A genome wide code to define cell-type specific CTCF binding and chromatin organization

Catherine Do[1], Guimei Jiang[1], Paul Zappile[3], Adriana Hegu[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

Abstract (125 words)

CTCF-mediated chromatin folding plays a key role in gene regulation, however the mechanisms underlying cell type-specific control are not fully elucidated. Comprehensive analyses reveal that CTCF binding stability and cohesin overlap in mice and humans is regulated by species specific differences in accessibility, the presence of CTCF extended Up and Downstream binding site sequences and motifs corresponding to expressed TFs enriched at most bound sites. CTCF repositions nucleosomes asymmetrically, with linker spacing altered by accessibility, while cohesin-mediated nucleosome phasing is affected by surrounding motifs which also impact chromatin insulation. Importantly, cell type-specific transcriptional programs determine motif enrichment at CTCF bound sites, reflecting the stabilizing/destabilizing effect of individual TFs. These studies identify mechanisms underlying cell type-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-type 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.

A number of studies have reported that Upstream and Downstream motifs, which engage peripheral CTCF ZFs can influence CTCF signal strength in mice. However, enrichment of these motifs has not been analyzed in the context of accessibility and cohesin overlap (*11-14*). Additionally, while previous research has 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, 15-19*). In this study, we demonstrate that the DNA sequences within a 35 bp window surrounding CBSs — including Up and Downstream motifs, as

well as 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 potential 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. In mice, U and D motifs are enriched at inaccessible binding sites where they respectively stabilize and destabilize CTCF binding. The latter are seldom found in human cells, which have proportionately fewer inaccessible sites compared to mice. Furthermore, 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.

Using long read, nano-NOME-seq single molecule analysis, we analyzed the impact of CTCF binding on nucleosome repositioning and DNA methylation (5mC and 5hmC) in different genomic contexts. The single molecule approach has the advantage of allowing (i) a comparison between the nucleosome occupancy of temporarily unoccupied (baseline) versus occupied (reorganized) molecules at CTCF bound sites, and (2) the relative proportion of occupied versus temporarily unoccupied sites, which reflects the residence time of individual proteins at different subsets of sites. This information has been missing from previous bulk analyses that average the two datasets. While it is known that CTCF binding remodels up and downstream nucleosomes (*20*), we found that temporarily unoccupied molecules at CTCF bound sites have nucleosomes closely positioned in a symmetrical manner around the CTCF binding site (CBS). Binding repositions nucleosomes asymmetrically in an N-terminal biased manner, such that the downstream nucleosome is pushed further away from the CBS than the upstream nucleosome at both accessible and inaccessible sites, however, the nucleosome free space is wider at accessible sites. As nucleosomes transition from the bound to the unbound state, the 5-methyl cytosine (5mC) levels partially revert to the baseline state at the CBS proximal free nucleosome spaces, while the memory of the bound state is retained at surrounding nucleosome positions, but only partially retained at the CTCF binding site. In contrast, repositioning alters the 5-hydroxymethyl cytosine (5hmC - the intermediate state between DNA methylation and demethylation) status at flanking nucleosomes, suggesting an intermediate epigenetic reprogramming event between the baseline and reorganized states.

Nucleosome asymmetry is mediated by CTCF binding, independent of the presence of cohesin, however cohesin is important for phasing of repositioned nucleosomes, particularly on the C terminal side. We also observed that the differential effects of Up and Downstream sequences as well as TF motifs that alter CTCF and cohesin signal strength, are reflected by similar positive and negative trends in nucleosome phasing and chromatin insulation effects. However, stronger CTCF and cohesin signals do not automatically translate into stronger insulation. Indeed, if phasing is disrupted by binding of peripheral CTCF ZFs and TFs to surrounding motifs, insulation is weakened. Thus, there is an additional layer of information that needs to be taken into account when considering how CTCF binding affects chromatin folding.

To examine the impact of transcriptional programming on CTCF binding, we looked at how variations in individual TF levels across different cell types are associated with changes in their enrichment at CTCF sites. In contrast to previous findings (*11, 15-19*) , we show that TFs can function to both stabilize and destabilize CTCF binding, as reflected by the direction of enrichment at increasing TF concentration. Allele specific validation highlights that SNPs affecting either CTCF or TF binding have reciprocal effects on the binding of each factor. Furthermore, 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 has a stabilizing or inhibitory effect 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. Together, these studies demonstrate that cell type-specific transcriptional programs, shape CTCF profiles linked to local and long-range chromatin organization and gene regulation.

Results

CTCF binding at differentially accessible sites is influenced by motif strength

To assess the impact of genomic context on CTCF binding, we examined motifs associated with accessible versus inaccessible CBSs in mESCs. This analysis revealed that accessible sites predominantly have motifs that closely resemble the consensus CTCF motif, compared to inaccessible sites which more often have a weaker motif in which the G in the triplet that ZF5 contacts, is replaced by an A, which we refer to here as an A_9 bias (**Figure 1A**). Analysis of the

human ENCODE IPSC WTC11 cell line reveals that accessible sites also predominantly resemble consensus motifs, however inaccessible sites do not have a bias towards a weaker motif which in humans is manifest as a G in the triplet that contacts ZF5, being replaced by a T (T_{10}). In both species, this bias is modest suggesting that a considerable number of accessible sites have a weak motif and many inaccessible sites harbor a strong motif, and that other factors affect binding stability at differentially accessible sites.

U and D motifs are enriched at inaccessible sites in mouse

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, as well as a less prevalent Downstream (D) motif can engage ZFs 9 to 11 and ZFs 1 and 2 to alter binding stability (*11-14*). In mESCs, we found that U and D motifs respectively located 7bps upstream and 11 bps downstream of the consensus CTCF motif, were more enriched at inaccessible sites associated predominantly with a weaker A_9 biased motif (**Figure 1B**). The percentage of weak and strong CBSs associated with U and D motifs at accessible and inaccessible sites for WT and CTCF mutants is shown in (**Figure S1A**). A comparable analysis of the human iPSC WTC11 cell line revealed that U motifs identified in our mESC data and other mouse-based studies, are rare in human cells and D motifs even more obsolete. The U window in human cells is more frequently comprised of a degenerate U motif (**Figure 1B**). In **Figure S1B** we show the percentage of weak and strong CBSs associated with U motifs at accessible and inaccessible sites for a range of different human tissues. Details of the ENCODE samples used in our analysis can be found in **Table S1A.**

U and D motifs have distinct impacts on CTCF and cohesin signal strength

The finding that U, D, and degenerate U motifs are found predominantly as an addendum to a weak CTCF motif with an A_9 bias at inaccessible sites in mESCs, suggests that the engagement of non-core binding ZFs could contribute to binding stability at these locations, increasing CTCF's capacity to block cohesin. To determine if this is the case, we analyzed the impact of U and D motifs on CTCF and cohesin (Smc3) profiles in mESCs expressing WT or mutant CTCF proteins with disrupted binding profiles (*9*). For this we used data previously generated from a CTCF complementation system in which endogenous CTCF tagged with an auxin inducible degron can be rapidly degraded, and CTCF expression rescued with either a Dox inducible FLAG-tagged wild-type or mutant CTCF expressed from the *Tigre* locus (*9*). The locations and nature of each

mutation are shown in **Figure 1C**, and the corresponding triplet to which each mutant ZF binds, is highlighted within the consensus motif on the left.

