The coordination between CTCF, cohesin and TFs impacts nucleosome repositioning and chromatin insulation to define state specific 3D chromatin folding

Catherine Do[1], Guimei Jiang[1], Giulia Cova[1], Paul Zappile[3], Adriana Heguy[3] & 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. Genome Technology Center, NYU Grossman School of Medicine, New York, NY, USA

* Corresponding author: Jane.Skok@nyulangone.org

Abstract (250 words)

CTCF-mediated chromatin folding plays a key role in gene regulation, however the mechanisms controlling chromatin organization across cell states are not fully elucidated. Comprehensive analyses reveal that CTCF binding stability and cohesin overlap in mice and humans, are regulated by species specific differences in CTCF binding site (CBS) accessibility and enrichment of motifs corresponding to expressed TFs. By analyzing TFs, we confirm the co-operativity and competitiveness of TF/CTCF binding, which we further validate by allele specific analysis of SNPs. TF motif enrichment at CTCF bound sites is determined by cell state-specific transcriptional programs, which either stabilize or destabilize CTCF binding, as reflected by changing TF concentration. To examine CTCF binding in the context of nucleosome positioning, we performed single molecule nano-NOMe-seq. Subsetting reveals a continuum of binding states for CTCF, which are differentially represented at accessible versus inaccessible CBSs. As expected, CTCF degradation leads to a progressive loss of binding and nucleosome repositioning giving profiles similar to CTCF free CBSs. We also observe a similar time dependent effect when the cohesin subcomponent, SSC1 is degraded although CTCF remains bound, indicating that cohesin mediates CTCF-associated nucleosome repositioning. Stratified analysis of CTCF signal strength and accessibility reveals that in the presence of cohesin, CTCF strength contributes to nucleosome repositioning and chromatin insulation independent of accessibility. However, cobound TFs can uncouple the relationship between signal strength and nucleosome repositioning, without affecting the connection between repositioning and insulation. These studies identify mechanisms underlying cell state-specific CTCF profiles, linked to local and long-range chromatin organization.

Introduction

CTCF is an eleven zinc finger protein that makes contact with a 12-15bp DNA motif through its central ZFs, 3-7. Together CTCF and the cohesin complex promote the formation of chromatin loops within highly self-interacting topologically associated domain (TAD) structures, via a loop extrusion mechanism [1,2]. The two factors are enriched at TAD boundaries where they function as insulators promoting interactions between enhancers and promoters within a domain, and restricting the action of regulatory elements to target genes outside a domain. CTCF mediated chromatin folding plays a key role in gene regulation as exemplified by numerous studies showing that disruption of insulated TAD boundaries by deletion of CTCF binding sites (CBSs) can alter gene expression leading to defects in developmental processes and cancer initiation as a result of illegitimate enhancer promoter contacts [3-7].

CTCF dependent chromatin interactions can be conserved, or cell-state specific [8], however the mechanisms underlying CTCF's ability to act as a chromatin organizer in a developmental stage, or disease specific manner have not been fully elucidated. We and others have found that CTCF can bind both accessible and inaccessible sites, with differential functional outcomes [9,10]. By analyzing the interplay between CTCF, cohesin and ATAC-seq data in CTCF WT and mutant expressing cells, we demonstrated that accessibility is a key determinant of CTCF binding stability, and CTCF signal strength in turn is the strongest predictor of cohesin overlap and chromatin insulation score [9]. Specifically, when CTCF binds accessible sites, it has a stronger signal and is more efficient at blocking cohesin and participating in chromatin loop formation than CTCF at inaccessible sites [9]. However, the relationship between accessibility and CTCF binding stability is not clear cut, because a subset of inaccessible CTCF bound sites have strong signals with cohesin overlap, and conversely, CTCF at accessible sites can exhibit weak signals devoid of cohesin. Furthermore, there are numerous accessible CBSs that remain unbound. Thus, other factors must positively and negatively contribute to binding.

While previous studies have identified a handful of transcription factor (TF) cofactors that stabilize CTCF binding in specific cell types, a comprehensive analysis exploring the role of TFs in cell type-specific binding is lacking [11-16]. In this study, we demonstrate that DNA sequences found within a 35 bp window surrounding CBSs motifs corresponding to expressed cell type-specific TFs can differentially affect CTCF signal strength and its capacity to block cohesin, in a manner that depends on whether they are located in accessible or inaccessible chromatin. In both mice and humans, we identified enriched TF motifs showing strong similarity within specific windows,

suggesting potential redundancy in TF-mediated regulation of CTCF binding and a mechanism for cell type-specific control. We also highlight differences between the two species: notably, mouse and human differ respectively in the proportion of inaccessible versus accessible CBSs in their genomes, which is reflected in the regulatory mechanisms that govern their binding. Although TF motifs are present at most binding sites in both species, in mice they are enriched at inaccessible sites, while in humans they are more abundant at accessible sites.

To examine the impact of transcriptional programming on CTCF binding, we looked at how variations in individual TF levels across different cell states and cell types are associated with changes in their motif enrichment at CTCF sites. In contrast to previous findings [11-16] , we show that TFs can function to both stabilize and destabilize CTCF binding, as reflected by the direction of enrichment at increasing TF concentration. By analyzing a handful of TFs, we could confirm both their co-operativity and competitiveness with CTCF binding at CBSs. Consistent with our ChIP-seq analysis, AlphaFold3 predicts that at individual locations, TFs can respectively compete or cobind with CTCF depending on whether TF binding overlaps the CBS or occurs up / downstream of it. Allele specific validation further highlights that SNPs affecting either CTCF or TF binding have reciprocal effects on the binding of each factor. Moreover, changes in binding of either factor can impact cohesin binding and to a lesser extent, ATAC-seq signals. Consequently, changes in TF expression levels can convert a weak CTCF-bound site into a strong site that halts the progression of cohesin, or vice versa, depending on whether the adjacent TF binds cooperatively or competitively with CTCF and whether its expression increases or decreases. This duality helps explain why many accessible CBSs are not occupied by CTCF itself, but rather by TFs that prevent its localization

Using long read, nano-NOME-seq single molecule footprinting, we analyzed the impact of CTCF binding on nucleosome repositioning in different genomic contexts. The single molecule approach allowed us to subset our data revealing a continuum of binding states for CTCF which are present at individual peaks, and differentially represented at accessible and inaccessible sites. This information has been missing from previous bulk analyses that average all the data. Using the clustered analysis, we demonstrate that CTCF has a reduced ability to reposition nucleosomes at inaccessible sites. Furthermore, we found that (i) the profiles of degraded CTCF and CTCF free CBSs were very similar, reflecting nucleosome occupied CBSs or binding of TFs that can still phase nucleosomes but cannot reposition them, and (ii) degradation of the SCC1 subcomponent of cohesin gave rise to a similar profile while CTCF remains bound, pointing to the interesting

finding that cohesin and loop extrusion mediate nucleosome positioning and phasing by creating or stabilizing the barrier effects at CTCF anchors.

A stratified analysis of CTCF signal strength and accessibility revealed that in the presence of cohesin, CTCF strength contributes to nucleosome repositioning and insulation independent of accessibility. However, cobound TFs can uncouple the relationship between signal strength and nucleosome repositioning without affecting the connection between nucleosome repositioning and insulation. These findings indicate that nucleosome repositioning is a more reliable predictor of chromatin insulation than CTCF signal strength and that TF cobinding can modulate loop extrusion at CTCF anchors. Together, these studies demonstrate that cell state-specific transcriptional programs, shape CTCF profiles linked to local and long-range chromatin organization and gene regulation.

Results

TF motifs enriched within 35 bps of CTCF bound sites stabilize CTCF binding

A number of studies have reported that CTCF binding can be stabilized by the cobinding of cell-type specific TFs adjacent to CTCF [11-16], but there have been no comprehensive, unbiased analyses to determine how cell type-specific cofactors regulate CTCF binding and its loop forming capacity. In addition, previous studies did not take into account the chromatin accessibility of CTCF binding sites, which can play an important role in CTCF cell-specific binding and function as previously shown by us and others[9,10].

To investigate, we used mESC data previously generated by our lab [9] to examine the enrichment of TFs at defined windows in the 35 bp region surrounding the CTCF binding site (**Figure 1A**). We focused on this region as we observed little enrichment of TF motifs beyond this point (**Figure S1A**), although a slight increase could be detected further upstream between 50-75bps, that could impact CTCF binding via a different mechanism. In **Figure 1A** the x-axis shows the gap/distance in bps between the end of the TF motif and the start of the CTCF motif for upstream regions, or the start of the TF motif and the end of the CTCF motif for downstream regions. CTCF has eleven zinc fingers of which ZFs 3-7 make up the core binding domain that contacts the 12-15bp consensus motif. It has previously been reported that at some sites an Upstream (U) CTCF motif, can engage ZFs 9 to 11 and a less prevalent Downstream (D) motif can impair the binding of ZFs 1 and 2 to alter binding stability [11,17-19]. We therefore defined CTCF upstream and downstream regions as the U motif window (U win) with binding ending at the 6 to 8 bp upstream gap, the U

144 spacer as binding ending at the 0 to 7 bp upstream gap, the D spacer as binding beginning at the
145 0 to 9 bp downstream gap, and the D motif window (D win) as binding beginning at the 10 to 11
146 downstream gap.
147
148 ***TF motifs have distinct enrichment profiles in mice and humans***
149 Overall, in the mESCs, the percentage of TF motif enrichment (excluding U motifs and U like
150 motifs) at both upstream and downstream regions highlights that TF motifs are predominantly
151 associated with inaccessible (71%) rather than accessible (59%) sites (**Figure 1A**). A similar
152 analysis in human WTC11 cells revealed that in contrast to mESCs, TF motifs are more enriched
153 overall at accessible (61%) rather than inaccessible (41%) sites (**Figure 1A**). The differences
154 between human and mouse could reflect the fact that there are proportionately more accessible
155 CTCF binding sites in human compared to mouse (**Figure S1B** and **Table S1 A-B**).
156
157 ***There is redundancy in the TFs that can bind a particular window***
158 In **Figure 1A**, TFs with distinct motifs (green stacked bars) can be distinguished from U and D
159 motifs (red and blue stacked bars, respectively). For each window, the TF motifs are shown in
160 **Figures S2-S3** for mESCs and **S4-S6** for human WTC11 cells. In the U window, TFs that
161 resemble U motifs are also detected, but the latter are less enriched in human compared to mouse
162 cells (2.79% and 1.73% in WTC11 versus 25.5% and 11.85% in mEScs at accessible and
163 inaccessible sites, respectively) as shown in **Figures S2A,** and **S4A-B**. Although CTCF is highly
164 conserved between human and mouse, this result suggests that there are differences in their
165 binding sites, such that in human there is less reliance on binding of ZFs 9-11 to the U motif.
166 Motifs of other TFs enriched in the U and D windows and spacers and their preference for
167 enrichment at accessible versus inaccessible sites are shown in **Figure 1B, S2-S3** for mESCs
168 and **Figure 1B, S4-S6** for human WTC11 cells. The frequency with which each TF motif is found
169 is shown as a percentage next to the motif in the figure. The majority of CTCF sites are enriched
170 in one or more TF motifs, with each factor having a small contribution. It is pertinent that motifs in
171 each window and group of TFs show similarities within and between TF families such as the ETS
172 family in the U spacer, or the HOX family in the U window, indicating there is redundancy in the
173 factors that could bind, thereby providing a potential mechanism for cell type-specific regulation.
174 TF enrichment for each window in WT mESCs is shown in **Table S2A**, while in **Table S2B,** TF
175 enrichment in WTC11 cells and other human tissues is shown to highlight that TF motif enrichment
176 at CTCF sites is detected across different cell types, but as expected, motif enrichment of TFs
177 varies, which could reflect each cell's transcription program.

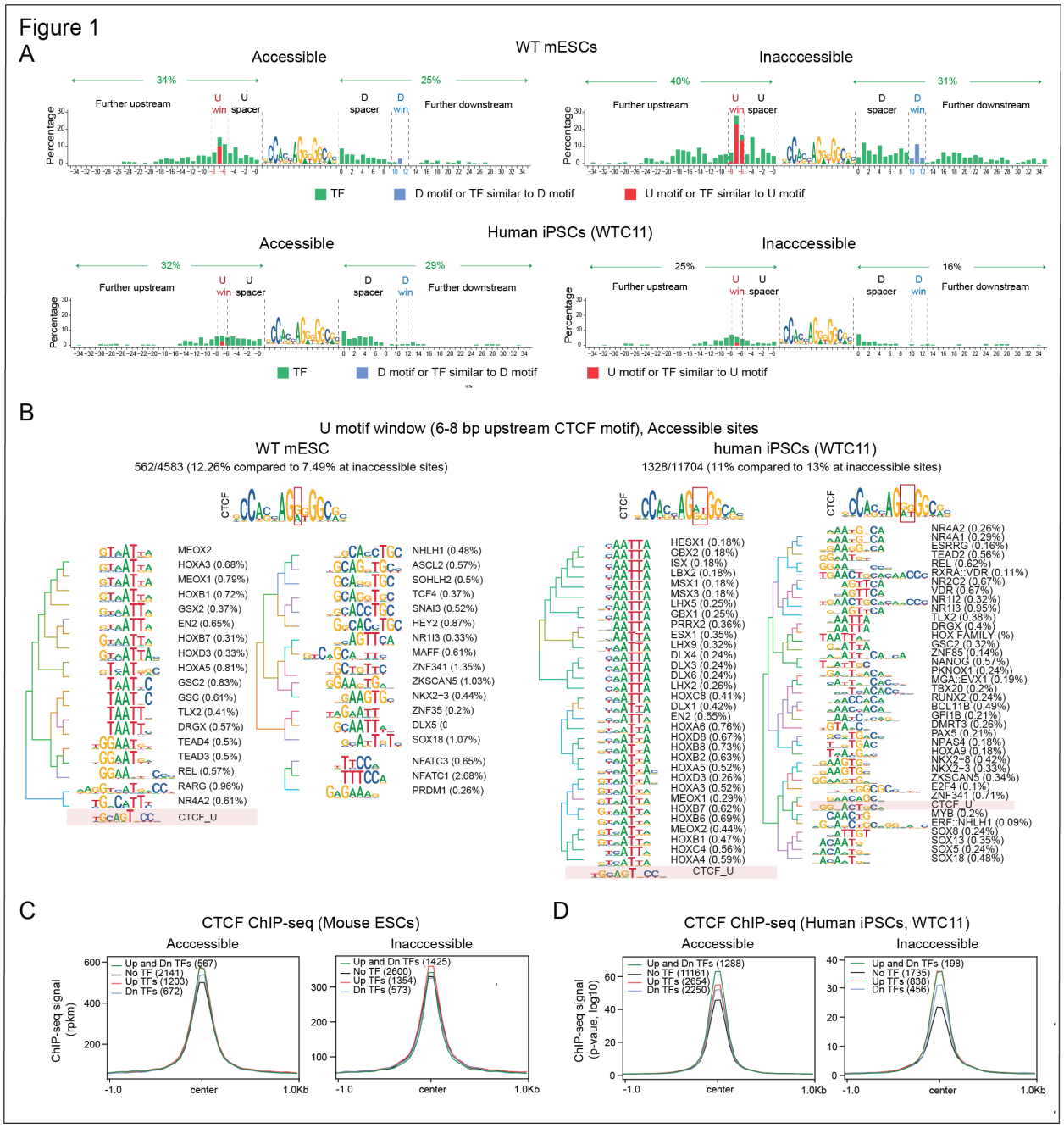


Figure 1: Bound CBSs are enriched in TF motifs within 35bps of the CTCF motif in mouse and human. A: Bar graphs showing the percentage of enriched TF motifs found using spacing motif enrichment analysis (SpaMo) at accessible and inaccessible sites in WT mESCs (top) and human iPSCs (bottom). The x-axis shows the gaps/distances in bp between the end of the TF motifs and the start of the CTCF motif for upstream gaps, or the start of the TF motif and the end of the CTCF motif for downstream gaps. For the U and D motif windows, the percentage of U and D motifs and TFs similar to the U and D motif are distinguished from other TF motifs (red or blue stacked bars, respectively). **B**: Motif logos of the enriched TF motifs found in the U motif window at accessible sites in mESCs (left) and human iPSCs (right). TF motifs ressembling the U motif and enrichment for the other windows are shown in **Figure S2-S6**,. **C-D**: CTCF binding profiles from ChIP-seq at accessible and inaccessible sites in mESCs (**C**) and human iPSCs (**D**) at CBSs containing at least one enriched TF motif within 35 bp upstream, downstream or both upstream and downstream compared to CBSs without any TF motifs.

179 ***The presence of TF motifs abutting CTCF binding sites is associated with stronger CTCF***
180 ***binding***
181 To determine whether enrichment of TF motifs is associated with changes in CTCF signal
182 strength, we compared the binding profiles of CTCF sites where we identified TF motifs at
183 upstream or downstream regions to sites without TF motifs. The presence of TF motifs at
184 accessible CTCF bound sites in the upstream and downstream regions, or both together were
185 associated with increases in CTCF signals (**Figure 1C**). The overall CTCF background profiles
186 were lower at inaccessible CTCF bound sites (**Figure 1C**) and the impact of TFs upstream or
187 downstream had a distinct effect compared to accessible sites. The differential effect sizes could
188 be explained by the predilection of TF classes found within specific windows as shown in **Figure**
189 **1B, S2-S3** and **S5-S6**. Importantly, enrichment of TF motifs is associated with a similar increase
190 in CTCF binding strength in human WTC11 cells as observed in mESCs, suggesting that TFs
191 play an equivalent role in stabilizing CTCF in the two species (**Figure 1D**). It is of note that motifs
192 in up and downstream windows contribute in a distinct manner, in the two species, likely due to
193 their distinct enrichment profiles.
194
195 In sum, the data in **Figure 1** shows that in human and mouse the majority of CTCF bound sites
196 are enriched in one or more TF motifs. Their enrichment appears to be functionally important for
197 increasing the stability of CTCF binding. Motifs within each window and group of TFs show
198 similarities within and between TF families indicating redundancy in the cofactors that can bind.
199 However, they are degenerate, thereby providing a mechanism for cell type-specific differences
200 in regulation. These findings suggest that cell type-specific binding of CTCF might be influenced
201 by a large pool of TFs that reflect the transcriptional program of each cell type, and our analyses
202 provide a resource for identifying which TFs have the potential to influence CTCF binding and
203 chromatin organization in different cell types (**Tables S2A, B**).
204
205 **Cobound TFs strengthen CTCF signals**
206 We have shown that enrichment of TF motifs in the 35bps surrounding CTCF bound sites is
207 associated with stronger CTCF signal (**Figure 1C** and **D**). To confirm the relationship between TF
208 binding and CTCF signal strength, we analyzed ChIP-seq binding of individual TFs (KLF4, OCT4,
209 MYC, SOX2 and YY1) in mESCs using data downloaded from O'Dwyer *et al*., and Laugwitz *et*
210 *al*., [20,21]. Our analysis, focused on CBSs with the consensus CTCF motif, highlighting sites at
211 which CTCF binds alone, CTCF and TFs cobind and TFs bind alone. Heatmaps and profiles for
212 accessible and inaccessible sites are shown in **Figure 2A** and **Figure S7**, respectively. The ChIP-

seq profiles validate our motif enrichment analysis, demonstrating that individual TFs stabilize CTCF binding, with an increased effect at inaccessible sites for most TFs.

