously knocked in at the Tigre locus
3X FLAG
CAGGS promoter
TRE3G
Ctcf mutant cDNA
mRuby2
rtTA3G
Tigre locus
Ctcf WT cDNA
B
Untreated (U)
Endogenous CTCF is expressed
+ IAA (I)
Endogenous CTCF is degraded
+IAA + Dox (ID)
Endogenous CTCF is degraded; WT/mutant transgenic CTCF is expressed
C
Untreated (U)
+ IAA (I)
+ IAA + Dox (ID)
99.7
0.19
0.08
0.044
0.16
0.29
95.8
0.74
9.17E-3
0.2
1.09
98.7
GFP
mRuby2
D
ZF1
ZF2
ZF3
ZF4
ZF5
ZF6
ZF7
ZF8
ZF9
ZF10
ZF11
N
C
CTCF
(MA0139.1)
E
GFP
mRuby
IAA + DOX (ID)
Mean Fluorescence Intensity
WT
R339Q
R342C
S354F
K365T
R377C
R377H
Q418R
R448Q
H455R
F
WT-U
WT
Anti-CTCF
Anti-GAPDH
+ IAA (I)
Anti-FLAG
Anti-GAPDH
+ IAA + Dox (ID)

184

Fig. 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 mRuby (transgenic WT CTCF) under the different conditions shown in (**B**). (**D**) Left, scheme showing the locations of the different types of CTCF mutations within a ZF. 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. Right, schematic representation of CTCF showing the locations of each mutation under investigation. (**E**) Bar graph showing the expression levels (mean fluorescence intensity) of endogenous CTCF (GFP labelled) and transgenic WT or mutant CTCF (mRuby2 labelled) after treatment with IAA and Dox (ID). The presence of the IAA leads to degradation of GFP labelled endogenous CTCF. The error bars represent the standard deviation between the two replicates. (**F**) 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. The second replicate is shown in Extended Data Fig. 2.

highly mutated residues in the region. Each class of mutation is color-coded as shown in **Fig. 1D** (left) **and Extended Data Fig. 1A, B** and the color code is maintained throughout the figures. Mutations were identified from cBioPortal [28-30] and COSMIC [31] (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 of R377, both of which (R377H and R377C) are equally represented in patients (**Extended Data Fig. 1B**). Selected mutations (R339Q, R342C, R448Q, R377H and R377C) have also been shown to be implicated in neurodevelopmental disorders (**Extended Data Fig. 1B**) [4,24]. To analyze the impact of each CTCF mutation, individual clones with comparable levels of transgene expression (mutant and WT) were selected based on flow cytometry and Western blot analysis (**Fig. 1E, F and Extended Data Fig. 2**).

Retained binding sites are enriched in accessible regions with strong CTCF signals

To investigate the impact of each mutation on CTCF binding, we analyzed the profiles of mutant versus WT CTCF at the molecular level by performing a FLAG ChIP-seq using cells expressing transgenic WT and mutant CTCF in the absence of endogenous CTCF (ID). Briefly, we degraded endogenous CTCF-mAID-GFP by IAA addition, induced WT or mutant transgene expression with Dox for 48 hours, and then performed FLAG ChIP-seq. We identified three groups of binding sites for each mutation: WT only (sites bound by WT CTCF that mutants can no longer bind), common

207 sites (sites that are bound by both WT and mutant proteins) and mutant only sites (*de novo* sites
208 that only mutant proteins bind)**.**
209
210 The location of each mutation and the complementary DNA triplet that the relevant ZF binds is
211 marked by a box in the consensus motif shown in **Fig. 2A** and **Extended Data Fig. 3A**. Below
212 that we highlight the motif identified by MEME [32] for the set of *de novo* sites that each mutant
213 binds**.** These motifs differ from that of the consensus sequence by either (i) a diminished
214 requirement for DNA bases that contact the zinc finger that is mutated (**Fig. 2A**), or (ii) an
215 alteration to the sequence of the consensus motif, which could be explained by the impact of
216 mutations on DNA-base interacting zinc finger residues crucial for DNA sequence recognition.
217 For example, K365 binds to the second guanine in the triplet GGC of the consensus motif and
218 the de novo motif found in K365T mutant is consistent with the loss of hydrogen bonds with the
219 central guanine due to the threonine substitution, as previously demonstrated by
220 crystallography[33]. Each mutant displays their own unique binding profiles including the two distinct
221 mutations of residue R377, which is located in the linker region between ZF3 and ZF4. R377C
222 binds to less common and more *de novo* sites compared to R377H and each mutant binds a motif
223 that is similar to the consensus motif but lacking bases associated with ZF3 (**Fig. 2A** and
224 **Extended Data Fig. 3A**).
225
226 The bar graphs show the percentage of the three groups of binding sites for each mutant, while
227 the heatmaps reveal the number of binding sites for each subset and highlight differences in
228 binding strength between WT and mutants at the three locations. CTCF signal strength is
229 strongest at common sites, however, it is of note that even at these retained sites, the strength of
230 CTCF binding is altered in a mutant specific manner, suggesting that a 'binding dosage effect'
231 could contribute to their functional impact. As expected, the H455R mutation in the zinc
232 coordinating residue of ZF7 loses the most binding sites and has a high percentage of WT only
233 sites.
234
235 To determine whether differences in peak intensity in the different groups could be explained by
236 differences in accessibility, we performed omniATAC [34] (**Fig. 2A and Extended Data Fig. 2A**).
237 Indeed, the profiles for the different binding groups indicate that CTCF binding strength is a
238 function of chromatin accessibility, with common sites that predominantly have the strongest
239 CTCF signal, being the most accessible. In sum, ChIP-seq analysis reveals that CTCF mutants
240 lose, retain and gain a subset of binding sites throughout the genome with mutant specific profiles.

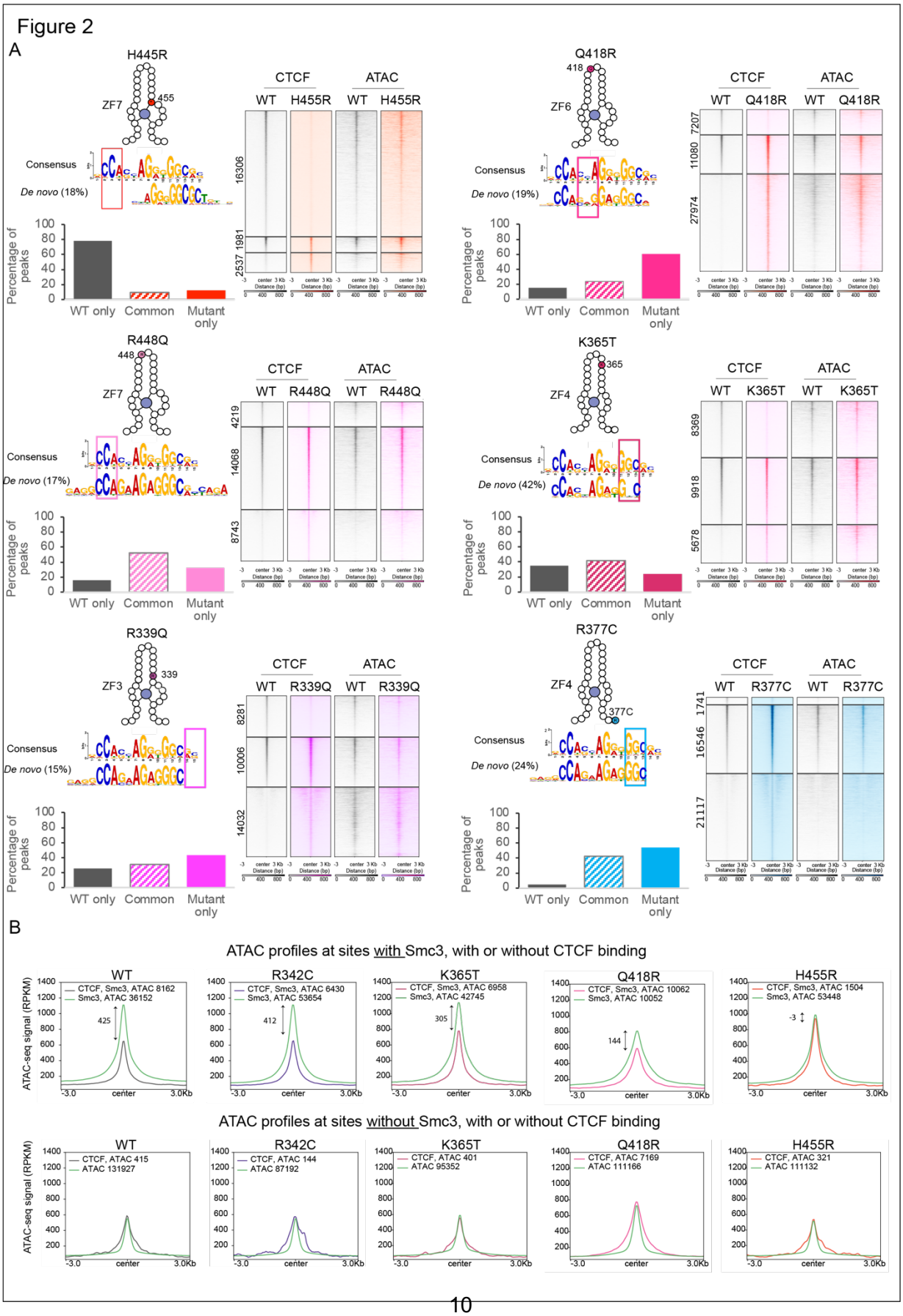


Figure 2
A
H445R
ZF7
455
Consensus
De novo (18%)
CTCF
ATAC
WT
H455R
16306
2537
1981
Percentage of peaks
WT only
Common
Mutant only
Q418R
418
ZF6
De novo (19%)
7207
11080
27974
R448Q
448
ZF7
De novo (17%)
4219
14068
8743
K365T
365
ZF4
De novo (42%)
8369
9918
5678
R339Q
339
ZF3
De novo (15%)
8281
10006
14032
R377C
377C
ZF4
De novo (24%)
1741
16546
21117
B
ATAC profiles at sites with Smc3, with or without CTCF binding
WT
CTCF, Smc3, ATAC 8162
Smc3, ATAC 36152
425
R342C
CTCF, Smc3, ATAC 6430
Smc3, ATAC 53654
412
K365T
CTCF, Smc3, ATAC 6958
Smc3, ATAC 42745
305
Q418R
CTCF, Smc3, ATAC 10062
Smc3, ATAC 10052
144
H455R
CTCF, Smc3, ATAC 1504
Smc3, ATAC 53448
-3
ATAC-seq signal (RPKM)
center
3.0Kb
ATAC profiles at sites without Smc3, with or without CTCF binding
CTCF, ATAC 415
ATAC 131927
CTCF, ATAC 144
ATAC 87192
CTCF, ATAC 401
ATAC 95352
CTCF, ATAC 7169
ATAC 111166
CTCF, ATAC 321
ATAC 111132

241

Fig. 2: **CTCF mutations have unique chromatin binding profiles**. (**A)** Scheme showing the locations of the different CTCF mutations within the ZF. The consensus CTCF motif highlights the triplet to which each mutant zinc finger binds (top) as well as the motif identified by MEME for the *de novo* mutant only binding sites (bottom). Each graph shows the percentage of WT and mutant CTCF binding at WT only, common and mutant only sites and heatmaps show the profile of CTCF binding and ATAC-seq signal for WT only, common and mutant only sites. The remaining mutants are shown in **Extended Data Fig. 3A.** (**B**) Profiles of ATAC-seq in WT and mutant CTCF expressing cells are shown for sites, with (color-coded for each mutant) or without (green) CTCF binding in the presence or absence of Smc3. The remaining mutants are shown in **Extended Data Fig. 3B.**

242 Loss and gain of CTCF binding occurs predominantly at weak inaccessible CTCF sites, while
243 binding is retained at the stronger common binding sites which have high accessibility, although
244 binding strength and accessibility can vary depending on the mutant.
245
246 **CTCF reduces ATAC-seq signal in a cohesin-dependent manner**
247 To further investigate the link between accessibility and CTCF binding, we compared the ATAC-
248 seq signal at CTCF-bound versus unbound sites in the presence or absence of Smc3, a subunit
249 of the cohesin complex (**Fig. 2B** and **Extended Data Fig. 3B**). Surprisingly, accessibility was
250 reduced in an Smc3-dependent manner at CTCF-bound sites. This was observed for both WT
251 and mutant proteins. Each CTCF mutant exhibited a differential ability to decrease the height and
252 breadth of ATAC-seq peaks, suggesting that CTCF is driving the change in accessibility and that
253 the relationship between CTCF and accessibility is bidirectional. We do not observe the same
254 phenomenon at sites without Smc3, which lends support to the idea that loop formation decreases
255 accessibility, suggesting that unique chromatin structures form at loop bases. The same
256 phenomenon is observed at enhancer and promoter sites, although the effect is weaker,
257 particularly at promoters (**Extended Data Fig. 3C**). The weakened effect may be due to
258 interference of TF binding at these sites and a competition between CTCF and TF-mediated loop
259 extrusion. These results highlight the usefulness of the mutants in distinguishing cause from
260 effect: without the mutants it would not be possible to conclude that binding of CTCF involved in
261 loop extrusion drives changes in accessibility.
262
263 **Mutations reduce the chromatin residence time and bound fraction of CTCF**
264 To elucidate the impact of the mutations on CTCF's binding dynamics, we performed fluorescence
265 recovery after photobleaching (FRAP) on the 9 mutants and WT CTCF [35,36], using the same auxin
266 and Dox conditions as described for the ChIP-seq and ATAC-seq experiments. We performed
267 FRAP by bleaching a circular 1 μm diameter circle, and monitored recovery for over 10 minutes,
268 recording 30 movies per condition over three replicates. While the recovery of WT CTCF largely
269 matched prior CTCF FRAP results [35], all mutants exhibited faster recovery

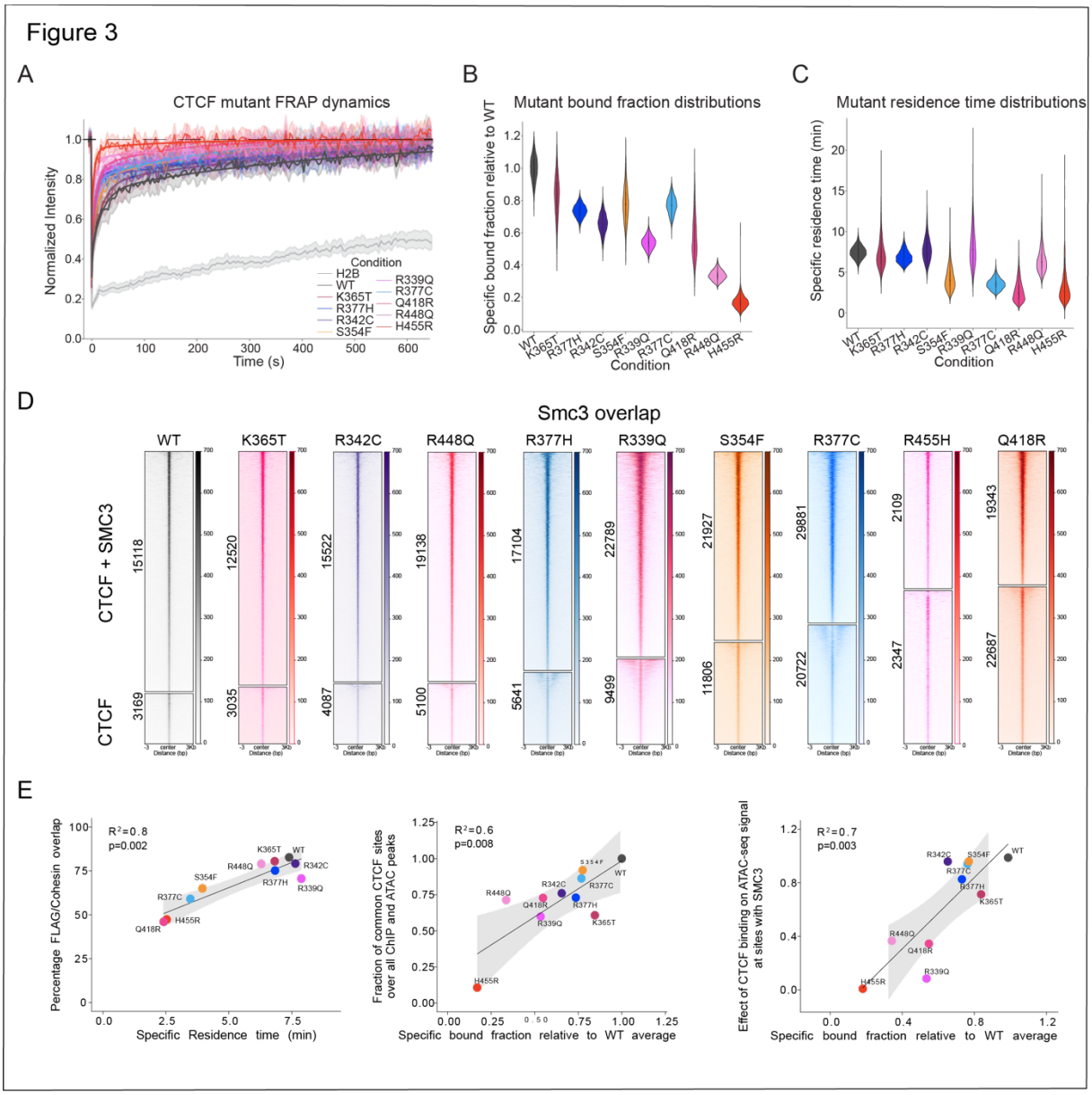


Fig. 3: **Each mutation uniquely impacts CTCF's chromatin bound fraction, residence time and interaction with DNA.** (**A**) Plots of FRAP dynamics for WT and mutant mRuby-CTCF overlaid with a two-state binding model. The scatterplots show the average recovery across all movies analyzed, and the outlines 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). (**D**) Heatmaps show the proportion and number of CTCF-cohesin versus CTCF only binding sites. All CTCF peaks correspond to the peaks found in mutant or WT CTCF expressing cells. (**E**) From left to right, correlation (R^2=0.8) between residence time and the percentage of CTCF-cohesin overlap for each mutant, correlation (R^2=0.6) 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, correlation (R^2=0.7) between the chromatin bound mutant versus WT fraction detected by FRAP and the effect of CTCF binding on ATAC-seq signal at Smc3 overlapping sites (shown in Figure 2**B**).

consistent with reduced and/or less stable DNA-binding (**Fig. 3A**). To quantitatively extract dynamic parameters, we fit the FRAP curves to a reaction-dominant kinetic model [35,36] and

273 estimated the residence time (duration of CTCF binding to DNA) and bound fraction (proportion
274 of total CTCF bound to DNA) of each mutant with subsequent comparison to the WT (**Fig. 3B,**
275 **C**). As shown in **Fig. 3B, C** there is a general reduction in the residence time and/or bound fraction
276 across each mutant, with WT CTCF exhibiting the largest specific bound fraction and residence
277 time and the mutants showing a lower or similar residence time. Each mutant displayed unique
278 parameters, regardless of the mutant category. The R377H and R377C mutants – both mutations
279 in the phosphate contacts – have similar bound fractions, but distinct residence times. The
280 mutations in the amino acids making direct contact with DNA (R339Q of ZF3, K365T of ZF4,
281 Q418R of ZF6, and R448Q of ZF7) have varied residence times and bound fractions, despite
282 belonging to the same group. The R448Q mutant of ZF7 showed the second lowest bound
283 fraction, despite exhibiting a residence time comparable to WT. The Q418R mutant of ZF6 had
284 the lowest residence time of all mutants examined with an intermediate bound fraction.
285 Conversely, the K365T mutant of ZF4 exhibited a bound fraction and residence time most closely
286 comparable to WT. The H455R mutation of ZF7, which assists in coordinating the zinc, had the
287 lowest bound fraction and the second lowest residence time of all examined mutants, accurately
288 reflecting the most deleterious mutation to canonical CTCF function. The R342C mutation of ZF3,
289 which occurs in a boundary residue between the two zinc-ligand histidine residues of ZF3, showed
290 a residence time similar to that of wild type, but had a lower bound fraction, possibly suggesting
291 a role for boundary residues in binding efficiency. Lastly, the S354F mutant of ZF4, located
292 between the two zinc-ligand cysteine residues of ZF4, which substitutes a bulky phenylalanine in
293 place of a serine in ZF4, showed both a moderate decrease in bound fraction as well as residence
294 time.
295
296 In summary, our FRAP results demonstrate that all mutants show changes in DNA binding either
297 through an altered residence time or bound fraction. Surprisingly, some mutations strongly affect
298 either the residence time or the bound fraction, with minimal effects on the other. This suggests
299 that the CTCF target search mechanism (ON-rate) can be decoupled from binding stability (OFF
300 rate) [37], such that some mutants affect the efficiency of CTCF search for binding sites, without
301 affecting the residence time of CTCF once bound.
302
303 **CTCF mutations have a graded impact on binding and function**
304 To further understand the relationship between FRAP and ChIP-seq data, we asked whether the
305 residence time detected by FRAP is linked to CTCF's binding stability and ability to block cohesin.