U and D motifs and the engagement of non-core binding ZFs have previously been shown to respectively stabilize and destabilize binding of CTCF in primary activated mouse splenic B cells (*12*), however their impact was not analyzed in the context of accessibility and nor was their effect

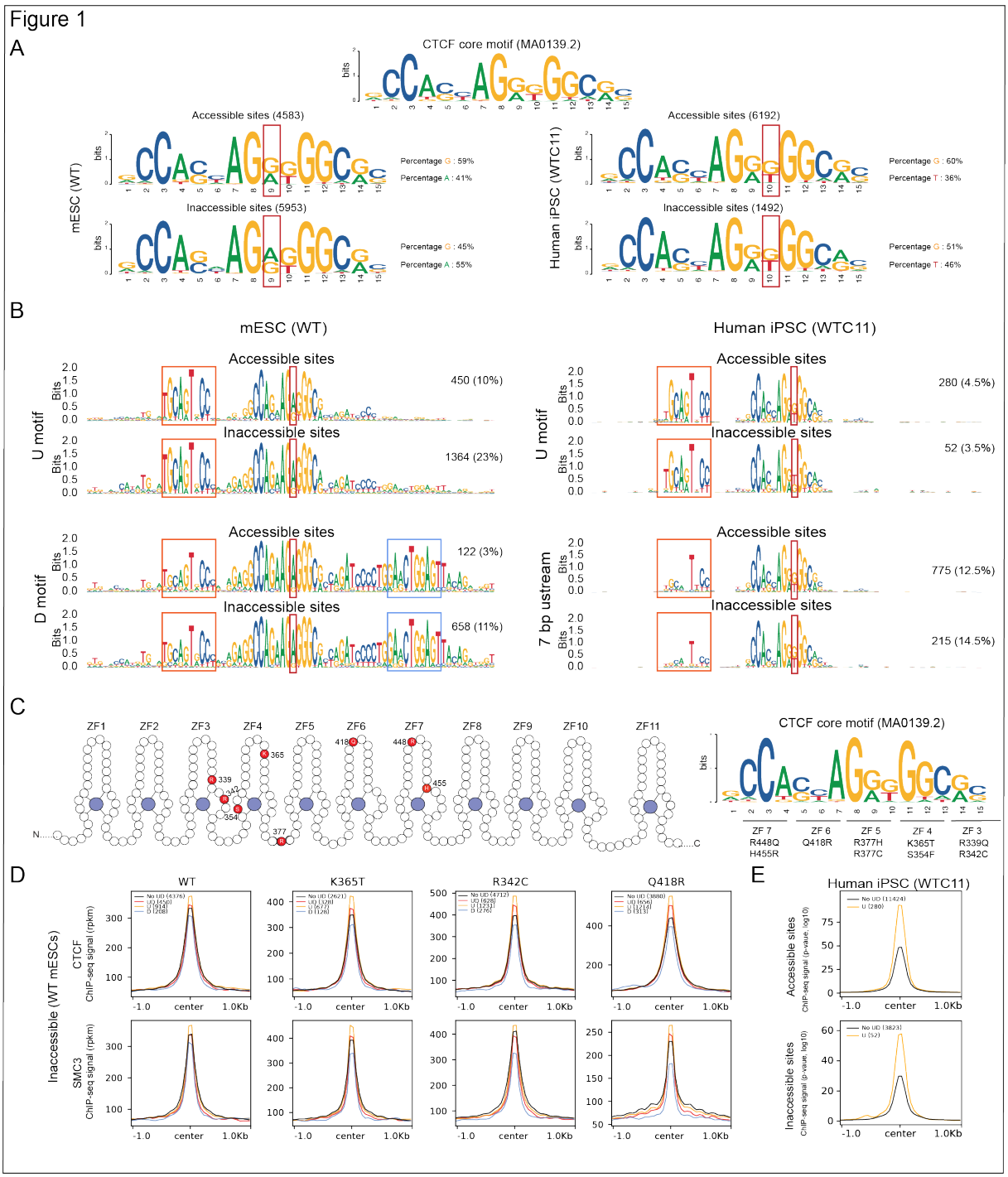


Figure 1: CTCF binding is stabilized by chromatin accessibility and the presence of CTCF Upstream motif (U) motif. A: The CTCF consensus motif (top) is compared with motifs found at accessible and inaccessible bound sites in WT mESCs (left panel) and human IPSCs (WTC11) (right panel). Inaccessible sites are enriched in weaker CTCF motifs with the A_9 bias in mouse, or a T_{10} bias in humans. **B**: The left panels show sequence logos of the 50 bp DNA sequences centered on the CTCF motif in the presence of the Upstream motif (U motif, top) or the U and Downstream motif (D motif, bottom) at accessible and inaccessible sites in WT mESCs. The presence of the U and D motifs are enriched at inaccessible sites, associated with a weaker CTCF motif (A_9 bias). The top right panels show the sequence logos of the 50 bp DNA sequences centered on the CTCF motif in the presence of the U motif at accessible and inaccessible sites in human IPSCs (WTC11). The U motif is associated with a weaker CTCF motif (T_{10} bias). In humans, fewer bound CTCF sites contain the U motif compared to mESC and no D motif was found. The sequence logos of the binding sites with any TF motif enrichment at the U motif position (7 bp upstream, bottom panels) suggest that the U motif is present in a degenerate form in humans. **C**: Illustration of the position of the CTCF mutations analyzed within CTCF (left). The DNA nucleotide triplets corresponding to each ZF that makes contact with the CTCF consensus motif are shown on the right. **D**: CTCF and Smc3 binding profiles from ChIP-seq at inaccessible sites in WT mESC and three CTCF mutants at CBSs containing the CTCF motif alone (no UD), CTCF and the U motif only (U), CTCF and the D motif only (D), or CTCF and both the U and D motifs (UD). The remaining mutants and binding profiles are shown for inaccessible and accessible sites in **Figure S1C and D**, respectively. **E**: CTCF binding profiles from ChIP-seq at accessible and inaccessible sites in WTC11 cells.

on cohesin overlap assessed. In WT CTCF expressing mESCs, consistent with previous studies that analyzed effects on CTCF signal strength (*12*), U motifs at accessible sites function to increase CTCF and cohesin peak strength and D motifs have a negative impact on their signals (**Figure S1C**). In the CTCF mutant mESCs the effect is analogous, with the impact of the U and D motifs enhanced in a mutation dependent manner (**Figure S1C**). A similar effect is seen at U and D enriched inaccessible sites, but with lower overall CTCF and Smc3 signal strength (**Figure S1D**). It is of note that WT and mutant CTCF proteins and corresponding cohesin profiles have similar signal strengths when U or UD motifs are present, indicating that the U motif is dominant and can counteract the negative effect of the D motif at accessible UD sites (**Figure S1C**). However, at inaccessible UD sites, while the presence of UD motifs generally has a positive impact on CTCF signal, the U motif is unable to fully compensate for the inhibitory effect of the D motif, and in WT and some mutant cells, the stabilizing effect of the combined UD motif is not sufficient to increase cohesin signal strength (**Figure 1D and S1D**).

Human cells have a paucity of U and D motifs

Analysis of CTCF binding in the human WTC11 cell line reveals that U motifs have a similar impact at accessible and inaccessible sites as in mESCs (**Figure 1E**), however there are far fewer U motifs (and almost no D motifs) compared to mouse cells, which we speculate could be because they have proportionately fewer inaccessible binding sites. The proportionally higher level of inaccessible binding sites in mouse is not restricted to mESCs and other mouse cell types display the same profile of binding (**Figure S1E** and **Table S1B**). Thus, although CTCF is highly conserved between human and mouse, our findings highlight binding site differences between the two species.