The heatmaps and profiles also demonstrate that CTCF free CBSs occupied by TFs are associated with a stronger TF signal than at cobound sites for most TFs analyzed (**Figure 2A and**

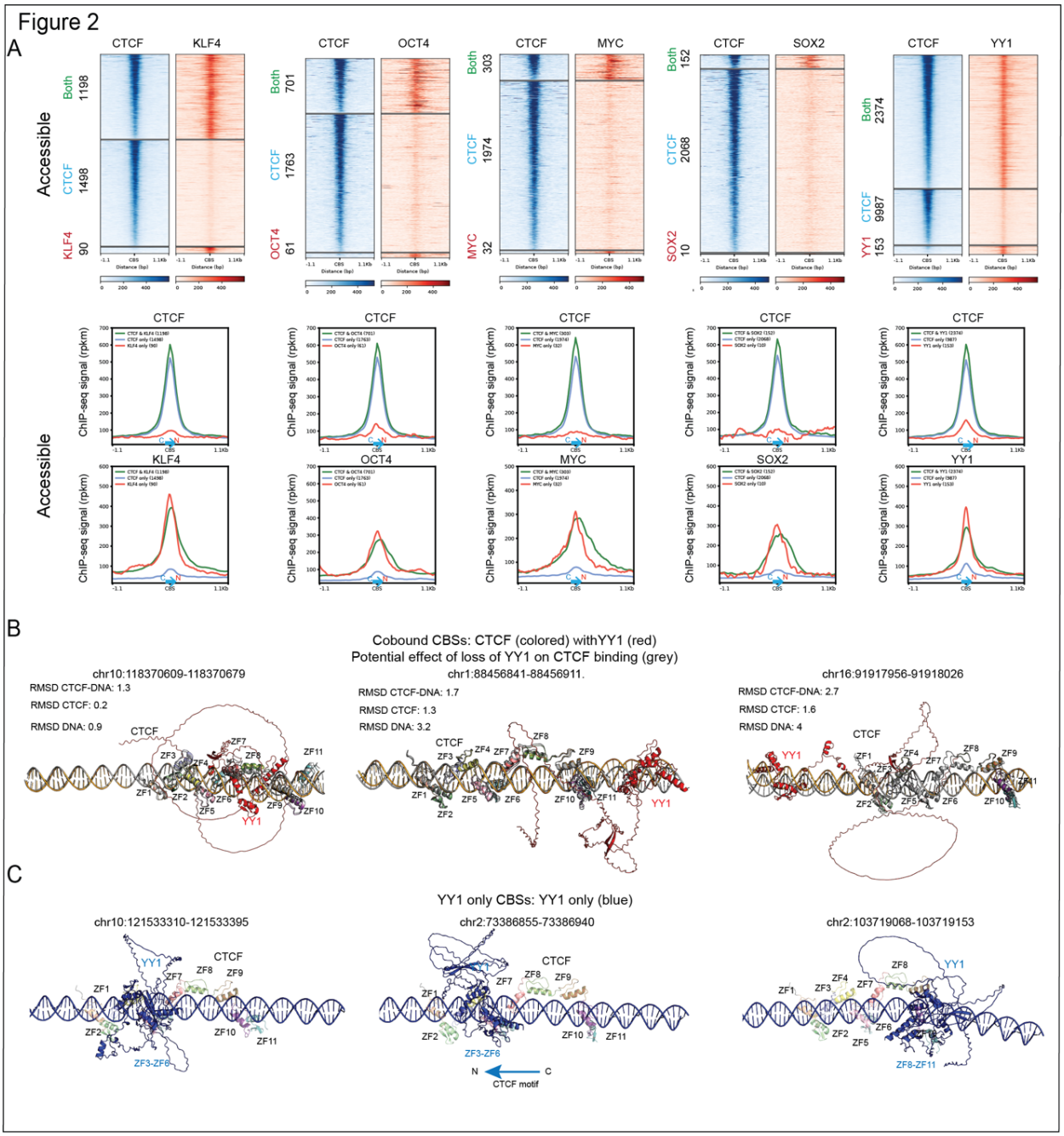


Figure 2: Individual TFs strengthen CTCF signal and disrupt CTCF-dependent nucleosome phasing. A: Heatmaps and profiles of CTCF (FLAG) and 5 TFs (including 4 pioneer TFs, KLF4, OCT4, MYC and SOX2) showing the effect of (i) TF binding on CTCF signal strength and (ii) CTCF binding on TF signal strength at sites where they are co-bound (green) compared to CTCF only (blue) and TF only (red) sites at accessible regions. This analysis is restricted to CTCF binding sites with the consensus CTCF motif and profiles are oriented on C to N terminal CTCF binding sites. The profiles at inaccessible sites are shown in **Figure S7**. **B**: AlphaFold3 predictions of the CTCF, YY1 and DNA complex at three cobound sites. The structures of CTCF with (colored) or without YY1 (grey) were aligned and compared by calculated the RMSD (root mean standard deviation) for the CTCF-DNA complex and separately for CTCF and DNA. **C**: AlphaFold3 predictions of the YY1 and DNA complex at three YY1 only bound CBSs. YY1 is shown in blue. CTCF binding predictions are overlayed to visualize the position of YY1 relative to CTCF. Predictions for the other TFs are shown in **Figure S8**.

S7). The peak summits of the TFs are almost centered on the C to N terminal oriented CTCF motif, suggesting a competition between CTCF and the TF. At cobound sites however, we observe binding of the TFs at sites overlapping as well as on either side of the CBS.

AlphaFold3 predicts CTCF/TF binding conformation at cobound and CTCF free CBSs

To get a clearer picture of how TFs and CTCF could affect each other's binding, we turned to AlphaFold3 and analyzed TF binding at cobound and TF bound CTCF free CBSs for each of the five TFs, KLF4, OCT4, MYC, SOX2 and YY1. In **Figure 2B**, we show three examples of YY1 binding with CTCF at cobound sites. YY1 can bind in the Upstream spacer between ZFs 7 and 9 as well as the upstream or downstream regions. At these locations, there is little change in the CTCF conformation as reflected by the small change in the route mean square deviation (RMSD), however cobinding can affect DNA conformation which could result in changes in the CTCF-DNA interaction.

When we examined AlphaFold3 predictions of YY1 binding (blue) at CTCF free CBSs, we observed a direct competition between YY1 and CTCF at their predicted binding sites (**Figure 2C**), which could explain the absence of bound CTCF. Examples showing AlphaFold3 predictions of how CTCF and the remaining TFs could affect each other's binding are shown in **Figure S8A-H**. As with YY1, each TF binds upstream, downstream or within the U spacer when cobound with CTCF, while at CTCF free CBSs, TFs tend to bind to sites overlapping the CBS, preventing CTCF binding. We also show examples of OCT4/SOX2 and MYC/MAX cobinding with CTCF. Interestingly, at the cobound locus shown in **Figure S8G,** the binding of SOX2 causes considerable DNA bending, which has been previously described [22]. At this locus, the predicted binding is in the Upstream spacer between ZFs 7 and 9 which is flexible [23], allowing ZFs 9-11 to be properly positioned on the DNA despite the DNA bending, highlighting the importance of ZF8 as a spacer.

Allele-specific analysis reveals both the cooperativity and competitiveness of CTCF-TF cobinding

To further validate the relationship between TFs and CTCF, we first downloaded the 145 ENCODE TF ChIP-seq sets available for the GM12878 cell line (**Table S3**) and found that 69% of the CTCF peaks overlap at least one TF, confirming that co-binding of CTCF with another TF is a frequent event. We then performed allele-specific analysis using ENCODE data from GM12878 to identify SNPs that impact CTCF or TF motifs (and therefore their binding affinity) (**Table S4**). In **Figure 3A, B** we show examples of heterozygous SNPs at two separate CTCF binding sites. In both cases CTCF binding is reduced on the allele with the base change, coupled with a loss of the cobound TF, overlapping RAD21 (a component of the cohesin complex) and to a lesser extent altered ATAC-seq signals. Thus, CTCF binding can affect binding of the TF, suggesting that in some cases CTCF binding might be required to open up the site for TF binding. Of note, the SNP in **Figure 3A** is linked in GWAS to colon cancer and cardiometabolic traits.

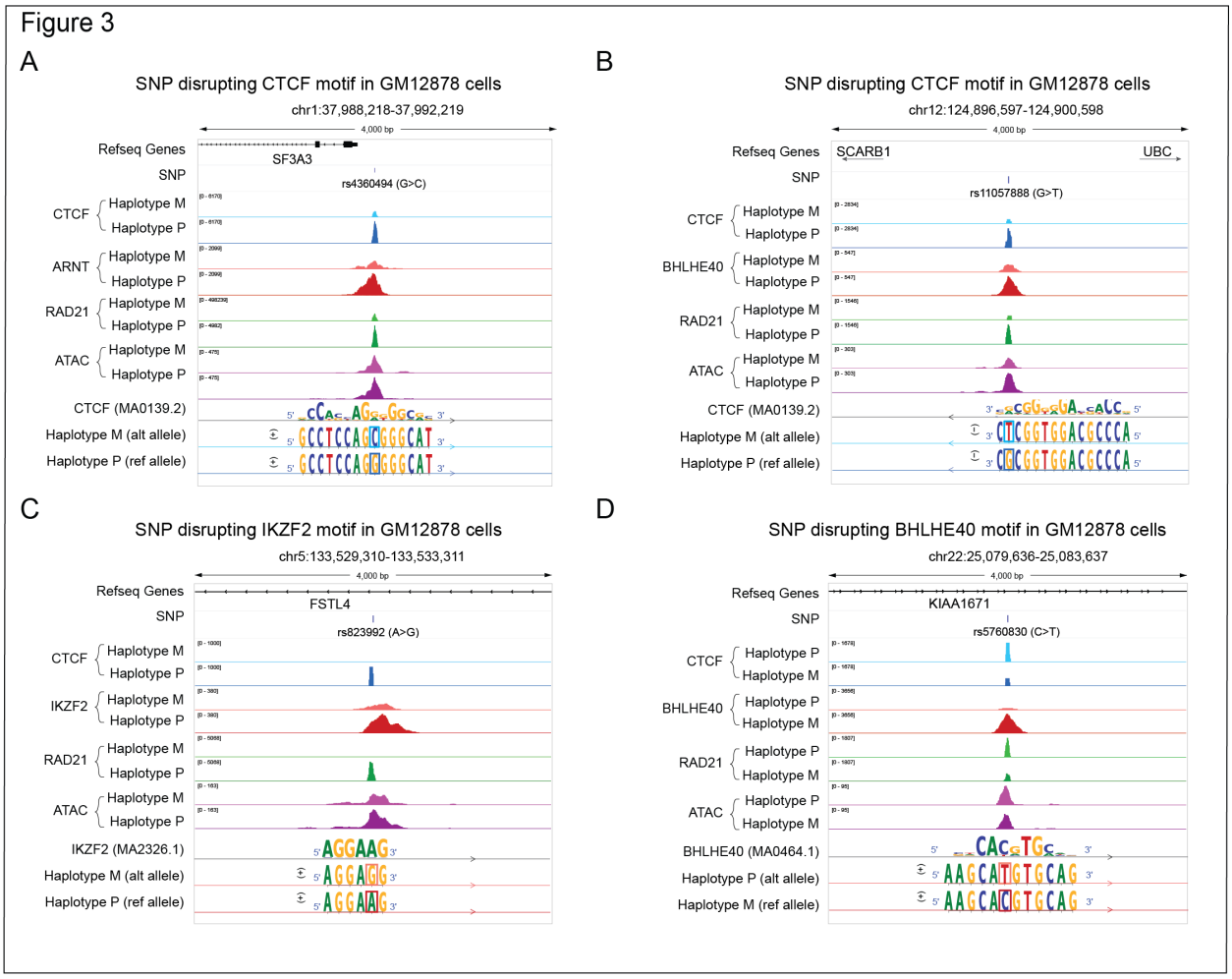


Figure 3. **SNPs reveal both the cooperativity and competitiveness of CTCF-TF cobinding**. **A**: Map showing allele-specific binding (ASB) of CTCF due to a SNP disrupting its motif on the maternal allele in the GM12878 cell line. RAD21 and ARNT binding are also disrupted on the CTCF unbound allele. The CTCF consensus motif and the sequence at this locus are shown on the bottom track along with the orientation. The rectangles highlight the heterozygous SNP. For visualization purpose, the sequence tracks are not to scale. M, maternal; P, paternal. **B**: Map showing ASB of CTCF due to a SNP disrupting its motif on the maternal allele, which is associated with the disruption of BHLHE40, cohesin binding and chromatin accessibility on the same allele. **C**: Map showing ASB of IKZF2 due to a SNP disrupting its motif on the maternal allele. CTCF, cohesin binding and to a lesser extent the chromatin accessibility are also disrupted on the IKZF2 unbound allele. **D**: Map showing ASB of BHLHE40 due to a SNP disrupting its motif on the paternal allele. At this locus, the unbound BHLHE40 allele is associated with CTCF and cohesin binding as well as more open chromatin. None of these regions overlap with known imprinting regions and the observed ASBs are likely be cis-dependent ASBs rather than parent-of-origin dependent.

Conversely, SNPs that affect binding of TFs on one allele can impact binding of CTCF at overlapping sites. In **Figure 3C**, the base change in the maternal haplotype impairs IKAROS binding, with a concomitant reduction in CTCF, RAD21 and ATAC-seq signals. Thus, in this setting binding of the TF is required for opening up the site to enable CTCF binding. In contrast, the SNP disrupting the BHLHE40 motif and binding on the paternal allele at the *KIAA1671* locus is permissive for CTCF binding and associated with concomitant binding of RAD21 and more open chromatin (**Figure 3D**). This highlights an example of a TF that binds competitively with CTCF and together with the data in **Figure 3B**, demonstrates that a single TF can have both a stabilizing and destabilizing effect as shown in **Figure 2A**.

Overall, we identified 131 regions with allele-specific binding of CTCF that are associated with allele-specific binding of a cofactor and where the presence of a SNP disrupting either the CTCF or the cofactor motif provides a causal explanation. In half of these cases, the perturbation introduced by the SNP indicates that TF co-binding impacts CTCF binding, while in the other half, the effect is reversed, confirming the reciprocal nature of TF-CTCF co-binding (**Table S4**). Additionally, we observed a destabilizing effect in 10% of the cases, suggesting that binding cooperativity is more prevalent than competitiveness.

Transcriptional reprogramming contributes to cell state-specific CTCF binding

Our analyses demonstrate that cofactor binding can influence CTCF's binding stability, however the impact of changes in transcriptional programs on CTCF binding have not been investigated. To address this question, we compared TF expression levels at accessible and inaccessible sites in human IPSCs. Overall, we found that TFs that bind motifs abutting CTCF were expressed at a higher level at accessible sites (**Figure 4A**) and that those sites are more sensitive to changes in TF expression levels compared to inaccessible sites, as shown by the significant difference in the correlation slopes in **Figure 4B**.

To examine the effect of increasing transcriptional output on CTCF, we analyzed the effect of variations in TF levels across 84 different ENCODE cell lines on their enrichment at CTCF binding sites (**Table S2B**). We identified a subset of TFs whose expression levels showed a significant positive or negative correlation with motif enrichment around bound CTCF sites (**Figure 4C and Table S5**). This suggests that, in addition to context-specific effects, TFs might exert intrinsic stabilizing or destabilizing influences on CTCF binding and that the enrichment or depletion of a particular TF at CTCF sites in each cell type is dependent on its expression level, highlighting a mechanism by which cell type-specific transcriptional programs can impact CTCF binding profiles.

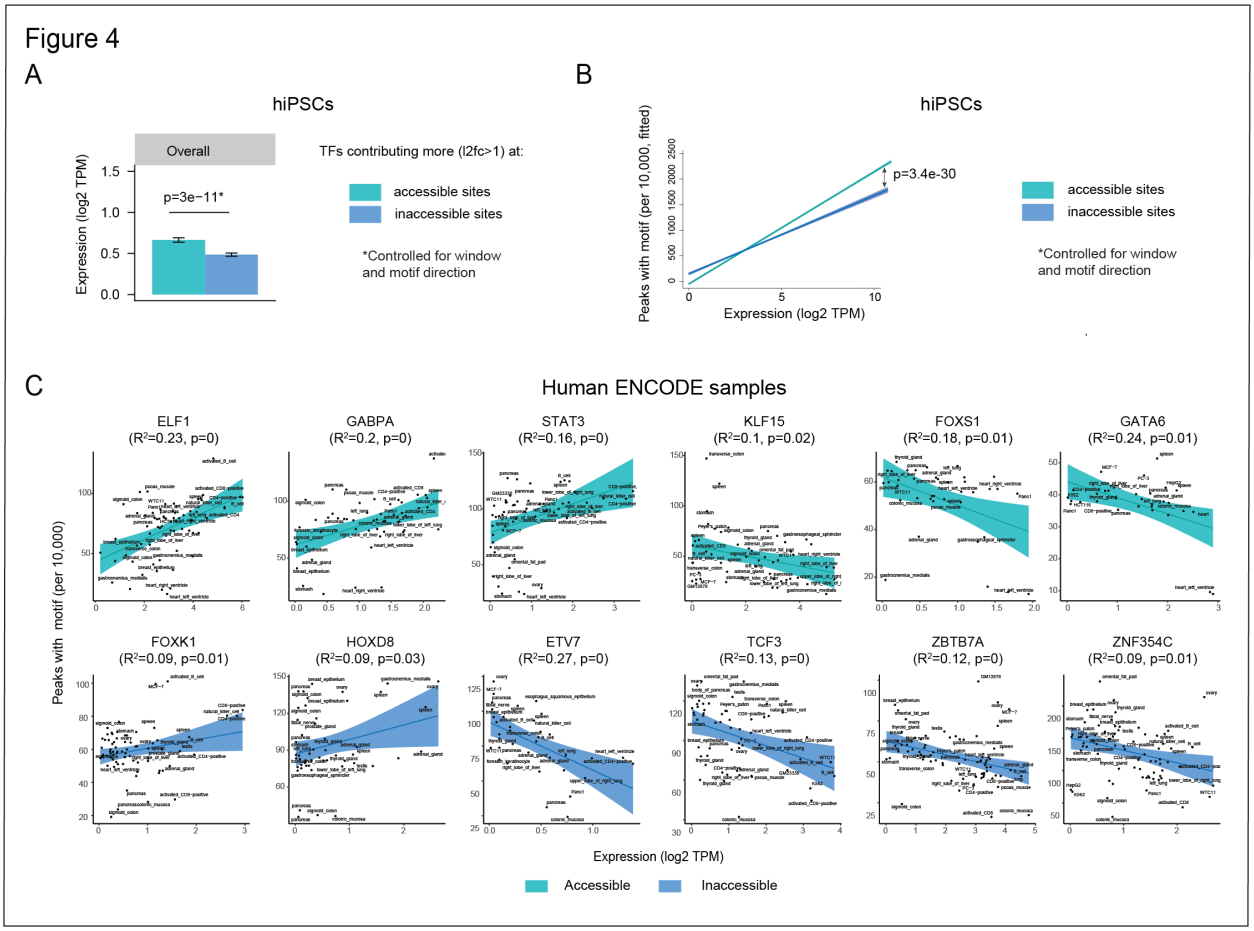


Figure 4: The percentage of TF motifs within 35 bp of CTCF consensus motif correlates with their expression level. A: Bar graph showing an increase in expression of TFs with a higher contribution at accessible versus inaccessible sites (≥2 fold differential contribution) in human iPSCs, after controlling for the differential effects of the gap window and direction relative to CTCF motif. The error bars correspond to the standard error computed using a multivariate linear model. **B**: Fitted correlation between overall TF expression and the percentage of CTCF sites found with motifs after controlling for the differential effect of windows and motif direction at accessible and inaccessible sites in human iPSCs. The shaded areas correspond to the 95%-confidence intervals. **C**: Examples of significant correlations between TF expression level and the percentage of CTCF peaks associated with the TF motif at accessible and inaccessible sites across human ENCODE cells and tissues. The shaded areas correspond to the 95%-confidence intervals.