As shown in the heatmaps (**Fig. 3D**), the proportion of global CTCF-cohesin versus CTCF only binding sites is graded across mutants, to some extent mirroring a gene dosage effect that could occur as a consequence of altered residence time (**Fig. 3E**). Indeed, analysis of Smc3 ChIP-seq revealed that the residence time was strongly associated (R^2=0.8, p=0.002) with the percentage of CTCF-cohesin overlap for each mutant (**Fig. 3D**) which reflects the impact of binding stability on CTCF's function in loop extrusion.

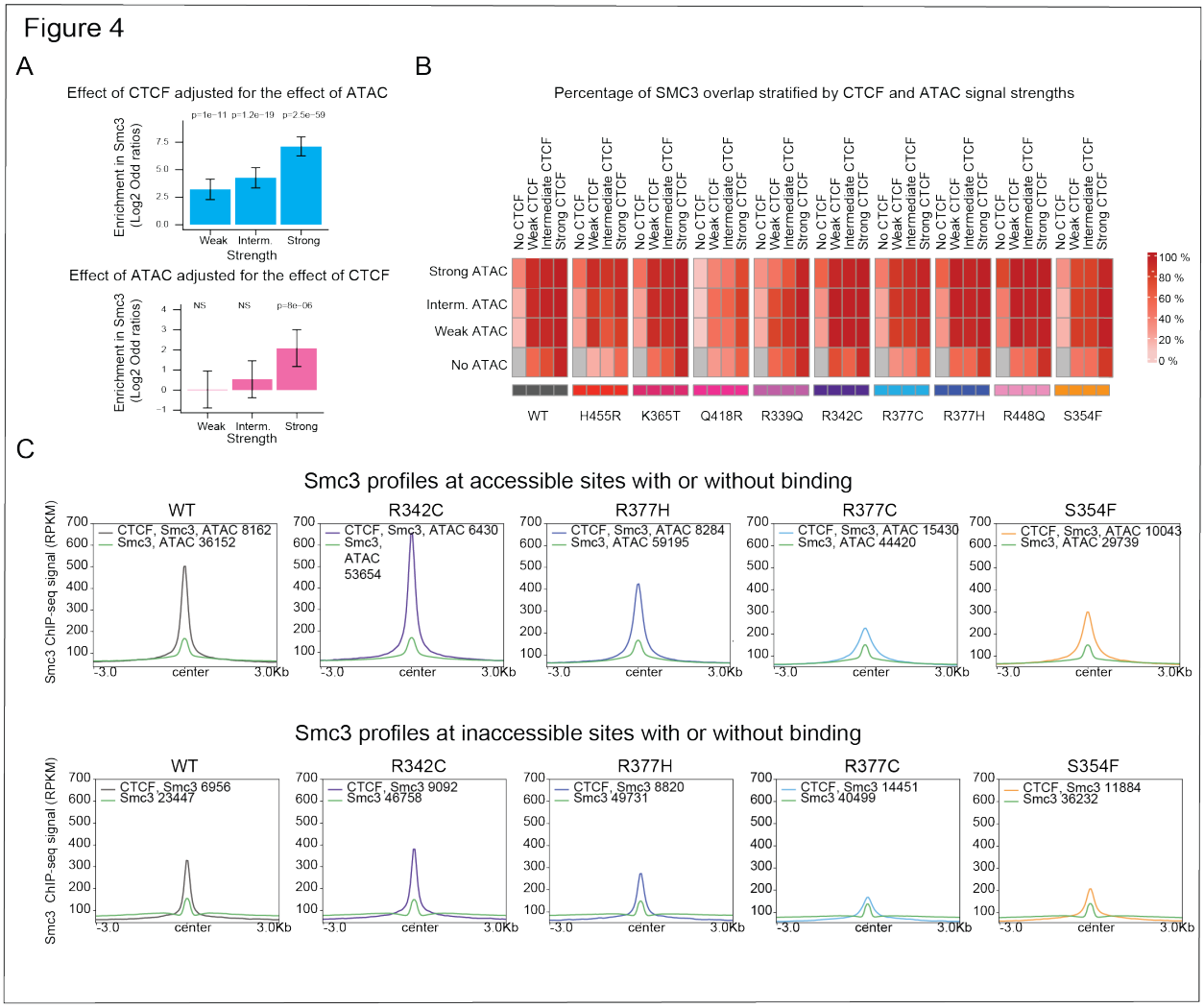


Fig. 4. Effect of CTCF binding and accessibility on Smc3 overlap. (A) Bar graph showing the independent effect of CTCF (top) and ATAC-seq (bottom) signal strength on Smc3 enrichment, estimated using a multivariate logistic model on data from WT cells. CTCF is associated with an enrichment in Smc3 in a dose dependent manner, while only strong accessible sites are associated with a significant enrichment (bottom). The enrichment in Smc3 is estimated using the log2 odd ratios. The error bars correspond to the 95% confidence interval. NS, Not significant. (**B**) Percentage of Smc3 overlap in WT and CTCF mutants, stratified by ATAC-seq and CTCF signal strengths. **(C)** 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 **Extended Data Fig. 4A**.

313 We next compared the specific bound fraction estimated by FRAP, to the bound fraction using
314 FLAG ChIP-seq and ATAC-seq peaks to define all potential binding sites. While we did not find a
315 good correlation when assessing the overall CTCF ChIP-seq bound fraction, we found a strong
316 and significant correlation (R^2=0.6, p=0.008) between the chromatin bound fraction of the mutants
317 detected by FRAP and the fraction of mutant binding at common CTCF sites (**Fig. 3E**). These
318 data suggest that the FRAP bound fraction mostly captures the effect of the mutations on the
319 strong, accessible binding sites. Furthermore, when we analyzed the overall ability of CTCF/
320 cohesin overlapping sites to reduce accessibility, we found a strong correlation with the fraction
321 bound (R^2=0.7, p=0.003), consistent with CTCF functioning in this context in a dosage dependent
322 manner (**Fig. 3E**).
323
324 **The relationship between cohesin and accessibility at CTCF-bound and CTCF-free sites**
325 The ChIP-seq and FRAP data demonstrate that accessibility contributes to CTCF binding, and
326 that CTCF binding stability/residence time is linked to cohesin overlap. To separate out the
327 contribution of CTCF and ATAC strength on Smc3 signals, we performed multivariate modelling
328 to determine (i) the effect of CTCF strength on Smc3 overlap, adjusted for the effect of ATAC
329 signal, and (ii) the effect of ATAC strength on Smc3 overlap, adjusted for the effect of CTCF
330 signal. **Figure 4A** (top) shows that Smc3 enrichment independent of ATAC-seq signal, is strongly
331 associated with CTCF signal strength in a dose-dependent manner. Only strong accessibility
332 significantly influences Smc3 binding independent of CTCF, however, accessibility has much less
333 contribution compared to CTCF binding stability **(Figure 4A**, bottom). Thus, CTCF signal strength
334 plays a dominant role in determining the probability of CTCF/cohesin overlap, consistent with the
335 FRAP and ChIP-data (**Figure 3E**).
336
337 The heatmap in **Figure 4B** and **Table S1**, further shows that, in contrast to CTCF-bound sites,
338 the percentage of Smc3 overlap at CTCF-free sites is increased with increasing accessibility,
339 however signal strength at highly accessible sites is much lower than sites overlapping even weak
340 CTCF-bound sites. The same effect is observed for the mutants, but in some cases Smc3 overlap
341 at CTCF-bound sites is weaker than WT CTCF, in line with their reduced residence times.
342

Figure 5

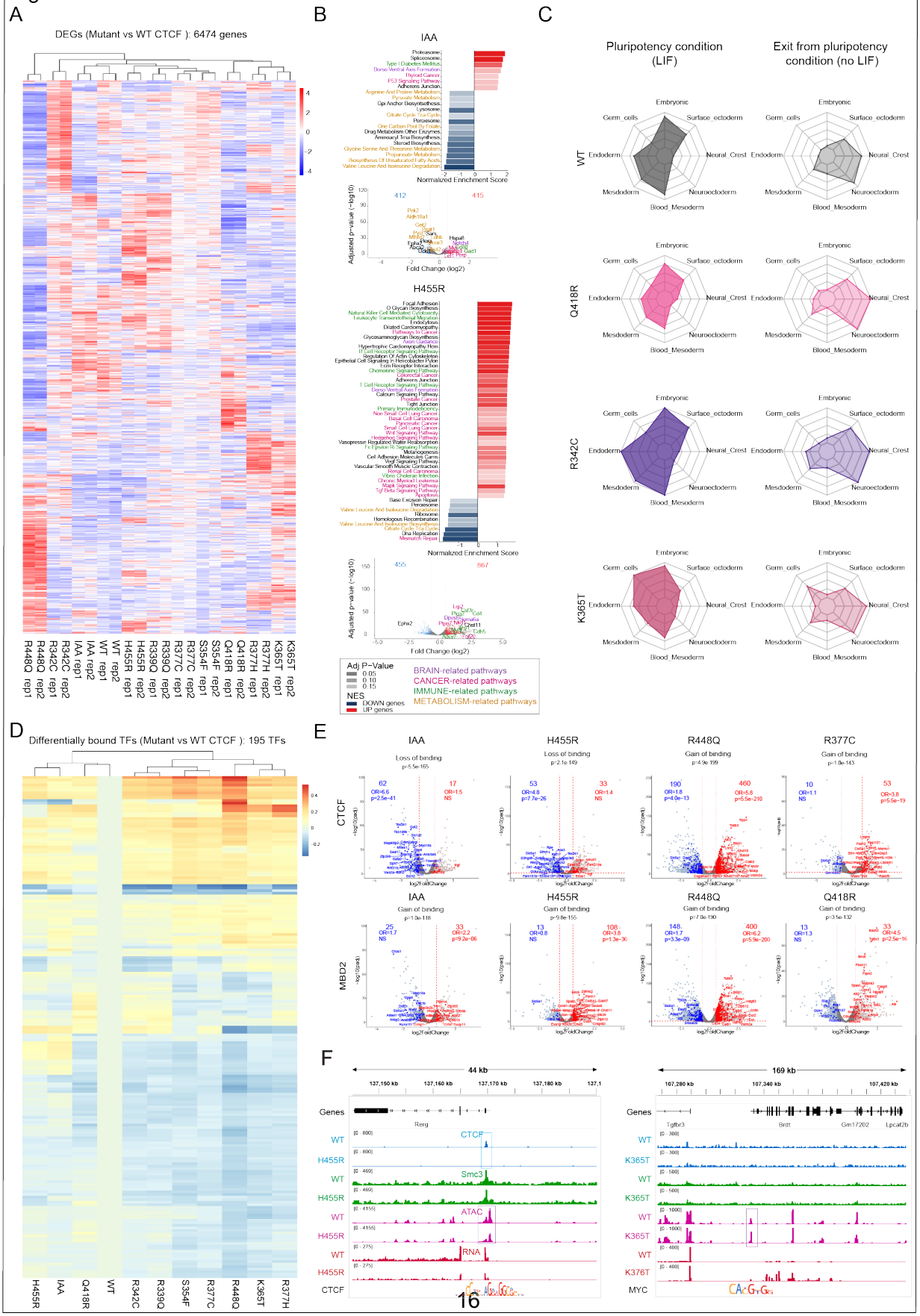

A
DEGs (Mutant vs WT CTCF): 6474 genes
R448Q rep1
R448Q rep2
R342C rep1
R342C rep2
IAA rep1
IAA rep2
WT rep1
WT rep2
H455R rep1
H455R rep2
R339Q rep1
R339Q rep2
R377C rep1
R377C rep2
S354F rep1
S354F rep2
Q418R rep1
Q418R rep2
R377H rep1
R377H rep2
K365T rep1
K365T rep2
B
IAA
Normalized Enrichment Score
Fold Change (log2)
H455R
Normalized Enrichment Score
Fold Change (log2)
Adj P-Value
NES
DOWN genes
UP genes
BRAIN-related pathways
CANCER-related pathways
IMMUNE-related pathways
METABOLISM-related pathways
C
Pluripotency condition (LIF)
Exit from pluripotency condition (no LIF)
WT
Q418R
R342C
K365T
Embryonic
Surface_ectoderm
Neural_Crest
Neuroectoderm
Blood_Mesoderm
Mesododerm
Endoderm
Germ_cells
D
Differentially bound TFs (Mutant vs WT CTCF): 195 TFs
H455R
IAA
Q418R
WT
R342C
R339Q
S354F
R377C
R448Q
K365T
R377H
E
IAA
H455R
R448Q
R377C
CTCF
MBD2
Loss of binding
Gain of binding
F
44 kb
169 kb
Genes
WT
H455R
K365T
K376T
CTCF
Smc3
ATAC
RNA
MYC

Fig. 5: **CTCF mutations alter gene expression, cellular reprogramming 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 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 transgene, H455R. All other mutants are shown in **Extended Data Fig. 5A**. 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**) Radar plots showing the average expression of developmental germ layer genes in WT and mESCs cultured in pluripotency (LIF) and exit from pluripotency (no LIF) conditions. (**D**) 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. (**E**) Volcano plots show examples of two differentially bound TFS (CTCF, MBD2) overlapping the promoters of DEGs in CTCF degraded (IAA) and CTCF mutant expressing cells (ID). The remaining mutant comparisons are shown in **Extended Data Fig. 5B**. 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). (**F**) Examples of altered CTCF binding at the promoter of differentially expressed *Rerg* and predicted differentially bound MYC at the promoter of differentially expressed *Brdt*.

Overall Smc3 profiles at CTCF-bound sites confirmed that CTCF is most effective at blocking cohesin at accessible sites enriched with strong CTCF signals, and they further highlight that each mutant CTCF performs this function with variable ability, depending on their residence time (**Fig. 3E, 4C and Extended Data Fig. 4A**). A comparison of Smc3 profiles at CTCF bound versus unbound sites, demonstrates that other factors such as TFs, contribute very little compared to CTCF, regardless of accessibility and likely because residence times of TFs on chromatin are much shorter than CTCF's [38]. Indeed, at CTCF-free sites, the Smc3 profiles are similar at accessible and inaccessible sites, suggesting that accessibility has little contribution to cohesin binding in the absence of CTCF. A similar trend was observed at enhancers and promoters where accessibility is higher and TFs are known to be enriched (**Extended Data Fig. 4B**). These data suggest that for a given cell type, the subset of strong and accessible CTCF binding sites are predominantly involved in loop extrusion. However, inaccessible CTCF binding sites might still have a role in loop formation. In support of this, a subset of inaccessible sites (16% in the WT) have strong CTCF signals. Furthermore, similar to the dynamic TF-mediated loops previously described between enhancers and promoters [39,40], weaker inaccessible CTCF binding sites could also be involved in more transient, dynamic looping events with a reduced loop lifetime [41,42], especially given that they exhibit an overall stronger Smc3 overlap than TFs.

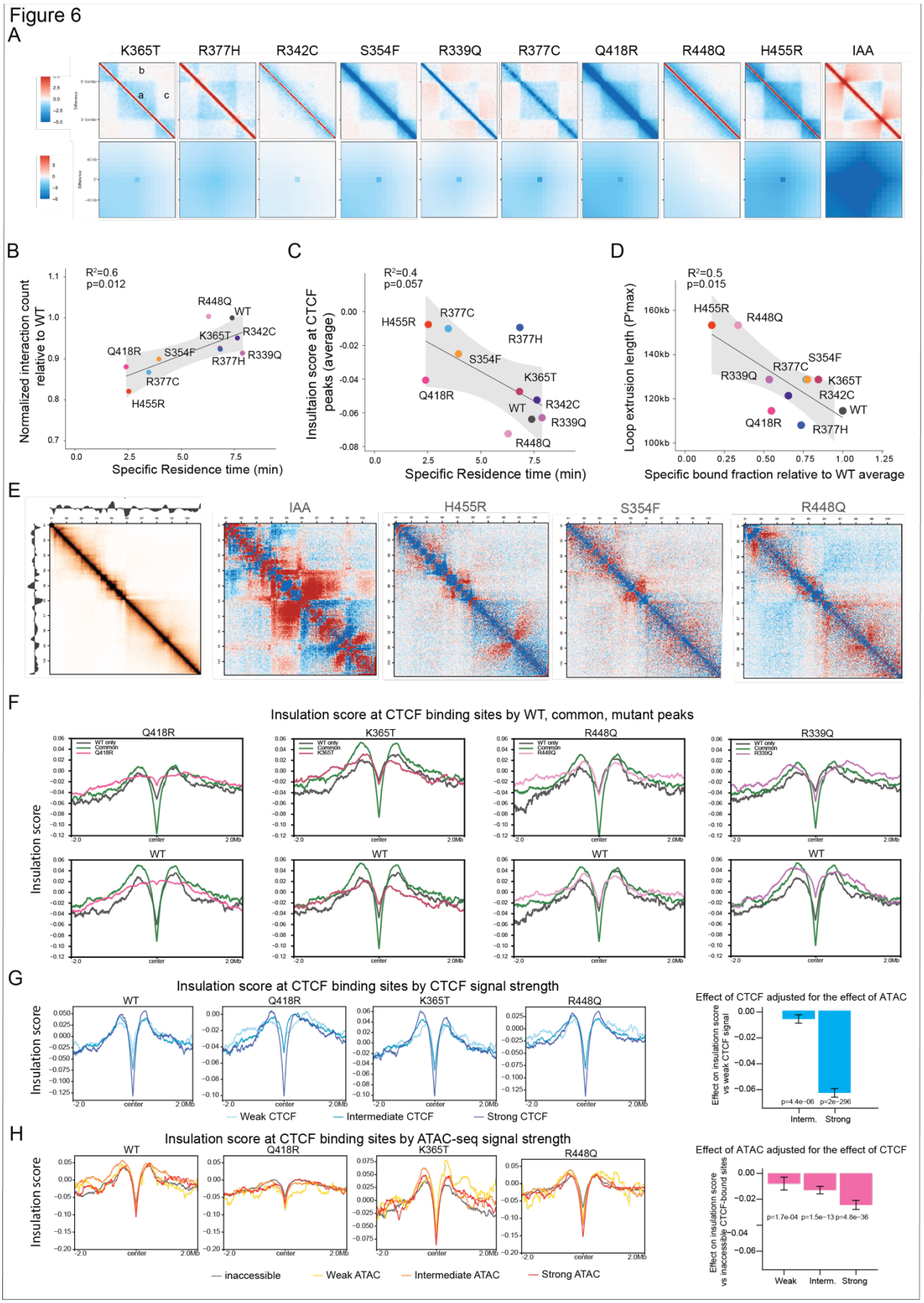
Each CTCF mutation alters gene expression and TF binding in a unique manner.