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, 15-19*), but there have been no comprehensive, unbiased analyses to determine how cell type-specific cofactors regulate CTCF binding and its loop forming capacity at accessible and inaccessible sites with weak and strong binding motifs. To investigate, we used spacing motif enrichment analysis (SpaMo (*21*)) to examine the enrichment of TFs at defined windows in the 35 bp region surrounding the CTCF binding site (**Figure 2A**). We focused on this region as we observed little enrichment of TF motifs beyond this point (**Figure S2A**), although a slight increase could be detected further upstream between 50-75bps, that could impact CTCF binding via a different mechanism. In **Figure 2A** 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. The U motif window (U win) was defined as binding ending at the 6 to 8 bp upstream gap, the U spacer as binding ending at the 0 to 7 bp upstream gap, the D spacer as binding beginning at the 0 to 9 bp downstream gap, and the D motif window (D win) as binding beginning at the 10 to 11 downstream gap.

TF motifs have distinct enrichment profiles in mice and humans

Overall, in the mESCs, the percentage of TF motif enrichment (excluding U motifs and U like motifs) at both upstream and downstream regions highlights that TF motifs are predominantly associated with inaccessible (71%) rather than accessible (59%) sites (**Figure 2A**). A similar analysis in human WTC11 cells revealed that in contrast to mESCs, TF motifs are more enriched overall at accessible (61%) rather than inaccessible (41%) sites (**Figure 2B**). The differences between human and mouse could reflect the fact that there are proportionately more accessible CTCF binding sites in human compared to mouse.

There is redundancy in the TFs that can bind a particular window

TFs with distinct motifs (green stacked bars) can be distinguished from U and D motifs (red and blue stacked bars, respectively). As shown here and in **Figure 1B**, human WTC11 cells have far fewer U motifs and almost no D motifs compared to mESCs. In the U window, TFs that resemble U motifs are also detected, but as with U motifs, the latter are less enriched in human compared to mouse cells (2.79% and 1.73% in WTC11 versus 25.5% and 11.85% in mEScs at accessible and inaccessible sites, respectively) as shown in **Figures S2C, and S3A, B**. We also detected a

small subset of U motifs associated with an Ebox motif in mESCs, that are enriched at inaccessible sites (**Figure S2B**). Motifs of other TFs enriched in the U and D windows and spacers and their preference for enrichment at accessible versus inaccessible sites are shown in **Figure**

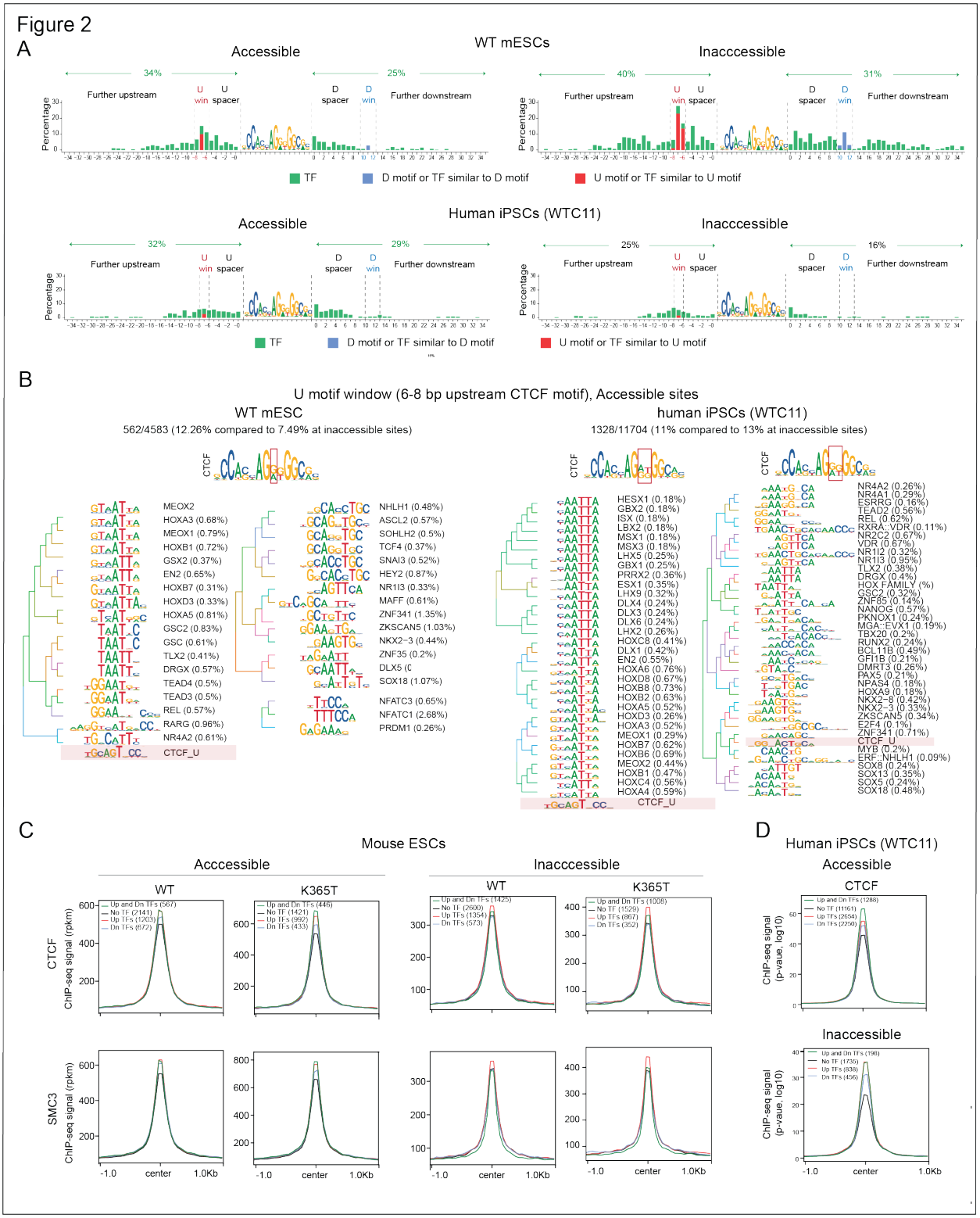


Figure 2: Bound CBSs are enriched in TF motifs within 35bps of the CTCF motif in mouse and human. A: Bar graphs showing the percentage of enriched TF motifs found using spacing motif enrichment analysis (SpaMo) at accessible and inaccessible sites in WT mESCs (top) and human iPSCs (bottom). The x-axis 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. The U motif window (U win) was defined as the 6 to 8 bp upstream gap, the U spacer as the 0 to 7 bp upstream gap, the D spacer as the 0 to 9 bp downstream gap and the D motif window (D win) as the 10 to 11 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). **C**: Motif logos of the enriched TF motifs found in the U motif window at accessible sites in mESCs (left) and human iPSCs (right). In mESCs, about half of the TFs in this window have a motif different from the U motif at accessible sites while at inaccessible sites most of them resemble the U motif (**Figure S2B**). In human, most of the TF motifs in this window are different from the U motif at both accessible and inaccessible (shown in **Figure S3A** - **C**), suggesting a less important role of ZF 9-11 in human. Members of the HOX family are highly represented in both species. **D**: CTCF and Smc3 binding profiles from ChIP-seq at accessible and inaccessible sites in WT mESC and one CTCF mutant 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. The remaining mutants CTCF and Smc3 binding profiles at accessible and inaccessible sites are shown in **Figure S2G and H**. **E**: CTCF binding profiles from ChIP-seq at accessible and inaccessible sites in human iPSCs 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.