298 **CTCF mediates nucleosome repositioning around the CBS**
299 To ask how CTCF binding affects nucleosome positioning in a context-specific manner, we
300 performed long-read Nano-NOMe-seq in mESCs, which allowed us to assess nucleosome
301 positioning using exogenous 5mC at HGC (GC excluding CGC) at the single molecule (SM) level.
302 To assess the effect of CTCF motif orientation and to discriminate molecules with occupied and
303 unoccupied CTCF binding sites, we restricted our analyses to CTCF peaks with GpC-containing
304 CTCF motifs after confirming that nucleosome occupancy at these sites were similar (**Figure**
305 **S9A**).
306
307 Unlike other similar analyses, we separated our profiles into occupied and unoccupied molecules.
308 At CTCF ChIP-seq peaks, a substantial fraction of molecules was temporarily unoccupied, which
309 is expected since DNA-protein interactions are dynamic. By comparing temporarily unoccupied
310 to occupied molecules at CTCF bound sites - thereby controlling for context-specific confounding
311 factors between bound and unbound CBSs - we confirmed previous findings demonstrating the
312 displacement and phasing of flanking nucleosomes after CTCF binding (**Figure 5A**).
313 Nucleosomes appeared less phased at inaccessible compared to accessible CTCF sites and
314 exhibited slightly shorter periodicity as well as a reduced CTCF flanking nucleosome-free region
315 (NFR) (**Figure 5A**).
316
317 **Subdivision of CTCF binding sites into 14 clusters reveals a continuum of binding states**
318 **at aggregated and individual peaks**
319 To further understand the binding dynamics of CTCF binding on chromatin, we performed
320 unsupervised clustering of the single molecules, incorporating data from all experimental
321 conditions in mESC across both accessible and inaccessible regions. This generated 14 clusters,
322 based on 50 bp-binned nucleosome occupancy within 1 kb of CBSs (**Figure 5B**). Differential
323 enrichment of each cluster in specific conditions and genomic contexts underscores the biological
324 relevance of each (**Figure S9B**). The UMAP projections of clustered SMs revealed that UMAP 1
325 correlates with global nucleosome occupancy whereas UMAP 2 partially captures differences in
326 nucleosome phasing patterns, despite only subtle variations in phasing scores (**Figure 5B,** right
327 and **S9C**).
328
329 Trajectory inference using Slingshot [24] helped define the relationships and dynamics among
330 clusters. For this analysis, we set cluster 1 as the starting point, as the molecules in this cluster -
331 along with clusters 2 and 3 (black clusters) - exhibit the characteristics of highly accessible, newly

332 replicated chromatin, previously shown to lack well-defined nucleosome organization [25].
333 Interestingly, cluster 1 displays a weak CTCF footprint with no surrounding nucleosome phasing,
334 while clusters 2 and 3 exhibit weak phasing either on the left or right of the CTCF footprint, a
335 distinction that could reflect the direction of replication. This observation suggests that CTCF can
336 act as a barrier during redeposition of nucleosomes on nascent chromatin and might help to
337 organize nucleosomes on newly replicated DNA.
338
339 Slingshot identified a bifurcation point corresponding to stable CTCF binding (red clusters 9-10)
340 with two divergent branches (green and blue lines) along the UMAP 2 axis, leading to violet
341 (clusters 4-8) and light blue (clusters 11-13) groups, respectively (**Figure 5B**). This branching
342 pattern may reflect alternative temporary nucleosome configurations associated with CTCF
343 binding. Indeed, the red clusters, enriched at strong CTCF peaks, show a pronounced CTCF
344 footprint and well-defined flanking NFR, consistent with CTCF-associated nucleosome
345 repositioning [10]. The **violet** clusters (4-8) show a CTCF footprint on the left, middle or right side
346 of a nucleosome, supporting findings that CTCF can bind its motifs regardless of nucleosome
347 position [26]. In these clusters, the close proximity of nucleosomes to the CTCF motif and the
348 presence of partial NFRs suggest a less stable CTCF binding – supported by the enrichment in
349 weak CTCF peaks - or intermediate states prior to, or during the action of nucleosome remodelers
350 by CTCF [10]. The **light blue** clusters (11-13) maintain nucleosomes in an organized, phased
351 manner but with a narrower NFR around unbound CTCF sites (lacking a footprint), possibly
352 reflecting temporally unbound CTCF. A third branch (orange) along UMAP1 leads to the **dark**
353 **blue cluster** (14), which is enriched in the CTCF-depleted condition following IAA treatment. This
354 might represent sites at which CTCF rebinds weakly or does not return at all, leading to disruption
355 of nucleosome phasing. Finally, the purple path reflects the cyclical process between nascent
356 and mature chromatin configurations.
357
358 Importantly when we focus on individual CTCF peaks we can detect all fourteen clusters, as
359 shown for three representative examples (**Figure 5C**). These analyses are important for
360 demonstrating that the clusters are not simply an aggregation of data from different binding sites
361 enriched for one cluster or another across the genome. Instead, they represent a continuum of
362 different states of CTCF binding at each peak. The proportion of each cluster is shown for twelve
363 individual binding sites (**Figure 5D**).

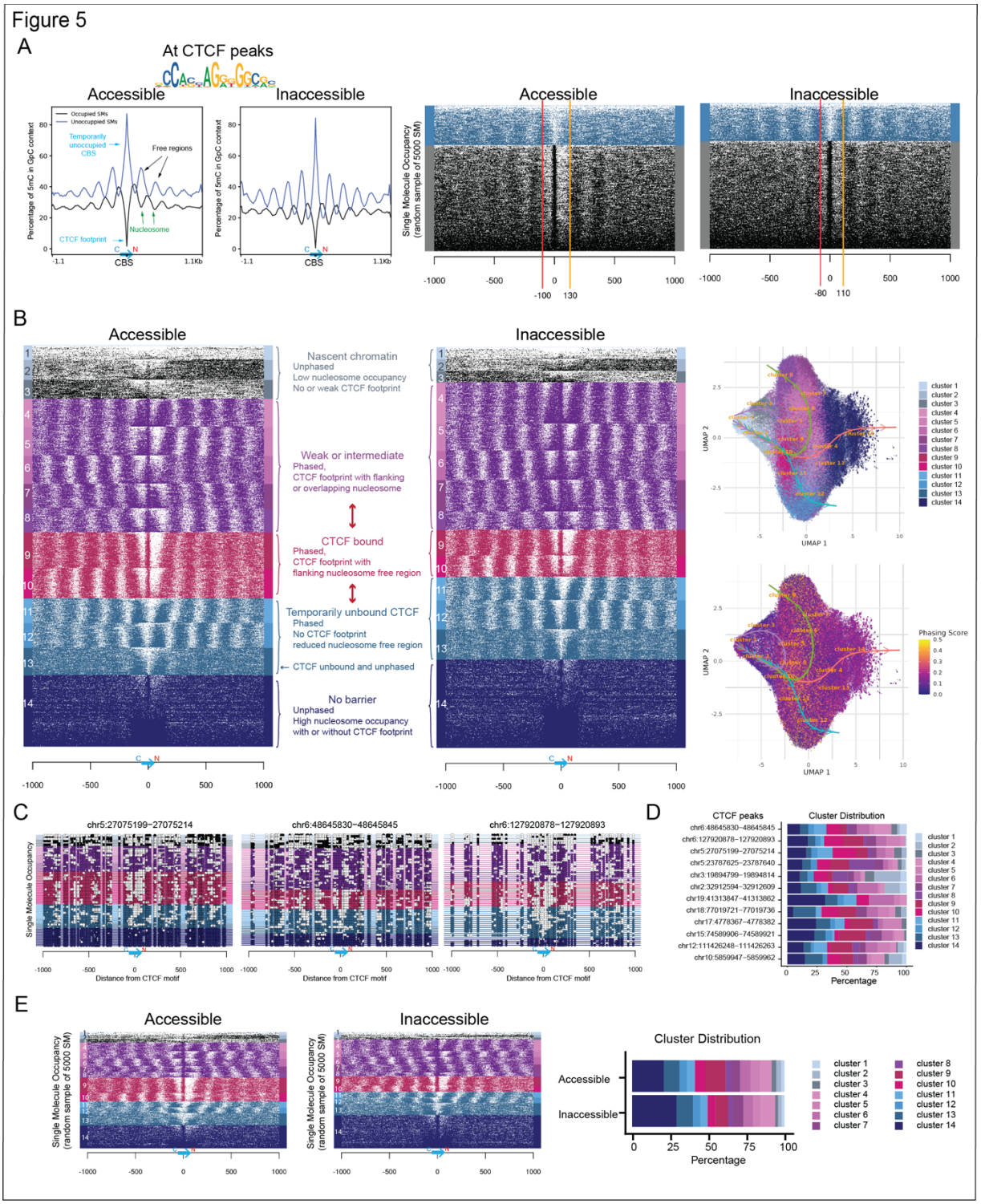


Figure 5
A
At CTCF peaks
Accessible
Inaccessible
Occupied SMs
Unoccuppied SMs
Temporarily unoccupied CBS
Free regions
Nucleosome
CTCF footprint
Percentage of 5mC in GpC context
CBS
Single Molecule Occupancy (random sample of 5000 SM)
B
Nascent chromatin
Unphased
Low nucleosome occupancy
No or weak CTCF footprint
Weak or intermediate
Phased,
CTCF footprint with flanking or overlapping nucleosome
CTCF bound
Phased,
CTCF footprint with flanking nucleosome free region
Temporarily unbound CTCF
Phased
No CTCF footprint
reduced nucleosome free region
CTCF unbound and unphased
No barrier
Unphased
High nucleosome occupancy with or without CTCF footprint
UMAP 1
UMAP 2
Phasing Score
C
chr5:27075199−27075214
chr6:48645830−48645845
chr6:127920878−127920893
Distance from CTCF motif
D
CTCF peaks
Cluster Distribution
Percentage
E
Accessible
Inaccessible
Cluster Distribution
Percentage

364

Figure 5: Subdivision of CTCF binding sites into 14 clusters reveals distinct phases of binding. A: Profiles and single molecule heatmaps showing the nucleosome occupancy (NO) and CTCF binding footprinting from nano-NOMe-seq for accessible and inaccessible CTCF sites. Both occupied (black) and temporarily occupied (blue) molecules at CTCF bound sites (ChIP-seq peaks) are shown. The data are centered on the 15-bp oriented consensus motif. The y-axis of the profiles shows the percentage of exogenous 5mC detected in an HGC context. High exogenous methylation reflects TF or nucleosome free chromatin accessible to the GpC methyltransferase, while lower methylation reflects occupied regions. The heatmaps show the nucleosome occupancy and CTCF footprinting at the single molecule (SM) level for a random sample of 5000 molecules. Black (or blue) and white dots represent occupied and free GpCs, respectively. The red and orange lines correspond to the stable position of the first upstream and downstream nucleosome flanking the CTCF motif. **B**: SM heatmaps of 5000 molecules per condition and genomic context showing the nucleosome occupancy and CTCF binding footprinting at accessible and inaccessible CTCF sites by cluster (left), and UMAP of the SMs colored by cluster and phasing scores (right). The arrowed lines show the SM trajectories identified by Slingshot. **C**: Example of heatmaps showing the nucleosome occupancy and CTCF binding footprinting at three individual CTCF peaks. **D**: Bar graph showing examples of cluster distribution at individual CTCF peaks. **E**: SM heatmaps and bar graphs showing the representative proportion of each cluster at accessible and inaccessible sites.

Analysis of clustering at accessible and inaccessible sites highlights that the narrowing of the NFR surrounding CTCF seen in **Figure 5A,** is reflected by a decrease in the red clusters (strongly bound CTCF) and an increase in Cluster 14 (unbound or weakly bound CTCF with no nucleosome repositioning (**Figure 5E**). Additionally, a reduced ability to reposition nucleosomes is reflected by a more robust nucleosome presence on either side or overlapping the CBS (violet clusters). In sum, single molecule nano-NOMe-seq data reveals the dynamics of CTCF binding states at both aggregated and individual peak level, highlighting CTCF's ability to reposition nucleosomes to create a NFR surrounding its binding sites and to act as a barrier that facilitates nucleosome phasing, both of which are impaired at inaccessible CBSs.

Unoccupied CBSs are bound by nucleosomes or other factors

We next analyzed CBSs that are not bound by CTCF in mESCs. Most of the molecules at those sites are occupied (59% and 72% respectively at accessible and inaccessible sites), either by a nucleosome or a TF as suggested by the persistence of a narrow footprint (**Figure 6A**). Nucleosomes are still phased around the CBS but there is a noticeable increase in the proportion of violet and light blue clusters and a decrease in the red cluster compared to CTCF bound sites (**Figure 5E**). These data indicate that binding of TFs to CTCF free sites can still phase nucleosomes, but these are not associated with a NFR flanking the CBS. As before, we see a more robust nucleosome presence at inaccessible sites, consistent with less TF binding at these regions.

CTCF is required to reposition and phase nucleosomes

To confirm the effect of CTCF on nucleosome phasing and repositioning, we used an mESC CTCF auxin inducible degron [9]. Loss of CTCF protein at 3, 18 and 48 hours resulted in an

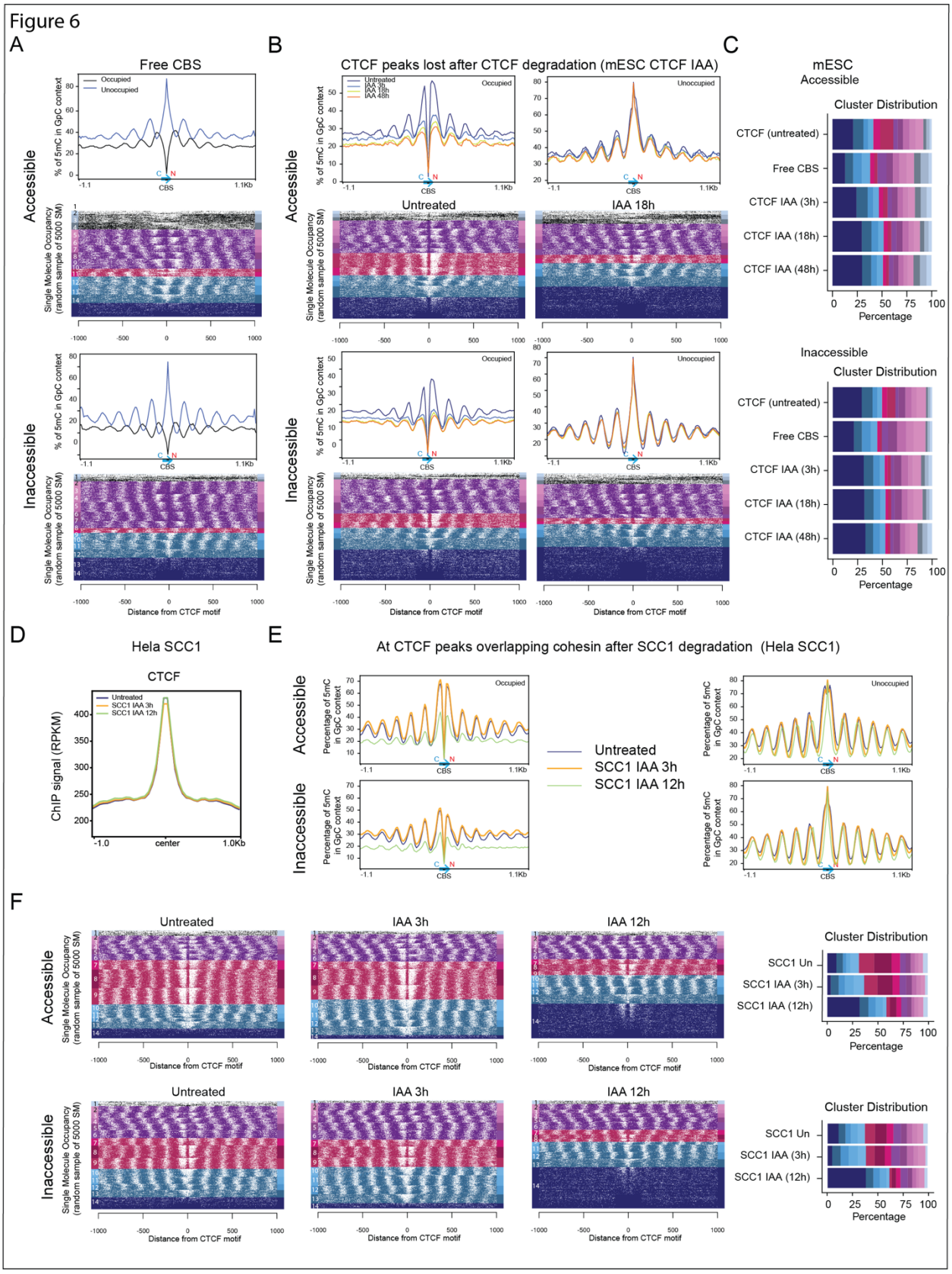


Figure 6
A
Free CBS
Occupied
Unoccupied
% of 5mC in GpC context
Accessible
Inaccessible
Single Molecule Occupancy (random sample of 5000 SM)
Distance from CTCF motif
B
CTCF peaks lost after CTCF degradation (mESC CTCF IAA)
Untreated
IAA 3h
IAA 18h
IAA 48h
Occupied
Unoccupied
Untreated
IAA 18h
C
mESC
Accessible
Cluster Distribution
CTCF (untreated)
Free CBS
CTCF IAA (3h)
CTCF IAA (18h)
CTCF IAA (48h)
Percentage
Inaccessible
D
Hela SCC1
CTCF
Untreated
SCC1 IAA 3h
SCC1 IAA 12h
ChIP signal (RPKM)
center
E
At CTCF peaks overlapping cohesin after SCC1 degradation (Hela SCC1)
Percentage of 5mC in GpC context
Untreated
SCC1 IAA 3h
SCC1 IAA 12h
F
Untreated
IAA 3h
IAA 12h
SCC1 Un
SCC1 IAA (3h)
SCC1 IAA (12h)