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

365 was degraded in the absence (IAA) or presence of the WT or mutant CTCF transgenes (ID).
366 Alterations in transcriptional output were determined by comparison with cells expressing WT
367 transgenic CTCF (ID) (**Fig. 5A**). Overall, 6474 genes had altered gene expression across all
368 mutants and including the IAA condition (**Table S2**). Of note is the difference in gene expression
369 changes between R377C and R377H, two distinct mutants of the same residue which give rise to
370 different sets of differentially expressed genes (DEGs).
371
372 Gene set enrichment analysis using the KEGG database [43] showed a strong enrichment in
373 functional pathways related to cancer, brain, immune and metabolic processes (color coded in
374 **Fig. 5B**, with genes outside these pathways represented in black**)** among DEGs across mutants.
375 This finding is compatible with the human diseases associated with CTCF mutations. The top
376 differentially expressed genes of those pathways are shown in volcano plots below with the same
377 color-code. Examples in **Fig. 5B** show changes in pathway and gene expression in the absence
378 of CTCF (IAA condition) and in cells expressing the mutant CTCF transgenes, H455R of ZF7,
379 S354F of ZF4, R448Q of ZF7 and R339Q of ZF3 (ID condition). All other mutants are shown in
380 **Extended Data Fig. 5A**.
381
382 The data in **Figures 5A, B** demonstrate that each mutant displays a distinct profile of gene
383 expression changes linked to alterations in a unique set of functional pathways. Interestingly, cells
384 in which endogenous CTCF is degraded (IAA condition) show fewer gene expression changes
385 compared to mutant expressing cells, except for the R377C mutation (**Extended Data Fig. 5A**).
386 Some of the mutants have alterations in many pathways (R399Q of ZF3), while others highlight
387 changes in only a few (Q418R of ZF6 and R377C of ZF4) as shown in **Fig. 5B and Extended**
388 **Data Fig. 5A**, suggesting that for those mutants, the alteration in gene expression may not
389 accumulate in specific pathways.
390
391 **Mutant specific alterations in transcriptional networks impact reprogramming**
392 The findings from **Figure 5A, B** indicate that each mutant has a distinct transcriptional network,
393 that could potentially affect exit from pluripotency and the ability to be reprogrammed, To
394 investigate, we compared expression profiles of WT and a selection of mutant CTCF expressing
395 cells before and after 72 hours of LIF withdrawal [44]. RNA-seq revealed that each mutant
396 demonstrated a distinct differentiation trajectory compared to WT cells, which are most skewed
397 towards a neuroectodermal fate (**Figure 5C** and **Table S3**). These data highlight the significance
398 of CTCF mutant expression changes in the context of cellular reprogramming and development,

399 findings that provide insight into their disease associated effects, especially in the context of
400 neurodevelopmental disorders.
401
402 **Transcriptional changes correspond to predicted alterations in TF binding at promoters**
403 To investigate whether mutant specific transcriptional changes could be explained by altered TF
404 binding, we performed a footprinting analysis of ATAC-seq data using the TOBIAS pipeline [45] in
405 cells which have endogenous CTCF degraded in the absence (IAA) or presence of WT or mutant
406 transgenes (ID). For this analysis replicates were merged and comparisons made with WT ID
407 cells. Overall, 195 TFs had predicted differential binding across all mutants and including the IAA
408 condition. As a proof of principle, we detected predicted differential binding for CTCF, which we
409 know from our analyses has altered binding in the mutants. Consistent with the RNA-seq
410 analyses, we found that the two R377 mutants give rise to distinct profiles of predicted differential
411 CTCF binding. Numerous other predicted differentially bound TFs (n = 194) were identified as
412 shown in **Fig. 5D** and **Table S4**.
413
414 We overlapped predicted differentially bound TFs with a region of 2kb around the promoters of
415 DEGs in cells expressing different CTCF mutants, to analyze the impact of predicted differential
416 binding of TFs on gene expression (**Fig. 5E**). Although TF binding at gene promoters regulates
417 gene expression, TFs can also bind distal and proximal enhancers to exert transcriptional control.
418 However, since we cannot connect enhancers to their target genes with any certainty, these
419 regulatory elements are excluded from our analysis and we focus only on DEGs with predicted
420 differentially bound TFs at their promoters, where we observed a strong enrichment among the
421 DEGs. It is of note, that the majority of the predicted changes in TF binding at the promoters of
422 DEGs do not overlap CTCF ChIP-seq peaks or predicted differential CTCF binding as shown
423 further below in the direct versus indirect effect analyses. This data suggests that many of the
424 transcriptional changes associated with CTCF mutations could be explained by indirect effects
425 resulting from the disruption of specific TF pathways.
426
427 **Fig. 5E and Extended Data Fig. 5B** highlight the impact of two of the top predicted differentially
428 bound TFs (CTCF and MBD2) at promoters of DEGs in different mutants. These are shown in
429 volcano plots for cells in which endogenous CTCF is degraded (IAA) and in cells expressing
430 different CTCF mutations (ID). The identification of CTCF among the top differentially bound TFs
431 highlights the biological relevance of the footprinting approach. Interestingly, MBD2 is a methyl-
432 sensitive factor that belongs to the NurD chromatin remodeling complex, that has been implicated

Figure 6


A
K365T
R377H
R342C
S354F
R339Q
R377C
Q418R
R448Q
H455R
IAA
B
R²=0.6
p=0.012
Normalized interaction count relative to WT
Specific Residence time (min)
C
R²=0.4
p=0.057
Insultaion score at CTCF peaks (average)
Specific Residence time (min)
D
R²=0.5
p=0.015
Loop extrusion length (P'max)
Specific bound fraction relative to WT average
E
IAA
H455R
S354F
R448Q
F
Insulation score at CTCF binding sites by WT, common, mutant peaks
Insulation score
G
Insulation score at CTCF binding sites by CTCF signal strength
Weak CTCF
Intermediate CTCF
Strong CTCF
Effect of CTCF adjusted for the effect of ATAC
Effect on insulationn score vs weak CTCF signal
p=4.4e−06
p=2e−296
Interm.
Strong
H
Insulation score at CTCF binding sites by ATAC-seq signal strength
inaccessible
Weak ATAC
Intermediate ATAC
Strong ATAC
Effect of ATAC adjusted for the effect of CTCF
Effect on insulationn score vs inaccessible CTCF-bound sites
p=1.7e-04
p=1.5e−13
p=4.8e−36
Weak
Interm.
Strong

433

Fig. 6: 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 **Fig. 6A**) 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 insulation at score at CTCF peaks and the specific residence time detected by FRAP. (**D**) 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. (**E**) 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. (**F**) Profiles show the aggregated insulation score at WT only, common and mutant only binding sites for four mutants. The profiles for the remaining mutants are shown in **Extended Data Fig. 6A.** (**G**) Profiles show the aggregated insulation score at CTCF binding sites with weak, intermediate and strong CTCF signals for three mutants. The profiles for the remaining mutants are shown in **Extended Data Fig. 6B.** The bar graph on the right shows the independent effect across samples of CTCF strength (compared to weak signal) on insulation score. (**H**) Profiles show the aggregated insulation score at CTCF binding sites with no, weak, intermediate and strong ATAC signals for three mutants. The profiles for the remaining mutants are shown in **Extended Data Fig. 6C.** The bar graph on the right shows the independent effect across samples of ATAC strength (compared to inaccessible sites) on insulation score. The estimates for the bar graphs in (**G**) and (**H**) were obtained using a mixed multivariate model. The error bars correspond to the 95th confidence intervals.

in cancer and neurodevelopmental disorders. NurD is also required for the exit from pluripotency of ESCs and their lineage commitment [46-49]. Consistent with our finding that each 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. Both CTCF and MBD2 exhibit an overall activator profile with a positive association between binding and expression (**Fig. 5E and Extended Data Fig. 5B**). However, for some mutants, a weaker but significant negative association was also observed suggesting that both factors can act as activators or repressors depending on the cell-specific genomic context, consistent with previous reports[48,50].

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 **Fig. 5F.** 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 the promoter of this gene. 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 and their impact on developmental processes, that are particularly relevant to brain disorders.

454

455 **Chromatin interactivity and insulation are linked to mutant binding property**

456 To determine how CTCF mutants impact chromatin interactions we performed Hi-C. We found
457 that each mutant has a distinct interaction profile and exhibits (i) variable loss of intra TAD
458 interactions and boundary strength as well as (ii) aggregated loop strength. The IAA condition, in
459 which CTCF is degraded shows a loss of intra TAD interactions and a gain of inter TAD
460 interactions as well as the most reduction in aggregated loop strength (**Fig. 6A**), while mutants
461 display their own unique effects on TAD interactions and loop strength.

462

463 Consistent with our finding that cohesin overlap is correlated with CTCF mutant residence time
464 as determined by FRAP analysis (**Fig. 3E**), we found that residence time is also associated with
465 chromatin interactivity and insulation score (**Fig. 6B, C**).The correlation with the latter was modest
466 and did not reach significance, probably because of the lower Hi-C resolution (10 kb) compared
467 to ChIP-seq. Gassler et al., showed that the first derivative of the relative contact probability curve
468 after log-log transformation (P'max) provides an estimate of the average loop size and cohesin
469 density [51]. Here we show that the loop extrusion length (P'max) for each mutant is associated with
470 the fraction that is bound to chromatin (**Fig. 6D**), and the longest loops are associated with
471 mutants that have the lowest bound fraction indicating a lower density of CTCF-cohesin anchors
472 in these mutants (**Fig. 3B**). The variability in the interaction profile of each mutant can be clearly
473 seen across a 10Mb region of chromosome (**Fig. 6E**).

474

475 To further examine the impact of each mutant on chromatin organization, we analyzed the
476 aggregated insulation score at WT only, common and mutant only binding sites using a 2Mb
477 window centered on CTCF peak summits (**Fig. 6F and Extended Data Fig. 6A**). The insulation
478 scores were generated from Hi-C data at a 10kb resolution. This analysis revealed that the strong,
479 common accessible binding sites with the highest cohesin overlap, have the highest insulation
480 (lowest insulation score) compared to WT only and mutant only sites. Whether insulation scores
481 at WT only and mutant only sites are lower or higher than each other appears to depend on the
482 residence time and cohesin overlap of each mutant, in line with what we observed in **Fig. 3C, E**
483 **and 6C**. These studies further demonstrate that imaging and molecular analyses can be
484 functionally integrated to provide new insight into mutant specific effects.

485

486 Since WT and mutant only sites are associated with weaker CTCF peaks and lower accessibility,
487 we examined the insulation score at CTCF binding sites, stratified by CTCF and ATAC-seq

488 strength. The profiles in **Fig. 6G** and **H** indicate that stronger CTCF signal is associated with

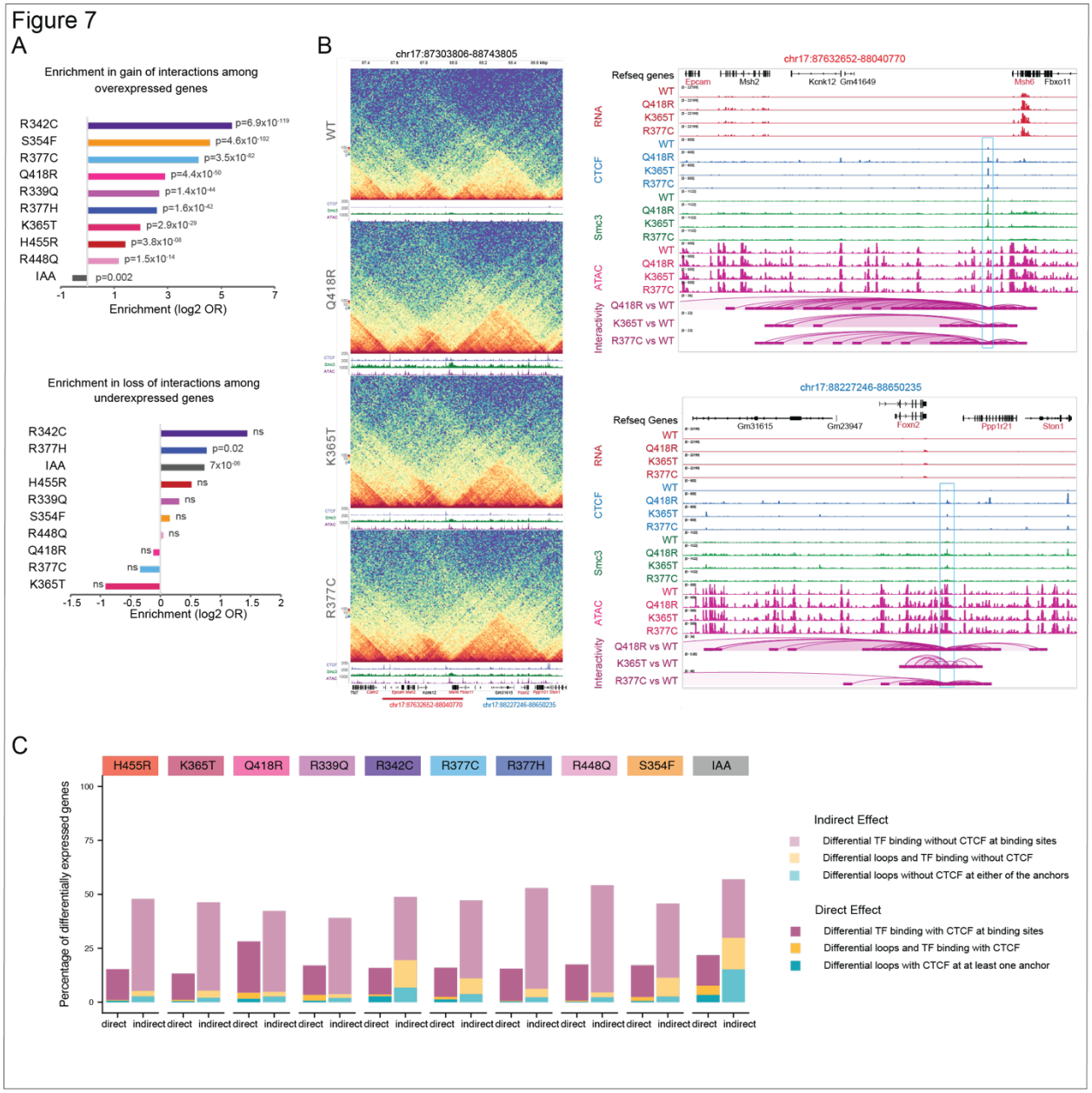


Fig. 7: Changes in gene expression are linked to changes in chromatin interactivity but to a lesser extent than changes in TF binding at gene promoters: (**A**) Bar graphs showing the enrichment in over-expressed (top) or under-expressed (bottom) genes among gained (top) or lost (bottom) loops with one anchor overlapping the promoter of the DEGs in IAA and CTCF mutants. (**B**) 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. **C**) Bar graph showing the percentage of DEGs associated with differential TF binding, differential loops or both, stratified by direct and indirect effect. Direct effect was defined as differential TF binding with a predicted differential CTCF binding or ChIP-seq peak at the gene promoter and/or differential loop with a CTCF peak at either anchor.