2B, S2C, D, E and F for mESCs and **Figure 2B, S3A-G** for human WTC11 cells. The frequency with which each TF motif is found is shown as a percentage next to the motif in the figure. The majority of CTCF sites are enriched in one or more TF motif, with each factor having a small contribution. It is pertinent that motifs in each window and group of TFs show similarities within and between TF families such as the ETS family in the U spacer, or the HOX family in the U window, indicating there is redundancy in the factors that could bind, thereby providing a potential mechanism for cell type-specific regulation. TF enrichment for each window in WT and mutant mESCs is shown in **Table S2A**, while in **Table S2B,** TF enrichment in WTC11 cells and other human tissues is shown to highlight that TF motif enrichment at CTCF sites is detected across different cell types, but as expected, motif enrichment of TFs varies, which could reflect each cell's transcription program.

Enrichment of TFs at CTCF bound sites stabilizes CTCF binding in mouse and human

To determine whether enrichment of TF motifs is associated with changes in CTCF and cohesin (Smc3) signal strength, we analyzed their binding profiles at upstream or downstream regions enriched for TF motifs in both WT and mutant CTCF expressing cells. Enrichment of TF motifs at accessible CTCF bound sites in the upstream and downstream regions, or both together were associated with increases in CTCF and Smc3 signals in WT cells (**Figure 2C**). We observed a similar change in the mutant CTCF mESCs, but with variable effects, depending on the mutation (**Figure 2C and S2G**). The overall CTCF and Smc3 background profiles were lower at inaccessible CTCF bound sites (**Figure 2C and S2H**) and the impact of TFs upstream or downstream had a distinct effect compared to accessible sites. Upstream TF motifs always have a stronger impact than downstream motifs when present alone, however, while at accessible sites

enrichment of both together increased the signal strength more than either motif alone, at inaccessible sites the presence of downstream TF motifs had a negative impact on the stabilizing effect of upstream motifs. Importantly, enrichment of TF motifs is associated with a similar increase in CTCF binding strength in human WTC11 cells as observed in mESCs, suggesting that TFs play an equivalent role in stabilizing CTCF and promoting its role in blocking cohesin in the two species (**Figure 2D**). It is of note that motifs in up and downstream windows contribute in a distinct manner, in the two species, likely due to their distinct enrichment profiles.

In sum, the data in **Figure 2** shows that in human and mouse the majority of CTCF bound sites are enriched in one or more TF motifs. As with U motifs, their enrichment appears to be functionally important for increasing the stability of CTCF binding and CTCF's capacity to block cohesin. Motifs within each window and group of TFs show similarities within and between TF families indicating redundancy in the cofactors that can bind, thereby providing a mechanism for cell type-specific differences in regulation. These findings suggest that cell type-specific binding of CTCF might be influenced by a large pool of TFs that reflect the transcriptional program of each cell type. The analyses performed here provide a resource for identifying which TFs have the potential to influence CTCF binding and chromatin organization in different cell types (**Tables S2A, B**).

CTCF repositions and phases surrounding nucleosomes in a context specific manner

To ask how CTCF binding affects nucleosome positioning and DNA methylation in a context-specific manner, we performed long-read Nano-NOMe-seq and WGBS-NOMe-seq in WT mESCs. Both assays allow us to assess nucleosome positioning using exogenous 5mC at HGC (GC excluding CGC) and endogenous 5mC/5hmC at WCG (CG excluding GCG and CCG) at the single molecule level. WGBS-NOMe-seq generates shorter informative molecules and in contrast to Nano-NOMe-seq is unable to distinguish 5mC from 5hmC. Both assays show similar results, although the baseline estimation of endogenous and exogenous methylation is lower in Nano-NOMe-seq compared to WGBS-NOMe-seq (**Figure S4A**). To assess the effect of CTCF motif orientation and to discriminate bound and unbound molecules at CTCF peaks, we restricted our analyses to CTCF peaks with GpC-containing CTCF motifs after confirming that nucleosome occupancy and DNA methylation at these sites were similar (**Figure S4B**).

CTCF-mediating repositioning of nucleosomes is asymmetric around the CBS

Our analyses demonstrate that at CTCF ChIP-seq peaks, a substantial fraction of molecules are temporarily unbound. This is expected since DNA-protein interactions are dynamic and the proportion of bound versus unbound fractions reflect CTCF's residence time at each subset of peaks analyzed. Rather than comparing CTCF-bound sites to the subset of permanently unbound CBSs in the mESCs, temporarily unbound molecules were used to control for unknown confounding factors when assessing the context-specific effects of CTCF binding. Our data confirm previous findings demonstrating the displacement of and phasing of flanking nucleosomes after CTCF binding, but we found nucleosomes are repositioned asymmetrically with the N terminal nucleosome being displaced further away from the CTCF binding site than the C terminal nucleosome (**Figure 3A**). On either side we found phased nucleosomes positioned at specific locations. The same is true for both accessible and inaccessible sites, however, at accessible versus inaccessible sites the downstream (CTCF N-terminal) and upstream (CTCF C-terminal) nucleosomes are stabilized at 130 bp and 110 bp versus 80 bp and 110 bp from the CBS, respectively. Thus, at inaccessible sites the nucleosome spacing is narrower. Furthermore, surrounding nucleosomes are less well phased at inaccessible sites (**Figure 3A**). Similarly, when we compare CBSs with strong versus weak CTCF signals, we find that while the positioning of asymmetric nucleosome is retained at accessible and inaccessible sites, weak CTCF peaks are associated with less well-organized phasing of flanking nucleosomes.

Smc3 is required to phase CTCF-mediated asymmetrically repositioned nucleosomes

We detected similar single molecule nucleosome occupancy patterns when comparing strong CTCF signals with bound sites overlapping strong Smc3 signals (**Figure 3B**). This is expected, as we previously showed that Smc3 signals overlap binding sites with strong CTCF signals, which are associated with longer CTCF residence times (*9*). However, when we compared the subset of CTCF peaks without Smc3 to those with strong Smc3 signal we observed a higher proportion of unbound molecules (33% versus 21%), indicating that those sites have lower residence times. Again, this is to be expected as Smc3 free CTCF bound sites have weak signals (*9*). Of note, in the absence of Smc3, the single molecule heatmap shows that nucleosomes are still asymmetrically displaced by CTCF binding, however surrounding nucleosomes are not well phased, especially on the C terminal side of CTCF. Thus, cohesin is required for phasing flanking nucleosomes.

Unoccupied CBSs are bound by other factors

We next analyzed CBSs that are not bound by CTCF in mESCs. Surprisingly, most of the molecules at those sites are occupied (59% and 72% respectively at accessible and inaccessible sites) with a narrow footprint reflecting TF binding rather than nucleosome occupancy. This result suggests that binding of TFs in the vicinity of a CTCF motif can displace CTCF and phase surrounding nucleosomes (**Figure 3C**). In contrast to when CTCF is bound, TF-mediated repositioning is symmetrical around the CTCF binding site, with the C and N terminal nucleosome being pushed equivalent distances. Although the nucleosome positioning of temporarily unbound molecules at CTCF peaks and CTCF-free sites were similar, TFs at bound sites appear to disrupt phasing of surrounding nucleosomes, with greater affects at accessible sites, suggesting that the more TFs present, the greater the impact on phasing.