390

Figure 6: ***SCC1 mediates CTCF associated nucleosome repositioning and phasing.*** **A:** Profiles and SM heatmaps stratified by cluster showing the nucleosome occupancy and CTCF binding footprinting for CTCF-free and motif-containing accessible and inaccessible CBSs. CTCF-free binding sites are defined as FIMO-detected CTCF motifs without CTCF or SMC3 ChIP-seq peaks. **B:** Profiles (for untreated, 3, 18 and 24 hours) and heatmaps of untreated and 18 hour IAA degraded mESCs at accessible and inaccessible sites. The heatmaps for 3 and 48 hours are shown in **Figure S10**. **C**: bar graphs showing the distribution of each cluster at CTCF peaks, CBS and after CTCF degradation (IAA). **D**: CTCF ChIP-seq profiles in untreated human Hela cells and after SCC1 degradation at 3 and 12 hours. **E:** Profiles of nucleosome occupancy in untreated human Hela cells and after SCC1 degradation (3 and 12 hours) at accessible and inaccessible sites. **F**: SM heatmaps and bar graphs showing the distribution of each cluster in untreated Hela cells and after IAA degradation.

increasing reduction in nucleosome repositioning at CBSs (profiles in **Figure 6B**). The more granular analyses of the heatmaps and barplot show an increased proportion of cluster14 (dark blue), with a less marked footprint at the CBS over time (**Figure 6B, C** and **Figure S10A**). As expected, there is a notable decrease in the representation of the strong, red clustered CTCF binding sites and at the violet clusters, we detect a weaker CTCF footprint and a more robust nucleosome presence (**Figure S10A**). Interestingly, after CTCF degradation, CBSs show similar proportions of violet, red and light blue clusters to CTCF free sites (**Figure 4A**). Given these similarities, we speculate that the weak footprint at CTCF degraded sites could be due to the presence of incoming, or previously bound TFs, although we cannot rule out that some of this footprint is due to leftover, undegraded or weak CTCF [27].

SCC1 mediates CTCF associated nucleosome repositioning and phasing

Since CTCF degradation leads to loss of both CTCF and cohesin at CBSs (**Figure S10B**), the CTCF degron experiment could not be used to determine which factor is responsible for mediating the changes we observed after CTCF degradation. To address this question, we used an auxin inducible degron of SCC1 (a component of the cohesin complex) in the human Hela cell line [28]. Since cohesin is required for mitosis, we assessed early degradation time points, 3 hours and 12 hours, where we could still collect live cells. Alongside nano-NOMe-seq we performed CTCF ChIP-seq. CTCF binding profiles remained the same as untreated cells after IAA treatment at 3 hours and 12 hours (**Figure 4D**). In contrast, the nucleosome profiles for occupied and unoccupied sites show that although there is little change in nucleosome clustering at the 3-hour time point, despite unchanged CTCF profiles, we could detect a reduction in nucleosome repositioning and phasing at CBSs by 12 hours (**Figure 4E**), similar to that observed after CTCF degradation (**Figure 4B)**. Looking at the heatmaps and barplots (**Figure 4F)** at both accessible and inaccessible sites there is a noteable reduction in the strongly bound CTCF fraction (red clusters) and an increase at weakly bound CTCF sites characterized by disorganized nucleosomes surrounding the CBS (dark blue clusters). Violet clusters are also reduced but more

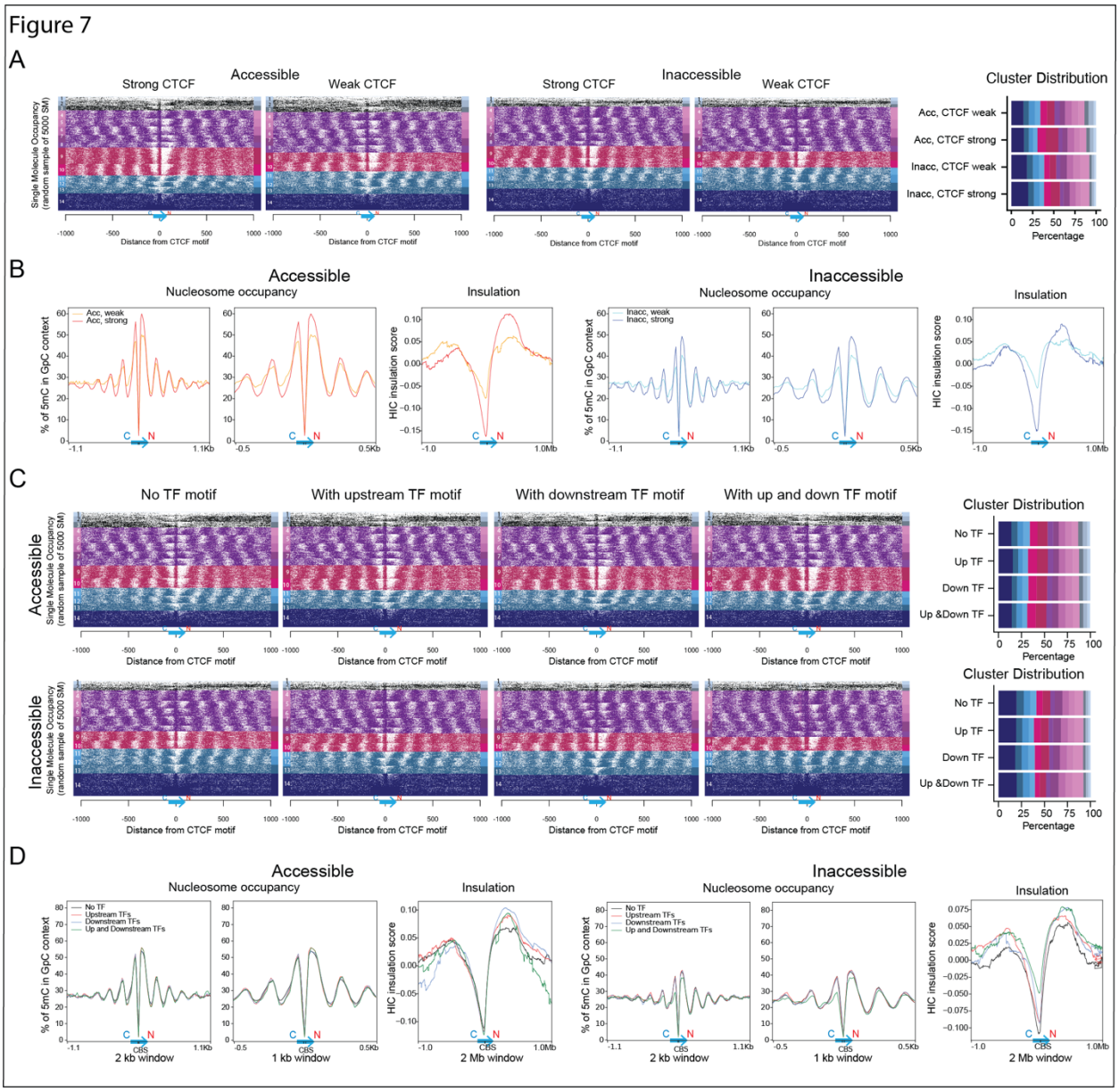


Figure 7: The CTCF flanking NFR is associated with insulation. A. Clustered heatmaps of SMs stratified by chromatin accessibility and CTCF signal strength. **B:** Profiles comparing the nucleosome occupancy and insulation score between weak versus strong CTCF sites stratified by chromatin accessibility. The averaged insulation score generated at 10 kb resolution from HIC data is shown for each genomic context. **C**: Heatmaps of SMs stratified by clusters at accessible and inaccessible CTCF peaks in the absence or presence of TF motifs. The bar graphs show the distribution of each cluster by genomic context. **D**: Profiles comparing the nucleosome occupancy and insulation score between CTCF-alone and cobound sites stratified by chromatin accessibility.

so at inaccessible regions, and at these sites there is a stronger nucleosome presence surrounding the CBS, obscuring the CTCF footprint in most places. That the 12-hour time point effect occurs in the background of unaltered CTCF binding, indicates that cohesin plays a dominant role in nucleosome repositioning and phasing. However, CTCF alone may exert an

425 additional weaker effect, as the red clusters are not reduced to the same extent as when CTCF
426 is depleted and there is a persistent NFR in the dark blue cluster. We speculate that the delay in
427 loss of nucleosome repositioning at 3 hours could be because nucleosome repositioning needs
428 to be reset during S phase [29].
429
430 In sum, we found that CTCF free CBS, CTCF and SCC1 degradation all give rise to similar
431 clustering profiles with a decreasing NFR over time, indicating that a reduction or absence of
432 CTCF binding can allow binding of TFs that can phase, but are unable to reposition surrounding
433 nucleosomes. Thus, cohesin and loop extrusion could mediate nucleosome positioning and
434 phasing by creating or stabilizing barrier effects at CTCF anchors.
435
436 **CTCF signal strength contributes more to nucleosome repositioning and insulation than**
437 **accessibility**
438 When we compare the clustering of strong versus weak CTCF signals and inaccessible versus
439 accessible sites, we observed that signal strength gives rise to a bigger difference in the
440 proportion of red clusters (strong CTCF binding in a NFR), compared to accessibility (**Figure 7A**).
441 Moreover, the differential effects of CTCF signal strength were similar at both accessible and
442 inaccessible sites, suggesting that CTCF signal strength contributes to nucleosome repositioning
443 independent of accessibility. We previously performed a stratified analysis showing that CTCF
444 signal strength, which was associated with cohesin stalling, also contributes more to insulation
445 than accessibility [9]. Here we highlight that a reduction in nucleosome repositioning at weak versus
446 strong sites is associated with a corresponding reduction in chromatin insulation compared to that
447 at inaccessible versus accessible sites (**Figure 7B**).
448
449 ***Surrounding motifs affect CTCF-mediated nucleosome repositioning and insulation***
450 ***independent of signal strength***
451 The presence of TF motifs are associated with differential effects on CTCF binding (**Figure 1C**).
452 Although CTCF signal strength is increased at accessible sites, we observe a minimal change in
453 nucleosome repositioning and chromatin insulation (**Figure 7C, D**). The same is true at
454 inaccessible sites, except when TF motifs are enriched on either side of CTCF. Despite a slight
455 increase in CTCF signal strength at these sites, both nucleosome repositioning and chromatin
456 insulation are reduced. Together, these data indicate that CTCF and cohesin signal strength can
457 be uncoupled from the size of the NFR surrounding CTCF, while the association between
458 nucleosome repositioning and insulation is maintained. This finding highlights that nucleosome

459 repositioning is a more reliable predictor of chromatin insulation than CTCF signal strength. The
460 data also demonstrate that TF binding can interfere with nucleosome repositioning and chromatin
461 insulation and we speculate that the presence of TFs could be a way to fine tune the effect of
462 CTCF signal strength on these two processes.
463
464 **Discussion**
465 Although it is known that a subset of CBSs is altered in a state-specific manner, the mechanisms
466 underlying this control have not been fully elucidated. Here we used CTCF ChIP-seq in mESCs
467 combined with CTCF ChIP-seq from ENCODE data in a wide range of human cell types to (i)
468 characterize accessible versus inaccessible sites, and (ii) determine what factors influence CTCF
469 binding stability and function within these two chromatin environments. Although CTCF is highly
470 conserved in mice and humans, we found different features associated with their respective
471 binding sites. Specifically, mouse cells have proportionately more inaccessible binding sites
472 compared to human cells. In parallel, TF motifs are enriched at inaccessible sites in mouse, which
473 we have previously shown generally have weaker CTCF signals and reduced cohesin overlap
474 compared to accessible sites [9]. Since TF cobinding stabilizes CTCF, we speculate that their
475 enrichment at inaccessible sites in mouse, restores the balance of strong versus weak
476 CTCF/cohesin signals in the two species, compensating for the relative reduction of accessible
477 sites in mouse.
478
479 CTCF has been shown by us and others to act as a pioneer factor, binding inaccessible as well
480 as accessible chromatin [9,30]. To follow up on previously published studies hinting at a role for TFs
481 in cell state specific binding of CTCF, we performed a comprehensive analysis of enrichment of
482 TF motifs in the 35bp region of the CBS in mESCs as well as 85 ENCODE cell lines. We identified
483 a wide range of motifs corresponding to expressed cell state-specific factors that could perform
484 this function. In human and mouse, the majority of CTCF bound sites are enriched in motifs
485 located within 35 bps of CTCF. Within a given window, enriched TF motifs show a strong similarity,
486 suggesting redundancy in the regulation of CTCF binding. Furthermore, motifs are degenerate
487 which could allow for differential effects of TFs from the same class or motif cluster. Given the
488 redundancy between factors that can bind, we speculate that at a single CTCF binding site, the
489 presence of distinct TFs in different cell types could fine-tune CTCF's stability and its ability to
490 block cohesin in a lineage or state specific manner. Indeed, we demonstrate that CTCF signals
491 are strengthened at sites enriched with TF motifs and by analyzing the effect of a handful of
492 cofactor TFs on CTCF binding profiles, we validate their stabilizing effect.

493
494 The significance of these findings is underscored by our analysis of human tissues from
495 ENCODE, all of which have unique transcriptional programs with individual TFs expressed at
496 different levels. Importantly, increases in the expression level of a TF can both increase or
497 decrease the enrichment of its motif at CTCF bound sites, depending on whether the TF has a
498 stabilizing or destabilizing effect. The relationship between TFs and CTCF were validated by allele
499 specific analysis which demonstrate that the two factors have a reciprocal effect on each other's
500 binding such that SNPs affecting either the binding of CTCF or the TF will impact enrichment of
501 the other factor, altering cohesin overlap and accessibility. SNPs in CBSs or TF motifs highlight
502 both the cooperativity and competitiveness of CTCF-TF cobinding. Our finding that cell type-
503 specific TFs can have a negative effect on CTCF binding addresses another long-standing
504 question about why many accessible CBSs remain unbound. Here we show that binding of other
505 factors can compete with CTCF to prevent its occupation.
506
507 AlphaFold3 predictions demonstrate that TF binding to CTCF free CBSs, overlaps the CBS,
508 providing an explanation for their inability to cobind. In contrast, at cobound sites, binding of the
509 TF occurs in the upstream or downstream regions, sites that do not interfere with CTCF's
510 occupation. Although AlphaFold3 cannot predict the impact of cobinding on CTCF stability, our
511 ChIP-seq analyses confirm that this leads to a stronger CTCF signal, despite TF-mediated
512 changes in DNA torsion and bending. However, binding of cofactors can alter nucleosome
513 phasing, and we speculate that changes in DNA torsion and bending could be responsible for
514 disrupting repositioning, and the outcome could be TF specific as previously shown [22]. Cobinding
515 of YY1 and CTCF occur at a subset of PAX5 binding sites in the *immunoglobulin heavy chain*
516 *(Igh)* locus [31]. This is of interest, because we previously showed that both PAX5 [32,33] and YY1 [34]
517 are important for mediating chromatin looping and locus contraction in early B cell development
518 to induce long range V-to-DJ rearrangements which is important for the generation of antibody
519 diversity. In the absence of either factor, locus contraction and rearrangement are impaired,
520 consistent with the idea that cofactors are important for CTCF mediated function.
521
522 Use of long read, nano-NOME-seq single molecule analysis allowed us to interrogate the impact
523 of CTCF binding on nucleosome repositioning at a resolution other assays such as MNase could
524 not achieve. By subsetting the data, we were able to show a continuum of CTCF binding states,
525 starting with binding to highly accessible chromatin devoid of nucleosomes, which has been
526 shown to reflect newly replicated chromatin [25]. Binding of CTCF occurs early and helps to organize

527 nucleosomes around the binding site leading to ordered nucleosome phasing and repositioning,
528 which likely requires recruitment of chromatin remodellers, such as members of the mSWI/SNF
529 family [35]**.** The clustering shows that CTCF can bind with either its N or C terminal abutting a
530 nucleosome or less frequently, overlapping a nucleosome. This is in contrast to the findings of
531 Pugacheva *et al*., who posit that CTCF binds only at N terminal nucleosome entry sites [35].
532
533 At CTCF free sites, CBSs are occupied either by a nucleosome or TFs that can phase
534 nucleosomes, but cannot reposition them. The effect of the TF on nucleosomes is more dynamic
535 as reflected by less stable nucleosome positioning around CTCF, in line with the low residence
536 time of of these factors compared to CTCF. Furthermore, there is a more robust nucleosome
537 presence at inaccessible sites, consistent with less TF binding at these regions. Interestingly, we
538 detect similar overall binding states when we degrade CTCF, suggesting that in the absence of
539 CTCF, TFs can come in and bind at CBSs. This data further highlights that CTCF and TFs can
540 bind competitively at CBSs, when their binding sites directly overlap.
541
542 While CTCF has been shown to act as a barrier that promotes nucleosome phasing and
543 positioning at its binding sites [36], SCC1 depletion data suggests that cohesin plays a critical and
544 dominant role in nucleosome repositioning and phasing at CTCF sites. Notably, the loss of
545 phasing and repositioning observed following SCC1 degradation - despite unchanged CTCF
546 binding strength - closely recapitulates the nucleosome configuration seen upon combined loss
547 of CTCF and SMC3 binding at CBSs after CTCF depletion. This finding suggests that cohesin
548 and loop extrusion might be responsible for the local nucleosome organization observed at CTCF
549 anchors.
550
551 Clustering demonstrates that the proportion of chromatin with repositioned nucleosomes depends
552 on accessibility as well as CTCF signal strength, both of which likely influence the recruitment of
553 nucleosome remodellers. However our studies highlight that signal strength has a greater impact
554 on nucleosome repositioning, and this is also reflected in a bigger impact on chromatin insulation
555 than accessibility, confirming our previous findings [9]. However, we also show that CTCF signal
556 strength and nucleosome repositioning can be uncoupled in the presence of TFs, while the
557 association between nucleosome repositioning and insulation is still maintained. Thus,
558 nucleosome repositioning is a more reliable predictor of insulation than CTCF signal strength and
559 we propose that use of single molecule nucleosome occupancy could be a powerful tool to predict
560 effects on insulation.