489 stronger insulation, while accessibility has a more modest impact. To quantify their independent
490 effects, we performed a mixed multivariate model, the results of which are consistent with the
491 analysis of Smc3 overlap shown in **Fig. 4A**.
492
493 **Changes in gene expression are linked to changes in chromatin interactivity**
494 To determine if changes in gene expression are linked to altered chromatin folding, we analyzed
495 interactions involving the promoters of genes that were differentially expressed. We found that
496 the promoters of overexpressed genes were enriched in gain of chromatin loops (**Fig. 7A**). This
497 enrichment was not observed among the under-expressed genes, suggesting that under-
498 expression observed in mESC expressing mutant CTCF might be an indirect effect, independent
499 of CTCF's function in chromatin organization, while overexpression might reflect a direct CTCF-
500 mediated effect on loop extrusion as depicted in the examples showing the gain of interactions
501 from CTCF binding sites toward the promoters of overexpressed genes, including *Msh6*, *Epcam*,
502 *Foxn2* and *Cox7a2l* (**Fig. 7B** and **Extended Data Fig. 7A)**. Interestingly, all these genes are
503 implicated in tumorigenesis. Indeed, MSH6 is a mismatch repair factor and its mutation is
504 associated with Lynch syndrome and cancer susceptibility [52]. Overexpression of MSH6 might also
505 exert an oncogenic function in glioblastoma by promoting cell proliferation [53]. *Epcam* encodes for
506 a membrane glycoprotein involved in epithelial cell adhesion and has been described as both an
507 oncogene and tumor suppressor depending on the microenvironment [54]. FOXN2 acts as a tumor
508 suppressor in multiple cancers, including breast, lung and liver [55-57]. Overexpression of COX7A2L
509 might promote hypoxia tolerance in breast and endometrial cancer [58].
510
511 **Most of the change in expression are the result of indirect effects.**
512 Our data showed that most CTCF mutations have more impact on gene expression than the IAA
513 condition. Indeed, CTCF degradation has been shown to induce limited gene expression changes
514 [3,8]. To ask whether this could be explained by indirect effects which depend on the cell-specific
515 transcriptomic context, we estimated the direct versus indirect effect of CTCF mutations on gene
516 expression. Since CTCF can act as a traditional transcription and chromatin regulator [50], we
517 defined indirect CTCF effects as (i) differentially expressed genes (DEGs) with a predicted
518 differentially bound TF at gene promoters not overlapping a CTCF ChIP-seq peak, or a
519 differentially bound CTCF site, and (ii) DEGs with differential looping without CTCF binding at any
520 loop anchors.
521

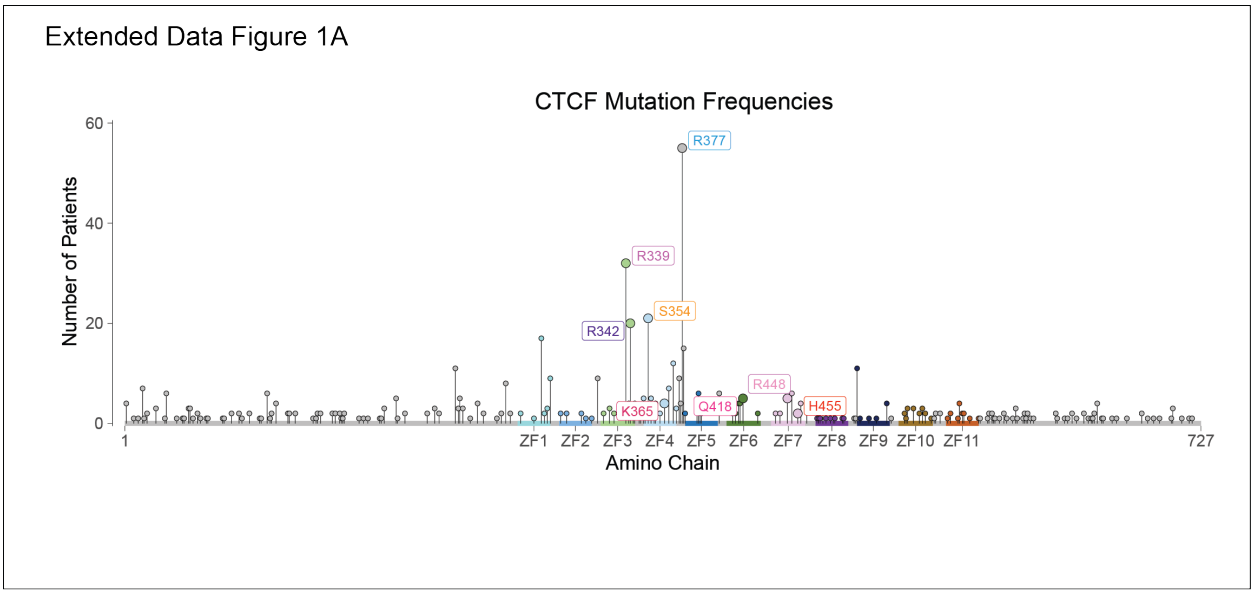
522 **Figure 7C** shows that the indirect effects account for approximately twice as many DEGs
523 compared to the direct effects, except for Q418R. This mutant shows a weak residence time but
524 a high number of peaks supporting the hypothesis that unstable peaks might be involved in
525 dynamic and functional looping. In addition, a higher percentage of DEGs can be explained by
526 changes in TF or CTCF binding at gene promoters, rather than by changes in chromatin
527 interactivity in all the mutants. For the IAA condition, the contribution of differential loops in
528 expression changes is higher. A similar result was observed when segregating up- and down--
529 regulated genes but with a higher proportion of direct versus indirect effects in the IAA condition
530 for down-regulated genes consistent with loss of loops upon CTCF degradation (**Extended data**
531 **Fig. 7B**). These findings suggest that although CTCF mutations induce changes in chromatin
532 looping, the primary mechanism driving alterations in gene expression is loop-independent and
533 is instead mediated by changes in transcription factor binding. This indicates that cell-specific
534 transcriptional programs may modulate the effects of CTCF mutations.
535
536 **DISCUSSION**
537 Using a combination of imaging, structural and molecular approaches we have examined the
538 impact of nine, high frequency cancer associated CTCF mutations in ZFs that make contact with
539 the core consensus binding motif. Collectively the graded mutant perturbations offer a platform
540 for investigating the molecular and epigenetic features that govern CTCF's choice of binding sites,
541 and the degree of stability with which CTCF binds at individual locations and is able to function
542 as a chromatin organizer. Since the mutants allow us to distinguish direct from indirect CTCF-
543 mediated effects and cause from consequence, they provide a deeper understanding of CTCF's
544 downstream impact on chromatin and gene expression.
545
546 Our studies reveal that the specificity of interaction between the ZFs in CTCFs binding domain
547 and its target DNA sequence is a key determinant of attachment. However, if the same sequence
548 is found in both accessible and inaccessible chromatin, CTCF will preferentially bind the
549 accessible sites as highlighted by the fact that mutants bind common accessible sites in
550 preference to less accessible WT only sites. Thus, sites in open chromatin dominate the search
551 space. This is underscored by the strong correlation between the fraction of chromatin bound
552 mutants detected by FRAP, and mutant binding at common, accessible CTCF sites. These
553 findings are consistent with studies from Dirk Schubeler's group showing that CTCF can bind both
554 accessible and inaccessible sites, but that the latter are less likely to be involved in insulation [59,60].
555 We also observed a good correlation between each mutant's residence time and CTCF-cohesin

556 overlap, suggesting a model in which binding stability is important for blocking cohesin movement
557 on chromatin. A similar, but more modest correlation was found between residence time and
558 chromatin interactivity as well as insulation score, linking binding stability with cohesin overlap as
559 well as loop and boundary formation.
560
561 By dissecting out the contribution of CTCF signal strength and accessibility to Smc3 overlap, we
562 were able to reveal the dominant effect of CTCF. Smc3 overlap at accessible and inaccessible
563 CTCF-bound sites is stronger than overlap at accessible, CTCF-free sites, including enhancers
564 and promoters, likely because TFs have a short residence time compared to CTCF [38]. Our data
565 demonstrate that strong accessible CTCF bound sites are associated with higher insulation,
566 consistent with higher Smc3 stalling and more stable loops. Previous studies showed that when
567 RAD21 (a component of the cohesin complex) levels are decreased upon degradation, CTCF-
568 mediated loop lifetime is reduced [41,42], suggesting that inaccessible CTCF-bound sites, where we
569 observed weaker cohesin signals, might participate in more transient looping events similar to the
570 CTCF-independent and Smc3-dependent dynamic loops described between distal enhancers
571 and promoters [39,40].
572
573 In principle, these features could have been uncovered by analyzing WT CTCF in the context of
574 cohesin overlap and accessibility. However, collectively the graded effects of the mutants act as
575 perturbations that lend support to our models, underscoring how different aspects of CTCFs
576 binding properties contribute. Moreover, they reveal that each mutant's search for binding sites
577 can be decoupled from its chromatin bound residence time, which reflects binding stability. For
578 example, the Q418R mutant binds an intermediate number of sites but at these sites it is has a
579 very low residence time and is less competent at blocking cohesin. In contrast, R448Q binds
580 fewer sites with a similar residence time compared to WT CTCF, and it is slightly better at blocking
581 cohesin than WT protein. Furthermore, H455R and R448Q, which have very different chromatin
582 residence times, have similarly low chromatin bound fractions, correlated to their P'-max. The
583 increased loop length of both mutants can be explained by the unobstructed passage of cohesin
584 past what would normally be a bound CTCF site. Uncoupling of residence time and fraction bound
585 was previously demonstrated for a mutant form of CTCF with ZF8 deleted [61].
586
587 The crosstalk between accessibility and CTCF binding goes in both directions, such that
588 accessibility affects binding and binding affects accessibility. Binding at accessible sites, narrows
589 and reduces the ATAC-seq peak. Evidence for this being a CTCF driven effect comes from

590 analysis of the CTCF mutants. Those with low fraction bound (H455R and Q418R) are unable to
591 function in this capacity, while the R342C mutant has a stronger effect compared to WT CTCF.
592 We speculate that these differences reflect decreased accessibility at the bases of CTCF-Smc3-
593 mediated loops, because in the absence of Smc3, when loops cannot form there is no reduction
594 in accessibility.
595
596 While loss or gain of CTCF binding at promoters can account for direct mutant specific effects on
597 accessibility and gene expression, we could only demonstrate a direct effect of gained CTCF
598 binding associated with gained loops involving the promoters of over-expressed genes. Under-
599 expression could, therefore, more often reflect indirect effects of CTCF binding alterations.
600 Indeed, all mutants have their own unique indirect impact in globally changing accessibility at sites
601 where CTCF is not bound. This effect links changes in TF footprinting with altered transcriptional
602 output, explaining most of the mutant specific effects that we observed. We also identified a
603 higher proportion of DEGs mediated by changes in TF or CTCF binding at the gene promoters
604 than by changes in chromatin looping, suggesting that the primary mechanism driving alterations
605 in gene expression might be loop independent. However, we cannot rule out that our estimation
606 of loop-dependent expression changes might be underestimated due to the lower resolution of
607 Hi-C data. Furthermore, the impact of each mutant on TF expression and altered TF binding at
608 the promoters of differentially expressed genes is less likely to give rise to detectable changes in
609 interactivity by Hi-C, consistent with our finding that Smc3 signals at CTCF-free sites are far
610 weaker than at CTCF-bound sites, and thus likely to be involved in weaker, more dynamic loops.
611
612 Aside from the R377C mutant, we observed fewer changes in gene expression in IAA treated
613 cells compared to cells expressing CTCF mutants, which could be explained by the predominant
614 contribution of indirect changes mediated by TF binding on the mutant-specific transcriptomic
615 profile. Although the mutants all uniquely affect gene expression pathways, we detected
616 enrichment in pathways affecting the brain, immune system, cancer and metabolism, which is
617 consistent with the clinical setting in which the mutations are found, namely cancer and brain
618 disorders. The added effect on immune and metabolic pathways provides insight into other
619 changes that could occur in CTCF mutant-mediated diseases. Furthermore, we showed that
620 CTCF mutations might also affect the trajectory of exit from pluripotency. Indeed, while upon LIF
621 withdrawal both WT and CTCF mutant cells tend to acquire a neuroectodermal expression profile,
622 we observed mutant-specific transcriptional signatures of other germ layers, suggesting a more
623 heterogenous differentiation trajectory, which could be confirmed in future works by scRNA-seq.

624 This finding provides a potential mechanism for CTCF mutation-associated neurodevelopmental
625 disorders. Indeed, while CTCF has been involved in neuro-progenitor maintenance and survival[62],
626 our analyses suggest that CTCF mutations might also alter brain development by affecting germ
627 layer commitment.
628
629 Taken together the binding domain mutants we have analyzed here, provide a new appreciation
630 of CTCF's bidirectional relationship with chromatin, its ability to bind and function in different
631 contexts as well as the potential impact of each mutation in clinical settings. Furthermore, our
632 analyses provide a better understanding of how any genetic or epigenetic disorder that alters the
633 landscape of chromatin can in turn impact CTCF binding and function.
634
635
636 **EXTENDED DATA FIGURES AND SUPPLEMENTARY TABLE:**
637
638 **Supplementary Table 1:** Number and percentage of Smc3 peaks overlapping a CTCF or ATAC-
639 seq peak, stratified by CTCF and ATAC signal strength.
640
641 **Supplementary Table 2:** Differentially expressed genes (DEGs) identified using DE-seq. DEGs
642 were defined as adjusted genes with p-value<0.05 and absolute log2 fold change $\geq$ log2(1.5)
643 compared to WT.
644
645 **Supplementary Table 3:** Expression of germ layer genes (TPM) in LIF and no LIF condition.
646
647 **Supplementary Table 4:** Differentially bound transcription factors predicted by footprinting
648 analyses.

649



Extended Data Fig. 1A. Prevalence of missense CTCF mutations on ZF3-ZF7 in human cancers. Schematic representation of CTCF showing the location and frequency of each mutation under investigation.

650

651

652

Extended Data Figure 1B

Mutations reported in > than 2 samples	Zinc Finger	No. of samples	Associated cancers	Associated neurological disorder
Mutations in phosphate contacts (positions -5, 4th position of the linker between adjacent ZFs)				
R377C*	5	27	Breast, Colorectal, Endometrial, Haematopoietic And Lymphoid, Skin Cancer	ASD, CRD
R377H*	5	26	Breast, Colorectal, Endometrial, Esophagogastric, Haematopoietic And Lymphoid, Urinary Tract	ASD, CRD
R320C	3	6	Colorectal, Endometrial, Prostate Cancer	
R320H	3	3	Colorectal, Esophagogastric, Head And Neck Cancer	
K349E	4	2	Endometrial Cancer	
R377L	5	2	Breast, Esophagogastric Cancer	
Y358C	4	2	Colorectal, Endometrial Cancer	
Mutations in the amino acid that make direct contact with DNA (positions -1, +2, +3 and +6)				
R339W	3	20	Colorectal, Endometrial, Haematopoietic And Lymphoid, Renal Cancer	
R339Q*	3	12	Endometrial, Esophagogastric, Haematopoietic And Lymphoid Cancer	CRD,ASD,DD
Q418R*	6	5	Embryonal Tumor, Glioma, Hepatobiliary Cancer	
R448Q*	7	5	Endometrial, Haematopoietic And Lymphoid, Renal Cancer	CRD,ASD,DD
K365T*	4	4	Endometrial Cancer	
R368C	4	4	Colorectal, Esophagogastric, Haematopoietic And Lymphoid Cancer	DD,NDD,CRD,ASD
D451H	7	3	Skin, Urinary Tract Cancer	
D451V	7	3	Haematopoietic And Lymphoid Cancer	
R368H	4	3	Breast, Ovary, Prostate Cancer	CRD,ASD,DD
E362K	4	2	Haematopoietic And Lymphoid Cancer	
D390N	5	2	Haematopoietic And Lymphoid Cancer	CRD,ASD
Mutations in the amino acids in the other regions of the zinc finger				
S354F*	4	21	Breast, Endometrial, Esophagogastric, Head And Neck, Hepatobiliary, Skin, Urinary Tract Cancer	
P378L	5	10	Endometrial, Haematopoietic And Lymphoid Cancer	NDD, CRD, DD
S388N	5	6	Haematopoietic And Lymphoid, Hepatobiliary, Urinary Tract Cancer	
S402L	5	6	Breast Cancer	
P378A	5	5	Breast, Urinary Tract Cancer	
F351L	4	5	Breast, Urinary Tract Cancer	
G375V	4	5	Endometrial, Haematopoietic And Lymphoid Cancer	
A359T	4	4	Colorectal, Ovary, Prostate Cancer	
A387V	5	4	Breast, Endometrial Cancer	
R415W	6	4	Colorectal, Endometrial, Penile Cancer	
F416L	6	4	Endometrial, Glioma Cancer	
E376K	4	4	Breast, Haematopoietic And Lymphoid, Urinary Tract Cancer	
K352N	4	3	Endometrial Cancer	
D328N	3	3	Colorectal, Endometrial Cancer	
S461F	7	3	Skin Cancer	
E348K	3	3	Colorectal, Haematopoietic And Lymphoid Cancer	
F379I	5	2	Haematopoietic And Lymphoid Cancer	
P440S	7	2	Skin Cancer	
I411V	6	2	Urinary Tract Cancer	
H413Y	6	2	Colorectal Cancer	
D443G	7	2	Hepatobiliary Cancer	
F331Y	3	2	Haematopoietic And Lymphoid Cancer	
R341C	3	2	Colorectal, Skin Cancer	
R341H	3	2	Head And Neck Cancer	
Q428E	6	2	Haematopoietic And Lymphoid Cancer	
K344E	3	2	Breast, Haematopoietic And Lymphoid Cancer	
G375A	4	2	Haematopoietic And Lymphoid, Urinary Tract Cancer	
G375R	4	2	Head And Neck, Lung Cancer	
Mutations in boundary residues (position -2 and +9)				
R342C*	3	13	Breast, Endometrial, Esophagogastric, Haematopoietic And Lymphoid, Skin, Urinary Tract Cancer	NDD,CRD,DD
R371H	4	8	Colorectal, Endometrial, Haematopoietic And Lymphoid, Lung Cancer	
R342H	3	7	Endometrial, Haematopoietic And Lymphoid Cancer	CRD
R371C	4	4	Endometrial, Skin, Urinary Tract Cancer	
R457Q	7	4	Colorectal, Esophagogastric Cancer	
R389S	5	2	Breast, Skin Cancer	
Mutations in the amino acid that co-ordinate the zinc (position -10, -7, +7, +11 or +12 depending on ZF)				
C356R	4	3	Esophagogastric Cancer	
H373L	4	3	Endometrial Cancer	NDD,ASD
H455R*	7	2	Haematopoietic And Lymphoid Cancer	
C324Y	3	2	Haematopoietic And Lymphoid Cancer	
C356F	4	2	Haematopoietic And Lymphoid, Head And Neck Cancer	
H345P	3	2	Breast Cancer	
H345R	3	2	Colorectal, Prostate Cancer	

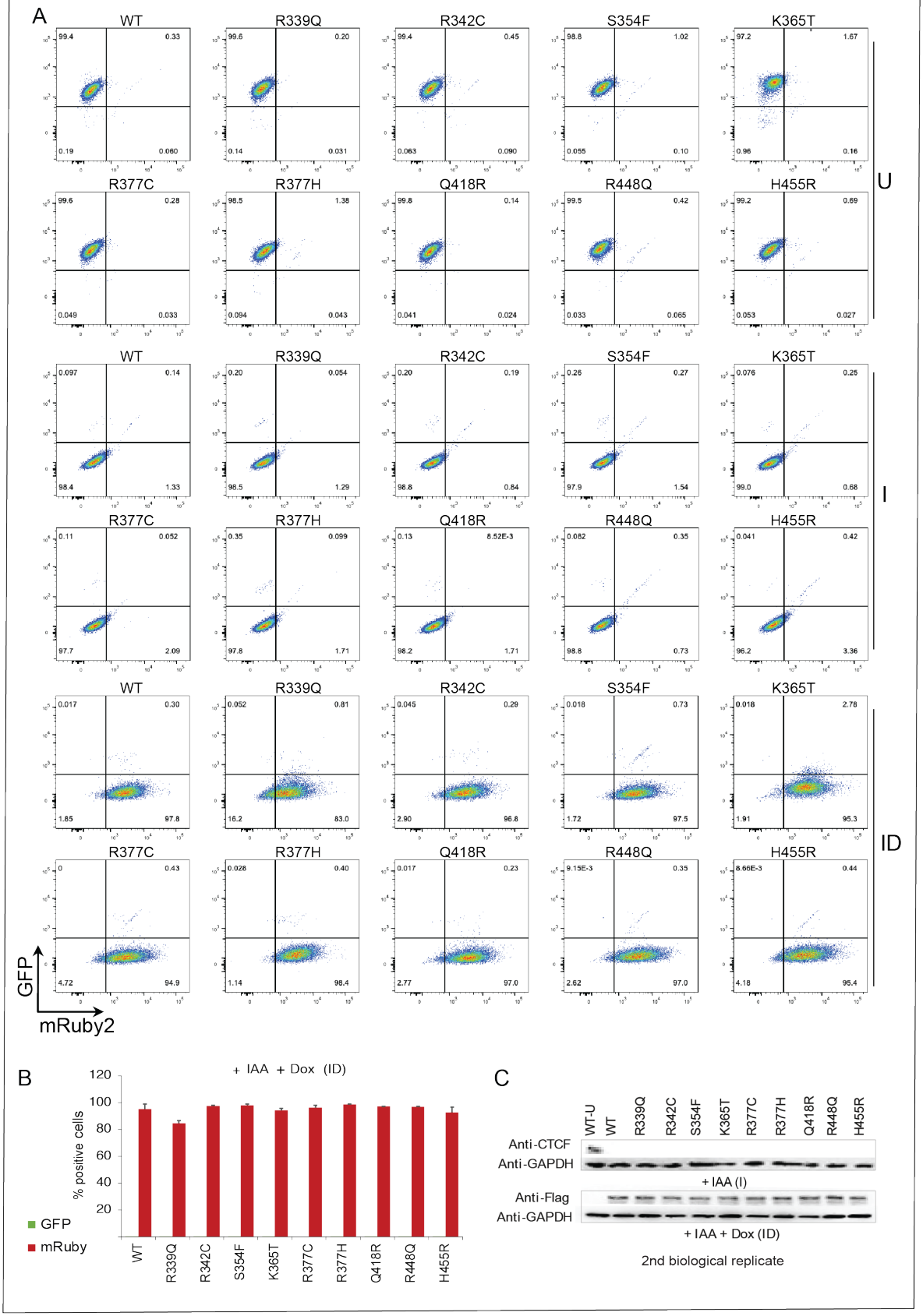
653

654

655

Extended Data Fig. 1B: Table of CTCF mutations. Table showing CTCF mutations identified from cBioPortal[27,29] and COSMIC (http://www.sanger.ac.uk/genetics/CGP/cosmic/). We selected a subset of mutations that occur at high frequency in cancer patients, focusing on those found in ZFs that make contact with the 12-15bp consensus CTCF binding motif. These are highlighted with an asterisk and color coded as in **Fig. 1D.** Mutations analyzed included those found 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: 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. The table identifies which zinc finger each mutation is found in, the number of patients it was identified in and the cancers each is associated with. Mutations were annotated for neurodevelopmental or neurological disorders using the CTCF variant catalog published in Price et al. [4]. ASD, autism spectrum disorder; CRD, CTCF related neurodevelopmental disorder; DD developmental disorder; NDD, neurodevelopmental disorder.