CTCF binding alters DNA methylation

CTCF binding has previously been described as methyl-sensitive due to the presence of CpGs in its motifs and its involvement in imprinting. However, several studies have pointed out that only a fraction of CTCF binding sites are affected by DNA methylation (*22*). In line with this, we found that the majority of CTCF-motif containing peaks (60%) do not contain a CpG. Furthermore, there was no detectable difference in nucleosome occupancy at CTCF peaks with or without CpG-containing motifs (**Figure S4C**). Although the baseline DNA methylation level of non CpG-containing sites is higher than CpG containing CBSs, the DNA methylation level at CTCF footprints and the directly flanking upstream and downstream nucleosome-free regions are similarly hypomethylated (**Figure S4C**). This suggests that the hypomethylation associated with CTCF binding, outside of imprinting regions, is more likely to be a consequence of CTCF binding rather than a prerequisite for occupation, similar to hypomethylation associated with TF binding. Indeed, DNA methylation (5mC and 5hmC) aligns with nucleosome positioning and CTCF footprinting, with higher methylation levels at unoccupied regions and demethylation at occupied regions (**Figure 3D)**. The 5mC partially reverts to the high baseline levels at the CBS proximal nucleosome free spaces, while the memory of the bound state is retained at surrounding phased nucleosomes and at the CTCF binding site. In contrast, the 5hmC methylation level of unbound molecules is decreased at the nucleosome-free spaces of the bound molecules, reflecting an early sign of demethylation associated with the repositioning of nucleosomes.

Surrounding motifs affect CTCF-associated nucleosome phasing

U and UD motifs are associated with more stable CTCF binding and D motifs with less stable binding, as reflected by the strength of associated CTCF ChIP-seq signals (**Figure 1D**) and the

proportion of bound molecules (**Figure S4D**). Similarly, enrichment of up and downstream TF motifs is associated with differential effects on CTCF binding, as reflected by the

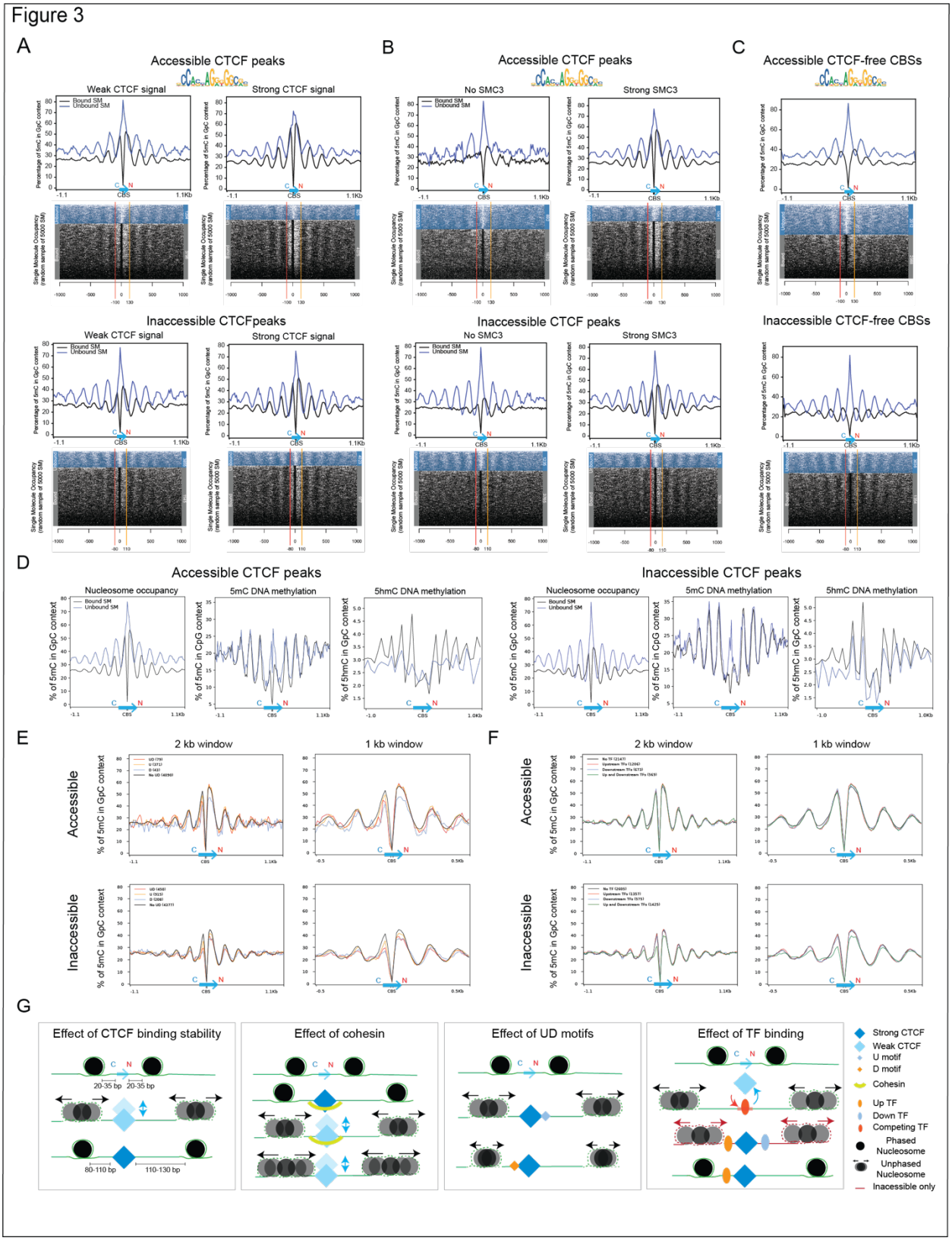


Figure 3: CTCF repositions and phases surrounding nucleosomes in a context specific manner. A: Profiles and single molecule heatmaps showing the nucleosome occupancy (NO) and CTCF binding footprinting from nano-NOMe-seq for weak and strong CTCF peaks at accessible and inaccessible sites. The data are centered on the 15-bp consensus motif oriented on the positive strand (motifs on the negative strand have been reoriented toward the positive strand). The CTCF core 15 bp-motif oriented on the positive strand is shown on top. 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. Both bound (black) and temporarily unbound (blue) molecules at CTCF bound sites (ChIP-seq peaks) are shown and provide an estimate of the CTCF bound fraction at a given time point for strong and weak CTCF ChIP-seq peaks. Black (or blue) and white circles 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:** Profiles and SM heatmaps showing the nucleosome occupancy and CTCF binding footprinting for weak and strong SMC3 signals at accessible and inaccessible CTCF peaks. **C**: Profiles and SM heatmaps 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. The narrow footprinting at CTCF binding sites corresponds to the binding of other TFs. **D**: Profiles showing the exogenous GpC 5mC level (nucleosome occupancy), endogenous CpG 5mC (DNA methylation and 5hmC, intermediate DNA methylation state in a WGC context) at accessible and inaccessible CTCF peaks for bound and unbound molecules. **E**: Profiles showing the nucleosome occupancy and CTCF binding footprinting of bound molecules at accessible and inaccessible CTCF peaks in the absence or presence of the U and/or D motifs. The heatmaps of bound and unbound molecules are shown in **Figure S5D. F**: Profiles showing the nucleosome occupancy and CTCF binding footprinting of bound molecules at accessible and inaccessible CTCF peaks in the absence or presence of upstream and/or downstream TF motifs. The heatmaps of bound and unbound molecules are shown in **Figure S5E**. **G**: Cartoon depicting the effect of CTCF, cohesin, UD motifs and TF binding on nucleosome occupancy at CBSs. The range of the nucleosome distance from the CTCF motif corresponds to the stable position of the first upstream and downstream nucleosomes in inaccessible (lower values) and accessible (higher values) regions in the presence of more stable CTCF binding. The CTCF motif is oriented on the positive strand, the TF motif orientation is unknown and therefore not oriented.