561
562 Here we have shown that enrichment of a wide range of TF motifs stabilize CTCF binding and
563 confirmed this by analyzing the ChIP-seq of a handful of factors at CBSs. Together, these studies
564 define a mechanism by which transcriptional reprogramming during development and in disease
565 specific states can determine cell type-specific CTCF profiles, linked to local and long-range
566 chromatin organization.
567
568 **METHODS**
569 ***Cell line generation and culture***
570 The construction of the vector for cloning transgenic, doxycycline-inducible expression of WT and
571 mutant mouse *Ctcf* cDNA and gene targeting in mouse embryonic stem cells E14Tg2a (karyotype
572 19, XY; 129/Ola isogenic background) were performed as described in Do *et al.,* [9]. Mouse
573 embryonic stem cells and all clones derived from these were cultured under feeder-free conditions
574 in 0.1% gelatin (Sigma ES-006-B) coated dishes (Falcon, 353003) as described in Do *et al.,* [9].
575 Briefly, the cells were grown in DMEM (Thermo Fisher, 11965-118) supplemented with 15% fetal
576 bovine serum (Thermo Fisher, SH30071.03), 100 U/ml penicillin - 100 µg/ml streptomycin (Sigma,
577 P4458), 1 X GlutaMax supplement (Thermo Fisher, 35050-061), 1 mM sodium pyruvate (Thermo
578 Fisher, 11360-070), 1 X MEM non-essential amino-acids (Thermo Fisher, 11140-50), 50 µM b-
579 mercaptoethanol (Sigma, 38171), 10^4 U/ml leukemia inhibitory factor (Millipore, ESG1107), 3 µM
580 CHIR99021 (Sigma, SML1046) and 1 µM MEK inhibitor PD0325901 (Sigma, PZ0162).
581 SCC1-mEGFP-AID HeLa cell line in which SCC1 is tagged with an auxin induced degron was
582 previously generated by Wutz et al.,[28] and was kindly provided by Dr. Peters. The cells were
583 grown in DMEM supplemented with 10% DBS (Thermo Fisher, 10371029), 0.2 M glutamine
584 (Thermo Fisher, 25030081), penicillin/streptomycin and 1X non-essential amino acids (Thermo
585 Fisher, 11140050)
586
587 ***Induction of auxin inducible degradation of CTCF, SCC1 and doxycycline induced***
588 ***expression***
589 For degradation of endogenous CTCF in mESCs, the auxin-inducible degron was induced by
590 adding 500 µM indole-3-acetic acid (IAA, the chemical analog of auxin) (Sigma, I5148) to the
591 media. Expression of WT CTCF transgene was achieved by the addition of doxycycline (Dox, 1
592 µg/ml) (Sigma, D9891) to the media. The cells were treated with IAA and/or Dox for 3h, 18h or 2
593 days. In Hela, SCC1 were degraded by addition of 500 µM IAA for 3h or 12h.
594

595 ***Nano-NOMe-seq sample and library preparation***

596 *Nuclei extractraction, GpC methyl-transferase treatment and DNA extraction*

597 After 2 days of IAA and Dox treatment on WT mESCs, 1 million cells were collected and washed
598 with cold PBS at 300g at 4°C. As described in Battaglia *et al*., [37], nuclei were extracted using cold
599 lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 0.5 mM spermidine, 1.5 mM MgCl2, 0.1 mM
600 EDTA, 0.25% IGEPAL CA-630) for 3-6 minutes after checking nuclei under a light microscope.
601 The sample was centrifuged for 5 minutes at 800x g at 4°C, washed with 1 ml ice-cold nuclei
602 wash buffer (10 mM Tris-HCl pH 7.4, 50 mM NaCl, 0.5 mM spermidine, 1.5 mM MgCl2, 0.1 mM
603 EDTA), and centrifuged again for 5 minutes at 800x g at 4°C. The nuclei were resuspended in
604 564 µL 1X GpC MTase reaction buffer (NEB, B0227) and equilibrated to 37°C. This was followed
605 by 6 minutes of enzyme treatment (300 µL 1 M sucrose, 34 µL 10X GpC methyltransferase
606 reaction buffer, 3 µL 32 mM S-Adenosylmethionine (SAM) (NEB, B9003), 100 µL GpC
607 methyltransferase (M.CviPI) (NEB, M0227L)), followed by another 6 minutes of enzyme treatment
608 (50 µL GpC methyltransferase (M.CviPI), 3 µL 32 mM S-Adenosylmethionine (SAM)). An equal
609 volume of pre-warmed 2X stop buffer (20 mM Tris-HCl pH 7.4, 600 mM NaCl, 10 mM EDTA, 1%
610 SDS) was added and mixed well. Proteinase K (200 µg/mL) was added to the GpC methylation
611 reaction and incubated for 4 hours at 56°C, followed by RNase A (100 µg/mL) for 30 minutes at
612 37°C, followed by an additional round of proteinase K (200 µg/mL) for 30 minutes at 37°C.
613
614 HMW DNA was extracted using phenol-chloroform and precipitated by isopropanol. Briefly, an
615 equal volume of phenol-chloroform was added to the samples and mixed vigorously. The samples
616 were centrifuged at 13,000 rpm for 10 minutes at room temperature. The aqueous phase was
617 transferred to a clean tube, and 0.7-1 times the volume of isopropanol was added, mixed well,
618 and incubated at room temperature for 30 minutes. The samples were then centrifuged at 13,000
619 rpm for 10 minutes at 4°C. The DNA pellet was washed with 1 ml pre-chilled 70% ethanol. After
620 removing all the ethanol, the DNA was dissolved in EB buffer and incubated at 37°C for 30
621 minutes. The final DNA concentration was quantified using Nanodrop 2000, and molecular weight
622 was measured by genomic DNA screentape (Agilent, 5067-5365).
623

624 *Nano-NOMe-seq library preparation*

625 For Nano-NOMe-seq, the DNA fragment was sheared to an average size of 20kb with the Covaris
626 G-Tube (#520079) in an Eppendorf 5415D tabletop microcentrifuge following the manufacture's
627 protocol. Sample size was confirmed on genomic screentape with the Agilent Tapestation system.
628 The sheared DNA (1.5ug) was then input into the SQK-LSK114 library prep from Oxford

629 Nanopore to create the Nanopore library. Components were used from the LSK114 kit along with
630 several buffers and enzymes from New England Biolab (NEB) following the LSK114 protocol from
631 Oxford Nanopore. Following library prep, the sample concentration was verified on the Qubit
632 system using the dsDNA High Sensitivity Kit. The concentration and basepair size were used to
633 calculate Femtomolar concentration for loading. The sample was resuspended with library loading
634 beads and loaded at 150-200 femtomolar concentration onto the Promethion R10.4.1 flow cell
635 according to the LSK protocol. Sample were sequenced on the R10.4.1 Promethion flow cell using
636 the Oxford Nanopore Promethion P24 system with a sequencing run time of 72 hrs. Standard
637 settings were applied during the run setup, special settings were applied and fragments less than
638 1kb rejected and methylation detection was set for both 5mC and 5hmC along with the super
639 accurate basecalling feature turned on.
640
641 *ChIpmentation*
642 ChIPmentation was performed in the Hela cell lines as described in Do *et al.* [9]. Briefly, 10 millions
643 cells were double cross linked using 25mM EGS (ethylene glycol bis(succinimidyl succinate);
644 Thermofisher #21565), followed by addition of 1% formaldehyde (Tousimis #1008A). Chromatin
645 was sheared using a bioruptor (Diagenode) for 15 minutes. Mouse anti-CTCF (Active Motif,
646 61311) and Rabbit IgG (abcam, ab37415) antibodies were used for immunoprecipitation.
647 Tagmentation were performed using 1 µl Tagment DNA Enzyme from the Tagment DNA Enzyme
648 and Buffer Kit (Illumina #20034198). Purified DNA (20 µl) was amplified and barcoded using
649 indexed and non-indexed primers and NEBNext High-Fidelity 2X PCR Master Mix (NEB M0541).
650 DNA was purified using Agencourt AMPure XP beads (Beckman, A63881) to remove fragments
651 larger than 700 bp as well as the primer dimers. Library quality and quantity were estimated using
652 Tapestation (Agilent High Sensitivity D1000 ScreenTape #5067-5584 and High Sensitivity D1000
653 reagents #5067-5585) and quantified by Qubit (Life Technologies Qubit™ 1X dsDNA High
654 Sensitivity (HS) #Q33230). Libraries were then sequenced with the Novaseq6000 Illumina
655 technology according to the standard protocols and with around 200 million 150bp paired-end
656 total per sample.
657
658 ***Datasets and Data Processing***
659 *ChIP-seq, RNA-seq and ATAC-seq dataset*
660 The CTCF, SMC3 ChIP-seq, RNA-seq and ATAC-seq data for WT mESCs were previously
661 generated and processed as described in Do *et al.*, [9] using the Seq-N-Slide pipeline [38]. Hela
662 ChIPmentation data were processed using the same pipeline.

663 For ENCODE human samples, processed data for samples having CTCF ChIP-seq, RNA-seq
664 and ATAC-seq performed on the same biosample were downloaded from ENCODE. The
665 accession numbers of the dataset used in this work are listed in **Table S1A, S1B, S3**. CTCF
666 peaks were scanned for CTCF consensus motif (JASPAR MA0139.2) [39] using FIMO [40] with a p-
667 value cutoff of $5e10^{-5}$. Only samples with at least 20% of motif containing CTCF peaks were kept
668 for downstream analyses. Accessible CTCF peaks were identified by overlapping CTCF with
669 ATAC peaks using bedtools [41].
670 For the GM12878 cell line, we downloaded an additional set of 145 ENCODE TF ChIP-seq narrow
671 peak files to assess the percentage of CTCF binding sites with TF co-binding (the accession
672 numbers are listed in **Table S3**. To analyze allele-specific binding the BAM files for CTCF, RAD21,
673 TFs and ATAC-seq were also downloaded (the accession numbers are listed in **Table S3**). The
674 phased genotyping data of GM12878 was retrieved from the Platinium Genomes project [42]
675 (accession number PRJEB3246).
676
677 To validate the effect of TF co-binding, we downloaded KLF4, OCT4, SOX2, MYC, YY1 and input
678 ChiP-seq data (fastq files) generated in mESCs from GEO [20 21]. The data were processed using
679 the Seq-N-Slide pipeline [38] as described in Do *et al*., [9]. Briefly, the reads were aligned to mm10
680 genome with Bowtie2 [43] and narrow peaks were called using MACS2 [44] in pair-end mode and with
681 the input as control.
682
683 The CTCF-oriented heatmaps and profiles were generated using the computeMatrix scale-
684 regions function from deeptools [45] using a 15 bp bin and a region body length of 15 bp which
685 corresponds to the length of CTCF core motif.
686
687 *Hi-C data.*
688 The HIC data were previously generated using Arima HiC kit (A510008) and processed using
689 HiC-Pro and HiCExplorer as described in Do *et al*., [9]. The HIC matrices were generated at a 10
690 kb resolution. The insulation scores were computed using GENOVA [46] and the CTCF-oriented
691 profiles generated using deeptools [45] with a 1 kb bin and a region body length.
692
693 *Nano-NOMe-seq processing*
694 Base modification calling were performed on pod5 files using dorado basecaller [47] using the "--
695 min-qscore 10 sup,5mC_5hmC" option, followed by alignment using dorado aligner against the
696 mm10 reference genome. Reads with less than 10 MAPQ score were filtered out using samtools

697 [48]. Methylation in HGC context was called using dorado pileup using the "--combine-strands"
698 option. Single molecule (SM) methylation was extracted using modkit extract with the "--read-
699 calls-path" option [49]. GpC overlapping the ENCODE blacklisted regions [50] were excluded as well
700 as sites with more than 100x coverage.
701
702 CTCF-oriented GpC methylation profiles were generated using deeptools and single molecule
703 heatmaps generated using R. Single molecules with occupied or unoccupied CTCF were defined
704 as molecule containing a unmethylated or methylated GpC in the motif, respectively. No bias was
705 observed in the nucleosome positioning of molecules with and without GpC containing motifs
706 (**Figure S4A**). SM Heatmaps were created in R at the single molecule and GpC level using the
707 GpC methylation status called by modkit extract after centering the reads on the middle of CTCF
708 consensus motif.
709
710 ***Downstream Analyses***
711 *Motif logos and Motif spacing enrichment analysis*
712 Motif logos of the aggregated sequences centered on CTCF motif detected by FIMO [40] were
713 created using the R package, ggseqlogo [51]. Significantly enriched spacings of TF motifs in the
714 vicinity of CTCF bound motifs were identified using SpaMo [52]. The motif positional weight matrices
715 (PWM) were downloaded from JASPAR 2024 (CORE vertebrates non-redundant dataset) [39] . The
716 PWM for the U and D motifs were derived from Nakahashi *et al.,* [17]. The analyses were performed
717 using the 500 bp window centered on each bound CTCF motif stratified by chromatin accessibility.
718 TF motif clustering and trees were performed using the R package, universalmotif [53] for each
719 window defined as follow: upstream spacer, U window and further upstream, 0 to 5bp, 6-8 bp and
720 more than 9bp upstream, respectively; downstream spacer, D window and further downstream,
721 0 to 9 bp, 10-12 bp and more than 13 bp downstream respectively.
722
723 *Association between TF motif enrichment and TF expression*
724 To ask whether TFs more enriched at accessible sites compared to inaccessible sites were more
725 expressed in human iPSCs, accessible enriched TFs were defined as TFs with a 2-fold increase
726 in the percentage of motif found within 35 bp of the CTCF motif identified by the SpaMo analysis.
727 A multivariate linear regression using a backward-stepwise selection was then performed to test
728 the association between TF expression level (log2 TPM) and chromatin accessibility controlled
729 for the epistatic effect of the motif orientation and window of enrichment (relative to the CTCF
730 motif). The adjusted expression level by accessibility was estimated by computing the model

731 margins and the p-value by performing the pairwise comparisons of predictive margins using
732 STATA v17.
733
734 To test the overall correlation between TF expression and motif enrichment, we performed an
735 aggregated analysis across ENCODE human samples (**Table S1A**). The expression level of TFs
736 and percentage of motif were summed by sample, accessibility, window of enrichment and motif
737 orientation. A multivariate linear regression using a backward-stepwise selection was then
738 performed to test the correlation between the aggregated percentage of enriched motifs and TF
739 expression level (log2 TPM) controlled for the epistatic effect of chromatin accessibility, the motif
740 orientation and window of enrichment (relative to the CTCF motif). The differential effect of these
741 covariates on the TF enrichment was estimated using the adjusted slope (interaction term
742 between the covariate and TF expression) by computing the model margins and the pairwise
743 comparisons of predictive margins using STATA v17.
744 Bivariate linear regressions between individual TF motif enrichment and expression level were
745 performed using R.
746
747 *AlphaFold3 prediction*
748 For each transcription factor (TF) — KLF4, OCT4, SOX2, MYC, and YY1 — we randomly selected
749 loci for each category of "CTCF-only," "CTCF-TF cobound," and "TF-only" sites. The DNA
750 sequences (positive and negative strands) of these loci, centered on the CTCF or TF motif and
751 including 35 bp flanking sequences, were extracted using Bedtools [41].
752 TF and CTCF ZF1-11 amino acid sequences were retrieved from UniProt [54]: Q61164 for CTCF,
753 Q00899 for YY1, Q60793 for KLF4, P20263 for OCT4, P48432 for SOX2, P01108 for MYC, and
754 P28574 for MAX. The CTCF ZF1-11 positions were obtained from Yang *et al.* [23].
755 AlphaFold3 predictions [55] were carried out using the DNA sequences, the CTCF sequence, with
756 or without the TF sequence. The predicted protein-DNA structures were visualized in PyMOL
757 3.10, and the root mean square deviation (RMSD) between the structures was calculated using
758 PyMOL 3.10.
759
760 *Allele-specific binding analyses*
761 Allele-specific binding (ASB) was performed by separating the ChIP-seq or ATAC-seq reads
762 overlapping phased heterozygous SNPs in the GM12878 cell line using R. Allele-specific BAM
763 and bigwig files were then generated using samtools [48] and bamCoverage [45]. To identify ASB

764 SNPs disrupting the TF motif, allele-specific motif affinity was computed using the R package,
765 atSNP [56].
766
767 *Nano-NOMe-seq single molecules clustering, trajectory and phasing score*
768 For each condition (untreated, IAA) and genomic context (based on accessibility, CTCF signal
769 strength, presence of TF motifs), 5000 molecules were randomly sampled. The level of 5mC
770 methylation in GpC context was binned by 50 bp for each molecule. The missing values were
771 imputed using k-Nearest Neighbors before performing k-means clustering. UMAP was applied
772 using the uwot R package with n_neighbors = 15 and min_dist = 0.2 to project the nucleosome
773 occupancy profiles into two dimensions. For SM trajectories, a principal curve was first fitted on
774 the UMAP-embedded molecules using the princurve R package [57]. All molecules were then
775 projected onto this curve to obtain pseudotime estimate and trajectory inferred using the slingshot
776 R package [24], specifying cluster 1 as the root. Nucleosome phasing scores were calculated by
777 applying a fast Fourier transform (FFT) to the binarized methylation patterns of each read. Two
778 separate phasing scores were computed for accessible and inaccessible sites, targeting
779 periodicities of 190 bp and 170 bp, respectively. For each, the power spectrum was computed
780 and the phasing score was defined as the proportion of signal intensity within a narrow frequency
781 band around the expected nucleosome spacing.
782
783 **Data availability**
784 All raw and processed sequencing data files are deposited at NCBI's Gene Expression
785 Omnibus (GEO) and will be available to public on publication of the manuscript.
786
787 **Code availability**
788 No custom code/software was used in the study. The publicly available software used is indicated
789 in the methods section.
790
791 **Acknowledgements**
792 This work was supported by 1R35GM122515 (J.S) and NIH P01CA229086 (J.S). GC were
793 supported by a fellowship from the NCC.
794
795 The authors thank Skok lab members for helpful scientific discussions, New York University
796 School of Medicine High Performance Computing Facility (HPCF) for computing technical
797 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.