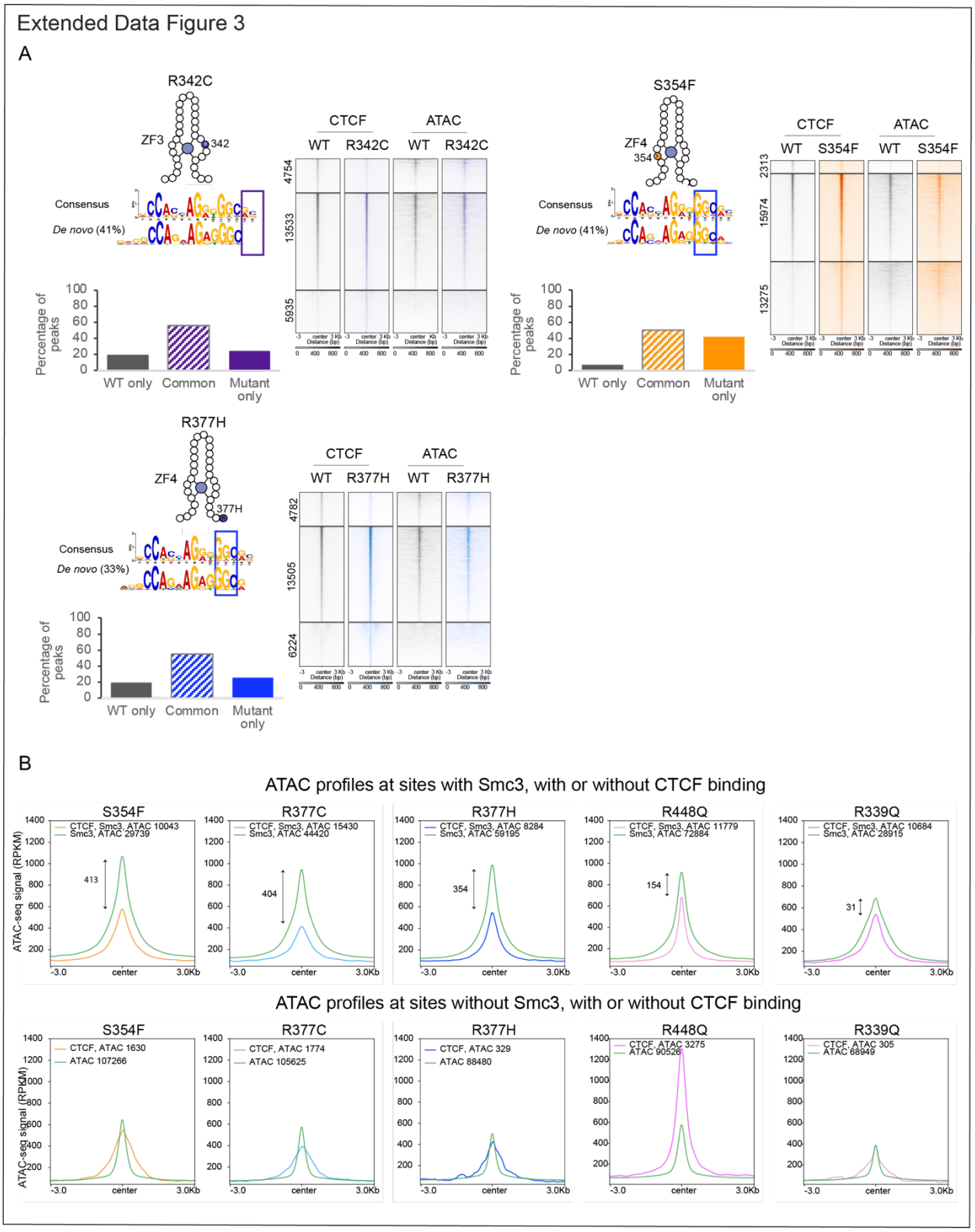
Extended Data Figure 2



657

Extended Data Fig. 2: **Flow cytometry analysis of mutant expression**. (**A**) Flow cytometry showing the level of GFP (endogenous CTCF) and mRuby (transgenic CTCF) in the IAA and Dox treated cells (ID). (**B**) Bar graph showing the percentage of GFP (endogenous CTCF) and mRuby2 (transgene WT and mutant CTCF) positive cells after treatment with IAA and Dox (ID). The presence of the IAA leads to degradation of GFP labelled endogenous CTCF. The error bars represent the standard deviation between the 2 replicates. (**C**) 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. The first replicate is shown in **Fig. 1**.

658

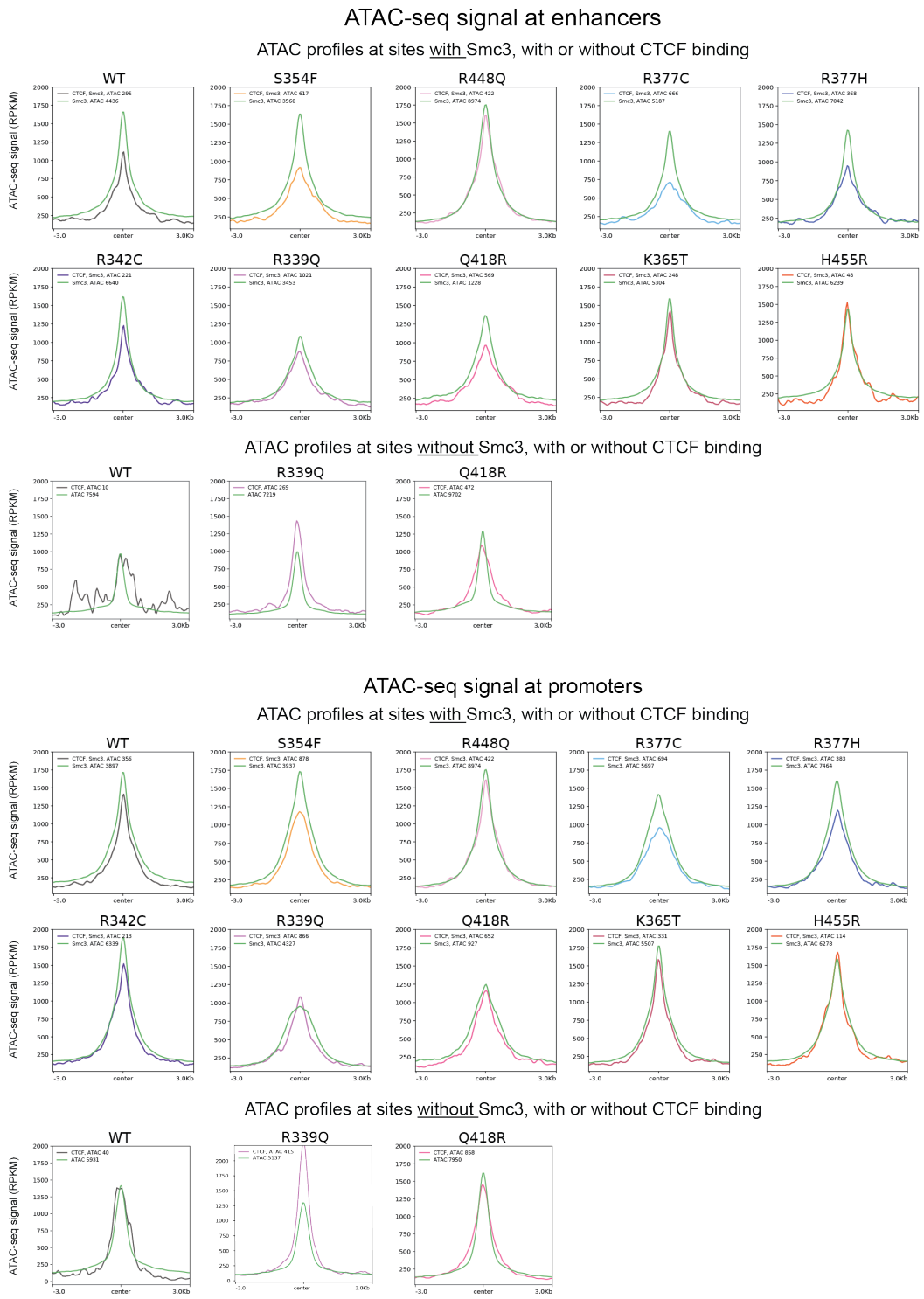

Extended Data Figure 3
A
R342C
ZF3
342
Consensus
De novo (41%)
CTCF
ATAC
WT
R342C
4754
13533
5935
Percentage of peaks
WT only
Common
Mutant only
S354F
ZF4
354
WT
S354F
2313
15974
13275
R377H
377H
4782
13505
6224
B
ATAC profiles at sites with Smc3, with or without CTCF binding
S354F
CTCF, Smc3, ATAC 10043
Smc3, ATAC 29739
413
R377C
CTCF, Smc3, ATAC 15430
Smc3, ATAC 44420
404
R377H
CTCF, Smc3, ATAC 8284
Smc3, ATAC 59195
354
R448Q
CTCF, Smc3, ATAC 11779
Smc3, ATAC 72884
154
R339Q
CTCF, Smc3, ATAC 10684
Smc3, ATAC 28915
31
ATAC-seq signal (RPKM)
-3.0
center
3.0Kb
ATAC profiles at sites without Smc3, with or without CTCF binding
CTCF, ATAC 1630
ATAC 107266
CTCF, ATAC 1774
ATAC 105625
CTCF, ATAC 329
ATAC 88480
CTCF, ATAC 3275
ATAC 90526
CTCF, ATAC 305
ATAC 68949

Extended Data Fig. 3: **CTCF mutations have unique chromatin binding profiles:** Scheme showing the locations of the different CTCF mutations within the ZF. The consensus CTCF motif highlights the triplet to which each mutant zinc finger binds (top) as well as the motif identified by MEME for the *de novo* mutant only binding sites (bottom). Each graph shows the percentage of WT and mutant CTCF binding at WT only, common and mutant only sites and heatmaps show the profile of CTCF binding and ATAC-seq signal for WT only, common and mutant only sites. The remaining mutants are shown in **Fig. 3**. (**B**) Profiles of ATAC-seq in WT and mutant CTCF expressing cells are shown for sites, with (color-coded for each mutant) or without (green) CTCF binding in the presence or absence of Smc3. The remaining mutants are shown in **Fig. 2B**.

661

Extended Data Figure 3

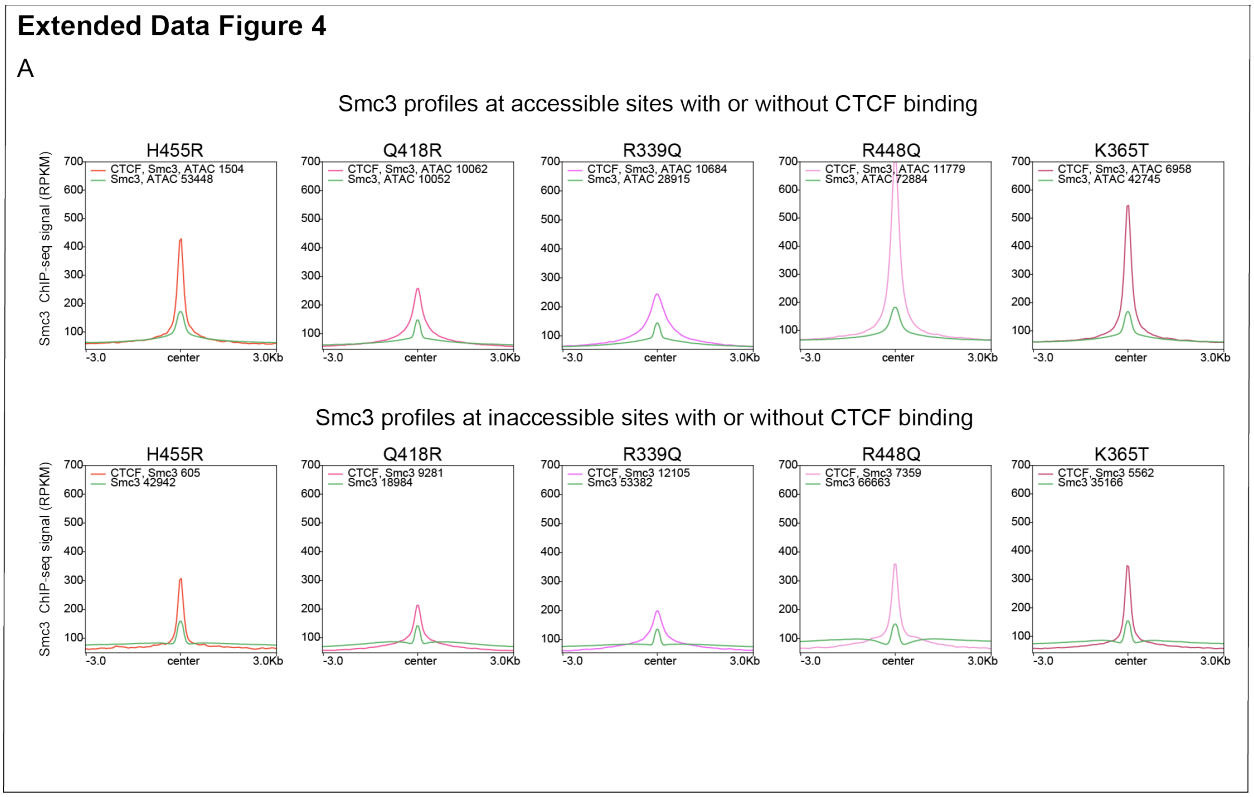
C



662

663

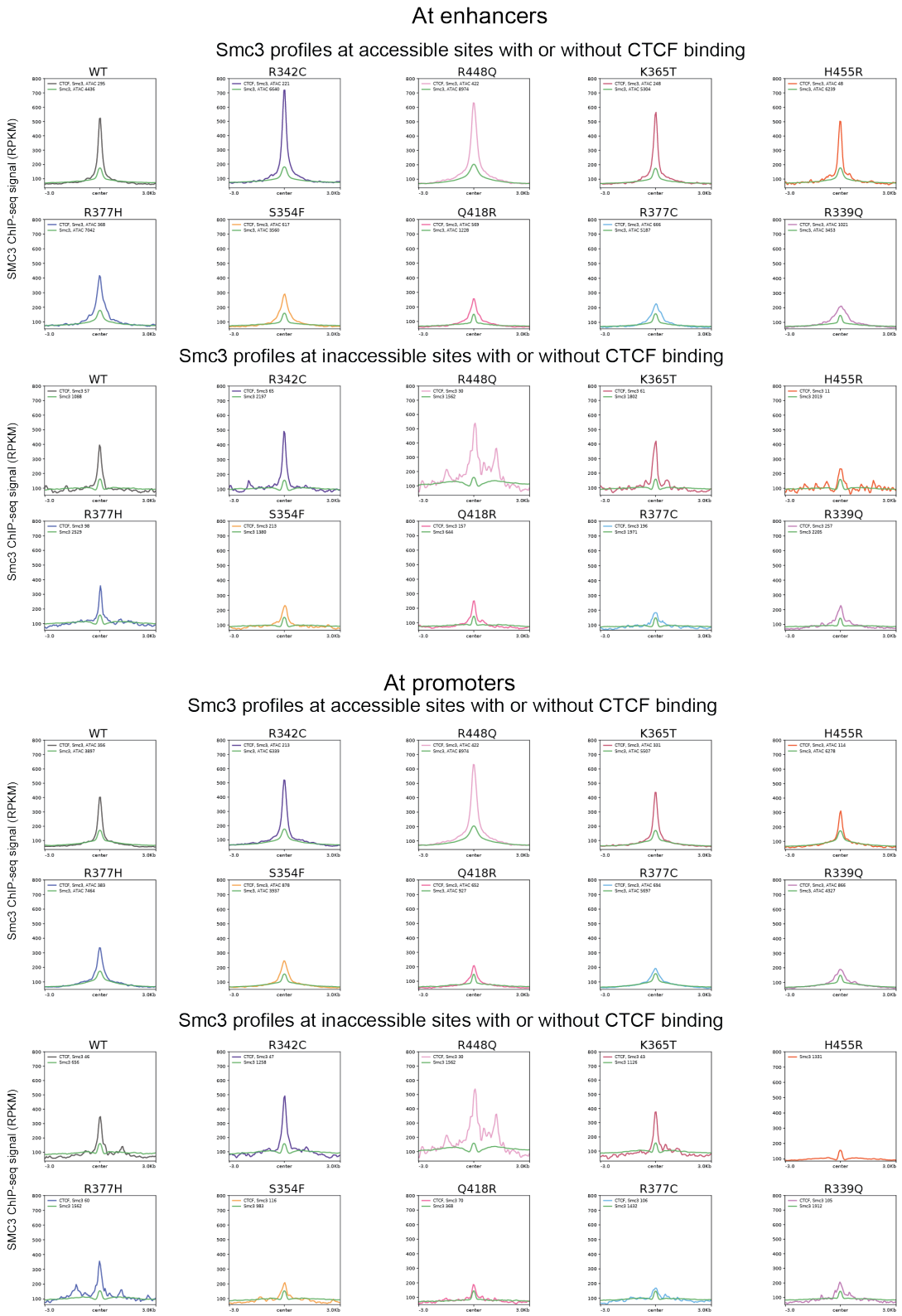
Extended Data Fig. 3: **CTCF mutations have unique chromatin binding profiles:** (**C**) Profiles of ATAC-seq in WT and mutant CTCF expressing cells are shown for sites, with (color-coded for each mutant) or without (green) CTCF binding in the presence or absence of Smc3 at gene enhancers (top) and gene promoters (bottom). In the absence of Smc3, only mutants with enough number of accessible CTCF peaks to generate an accurate signal profile at enhancers or promoters are shown. Of note, the WT profile is shown for reference although very few accessible CTCF peaks were identified at enhancers and promoters.