strength of CTCF ChIP-seq signals (**Figure 2C**) and the proportion of bound versus unbound molecules (**Figure S4D, E**). The presence of surrounding motifs does not impact the asymmetric repositioning of N and C terminal nucleosomes at accessible and inaccessible sites, but their enrichment alters phasing of surrounding nucleosomes with a similar positive and negative trend (**Figure 3E, F**) to the way they impact CTCF and cohesin binding (**Figure 1D and 2C**). However, for U and D sequences as well as TF motifs, we observe that while destabilizing motif combinations at differentially accessible sites disrupt phasing, stabilizing effects do not always translate into more organized nucleosomes (**Figure 3E and F**). Thus, CTCF and cohesin signal strength can be uncoupled from nucleosome phasing. In **Figure 3G** we show a cartoon representing the effects of CTCF binding in different genomic contexts on nucleosome repositioning and stabilization.

CTCF mutations are associated with altered nucleosome phasing at lost sites

WT CTCF and mutant R342C have similar residence times, but R342C has a substantial number of lost 'WT only' CTCF peaks and 'gained *de novo*' CTCF peaks (*9*). Comparing the effect of WT and mutant proteins by single molecule Nano-NOME-seq analysis, we observed similar effects on nucleosome positioning at weak and strong peaks, with the R342C CTCF mutant having

slightly less stable nucleosome repositioning overall (**Figure S4F**). A stratified analysis of 'WT only' weak peaks, 'common' strong peaks and 'mutant only' weak peaks showed a loss of CTCF-dependent nucleosome repositioning at WT only sites, while binding of mutant protein to common

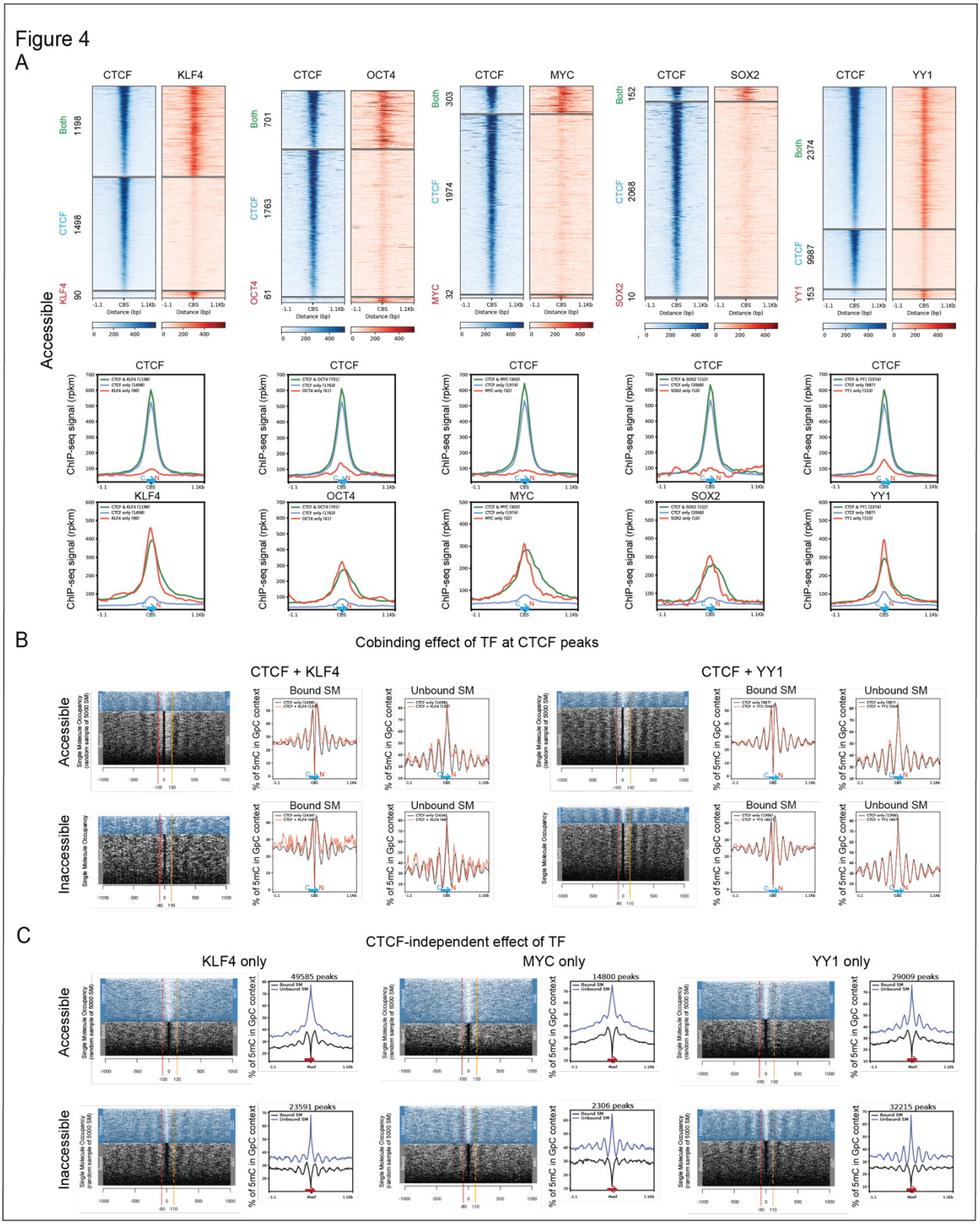


Figure 4: Individual TF strengthen CTCF signal and disrupt CTCF-dependent nucleosome phasing. A: Heatmaps (top) and profiles (bottom) of CTCF (FLAG) and 5 TFs (including 4 pioneer TFs, KLF4, OCT4, MYC and SOX2) showing the effect of (i) TF binding on CTCF signal strength and (ii) CTCF binding on TF signal 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 S5A**. **B**. TF data were downloaded from GEO (GSE201852 and GSE236376). **B**: Single molecule heatmaps of the nucleosome positioning centered on CBSs oriented on the positive strand at sites co-bound by CTCF and KLF4, or YY1 for bound (black) and unbound (blue) molecules. The profiles compare the nucleosome position at co-bound (red) versus CTCF only bound sites (without SpaMo enriched TF motifs, or KLF4, or YY1 binding detected by ChIP-seq) (black) for bound and unbound SM. The effect of MYC, OCT4 and SOX2 are shown in **Figure S5B. C:** Single molecule heatmaps and profiles demonstrate the CTCF-independent effect of each TF on nucleosome positioning. Each heatmap is centered on the KLF4, MYC or YY1 motif (oriented on the positive strand) at sites with no bound CTCF or CTCF motif. The data for OCT4 and SOX2 are shown in **Figure S5C.**

and *de novo* sites had no effect (**Figure S4G**). These findings also highlight that nucleosomes at *de novo* CTCF binding sites do not have the same repositioning and phasing as those found at *bona fide* CTCF binding sites.