The authors thank Sofia Battaglia from the Bernstein's lab for her advice and protocols for the Nano-NOMe-seq library preparation.

Author contributions

These studies were designed by Jane Skok and Catherine Do. The Nano-NOME-seq experiments were performed by Guimei Jiang and Paul Zappile under the supervision of Adriana Heguy. The analysis was performed by Catherine Do The paper was written by Jane Skok and Catherine Do.

Declaration of interests

The authors declare no competing interests.

SUPPLEMENTARY INFORMATION

Supplemental Tables

Table S1. Tab A: Accession numbers of the CTCF ChIP-seq, ATAC-seq and RNA-seq for human ENCODE samples used in the motif analyses. Only samples with data in all three assays and with at least 35% of CTCF motif-containing peaks were included. The percentage of CTCF motif found using FIMO are listed. **Tab B**: Number and percentage of accessible CTCF peaks in human and mouse including the CTCF ChIP-seq and ATAC-seq accession numbers for the mouse ENCODE samples. This data was used in **Figure S1B.** This table relates to **Figure 1.**

Table S2. Tab A: Proportions and counts of the sequences with enriched TF motifs within 35 bp of the CTCF consensus motif at accessible and inaccessible CBSs in WT mESCs. **Tab B:** Proportions and counts of the sequences with enriched TF motifs within 35 bp of the CTCF consensus motif at accessible and inaccessible binding sites in human ENCODE samples with at least 20% of CTCF motif-containing peaks. This table relates to **Figure 1**.

Table S3. ENCODE accession numbers of the TF ChIP-seq narrow peak and BAM files in GM12878 to assess the overall percentage of cobound CTCF and TF sites. This table relates to **Figure 3**.

832

833 **Table S4.** Allele-specific binding of CTCF associated with allele-specific of TF with SNPs
834 disrupting either CTCF or the TF motifs. This tables relates to **Figure 3**.

835

836 **Table S5**. Significant correlations between the percentage of TF motifs within 35 bp of CTCF
837 consensus motif and their expression level in ENCODE human samples. This relates to **Figure**
838 **4**.

839

840 **Supplemental Figures**

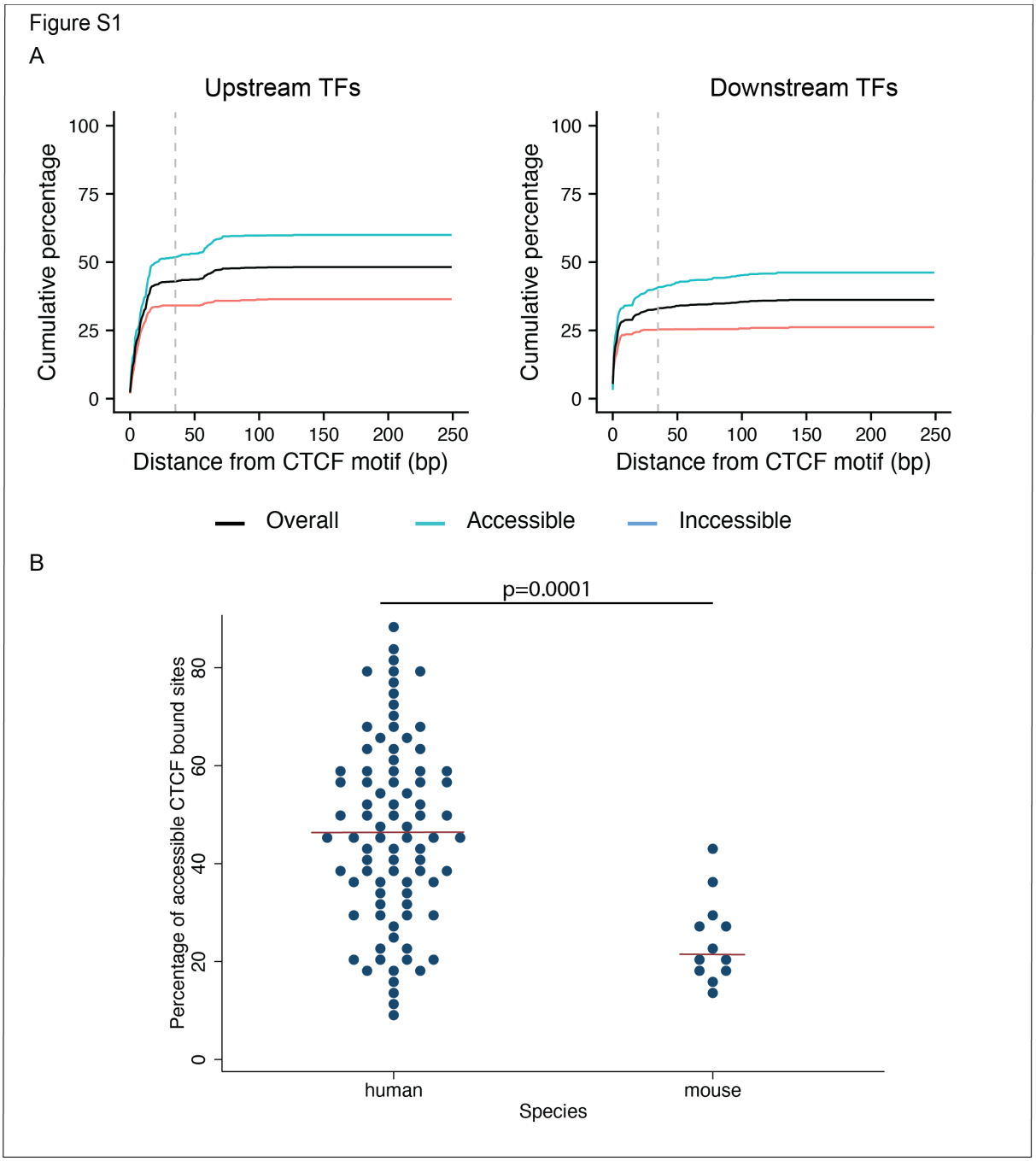
841

842

843

844

845



846

Figure S1. A: Cumulative percentage of CTCF peaks containing a TF motif within 250 bp at accessible and inaccessible sites. The dashed lines correspond to 35 bp upstream or downstream of the CTCF motif. This figure relates to **Figure 1**. **B:** Dotplot showing the higher percentage of accessible CTCF bound sites in human versus mouse. The red line represents the median. The p-value has been computed using a t-test. This figure is related to **Figure 1.**

847

848

849

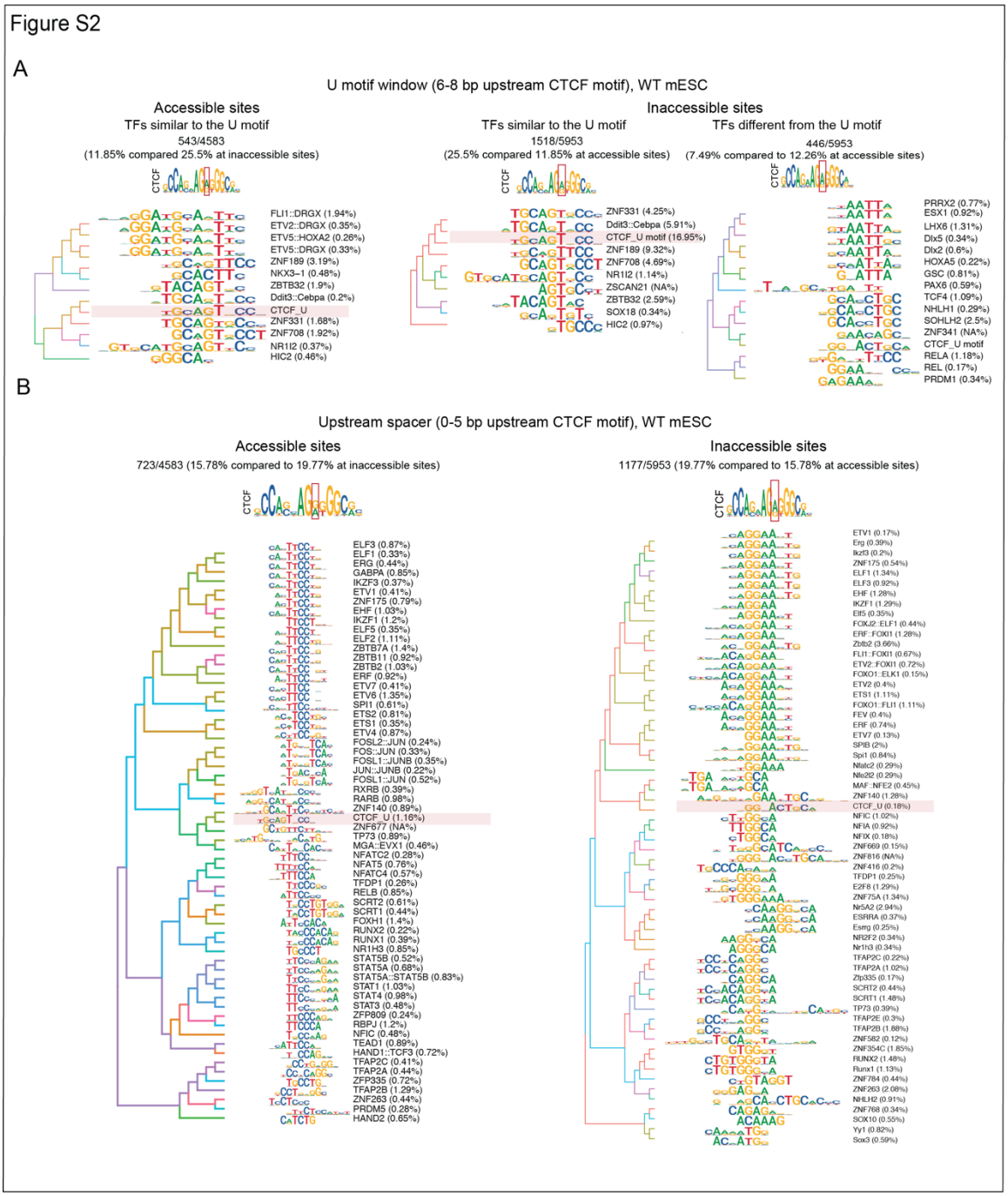
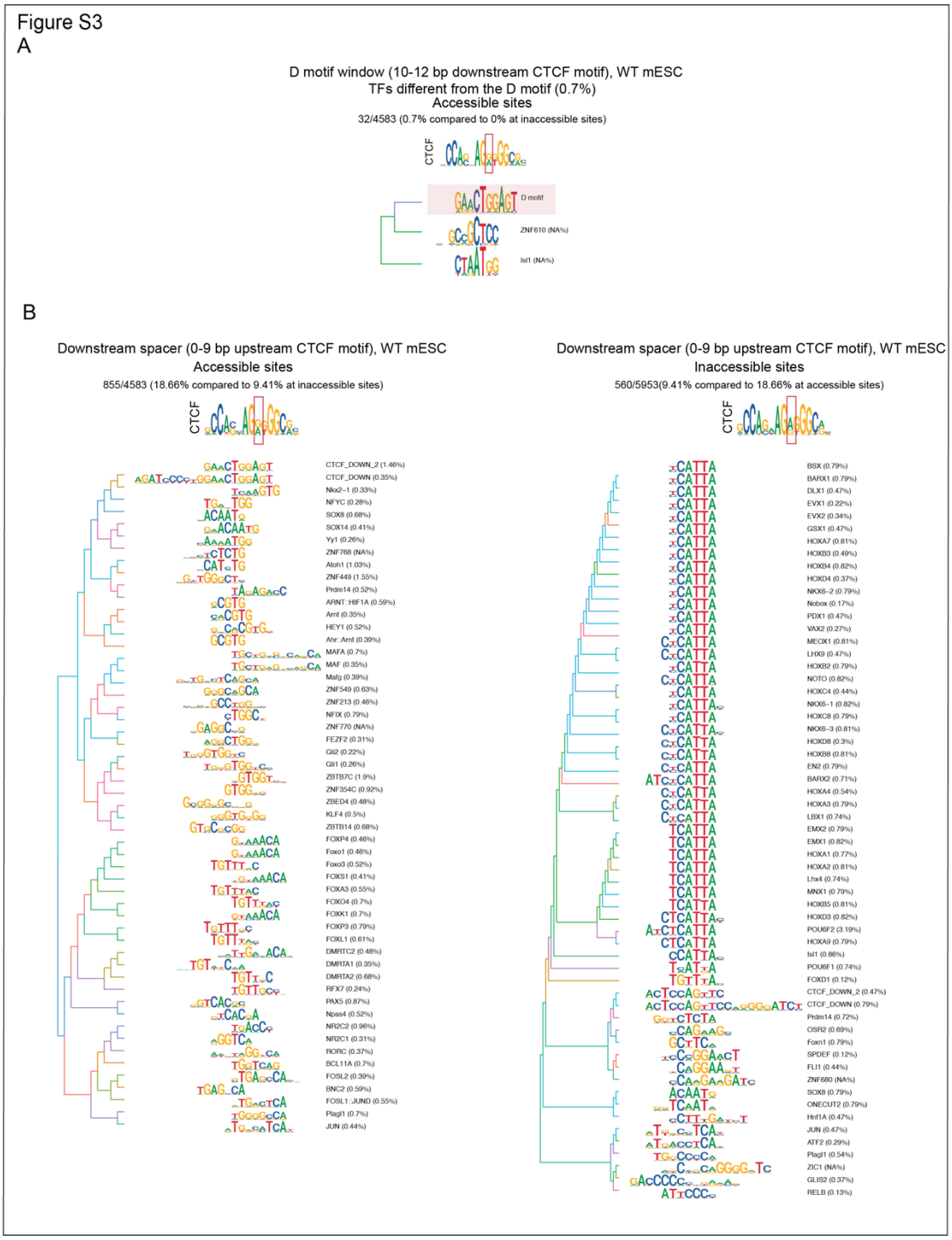


Figure S2. A: Motif logos of the enriched TF motifs found in the U motif window at accessible and inaccessible sites in WT mESCs. **B**: Motif logos of the enriched TF motifs found in the U spacer window at accessible and inaccessible sites in WT mESCs. This window is enriched in ETS-family motifs. This figure relates to **Figure 1**.



855

Figure S3. A: Motif logos of the enriched TF motifs found in the D window at accessible sites in WT mESCs. No motifs were significantly enriched at inaccessible sites. **B**: Motif logos of the enriched TF motifs found in the downstream spacer window at accessible and inaccessible sites in WT mESCs. This figure relates to **Figure 1**.