Extended Data Fig. 4: **The relationship between CTCF binding, cohesin overlap and accessibility. (A)** 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 **Fig. 4C**.

Extended Data Figure 4

B



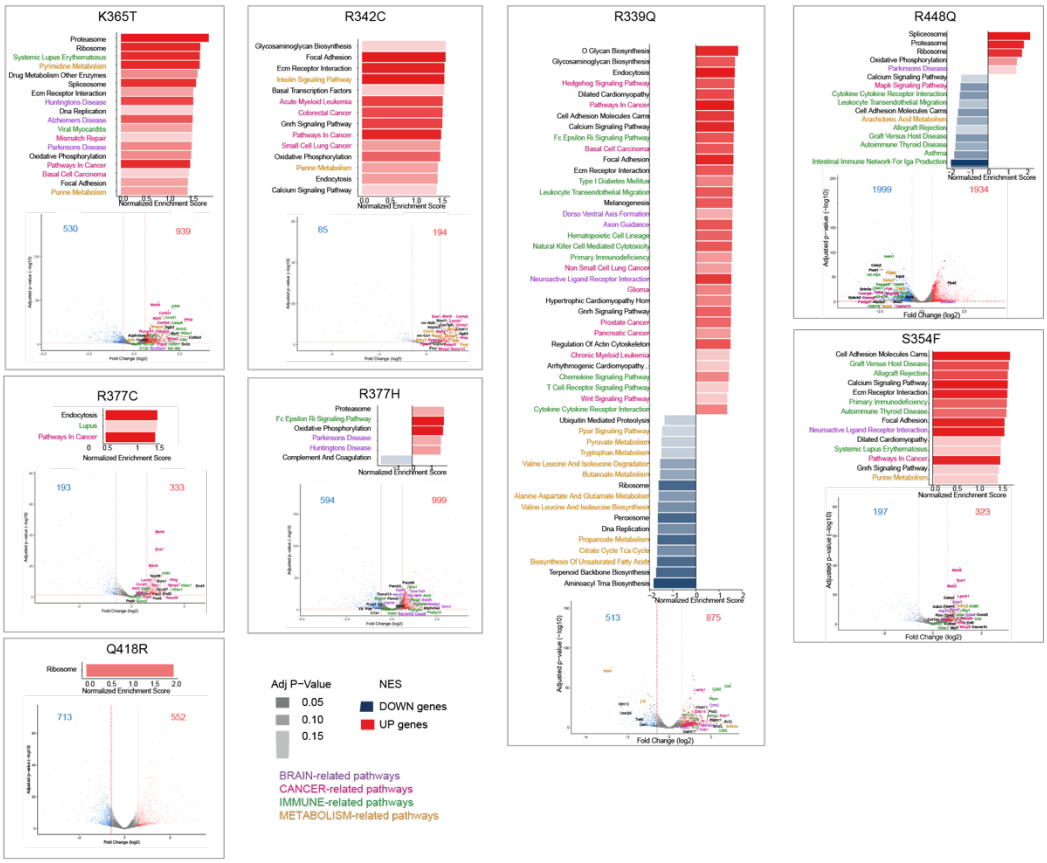
665

Extended Data Fig. 4: **The relationship between CTCF binding, cohesin overlap and accessibility. (B)** Profiles of Smc3 binding in WT and mutant CTCF expressing cells is shown for accessible and inaccessible sites with (color-coded for each mutant) or without (green) CTCF binding at enhancers (top) and gene promoters (bottom)

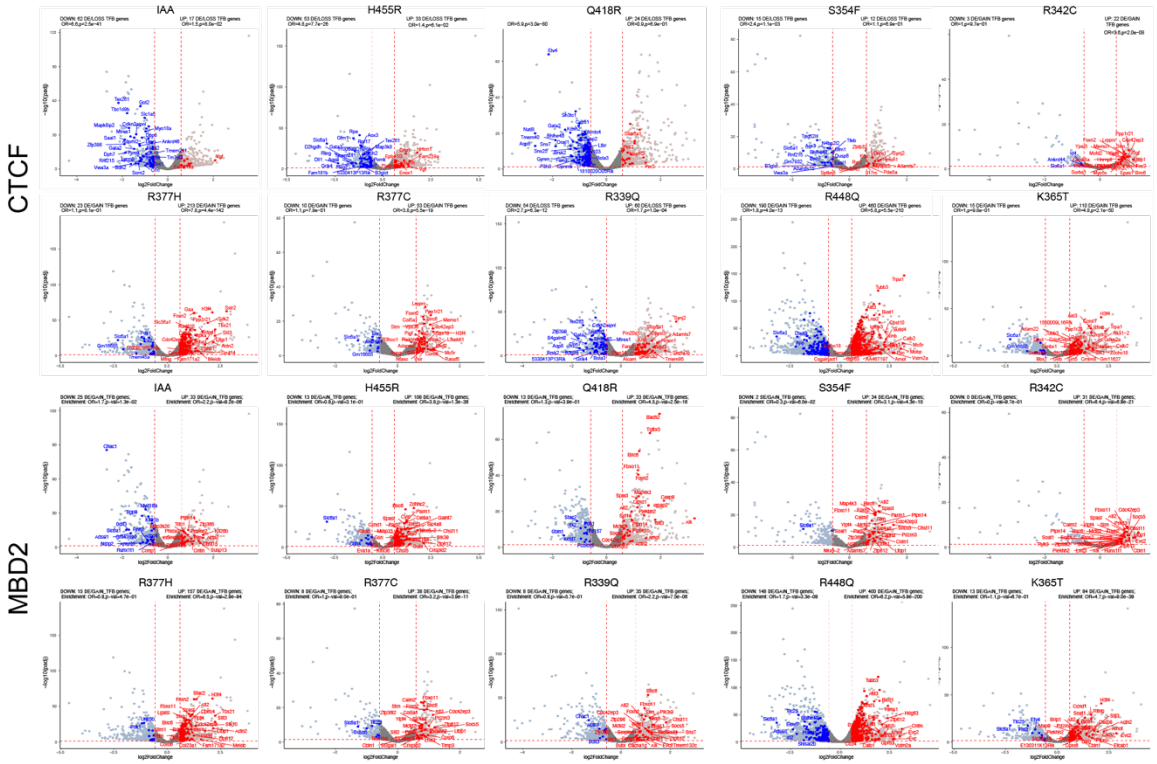
666

Extended Data Figure 5

A

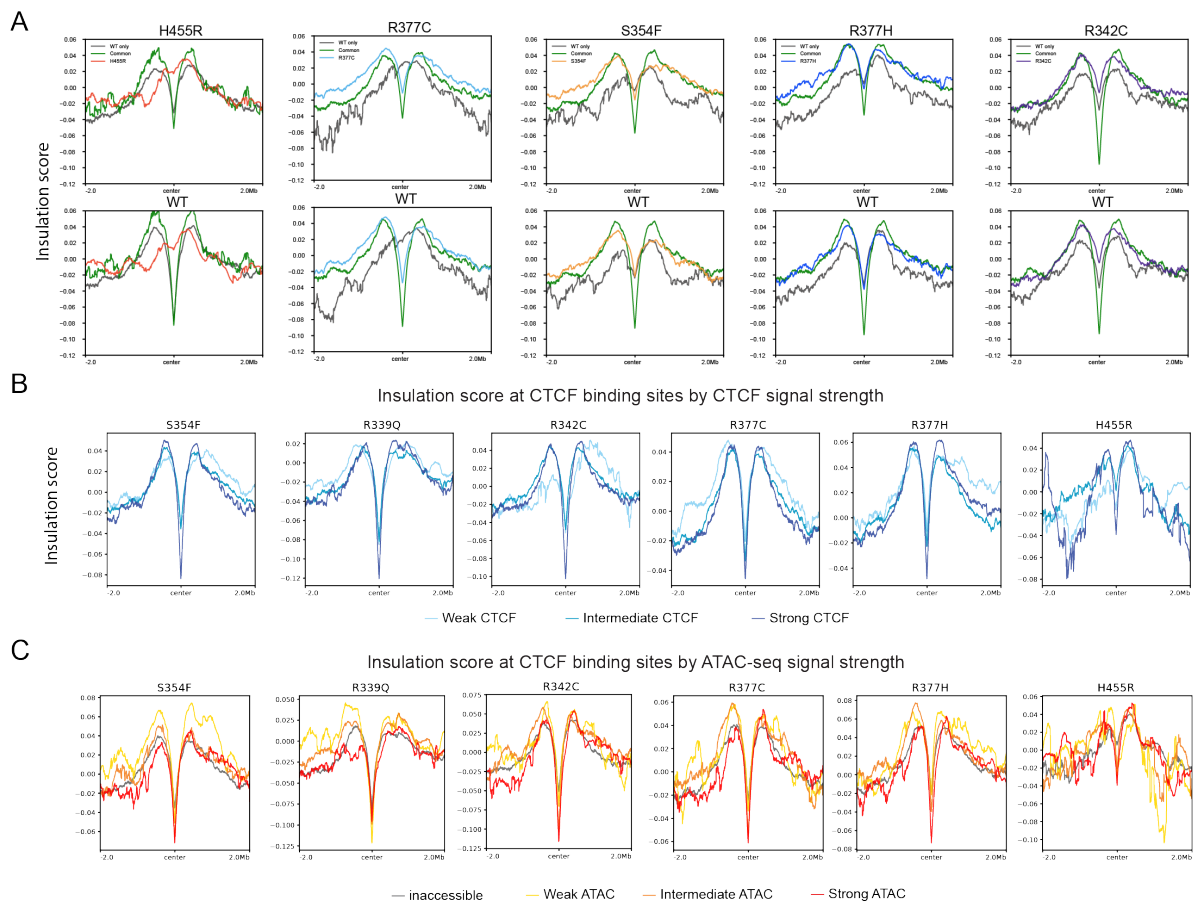


B



Extended Data Fig. 5: **CTCF mutations alter gene expression and TF binding.** (**A**) Gene set enrichment analysis of DEGs. All other mutants are shown in **Fig. 5B**. 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. (**B**) Volcano plots show examples of two differentially bound TFS (CTCF, MBD2) overlapping the promoters of DEGs in CTCF mutant expressing cells (ID). The remaining mutant comparisons are shown in **Fig. 5E**. 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).

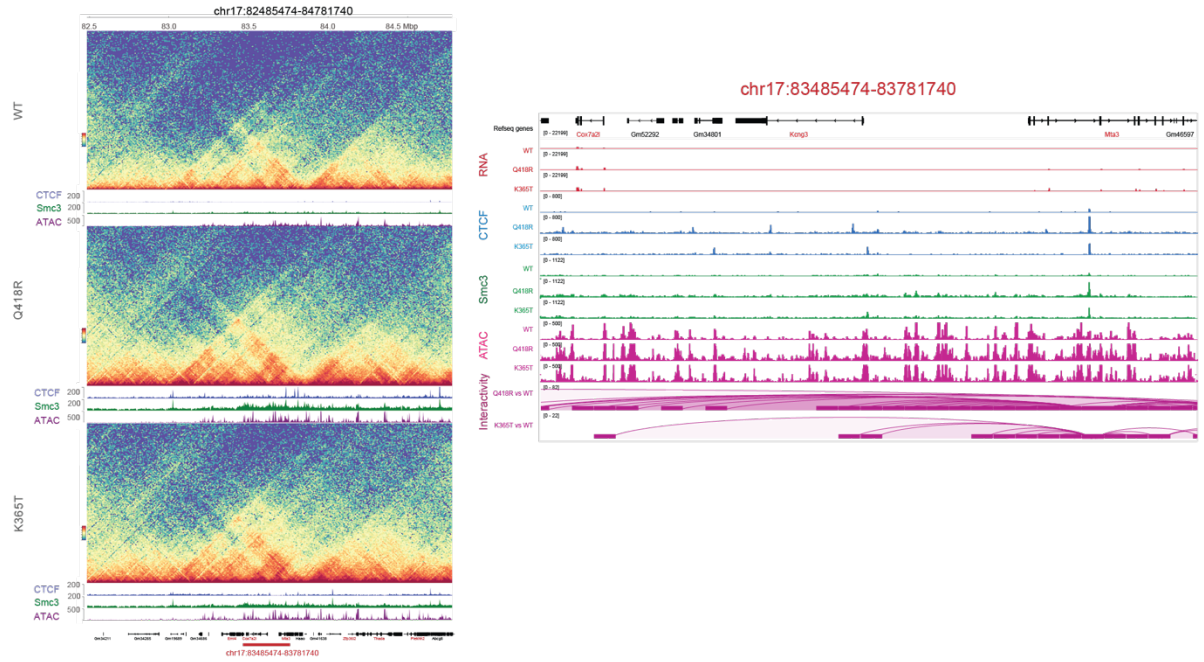
Extended Data Figure 6



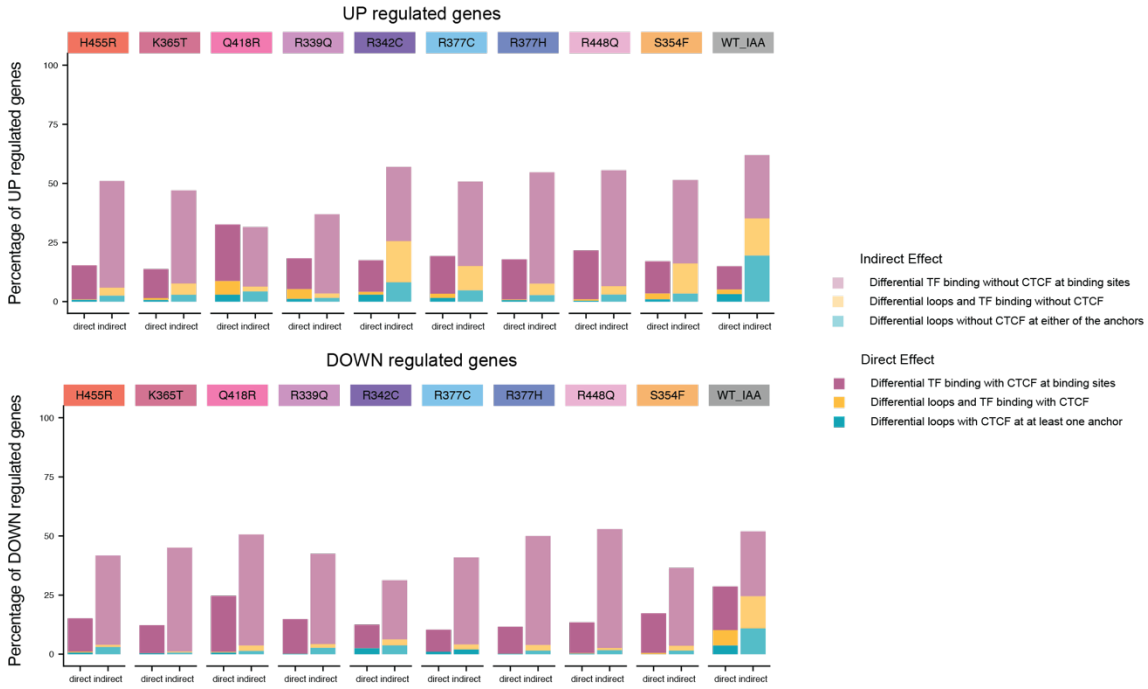
Extended Data Fig. 6: **CTCF mutations alter insulation and chromatin interactivity.** (**A**) Profiles show the aggregated insulation score at WT only, common and mutant only binding sites for four mutants. The profiles for the remaining mutants are shown in **Fig. 6F.** (**B**) Profiles show the aggregated insulation score at CTCF binding sites with weak, intermediate and strong CTCF signals for different mutants. The profiles for the remaining mutants are shown in **Fig. 6G.** (**C**) Profiles show the aggregated insulation score at CTCF binding sites with no, weak, intermediate and strong ATAC signals for different mutants. The profiles for the remaining mutants are shown in **Fig. 6H.**

Extended Data Figure 7

A



B



675

Extended Data Fig. 7: **Indirect effects and change in TF binding at gene promoters account for the majority of CTCF mutation-mediated changes in expression.** (**A**) Example of a locus on chromosome 7 (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 and K365T 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. Overexpressed genes are highlighted in red. Of note, only gained interactions were detected at these loci. This analysis was performed at 10 kb resolution. (**B**) Bar graphs showing the percentage of UP and DOWN DEGs associated with differential TF binding, differential loops or both, stratified by direct and indirect effect. Direct effect was defined as differential TF binding with a predicted differential CTCF binding or ChIP-seq peak at the gene promoter and/or differential loops with a CTCF peak at either anchor.

678 **Methods**

679 **Cell lines**

680 Mouse embryonic stem cells E14Tg2a (karyotype 19, XY; 129/Ola isogenic background) and all
681 clones derived from these were cultured under feeder-free conditions in 0.1% gelatin (Sigma ES-
682 006-B) coated dishes (Falcon, 353003) at 37°C and 5% CO_2 in a humidified incubator. The cells
683 were grown in DMEM (Thermo Fisher, 11965-118) supplemented with 15% fetal bovine serum
684 (Thermo Fisher, SH30071.03), 100 U/ml penicillin - 100 µg/ml streptomycin (Sigma, P4458), 1 X
685 GlutaMax supplement (Thermo Fisher, 35050-061), 1 mM sodium pyruvate (Thermo Fisher,
686 11360-070), 1 X MEM non-essential amino-acids (Thermo Fisher, 11140-50), 50 µM b-
687 mercaptoethanol (Sigma, 38171), 10^4 U/ml leukemia inhibitory factor (Millipore, ESG1107), 3 µM
688 CHIR99021 (Sigma, SML1046) and 1 µM MEK inhibitor PD0325901 (Sigma, PZ0162). The cells
689 were passaged every alternate day by dissociation with TrypLE (Thermo Fisher, 12563011). For
690 the exit from pluripotency experiments, mESCs were cultured for 3 days in the same medium but
691 without the addition of LIF (leukemia inhibitory factor, CHIR99021 (Sigma, SML1046)) and MEK
692 inhibitor (PD0325901).