Individual TFs strengthen CTCF signals and disrupt CTCF-dependent nucleosome phasing

We have shown that enrichment of TF motifs in the 35bps surrounding CTCF bound sites can strengthen CTCF signals and cohesin overlap and impact nucleosome phasing (**Figure 2C and D**). To further investigate, the relationship between TFs, CTCF signal strength and nucleosome repositioning we analyzed ChIP-seq binding of individual TFs (KLF4, OCT4, MYC, SOX2 and YY1) in mESCs using data downloaded from GEO (GSE201852 and GSE236376). Our analysis, focused on CBSs with the consensus CTCF motif, highlighting accessible (**Figure 4A**) and inaccessible sites (**Figure S5A**) at which CTCF binds alone, CTCF and individual TFs cobind and TFs bind alone. The ChIP-seq profiles validate our motif enrichment analysis, demonstrating that the presence of individual TFs stabilize CTCF binding, with an increased effect at inaccessible sites for most TFs.

The heatmaps also demonstrate that at accessible sites, CTCF free CBSs occupied by TFs are under-represented but associated with a stronger TF signal than that found at cobound sites. (**Figure 4A**). This is true for most TFs analyzed, even at the less frequent inaccessible sites (**Figure S5A**). 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 when motifs are in close proximity, or partially overlapping the CBS. At cobound sites however, with the exception of YY1, whose binding profile almost overlaps that of CTCF, we observe preferential binding of the TFs within the CTCF-dependent N-terminal nucleosome free space at accessible sites.

TF-mediated stable binding can be uncoupled from nucleosome phasing

Single molecule footprinting of cobound CTCF sites with KLF4 or YY1 (**Figure 4B**) and CTCF with MYC, OCT4 or SOX2 (**Figure S5B**) reposition nucleosomes in a manner that maintains the asymmetrical N terminal bias. However, when profiles of CTCF binding alone (black) are compared with cobound CTCF plus TF (red), we found that KLF4, MYC, OCT4 or SOX2 (which bind within the CTCF-dependent nucleosome free space) disrupt the nucleosome phasing around CTCF at both accessible and inaccessible sites. In contrast, YY1, whose binding almost overlaps that of CTCF, has a minimal effect. Together these findings indicate that, while binding of TFs can stabilize CTCF, their presence can also disrupt nucleosome phasing, depending on where they bind relative to CTCF in the nucleosome free space.

Single molecule footprints of each TF independent of CTCF (sites with no bound CTCF or CTCF motif) induce a symmetrical nucleosome-free space around their footprints while YY1 shows a modest upstream bias at bound versus unbound sites (**Figure 4C and S5C**). Like CTCF, the footprinting sites remain nucleosome free on temporarily unbound molecules. The proportionately larger unbound fraction of TFs suggests a shorter residence time compared to CTCF, which could also explain the weaker nucleosome phasing for both bound and unbound molecules. That phasing at accessible sites is weaker than at inaccessible sites, suggests co-binding of other TFs at these locations, which further disrupt nucleosome organization. This data demonstrates that each of the TFs analyzed can reposition nucleosomes in a TF specific manner, which is reflected by effects on nucleosome phasing at CTCF co-bound sites as shown in **Figure 4B and S5B**. The results were similar when analyzing TF bound sites at CTCF free CBSs (**Figure S5D**).

Surrounding motifs affect CTCF insulation mimicking their effects on nucleosome phasing

We have shown that the presence of U, D and TF motifs can stabilize and destabilize CTCF and Smc3 binding and alter nucleosome phasing with similar trends. We also demonstrate that stronger signals do not automatically translate into more organized phasing. To investigate whether disruption of CTCF/cohesin-mediated phasing has an impact at the level of 3D chromatin organization, we analyzed the effect of U, D and TF motifs on insulation, using Hi-C data from WT mESCs. Intriguingly, we found that enrichment of U, UD and D motifs alters insulation at differentially accessible sites in a manner that mirrors their effect on nucleosome phasing at differentially accessible sites (**Figure 5A and 3E**). Specifically, if nucleosome phasing is disrupted, insulation is weakened and vice versa. We observed a similar effect for TF motifs. At accessible sites, the enrichment of TF motifs in the vicinity of CBSs is associated with a modest increase in insulation in both WT and mutant mESCs, while at inaccessible sites, their presence

is associated with a decrease in insulation, especially in the presence of both upstream and downstream TF motifs (**Figure 5B**) that have a strong negative impact on phasing effect (**Figure 3F**).