856
857
858
859

Figure S4

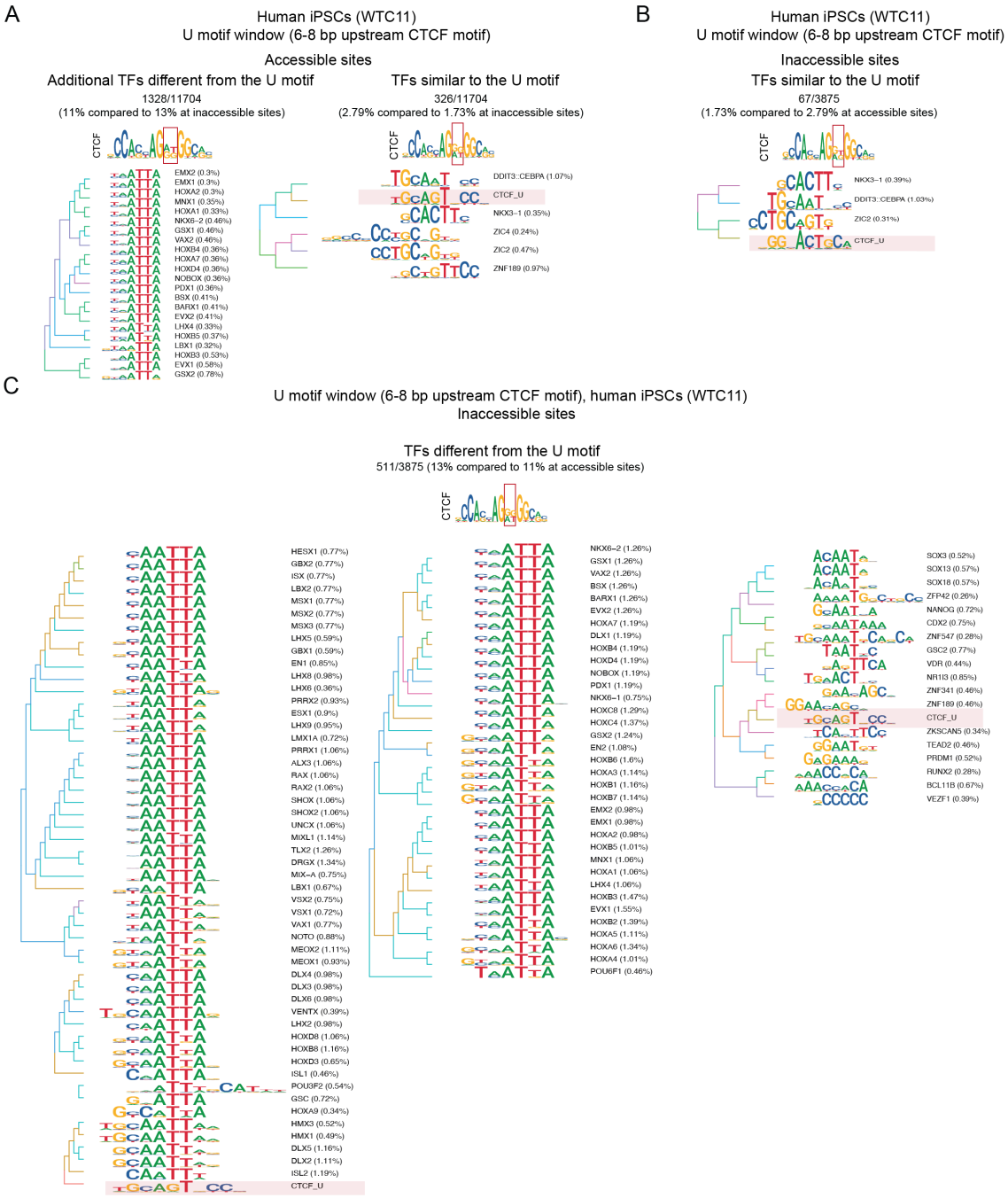

A
Human iPSCs (WTC11)
U motif window (6-8 bp upstream CTCF motif)
Accessible sites
Additional TFs different from the U motif
1328/11704
(11% compared to 13% at inaccessible sites)
TFs similar to the U motif
326/11704
(2.79% compared to 1.73% at inaccessible sites)
CTCF
EMX2 (0.3%)
EMX1 (0.3%)
HOXA2 (0.3%)
MNX1 (0.35%)
HOXA1 (0.33%)
NKX6−2 (0.46%)
GSX1 (0.46%)
VAX2 (0.46%)
HOXB4 (0.36%)
HOXA7 (0.36%)
HOXD4 (0.36%)
NOBOX (0.36%)
PDX1 (0.36%)
BSX (0.41%)
BARX1 (0.41%)
EVX2 (0.41%)
LHX4 (0.33%)
HOXB5 (0.37%)
LBX1 (0.32%)
HOXB3 (0.53%)
EVX1 (0.58%)
GSX2 (0.78%)
DDIT3::CEBPA (1.07%)
CTCF_U
NKX3−1 (0.35%)
ZIC4 (0.24%)
ZIC2 (0.47%)
ZNF189 (0.97%)
B
Human iPSCs (WTC11)
U motif window (6-8 bp upstream CTCF motif)
Inaccessible sites
TFs similar to the U motif
67/3875
(1.73% compared to 2.79% at accessible sites)
NKX3−1 (0.39%)
DDIT3::CEBPA (1.03%)
ZIC2 (0.31%)
CTCF_U
C
U motif window (6-8 bp upstream CTCF motif), human iPSCs (WTC11)
Inaccessible sites
TFs different from the U motif
511/3875 (13% compared to 11% at accessible sites)
HESX1 (0.77%)
GBX2 (0.77%)
ISX (0.77%)
LBX2 (0.77%)
MSX1 (0.77%)
MSX2 (0.77%)
MSX3 (0.77%)
LHX5 (0.59%)
GBX1 (0.59%)
EN1 (0.85%)
LHX8 (0.98%)
LHX6 (0.36%)
PRRX2 (0.93%)
ESX1 (0.9%)
LHX9 (0.95%)
LMX1A (0.72%)
PRRX1 (1.06%)
ALX3 (1.06%)
RAX (1.06%)
RAX2 (1.06%)
SHOX (1.06%)
SHOX2 (1.06%)
UNCX (1.06%)
MIXL1 (1.14%)
TLX2 (1.26%)
DRGX (1.34%)
MIX−A (0.75%)
LBX1 (0.67%)
VSX2 (0.75%)
VSX1 (0.72%)
VAX1 (0.77%)
NOTO (0.88%)
MEOX2 (1.11%)
MEOX1 (0.93%)
DLX4 (0.98%)
DLX3 (0.98%)
DLX6 (0.98%)
VENTX (0.39%)
LHX2 (0.98%)
HOXD8 (1.06%)
HOXB8 (1.16%)
HOXD3 (0.65%)
ISL1 (0.46%)
POU3F2 (0.54%)
GSC (0.72%)
HOXA9 (0.34%)
HMX3 (0.52%)
HMX1 (0.49%)
DLX5 (1.16%)
DLX2 (1.11%)
ISL2 (1.19%)
CTCF_U
NKX6−2 (1.26%)
GSX1 (1.26%)
VAX2 (1.26%)
BSX (1.26%)
BARX1 (1.26%)
EVX2 (1.26%)
HOXA7 (1.19%)
DLX1 (1.19%)
HOXB4 (1.19%)
HOXD4 (1.19%)
NOBOX (1.19%)
PDX1 (1.19%)
NKX6−1 (0.75%)
HOXC8 (1.29%)
HOXC4 (1.37%)
GSX2 (1.24%)
EN2 (1.08%)
HOXB6 (1.6%)
HOXA3 (1.14%)
HOXB1 (1.16%)
HOXB7 (1.14%)
EMX2 (0.98%)
EMX1 (0.98%)
HOXA2 (0.98%)
HOXB5 (1.01%)
MNX1 (1.06%)
HOXA1 (1.06%)
LHX4 (1.06%)
HOXB3 (1.47%)
EVX1 (1.55%)
HOXB2 (1.39%)
HOXA5 (1.11%)
HOXA6 (1.34%)
HOXA4 (1.01%)
POU6F1 (0.46%)
SOX3 (0.52%)
SOX13 (0.57%)
SOX18 (0.57%)
ZFP42 (0.26%)
NANOG (0.72%)
CDX2 (0.75%)
ZNF547 (0.28%)
GSC2 (0.77%)
VDR (0.44%)
NR1I3 (0.85%)
ZNF341 (0.46%)
ZNF189 (0.46%)
CTCF_U
ZKSCAN5 (0.34%)
TEAD2 (0.46%)
PRDM1 (0.52%)
RUNX2 (0.28%)
BCL11B (0.67%)
VEZF1 (0.39%)

Figure S4. A: Motif logos of the enriched TF motifs found in the U motif window at accessible sites in human iPSCs (WCT11)**. B:** Motif logos of the enriched TF motifs found in the U motif window and resembling the U motif at inaccessible sites in human iPSCs (WCT11)**. C**: Motif logos of the enriched TF motifs found in the U motif window and distinct from the U motif at inaccessible sites in human iPSCs (WCT11)**.** This figure relates to **Figure 1**.

867
868

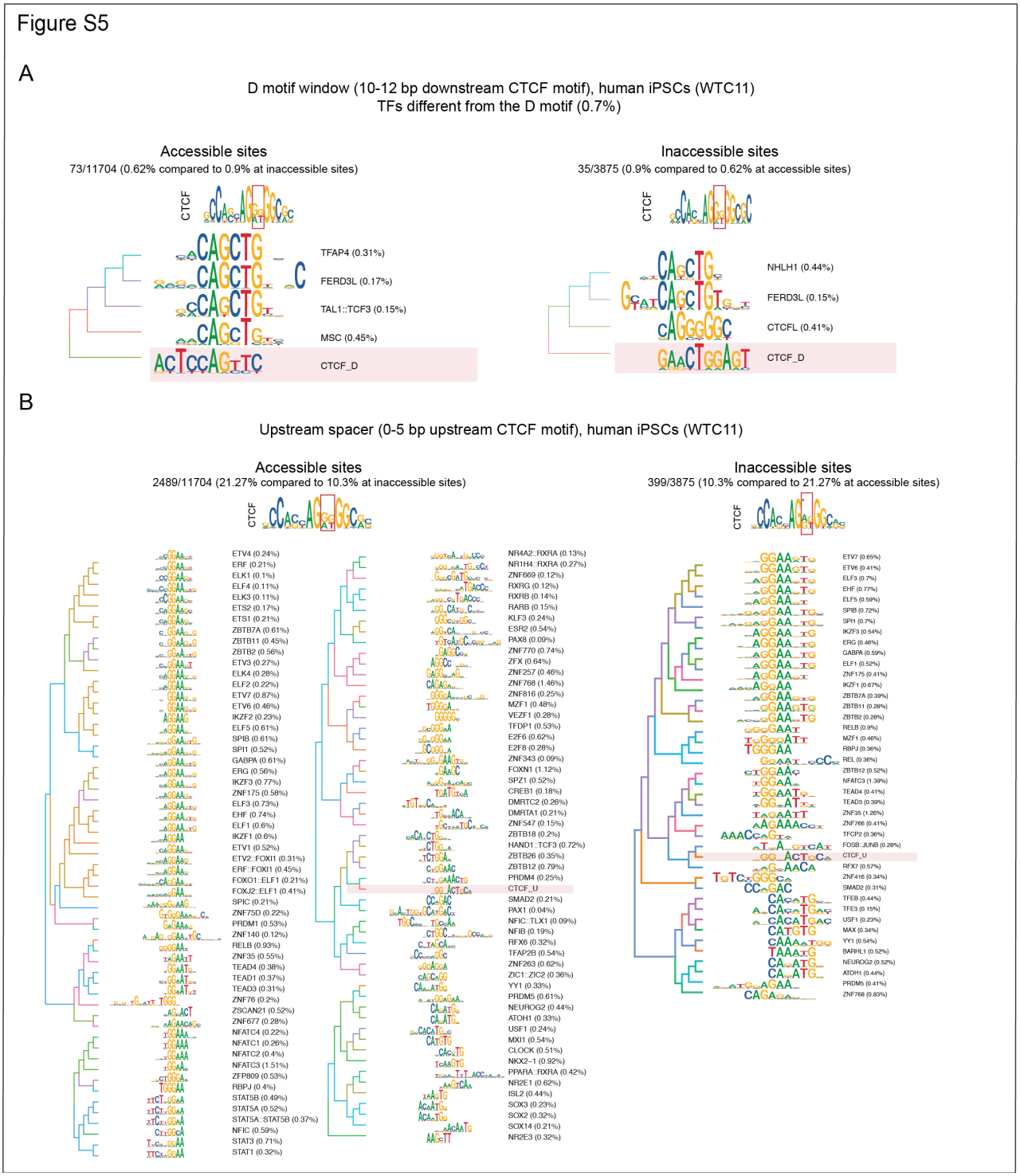


Figure S5. A: Motif logos of the enriched TF motifs found in the D motif window at accessible and inaccessible sites in human iPSCs (WCT11). **B**: Motif logos of the enriched TF motifs found in the upstream spacer at accessible and inaccessible sites in human iPSCs (WCT11)**.** This window is enriched in ETS-family motifs. This figure relates to **Figure 1**.

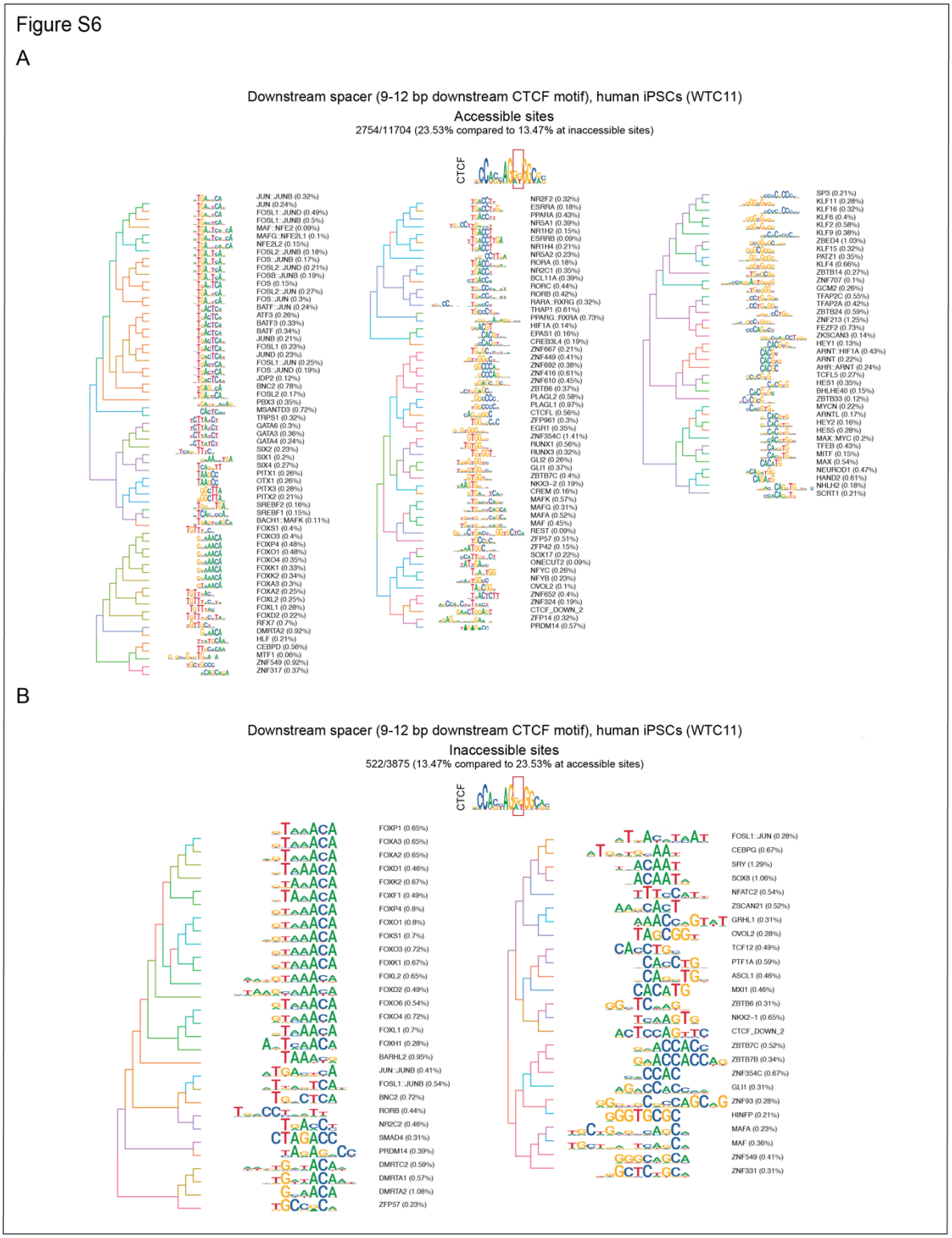


Figure S6. A: Motif logos of the enriched TF motifs found in the downstream spacer at accessible and inaccessible (**B**) sites in human iPSCs (WCT11)**.** This figure relates to **Figure 1**.

Figure S7

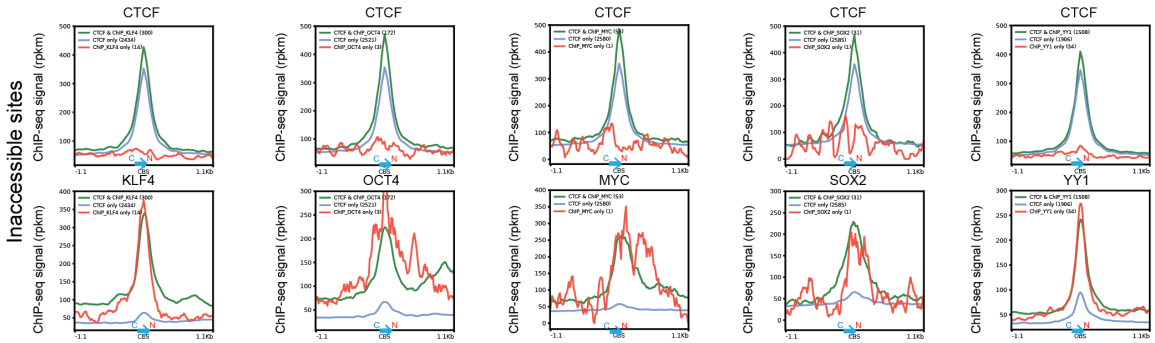
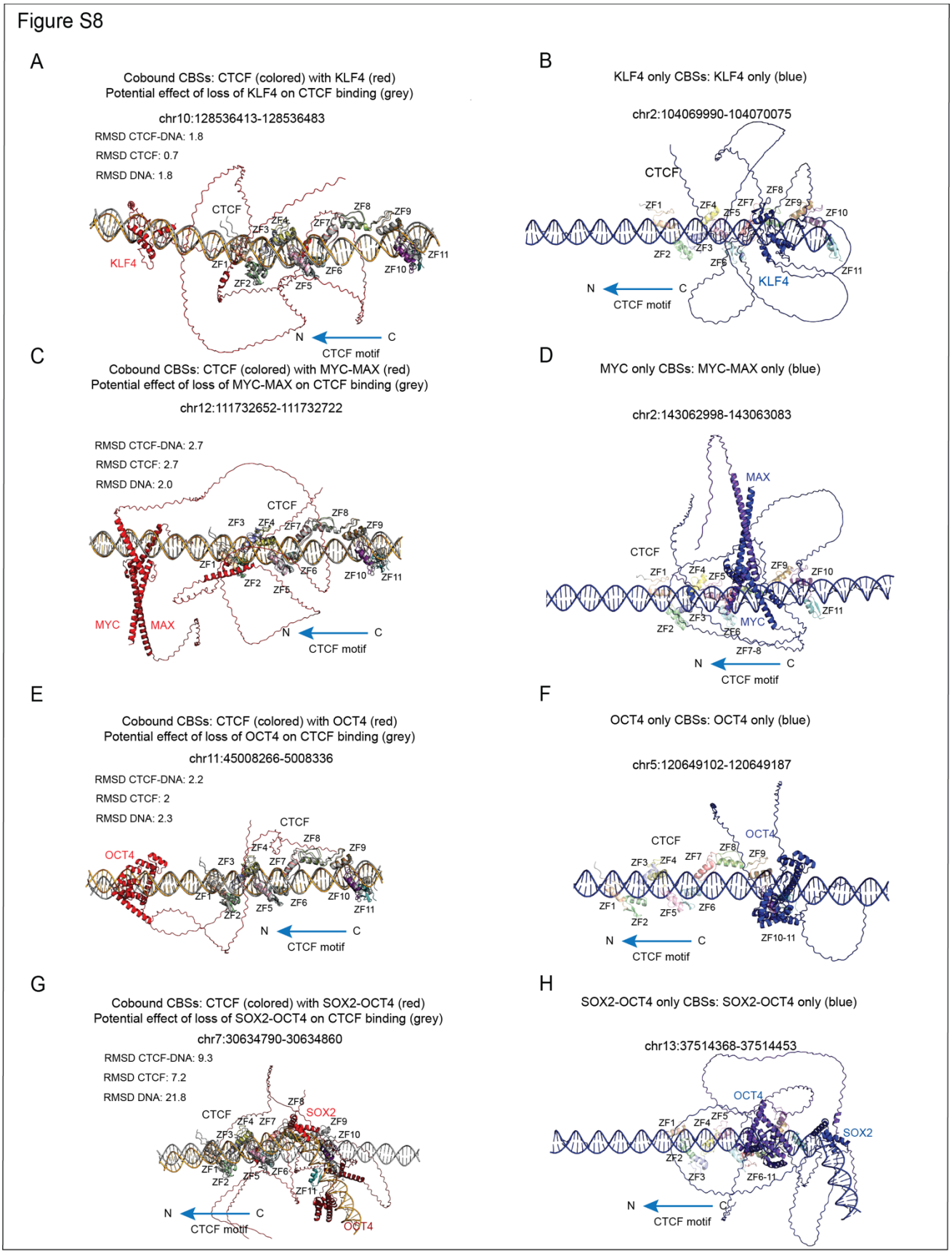


Figure S7: Individual TFs strengthen CTCF signal. Profiles of CTCF (FLAG) and 5 TFs (including 4 pioneer TFs, KLF4, OCT4, MYC and SOX2) showing the effect of TF binding on CTCF signal strength (top) and TF signal strength (bottom) at inaccessible sites where they are co-bound (green) compared to CTCF only (blue) and TF only sites (red). This analysis is restricted to CTCF binding sites with the consensus CTCF motif. The graphs are oriented on the C to N terminal CTCF core motif. This figure relates to **Figure 2**.