693

694 **DNA constructs**

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

701

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

712 EN1201 [3] harboring spCas9 nuclease and *Tigre*-targeting sgRNA was used for targeting of *Tigre*
713 locus (Addgene #92144).
714
715 **Gene targeting**
716 Mouse embryonic stem cell E14Tg2a harboring *Ctcf-AID-eGFP* on both alleles and a knock-in of
717 pEN114 - *pCAGGS-Tir1-V5-BpA-Frt-PGK-EM7-PuroR-bpA-Frt-Rosa26* at *Rosa26* locus was
718 used as the parental cell line for making all the transgenes [3]. pEN366 derived vectors harboring
719 the rescue transgenes (WT and mutant *Ctcf*) were used for targeting transgenes to the *Tigre* locus
720 (clone ID# EN156.3.5) [8]. For nucleofections, 15 µg each of plasmids harboring the transgenes
721 and 2.5 µg of those with sgRNA targeting the *Tigre* locus was used. Nucleofection were performed
722 using Amaxa P3 Primary Cell kit (Lonza, V4XP-3024) and 4D- transfector. 2 million cells were
723 transfected with program CG-104 in each case. The cells were recovered for 48 h with no
724 antibiotic followed by selection in puromycin (1 µg/mL) (Thermo Fisher, A1113803). Single
725 colonies were manually picked and expanded in 96 well plates. Clones were genotyped by PCR
726 and flow cytometry was performed to confirm that the level of expression of transgenes were
727 comparable. All the clones that were used for the analyses were homozygous for the integration
728 of the transgenes and their levels of expression were comparable.
729
730 **Induction of auxin inducible degradation of CTCF and doxycycline induced expression**
731 For degradation of endogenous CTCF, the auxin-inducible degron was induced by adding 500
732 µM indole-3-acetic acid (IAA, chemical analog of auxin) (Sigma, I5148) to the media. Expression
733 of transgenes was achieved by the addition of doxycycline (Dox, 1 µg/ml) (Sigma, D9891) to the
734 media. The cells were treated with IAA and/or Dox for 2 days.
735
736 **Western Blotting**
737 mESCs were dissociated using TrypLE, washed in PBS, pelleted and used for western blotting.
738 Approximately 2 million cells were used to prepare cell extract. Cell pellets were resuspended in
739 RIPA lysis buffer (Thermo Fisher, 89900) with 1X HALT protease inhibitors (Thermo Fisher,
740 78430), incubated on ice for 30 min, spun at 4°C at 13,000 rpm for 10 min and supernatant was
741 collected. For the western blot of CTCF, low salt lysis buffer (0.1 M NaCl, 25 mM HEPES, 1 mM
742 $MgCl_2$, 0.2 mM EDTA and 0.5% NP40) was used supplemented with 125 U/ml of benzonase
743 (Sigma E1014). Protein concentration was measured using the Pierce BCA assay kit (Thermo
744 Fisher, 23225). 20 µg of protein were mixed with Laemmli buffer (Biorad, 1610737) and β-
745 mercaptoethanol, heated at 95°C for 10 min and run on a Mini-protean TGX 4%-20%

746 polyacrylamide gel (Biorad, 456-1095). The proteins were transferred onto PVDF membranes
747 using the Mini Trans-Blot Electrophoretic Transfer Cell (Bio-Rad, 170-3930) at 80 V, 120 mA for
748 90 min. PVDF membranes were blocked with 5% BSA in 1 X TBST prior to the addition of
749 antibody. The membranes were probed with appropriate antibodies overnight at 4°C, anti-mouse
750 FLAG antibody (Sigma, F1804; 1: 1,000 dilution), anti GAPDH (BioLegend,W17079A; 1:5,000
751 dilution), anti CTCF (active motif, 61311; 1: 2,000 dilution). Membranes were washed five times
752 in PBST (1 × PBS and 0.1% Tween 20) for 5 min each and incubated with respective secondary
753 antibodies in 5% BSA at room temperature for 1 h. The blots were rinsed in PBST and developed
754 using enhanced chemiluminescence (ECL) and imaged by Odyssey LiCor Imager (Kindle
755 Biosciences).
756
757 **Flow cytometric analysis**
758 Cells were dissociated with TrypLE, washed and resuspended in MACS buffer for flow cytometric
759 analysis on LSRII UV (BD Biosciences). Analysis was performed using the FlowJo software.
760
761 **Cell culture – FRAP**
762 Control cells expressing H2B-mRuby were engineered using the same background as the CTCF
763 transgene expressing cells, E14Tg2a mESCs. Briefly, pEN114 - *pCAGGS-Tir1-V5-BpA-Frt-PGK-*
764 *EM7-PuroR-bpA-Frt-Rosa26* was used as the parental cell line to be nucleofected by the
765 pASH40-mRuby-neo with the PiggyBac transposase vector. For the FRAP experiment, CTCF
766 and H2B expressing cells were cultured for two days on 35 mm no 1.5H glass-bottom imaging
767 dishes (MatTek, Ashland, MA, P35G-1.5-14C) coated with Geltrex (Gibco, A1413201) according
768 to manufacturer's instructions. 48 hours prior to the experiment, the culture medium was changed
769 to medium supplemented with 500 μM 3-Indoleacetic acid (IAA; Sigma-Aldrich, I2886-5G) and 1
770 μg/mL doxycycline (Sigma-Aldrich, D3072-1ML). Just prior to imaging, the medium was changed
771 to imaging medium: phenol red free DMEM with all other aspects of the medium the same.
772
773 **Fluorescence recovery after photobleaching (FRAP)**
774 FRAP experiments were performed on an LSM900 confocal microscope (Zeiss, Germany)
775 equipped with a full, humidified incubation chamber that was maintained at 37°C with 5.5% CO_2.
776 The time series were acquired in confocal mode using two-stage acquisition with the following
777 excitation parameters: 561 nm excitation laser at 0.8% power, 53 μm pinhole size corresponding
778 to 1 AU, scan speed 7, and 3x crop factor corresponding to a pixel dwell time of 3.06 μs. During
779 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 – the average nuclear intensity – and – 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 – – 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 0.125 M followed by incubations of 5 min at room temperature in rotation. Fixed cells were washed twice with 5 ml of 1 X PBS containing 0.5% BSA and centrifuged at 3000 rpm for 5 min at 4°C. Pellets were finally resuspended in 500 µl 1X PBS containing 0.5% BSA, transferred to 1.5 ml 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). 100 µl of cold lysis buffer and 50 µl of 10% Triton X-100 (final 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, SMC3 (abcam, ab9263) and Rabbit IgG (abcam ab37415), 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

847 ChIPmentation protocol (Schmidl et al., 2015). In short, the beads were washed successively
848 twice in 500 µl cold low-salt wash buffer (20 mM Tris-HCl (pH 7.5), 150 mM NaCl, 2 mM EDTA
849 (pH 8.0), 0.1% SDS, 1% tritonX-100), twice in 500 µl cold LiCl-containing wash buffer (10 mM
850 Tris-HCl (pH 8.0), 250 mM LiCl, 1 mM EDTA (pH 8.0), 1% triton X-100, 0.7% sodium
851 deoxycholate) and twice in 500 µl cold 10 mM cold Tris-Cl (pH 8.0) to remove detergent, salts
852 and EDTA. Subsequently, the beads were resuspended in 25 µl of the freshly prepared
853 tagmentation reaction buffer (10 mM Tris-HCl (pH 8.0), 5 mM MgCl2, 10% dimethylformamide)
854 and 1 µl Tagment DNA Enzyme from the Nextera DNA Sample Prep Kit (Illumina #20034198)
855 and incubated at 37°C for 10 min in a thermomixer. Following tagmentation, the beads were
856 washed successively twice in 500 µl cold low-salt wash buffer (20 mM Tris-HCl (pH 7.5), 150 mM
857 NaCl, 2 mM EDTA (pH8.0), 0.1% SDS, 1% triton X-100) and twice in 500 µl cold Tris-EDTA-
858 Tween buffer (0.2% tween, 10 mM Tris-HCl (pH 8.0), 1 mM EDTA (pH 8.0)). Chromatin was eluted
859 and de-crosslinked by adding 70 µl of freshly prepared elution buffer (0.5% SDS, 300 mM NaCl,
860 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
861 850rpm and overnight at 65°C 850rpm. Next day, the supernatant was collected, transferred to
862 new DNA Low Binding tubes and supplemented with an additional 30 µl of elution buffer. DNA
863 was purified using MinElute Reaction Cleanup Kit (Qiagen #28204) and eluted in 20 ul. Purified
864 DNA (20 µl) was amplified as per the ChIPmentation protocol (Schmidl, Rendeiro et al. 2015)
865 using indexed and non-indexed primers and NEBNext High-Fidelity 2X PCR Master Mix (NEB
866 M0541) in a thermomixer with the following program: 72°C for 5 m; 98°C for 30 s; 14 cycles of
867 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
868 using Agencourt AMPure XP beads (Beckman, A63881) to remove fragments larger than 700 bp
869 as well as the primer dimers. Library quality and quantity were estimatred using Tapestation
870 (Agilent High Sensitivity D1000 ScreenTape #5067-5584 and High Sensitivity D1000 reagents
871 #5067-5585) and quantified by Qubit (Life Technologies Qubit™ 1X dsDNA High Sensitivity (HS)
872 #Q33230). Libraries were then sequenced with the Novaseq6000 Illumina technology according
873 to the standard protocols and with around 200 million 150bp paired-end total per sample.
874
875 **Total RNA-seq**
876 mESCs were dissociated using TrypLE, washed in 1X PBS, pelletted and 2.5 million cells were
877 used for extracting RNA with RNeasy plus kit (Qiagen #74134) in each case. RNA quality was
878 checked in Tapestation (Agilent High Sensitivity RNA ScreenTape #5067-5579, High Sensitivity
879 RNA Sample Buffer #5067-5580 and High Sensitivity RNA Ladder #5067-5581) and quantified
880 by Nanodrop. 500 ng were used for Total RNA libraries preparation using Illumina kit (Illumina

881 Stranded Total RNA Prep, Ligation with Ribo-Zero Plus #20040529). Library concentrations were
882 estimated using Tapestation (Agilent High Sensitivity D1000 ScreenTape #5067-5584 and High
883 Sensitivity D1000 reagents #5067-5585) and quantified by Qubit (Life Technologies Qubit™ 1X
884 dsDNA High Sensitivity (HS) #Q33230). Libraries were then sequenced with the Novaseq6000
885 Illumina technology according to the standard protocols and with around 100 million 150bp paired-
886 end total per sample.
887
888 **Hi-C**
889 Hi-C was performed in duplicates using around 1 million cells each. mESCs were dissociated
890 using TrypLE (Thermofishe #12605010), washed once in 1X PBS and resuspended in fresh 1X
891 PBS (1 million cells/ 1 ml). For double cross linking 25 mM EGS (ethylene glycol bis(succinimidyl
892 succinate); Thermofisher #21565) were added and cells were put in rotation for 30 min at room
893 temperature, followed by addition of 1% formaldehyde (Tousimis #1008A) for 10 min also in
894 rotation at room temperature. Quenching was performed by adding glycine to a final concentration
895 of 0.125M followed by incubations of 5 min at room temperature in rotation. Fixed cells were
896 washed twice with 5 ml of 1X PBS containing 0.5% BSA and centrifuged at 3000 rpm for 5 min at
897 4°C. Pellets were finally resuspended in 500 µl 1X PBS containing 0.5% BSA, transferred to 1.5ml
898 Eppendorf and centrifuged at 3000 rpm for 3 min at 4°C. Supernatant was completely removed,
899 pellets were snap-frozen in liquid nitrogen and stored at -80°C. Samples were subsequently
900 processed using the Arima Hi-C kit as per the manufacturer's protocol and sequenced with the
901 Novaseq6000 Illumina technology according to the standard protocols and with around 600 million
902 150bp paired-end reads per sample.
903
904 **ATAC-seq**
905 We used an improved ATAC-seq protocol (Omni-ATAC) adopted from Corces et al., [34] with small
906 adaptations. Briefly, cells in culture were treated with 200 U/ml DNase (Worthington # LS002007)
907 for 30 min at 37°C to remove free-floating DNA and any DNA from dead cells. Cells were then
908 harvested via trypsinization and resuspended in regular medium. After counting, 500,000 cells
909 were collected, resuspended in 1X cold PBS and spin down at 500 g for 5 min at 4°C in a fixed-
910 angle centrifuge. After centrifugation, the supernatant was removed and the pellet resuspended
911 in 500 µl of cold ATAC-seq resuspension buffer 1 (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM
912 $MgCl_2$, 0.1% NP-40 and 0.1% Tween-20 in water). 50 µl (50.000 cells) were transferred to a new
913 1.5 ml Eppendorf tube and 0.5 µl l 1% Digitonin (Promega #G9441) was added by pipetting well
914 few times. The cell lysis reaction was incubated on ice for 3 min. After lysis, 1 ml of cold ATAC-

915 seq resuspension buffer 2 (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$ and 0.1% Tween-
916 20 in water) was added. Tube was inverted three times to mix and nuclei were pelleted by
917 centrifugation for 10 min at 500 g at 4°C. Supernatant was very carefully removed (pellet could
918 be almost invisible at this step) and nuclei resuspended in 45 µl of transposition mix (25 µl 2X TD
919 buffer (Illumina #20034198), 16.5 µl 1X PBS, 0.5 µl 1% digitonin, 0.5 µl 10% Tween-20 and 2.5
920 µl water) by pipetting up and down few times. 5 µl of transposase enzyme (Illumina #20034198)
921 was then added and the transposition reaction was incubated at 37°C for 30 min in a thermomixer
922 with shaking at 1000 rpm. Reaction was cleaned up with Zymo DNA Clean and Concentrator-5
923 kit (Zymo#11-302C). Transposed DNA fragments were then amplified as described previously in
924 Buenrostro et al., [63]. Final libraries were purified with with Zymo DNA Clean and Concentrator-5
925 kit (Zymo#11-302C; Note: better to use two separate kits for pre- and post-amplification clean up),
926 checked by Tapestation (Agilent High Sensitivity D1000 ScreenTape #5067-5584 and High
927 Sensitivity D1000 reagents #5067-5585) and quantified by Qubit (Life Technologies Qubit™ 1X
928 dsDNA High Sensitivity (HS) #Q33230). Libraries were then sequenced with the Novaseq6000
929 Illumina technology according to the standard protocols and with around 50 million 150bp paired-
930 end reads per sample.
931
932 **QUANTIFICATION AND STATISTICAL ANALYSIS**
933
934 **ChIP-seq data processing and quality control**
935 Data were processed using Seq-N-Slide pipeline [64]. Briefly, after trimming for NEXTERA adaptor
936 sequence using trimGalore [65]. The reads were aligned to mm10 genome with Bowtie2 [66].
937 Ambiguous reads were filtered to use uniquely mapped reads in the downstream analysis. PCR
938 duplicates were removed using Sambamba [67]). Narrow peaks were called using MACS2 [68] in pair-
939 end mode and with IgG as control for Smc3 ChIP-seq or WT untreated condition (in which FLAG
940 CTCF is not expressed) for CTCF FLAG to control for unspecific binding, Peaks overlapping
941 ENCODE blacklisted regions were filtered out [69]. These procedures allowed us to minimize the
942 false positive calls, as reflected by the high percentage of called peaks containing CTCF
943 consensus motifs as detected by MEME [32] with 63% for WT. To ensure comparability between
944 conditions, FLAG and SMC3 chip-seq samples were down-sampled to the same number of
945 aligned (100 millions reads per replicates), except for 2 samples R377H-rep1 and R339Q-rep
946 which were under-sequenced (with 68 and 90 millions reads, respectively) and good quality reads
947 before peak calling using samtools [70]. Bigwigs were obtained for visualization on individual as well
948 as merged bam files using Deeptools [71]. For CTCF FLAG visualization, the differential signals

949 were generated after subtracting non-specific signal present in the WT untreated condition.
950 Heatmaps and average profiles were performed on merged bigwig files using Deeptools.
951
952 **RNA-seq data processing and quality control**
953 Data were processed using Seq-N-Slide pipeline [64]. Briefly, after trimming for the Illumina adaptor
954 and the T overhang as per Illumina recommendation using TrimGalore [65], reads were aligned
955 against the mouse reference genome (mm10) using the STAR [72] aligner and differentially
956 expressed genes were called using DESeq2 [73] with an adjusted p-value of 0.05 and a fold change
957 cutoff of 1.5. Heatmap was performed using pheatmap R package [74]and GSEA using fgsea R
958 package [75]and KEGG genesets [43]. Developmental germ layer gene sets were downloaded from
959 Hutchins et al.[76]. Radarplots were generated using the R package fsmb after averaging the
960 expression level of the genes of each germ layers and of the 2 replicates.
961
962 **ATAC-seq data processing and quality control**
963 Data were processed using Seq-N-Slide pipeline [64]. Briefly, Reads were aligned to mm10 genome
964 with Bowtie2 [66]. Ambiguous reads were filtered to use uniquely mapped reads in the downstream
965 analysis. PCR duplicates were removed using Sambamba [67]. Mitochondrial contamination was
966 assessed using samtools [70]. All the samples showed less than 0.6% of mitochondrial DNA, Peaks
967 were called using MACS2 [68]and peaks overlapping ENCODE blacklisted regions were filtered out
968 [69]. Bigwigs were obtained for visualization on individual as well as merged bam files using
969 Deeptools [71]. Heatmaps and average profiles were performed on merged bigwig files using
970 Deeptools.
971
972 **Hi-C Processing and Quality Control.**
973 HiC-Pro [77] was used to align and filter the Hi-C data. To generate Hi-C filtered contact matrices,
974 the Hi-C reads were aligned against the mouse reference genome (mm10) by bowtie2 (version
975 2.3.1) using local mode for the second step of HiC-Pro alignment (--very-sensitive -L 20 --score-
976 min G,20,8 --local) to ignore the molecular barcode UMI sequences and dark bases introduced
977 by Arima library prep kit at the 5' end of the read. Singleton, ambiguous and duplicated reads
978 were filtered out. Ligation sites for Arima HiC kit were set up using
979 GATCGATC,GANTGATC,GANTANTC,GATCANTC. Low mapping quality (MAPQ<30), self
980 circle, dangling ends and re-ligation reads were filtered out through the HiC-pro pipeline. Valid
981 pairs represented more than 80% for all the samples. Samples were downscaled to the same
982 number of good quality uniquely aligned reads using HiCExplorer [78]. Of note, we did not use the

983 number of valid pairs to downscale the samples to not overcorrect some conditions such as IAA
984 and mutant H455R which were expected to show a dramatic biological decrease in valid
985 interactions.
986
987 Downscaled interaction matrices were generated separately for each replicate and merged for
988 visualization and downprocessing as mcool files using cooler [79]. Samples were balanced using
989 the ICE correction method [80] with HiCExplorer [78]. HiC maps were generated from the merged
990 downscaled and corrected matrices using GENOVA and HiCExplorer [54,78]. RCP (relative contact
991 probabilities) and insulation score were generated using GENOVA [54]. After log-log transformation
992 and Loess smoothing using R, the first derivative (P'max) was estimated as described in Gassler
993 et al., [51].
994
995 **DOWNSTREAM ANALYSIS:**
996
997 **CTCF peak motif analysis**
998 Motifs were called using MEME [32] with a 1st order background model on CTCF peaks called using
999 MACS2 after separating the peaks into 3 groups for each pairwise comparison between WT and
1000 mutant CTCF: WT only, common and mutant only peaks. Peak intersection sets were generated
1001 using the intervene package [81]and motifs were searched within 100 bp of each peak summit.
1002
1003 **Stratified analysis at enhancers and promoters**
1004 H3K27ac peaks from 120/Ola mESCs were downloaded from ENCODE (ENCFF519QMV).
1005 CTCF, cohesin and ATAC-seq peaks were classified as active promoter sites when they overlap
1006 with an H3K27ac peak and are located in a gene promoter region, defined as 2kb upstream and
1007 1 kb downstream gene TSSs. Active enhancer sites when they overlap with an H3K27ac peak
1008 and are located outside gene promoter regions.
1009
1010 **Multivariate CTCF/Smc3 overlap and insulation score analysis**
1011 To test the independent of CTCF and Accessibility on Smc3 overlap, CTCF ChIP-seq and ATAC-
1012 seq signal strengths were categorized into weak, intermediate and strong signal using the inter-
1013 quartile values of the signal across WT and mutant samples. Weak binding or accessibility was
1014 defined as sites with signals inferior to the 1st quartile, intermediate binding or accessibility as
1015 sites with signals between the 2nd and 3rd quartiles and strong binding or accessibility as sites with
1016 signals superior to the 3rd quartile. For Smc3 overlap, a multivariate logistic model testing the

1017 effect of CTCF strength, ATAC strength and their interaction was performed in WT mESCs after
1018 averaging the replicates. The log2 odd ratios were computed to estimate the enrichment in Smc3.
1019 For insulation score a mixed multivariate model testing the effect of CTCF strength, ATAC
1020 strength and their interaction was performed across sample. For both models the marginal effects
1021 and 95th confidence intervals were computed. These analyses was performed using STATA v17.
1022

1023 **ATAC-seq footprinting analysis**

1024 Footprinting analysis was performed on ATAC-seq signal using TOBIAS [45]. Briefly, ATAC-seq
1025 signals were first corrected for Tn5 cutting sequence bias using the ATACorrect function and
1026 footprinting score calculated using FootprintScores at peaks detected by MACS. Finally
1027 differential footprinting estimated pariwise between WT and mutant CTCF using BINDetect and
1028 HOCOMOCO v11 core motif probability weight matrices [82]. Peaks were annotated for gene
1029 promoter using UROPA [83]and GENCODE version M25 [84]. Gene promoters were defined as
1030 regions within 2kb upstream and 1 kb downstream the gene transcription starting site.
1031

1032 **Loop calling and annotation**

1033 Significant loops (q-value<0.05) were called on the downscaled and corrected interaction
1034 matrices using FitHIC at 10kb resolution [85] and annotated for CTCF peaks and gene TSS using
1035 pairToBed [86].
1036

1037 **Differential aggregated peak and TAD analyses**

1038 Aggregated peak and TAD analysis on downscaled and corrected interaction matrices was
1039 performed using GENOVA at 10kb resolution [54]. For aggregated peaks analyses, the union of
1040 loops called by FitHIC in WT and mutant CTCF for each pairwise comparison was used to
1041 generate and compare aggregated interaction at loop anchors. For aggregated TAD analyses,
1042 the insulation score was first calculated to identify TAD and TAD boundaries. The union of TADs
1043 called by GENOVA in WT and mutant CTCF was used to generate and compare aggregated intra
1044 and inter-TAD interactions.
1045

1046 **Differential loop analysis**

1047 Differential loops (adjusted p-value<0.05) were called for each pairwise comparison between WT
1048 and mutant CTCF using MultiHiCCompare which performed a cross-sample normalization before
1049 testing for differential loops using an exact test [87]. Enrichment for differentially expressed genes

among differential loops was tested using logistic regression after annotating loop anchors with gene TSS.

Direct vs indirect effect analysis.

Direct effect was defined as differential TF binding with a predicted differential CTCF binding identified using TOBIAS [45] or ChIP-seq peak at the gene promoter and/or differential loops identified using MultiHiCCompare [87] with a CTCF peak at either anchor. Gene promoters were defined as regions spanning 2 kb upstream and 1 kb downstream gene TSSs. Overlap between gene promoters and differential loop were performed using pairToBed [86].

Data availability

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.

Code availability

No custom code/software was used in the study. The publicly available software used is indicated in the Methods above.

Ethics approval and consent to participate

Not applicable

Competing interests

The authors declare no competing interests.

Author' contributions

These studies were designed by Jane Skok and Catherine Do. The analysis was performed by Catherine Do, Theodore Sakellaropoulos. Experiments were performed by Guimei Jiang and Giulia Cova. Aristotelis Tsirigos reviewed and made revisions to the manuscript. 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

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 and a Pew-Stewart Cancer Research Scholar grant. GC and GJ were supported by fellowships from the NCC.

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.

References:

1. Parelho, V. *et al.* Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132**, 422-33 (2008).

2. Wendt, K.S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**, 796-801 (2008).

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

4. Price, E. *et al.* An updated catalog of CTCF variants associated with neurodevelopmental disorder phenotypes. *Front Mol Neurosci* **16**, 1185796 (2023).

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

6. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep* **15**, 2038-49 (2016).

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

1113 8. Nishana, M. *et al.* Defining the relative and combined contribution of CTCF and CTCFL to
1114 genomic regulation. *Genome Biol* **21**, 108 (2020).

1115 9. Nora, E.P. *et al.* Molecular basis of CTCF binding polarity in genome folding. *Nat Commun*
1116 **11**, 5612 (2020).

1117 10. Pugacheva, E.M. *et al.* CTCF mediates chromatin looping via N-terminal domain-
1118 dependent cohesin retention. *Proc Natl Acad Sci U S A* **117**, 2020-2031 (2020).

1119 11. Hansen, A.S. CTCF as a boundary factor for cohesin-mediated loop extrusion: evidence
1120 for a multi-step mechanism. *Nucleus* **11**, 132-148 (2020).

1121 12. Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication.
1122 *Cell* **164**, 1110-1121 (2016).

1123 13. Flavahan, W.A. *et al.* Insulator dysfunction and oncogene activation in IDH mutant
1124 gliomas. *Nature* **529**, 110-4 (2016).

1125 14. Guo, C. *et al.* CTCF-binding elements mediate control of V(D)J recombination. *Nature*
1126 **477**, 424-30 (2011).

1127 15. Guo, Y. *et al.* CRISPR Inversion of CTCF Sites Alters Genome Topology and
1128 Enhancer/Promoter Function. *Cell* **162**, 900-10 (2015).

1129 16. Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome
1130 neighborhoods. *Science* **351**, 1454-1458 (2016).

1131 17. Narendra, V. *et al.* Transcription. CTCF establishes discrete functional chromatin domains
1132 at the Hox clusters during differentiation. *Science* **347**, 1017-21 (2015).

1133 18. Xiang, Y., Zhou, X., Hewitt, S.L., Skok, J.A. & Garrard, W.T. A multifunctional element in
1134 the mouse Igkappa locus that specifies repertoire and Ig loci subnuclear location. *J*
1135 *Immunol* **186**, 5356-66 (2011).

1136 19. Filippova, G.N. *et al.* A widely expressed transcription factor with multiple DNA sequence
1137 specificity, CTCF, is localized at chromosome segment 16q22.1 within one of the smallest
1138 regions of overlap for common deletions in breast and prostate cancers. *Genes*
1139 *Chromosomes Cancer* **22**, 26-36 (1998).

1140 20. Debaugny, R.E. & Skok, J.A. CTCF and CTCFL in cancer. *Curr Opin Genet Dev* **61**, 44-
1141 52 (2020).

1142 21. Kemp, C.J. *et al.* CTCF haploinsufficiency destabilizes DNA methylation and predisposes
1143 to cancer. *Cell Rep* **7**, 1020-9 (2014).

1144 22. Gregor, A. *et al.* De novo mutations in the genome organizer CTCF cause intellectual
1145 disability. *Am J Hum Genet* **93**, 124-31 (2013).

1146 23. Konrad, E.D.H. *et al.* CTCF variants in 39 individuals with a variable neurodevelopmental
1147 disorder broaden the mutational and clinical spectrum. *Genet Med* **21**, 2723-2733 (2019).

1148 24. Valverde de Morales, H.G. *et al.* Expansion of the genotypic and phenotypic spectrum of
1149 CTCF-related disorder guides clinical management: 43 new subjects and a
1150 comprehensive literature review. *Am J Med Genet A* **191**, 718-729 (2023).

1151 25. Chen, H., Tian, Y., Shu, W., Bo, X. & Wang, S. Comprehensive identification and
1152 annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome.
1153 *PLoS One* **7**, e41374 (2012).

1154 26. Maurano, M.T. *et al.* Role of DNA Methylation in Modulating Transcription Factor
1155 Occupancy. *Cell Rep* **12**, 1184-95 (2015).

1156 27. Saldana-Meyer, R. *et al.* RNA Interactions Are Essential for CTCF-Mediated Genome
1157 Organization. *Mol Cell* **76**, 412-422 e5 (2019).

1158 28. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring
1159 multidimensional cancer genomics data. *Cancer Discov* **2**, 401-4 (2012).

1160 29. de Bruijn, I. *et al.* Analysis and Visualization of Longitudinal Genomic and Clinical Data
1161 from the AACR Project GENIE Biopharma Collaborative in cBioPortal. *Cancer Res* **83**,
1162 3861-3867 (2023).

1163 30. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using
1164 the cBioPortal. *Sci Signal* **6**, pl1 (2013).

1165 31. Tate, J.G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids*
1166 *Res* **47**, D941-D947 (2019).

1167 32. Bailey, T.L., Johnson, J., Grant, C.E. & Noble, W.S. The MEME Suite. *Nucleic Acids Res*
1168 **43**, W39-49 (2015).

1169 33. Yang, J. *et al.* Structures of CTCF-DNA complexes including all 11 zinc fingers. *Nucleic*
1170 *Acids Res* **51**, 8447-8462 (2023).

1171 34. Corces, M.R. *et al.* An improved ATAC-seq protocol reduces background and enables
1172 interrogation of frozen tissues. *Nat Methods* **14**, 959-962 (2017).

1173 35. Hansen, A.S., Pustova, I., Cattoglio, C., Tjian, R. & Darzacq, X. CTCF and cohesin
1174 regulate chromatin loop stability with distinct dynamics. *Elife* **6**(2017).

1175 36. Sprague, B.L., Pego, R.L., Stavreva, D.A. & McNally, J.G. Analysis of binding reactions
1176 by fluorescence recovery after photobleaching. *Biophys J* **86**, 3473-95 (2004).

1177 37. Hansen, A.S., Amitai, A., Cattoglio, C., Tjian, R. & Darzacq, X. Guided nuclear exploration
1178 increases CTCF target search efficiency. *Nat Chem Biol* **16**, 257-266 (2020).

1179 38. de Jonge, W.J., Patel, H.P., Meeussen, J.V.W. & Lenstra, T.L. Following the tracks: How
1180 transcription factor binding dynamics control transcription. *Biophys J* **121**, 1583-1592
1181 (2022).

1182 39. Cheng, L., De, C., Li, J. & Pertsinidis, A. Mechanisms of transcription control by distal
1183 enhancers from high-resolution single-gene imaging. *bioRxiv* (2023).

1184 40. Zuin, J. *et al.* Nonlinear control of transcription through enhancer-promoter interactions.
1185 *Nature* **604**, 571-577 (2022).

1186 41. Mach, P. *et al.* Cohesin and CTCF control the dynamics of chromosome folding. *Nat Genet*
1187 **54**, 1907-1918 (2022).

1188 42. Gabriele, M. *et al.* Dynamics of CTCF- and cohesin-mediated chromatin looping revealed
1189 by live-cell imaging. *Science* **376**, 496-501 (2022).

1190 43. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids*
1191 *Res* **28**, 27-30 (2000).

1192 44. Hirai, H., Karian, P. & Kikyo, N. Regulation of embryonic stem cell self-renewal and
1193 pluripotency by leukaemia inhibitory factor. *Biochem J* **438**, 11-23 (2011).

1194 45. Bentsen, M. *et al.* ATAC-seq footprinting unravels kinetics of transcription factor binding
1195 during zygotic genome activation. *Nat Commun* **11**, 4267 (2020).

1196 46. Lax, E. *et al.* Methyl-CpG binding domain 2 (Mbd2) is an epigenetic regulator of autism-
1197 risk genes and cognition. *Transl Psychiatry* **13**, 259 (2023).

1198 47. Mahmood, N., Arakelian, A., Szyf, M. & Rabbani, S.A. Methyl-CpG binding domain protein
1199 2 (Mbd2) drives breast cancer progression through the modulation of epithelial-to-
1200 mesenchymal transition. *Exp Mol Med* **56**, 959-974 (2024).

1201 48. Menafra, R. & Stunnenberg, H.G. MBD2 and MBD3: elusive functions and mechanisms.
1202 *Front Genet* **5**, 428 (2014).

1203 49. Reynolds, N. *et al.* NuRD suppresses pluripotency gene expression to promote
1204 transcriptional heterogeneity and lineage commitment. *Cell Stem Cell* **10**, 583-94 (2012).

1205 50. Holwerda, S.J. & de Laat, W. CTCF: the protein, the binding partners, the binding sites
1206 and their chromatin loops. *Philos Trans R Soc Lond B Biol Sci* **368**, 20120369 (2013).

1207 51. Gassler, J. *et al.* A mechanism of cohesin-dependent loop extrusion organizes zygotic
1208 genome architecture. *EMBO J* **36**, 3600-3618 (2017).

1209 52. Edelmann, W. *et al.* Mutation in the mismatch repair gene Msh6 causes cancer
1210 susceptibility. *Cell* **91**, 467-77 (1997).

1211 53. Chen, Y. *et al.* Oncogenic MSH6-CXCR4-TGFB1 Feedback Loop: A Novel Therapeutic
1212 Target of Photothermal Therapy in Glioblastoma Multiforme. *Theranostics* **9**, 1453-1473
1213 (2019).

1214 54. van der Weide, R.H. *et al.* Hi-C analyses with GENOVA: a case study with cohesin
1215 variants. *NAR Genom Bioinform* **3**, lqab040 (2021).

1216 55. Liu, X.H. *et al.* FOXN2 suppresses the proliferation and invasion of human hepatocellular
1217 carcinoma cells. *Eur Rev Med Pharmacol Sci* **25**, 731-737 (2021).

1218 56. Ma, J. *et al.* beta-Trcp ubiquitin ligase and RSK2 kinase-mediated degradation of FOXN2
1219 promotes tumorigenesis and radioresistance in lung cancer. *Cell Death Differ* **25**, 1473-
1220 1485 (2018).

1221 57. Ye, H. & Duan, M. FOXN2 is downregulated in breast cancer and regulates migration,
1222 invasion, and epithelial- mesenchymal transition through regulation of SLUG. *Cancer*
1223 *Manag Res* **11**, 525-535 (2019).

1224 58. Ikeda, K. *et al.* Mitochondrial supercomplex assembly promotes breast and endometrial
1225 tumorigenesis by metabolic alterations and enhanced hypoxia tolerance. *Nat Commun* **10**,
1226 4108 (2019).

1227 59. Iurlaro, M. *et al.* Systematic assessment of ISWI subunits shows that NURF creates local
1228 accessibility for CTCF. *Nat Genet* **56**, 1203-1212 (2024).

1229 60. Do, C. & Skok, J.A. Factors that determine cell type-specific CTCF binding in health and
1230 disease. *Curr Opin Genet Dev* **88**, 102244 (2024).

1231 61. Soochit, W. *et al.* CTCF chromatin residence time controls three-dimensional genome
1232 organization, gene expression and DNA methylation in pluripotent cells. *Nat Cell Biol* **23**,
1233 881-893 (2021).

1234 62. Watson, L.A. *et al.* Dual effect of CTCF loss on neuroprogenitor differentiation and
1235 survival. *J Neurosci* **34**, 2860-70 (2014).

1236 63. Buenrostro, J.D., Wu, B., Chang, H.Y. & Greenleaf, W.J. ATAC-seq: A Method for
1237 Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* **109**, 21 29 1-21 29
1238 9 (2015).

1239 64. Dolgalev, I. Seq-N-Slide [Computer software]. (2022).

1240 65. Krueger, F. TrimGalore. (2021).

1241 66. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods*
1242 **9**, 357-9 (2012).

1243 67. Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J. & Prins, P. Sambamba: fast processing
1244 of NGS alignment formats. *Bioinformatics* **31**, 2032-4 (2015).

1245 68. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).

1246 69. Amemiya, H.M., Kundaje, A. & Boyle, A.P. The ENCODE Blacklist: Identification of
1247 Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).

1248 70. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**(2021).

1249 71. Ramirez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data
1250 analysis. *Nucleic Acids Res* **44**, W160-5 (2016).

1251 72. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21
1252 (2013).

1253 73. Love, M.I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion
1254 for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

1255 74. Kolde, R. pheatmap: Pretty Heatmaps. R package version 1.0.12. (2019).

1256 75. Sergushichev, A. An algorithm for fast preranked gene set enrichment analysis using
1257 cumulative statistic calculation. *bioRxiv. doi: 10.1101/060012* (2016).

1258 76. Hutchins, A.P. *et al.* Models of global gene expression define major domains of cell type
1259 and tissue identity. *Nucleic Acids Res* **45**, 2354-2367 (2017).

1260 77. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing.
1261 *Genome Biol* **16**, 259 (2015).

1262 78. Wolff, J. *et al.* Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis,
1263 quality control and visualization. *Nucleic Acids Res* **46**, W11-W16 (2018).

1264 79. Abdennur, N. & Mirny, L.A. Cooler: scalable storage for Hi-C data and other genomically
1265 labeled arrays. *Bioinformatics* **36**, 311-316 (2020).

1266 80. Imakaev, M. *et al.* Iterative correction of Hi-C data reveals hallmarks of chromosome
1267 organization. *Nat Methods* **9**, 999-1003 (2012).

1268 81. Khan, A.M., A. Intervene: a tool for intersection and visualization of multiple genomic
1269 region sets. (2017).

1270 82. Kulakovskiy, I.V. *et al.* HOCOMOCO: towards a complete collection of transcription factor
1271 binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids*
1272 *Res* **46**, D252-D259 (2018).

1273 83. Kondili, M. *et al.* UROPA: a tool for Universal RObust Peak Annotation. *Sci Rep* **7**, 2593
1274 (2017).

1275 84. Frankish, A. *et al.* Gencode 2021. *Nucleic Acids Res* **49**, D916-D923 (2021).

1276 85. Ay, F., Bailey, T.L. & Noble, W.S. Statistical confidence estimation for Hi-C data reveals
1277 regulatory chromatin contacts. *Genome Res* **24**, 999-1011 (2014).

1278 86. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic
1279 features. *Bioinformatics* **26**, 841-2 (2010).

1280 87. Stansfield, J.C., Cresswell, K.G. & Dozmorov, M.G. multiHiCcompare: joint normalization
1281 and comparative analysis of complex Hi-C experiments. *Bioinformatics* **35**, 2916-2923
1282 (2019).
1283