Figure S8

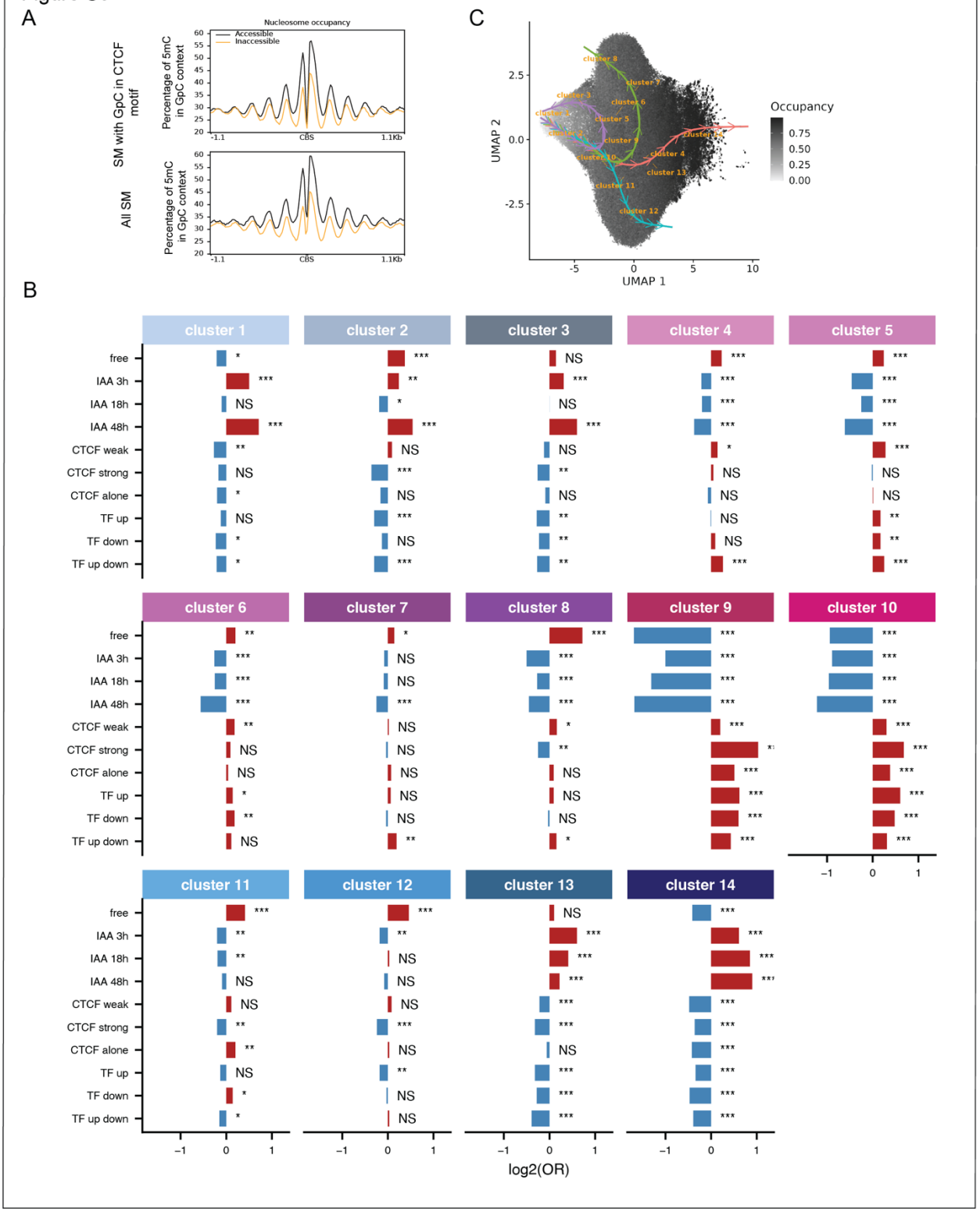


879

Figure S8: TFs can cobind or compete with CTCF at CBSs. AlphaFold3 predictions at cobound sites for KLF4 (**A**), MYC-MAX (**C**), OCT4 (**E**) and SOX2-OCT4 (**G**). To assess the effect of the cobound TF, the structures of CTCF with (colored) or without TF (grey) were aligned and the RMSD (root mean standard deviation) calculated for the CTCF-DNA complex and separately for CTCF and DNA (top). At cobound sites, TFs tend to bind outside CBSs or in the U spacer (between ZF7 and ZF9). SOX2 binding can induce a major bend in the DNA (**G**). AlphaFold3 predictions at TF only bound CBSs for KLF4 (**B**), MYC-MAX (**D**), OCT4 (without SOX2) (**F**) and SOX2-OCT4 (most of the SOX2 peaks overlap with OCT4) (**H**). TFs are shown in in blue. CTCF binding predictions are overlayed to visualize the position of the TF relative to CTCF. At those sites, the TFs tend to bind at CTCF core or extended binding sites and as a result they directly compete with its binding. This figure relates to **Figure 2**.

880

Figure S9



881

Figure S9: A: Profiles showing similar nucleosome occupancy phasing of occupied SM for GpC-containing CBSs compared to non GpC-containing CBSs. **B**: Enrichment of each SM cluster for the condition and genomic context under study. The odd ratios and p-values were computed using logistic regressions**. C:** UMAP of SMs colored by the molecule overall nucleosome occupancy. This figure relates to **Figure 5**.

882

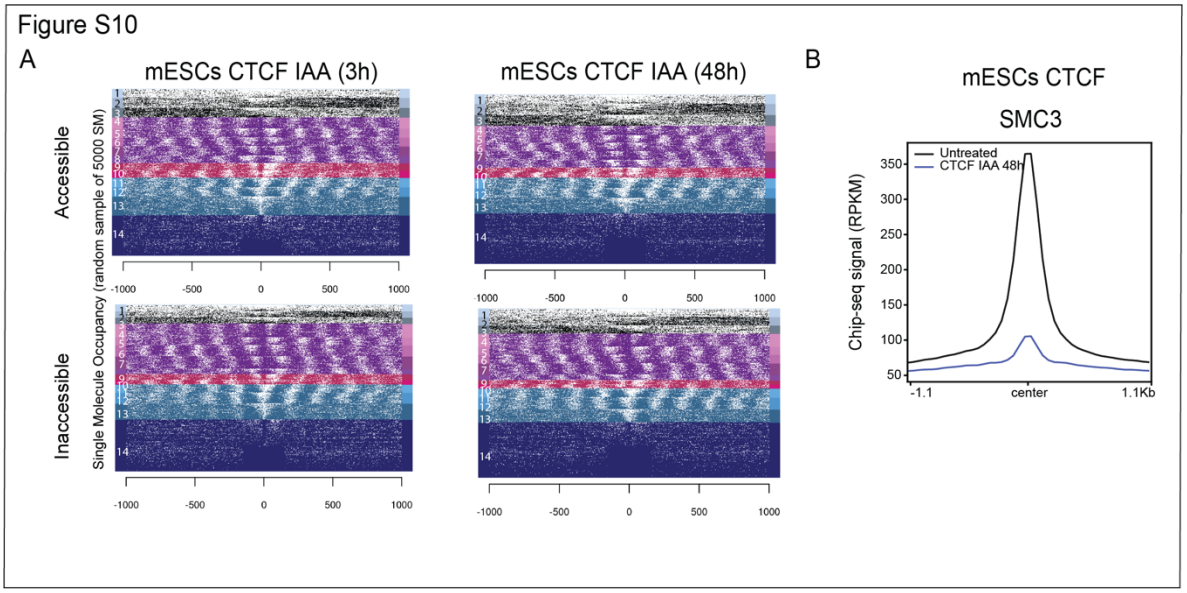


Figure S10: A: Heatmaps showing the effect of CTCF degradation on nucleosome phasing after 3h and 48h of IAA treatment in mESCs. **B:** Profiles of SMC3 at lost CTCF peaks after CTCF degradation (48h IAA) at accessible and inaccessible sites in mESCs. This figure relates to **Figure 6**.

References

1 Davidson, I. F. *et al.* DNA loop extrusion by human cohesin. *Science* **366**, 1338-1345 (2019). https://doi.org/10.1126/science.aaz3418

2 Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep* **15**, 2038-2049 (2016). https://doi.org/10.1016/j.celrep.2016.04.085

3 Dehingia, B., Milewska, M., Janowski, M. & Pekowska, A. CTCF shapes chromatin structure and gene expression in health and disease. *EMBO Rep* **23**, e55146 (2022). https://doi.org/10.15252/embr.202255146

4 Guo, C. *et al.* CTCF-binding elements mediate control of V(D)J recombination. *Nature* **477**, 424-430 (2011). https://doi.org/10.1038/nature10495

5 Hnisz, D. *et al.* Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454-1458 (2016). https://doi.org/10.1126/science.aad9024

6 Narendra, V. *et al.* Transcription. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* **347**, 1017-1021 (2015). https://doi.org/10.1126/science.1262088

7 Xiang, Y., Zhou, X., Hewitt, S. L., Skok, J. A. & Garrard, W. T. A multifunctional element in the mouse Igkappa locus that specifies repertoire and Ig loci subnuclear location. *J Immunol* **186**, 5356-5366 (2011). https://doi.org/10.4049/jimmunol.1003794

8 Hansen, A. S. CTCF as a boundary factor for cohesin-mediated loop extrusion: evidence for a multi-step mechanism. *Nucleus* **11**, 132-148 (2020). https://doi.org/10.1080/19491034.2020.1782024

9 Do, C. *et al.* Binding domain mutations provide insight into CTCF's relationship with chromatin and its contribution to gene regulation. *Cell Genom* **5**, 100813 (2025). https://doi.org/10.1016/j.xgen.2025.100813

10 Iurlaro, M. *et al.* Systematic assessment of ISWI subunits shows that NURF creates local accessibility for CTCF. *Nat Genet* **56**, 1203-1212 (2024). https://doi.org/10.1038/s41588-024-01767-x

11 Essien, K. *et al.* CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. *Genome Biol* **10**, R131 (2009). https://doi.org/10.1186/gb-2009-10-11-r131

12 Hu, G. *et al.* Systematic screening of CTCF binding partners identifies that BHLHE40 regulates CTCF genome-wide distribution and long-range chromatin interactions. *Nucleic Acids Res* **48**, 9606-9620 (2020). https://doi.org/10.1093/nar/gkaa705

13 Jerkovic, I. *et al.* Genome-Wide Binding of Posterior HOXA/D Transcription Factors Reveals Subgrouping and Association with CTCF. *PLoS Genet* **13**, e1006567 (2017). https://doi.org/10.1371/journal.pgen.1006567

14 Liu, Y. *et al.* CTCF coordinates cell fate specification via orchestrating regulatory hubs with pioneer transcription factors. *Cell Rep* **42**, 113259 (2023). https://doi.org/10.1016/j.celrep.2023.113259

15 Ross-Innes, C. S., Brown, G. D. & Carroll, J. S. A co-ordinated interaction between CTCF and ER in breast cancer cells. *BMC Genomics* **12**, 593 (2011). https://doi.org/10.1186/1471-2164-12-593

16 Sonmezer, C. *et al.* Molecular Co-occupancy Identifies Transcription Factor Binding Cooperativity In Vivo. *Mol Cell* **81**, 255-267 e256 (2021). https://doi.org/10.1016/j.molcel.2020.11.015

17 Nakahashi, H. *et al.* A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep* **3**, 1678-1689 (2013). https://doi.org/10.1016/j.celrep.2013.04.024

18 Plasschaert, R. N. *et al.* CTCF binding site sequence differences are associated with unique regulatory and functional trends during embryonic stem cell differentiation. *Nucleic Acids Res* **42**, 774-789 (2014). https://doi.org/10.1093/nar/gkt910
19 Soochit, W. *et al.* CTCF chromatin residence time controls three-dimensional genome organization, gene expression and DNA methylation in pluripotent cells. *Nat Cell Biol* **23**, 881-893 (2021). https://doi.org/10.1038/s41556-021-00722-w
20 O'Dwyer, M. R. *et al.* Nucleosome fibre topology guides transcription factor binding to enhancers. *Nature* **638**, 251-260 (2025). https://doi.org/10.1038/s41586-024-08333-9
21 Laugwitz, L. *et al.* ZSCAN10 deficiency causes a neurodevelopmental disorder with characteristic oto-facial malformations. *Brain* **147**, 2471-2482 (2024). https://doi.org/10.1093/brain/awae058
22 Malaga Gadea, F. C. & Nikolova, E. N. Structural Plasticity of Pioneer Factor Sox2 and DNA Bendability Modulate Nucleosome Engagement and Sox2-Oct4 Synergism. *J Mol Biol* **435**, 167916 (2023). https://doi.org/10.1016/j.jmb.2022.167916
23 Yang, J. *et al.* Structures of CTCF-DNA complexes including all 11 zinc fingers. *Nucleic Acids Res* **51**, 8447-8462 (2023). https://doi.org/10.1093/nar/gkad594
24 Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, 477 (2018). https://doi.org/10.1186/s12864-018-4772-0
25 Ostrowski, M. S. *et al.* The single-molecule accessibility landscape of newly replicated mammalian chromatin. *Cell* **188**, 237-252 e219 (2025). https://doi.org/10.1016/j.cell.2024.10.039
26 Grand, R. S. *et al.* Genome access is transcription factor-specific and defined by nucleosome position. *Mol Cell* **84**, 3455-3468 e3456 (2024). https://doi.org/10.1016/j.molcel.2024.08.009
27 Luan, J. *et al.* Distinct properties and functions of CTCF revealed by a rapidly inducible degron system. *Cell Rep* **34**, 108783 (2021). https://doi.org/10.1016/j.celrep.2021.108783
28 Wutz, G. *et al.* Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J* **36**, 3573-3599 (2017). https://doi.org/10.15252/embj.201798004
29 Owens, N. *et al.* CTCF confers local nucleosome resiliency after DNA replication and during mitosis. *Elife* **8** (2019). https://doi.org/10.7554/eLife.47898
30 Pop, R. T. *et al.* Identification of mammalian transcription factors that bind to inaccessible chromatin. *Nucleic Acids Res* **51**, 8480-8495 (2023). https://doi.org/10.1093/nar/gkad614
31 Ebert, A. *et al.* The distal V(H) gene cluster of the Igh locus contains distinct regulatory elements with Pax5 transcription factor-dependent activity in pro-B cells. *Immunity* **34**, 175-187 (2011). https://doi.org/10.1016/j.immuni.2011.02.005
32 Fuxa, M. *et al.* Pax5 induces V-to-DJ rearrangements and locus contraction of the immunoglobulin heavy-chain gene. *Genes Dev* **18**, 411-422 (2004). https://doi.org/10.1101/gad.291504
33 Roldan, E. *et al.* Locus 'decontraction' and centromeric recruitment contribute to allelic exclusion of the immunoglobulin heavy-chain gene. *Nat Immunol* **6**, 31-41 (2005). https://doi.org/10.1038/ni1150
34 Liu, H. *et al.* Yin Yang 1 is a critical regulator of B-cell development. *Genes Dev* **21**, 1179-1189 (2007). https://doi.org/10.1101/gad.1529307
35 Otto, J. E. *et al.* Structural and functional properties of mSWI/SNF chromatin remodeling complexes revealed through single-cell perturbation screens. *Mol Cell* **83**, 1350-1367 e1357 (2023). https://doi.org/10.1016/j.molcel.2023.03.013
36 Wiechens, N. *et al.* The Chromatin Remodelling Enzymes SNF2H and SNF2L Position Nucleosomes adjacent to CTCF and Other Transcription Factors. *PLoS Genet* **12**, e1005940 (2016). https://doi.org/10.1371/journal.pgen.1005940

37 Battaglia, S. *et al.* Long-range phasing of dynamic, tissue-specific and allele-specific regulatory elements. *Nat Genet* **54**, 1504-1513 (2022). https://doi.org/10.1038/s41588-022-01188-8

38 Dolgalev, I. Seq-N-Slide [Computer software]. (2022). https://doi.org/https://doi.org/10.5281/zenodo.5550459

39 Rauluseviciute, I. *et al.* JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **52**, D174-D182 (2024). https://doi.org/10.1093/nar/gkad1059

40 Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-208 (2009). https://doi.org/10.1093/nar/gkp335

41 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010). https://doi.org/10.1093/bioinformatics/btq033

42 Eberle, M. A. *et al.* A reference data set of 5.4 million phased human variants validated by genetic inheritance from sequencing a three-generation 17-member pedigree. *Genome Res* **27**, 157-164 (2017). https://doi.org/10.1101/gr.210500.116

43 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012). https://doi.org/10.1038/nmeth.1923

44 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008). https://doi.org/10.1186/gb-2008-9-9-r137

45 Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**, W187-191 (2014). https://doi.org/10.1093/nar/gku365

46 van der Weide, R. H. *et al.* Hi-C analyses with GENOVA: a case study with cohesin variants. *NAR Genom Bioinform* **3**, lqab040 (2021). https://doi.org/10.1093/nargab/lqab040

47 nanoporetech. Dorado. *https://github.com/nanoporetech/dorado* (2024).

48 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009). https://doi.org/10.1093/bioinformatics/btp352

49 nanoporetech. Modkit. *https://github.com/nanoporetech/modkit* (2023).

50 Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019). https://doi.org/10.1038/s41598-019-45839-z

51 Wagih, O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645-3647 (2017). https://doi.org/10.1093/bioinformatics/btx469

52 Whitington, T., Frith, M. C., Johnson, J. & Bailey, T. L. Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res* **39**, e98 (2011). https://doi.org/10.1093/nar/gkr341

53 Tremblay, B. J. universalmotif: An R package for biological motif analysis. *The Journal of Open Source Software* **9(100):7012** (2024). https://doi.org/10.21105/joss.07012

54 Apweiler, R. *et al.* UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**, D115-119 (2004). https://doi.org/10.1093/nar/gkh131

55 Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500 (2024). https://doi.org/10.1038/s41586-024-07487-w

56 Zuo, C., Shin, S. & Keles, S. atSNP: transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics* **31**, 3353-3355 (2015). https://doi.org/10.1093/bioinformatics/btv328

57 Hastie, T. & Stuetzle, W. Principal Curves. *ournal of the American Statistical Association, 84(406), 502–516.* (1989). https://doi.org/https://doi.org/10.1080/01621459.1989.10478797