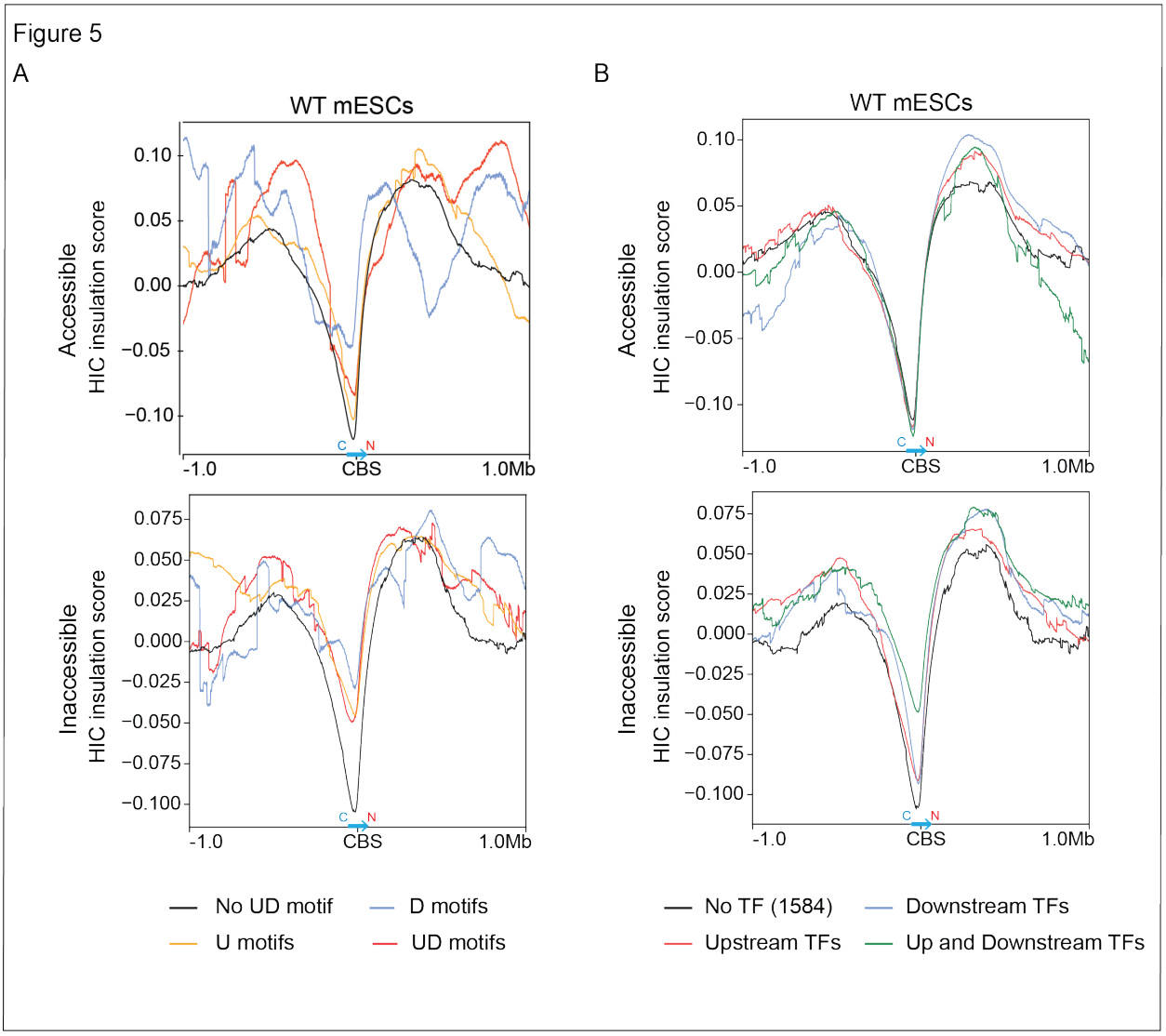


Figure 5. Surrounding motifs affect CTCF insulation mimicking their effects on nucleosome phasing. A: Profiles show reduced CTCF-mediated insulation at CTCF peaks in the presence or absence of Upstream and/or Downstream motifs at accessible and inaccessible sites in mESCs. A higher insulation score reflects lower insulation. The insulation scores are centered on the CTCF motif oriented on the positive strand. This data was generated at 10 kb resolution. **B**: Profiles show changes in CTCF insulation at CTCF peaks with or without enriched TF motifs identified by SpaMo, at accessible and inaccessible sites in WT mESCs.

Transcriptional reprogramming contributes to cell type-specific CTCF binding

Our analyses demonstrate that cofactor motif enrichment can influence CTCF's binding stability and ability to block cohesin, 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 (i) WT mESCs, and (ii) in human IPSCs. Overall, we found that TFs that bind motifs abutting CTCF were expressed at a higher level at accessible sites in both species (**Figure 6A**). However, in mESCs inaccessible sites are more sensitive to changes in TF expression levels, while in hiPSCs it is the reverse way round, as shown by the significant difference in the correlation slopes in the graphs in **Figure 6B**. This effect is independent from the differential effects of the window where the TF is found and the TF orientation (**Figure S6A-**

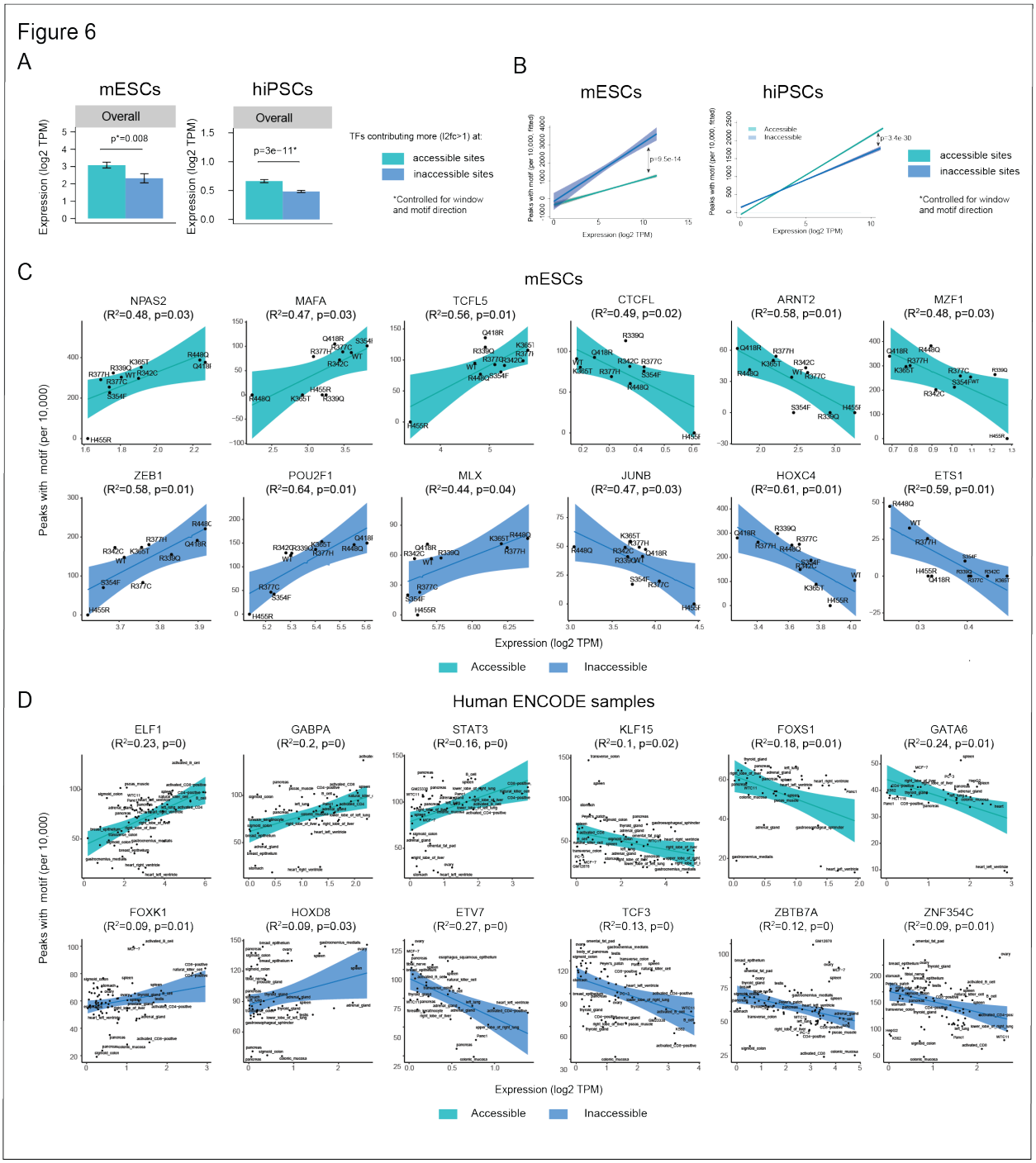


Figure 6: The percentage of TF motifs within 35 bp of CTCF consensus motif correlates with their expression level in WT and mutant mESCs. A: Bar graph showing a 1.9-fold (corresponding to a 0.92 difference in log2) increase in expression of TFs with a higher contribution at accessible versus inaccessible sites (≥2 fold differential contribution) in WT mESCs, 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 by the multivariate linear model. In human, the increase is more modest (1.13 fold change, 0.18 difference in log2). **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 mESCs and human iPSCs. The shaded areas correspond to the 95%-confidence intervals. The independent effects of the windows and TF orientation are shown in Figure **S4A, B**. **C**: Examples of significant correlation between TF expression level and the percentage of CTCF peaks associated with an individual TF motif at accessible and inaccessible sites in mESCs. The shaded areas correspond to the 95%-confidence intervals. **E**: Examples of significant correlations between TF expression level and the percentage of CTCF peaks associated with the TF motif at accessible and inaccessible sites in human iPSCs.

B). We speculate that differences in the response to changes in TF expression at inaccessible versus accessible sites in mouse and human may be explained by the fact that there are proportionately more inaccessible sites in mouse compared to humans.

To examine the effect of increasing transcriptional output on CTCF, we analyzed the impact of individual TFs in WT and mutant CTCF expressing cells, which all have their own unique transcriptional program (*9*). Each TF has an increasing or decreasing association with CTCF bound sites depending on whether their effect is stabilizing or destabilizing (**Figure 6C**). A similar analysis was done in humans, examining the effect of variations in TF levels across 47 different ENCODE cell lines on their enrichment at CTCF binding sites. Our findings again demonstrate that variations in TF output in each cell type, impacts the enrichment of individual TFs at CTCF bound sites, but with higher significance given the larger sample size (**Figure 6D**). The finding that TFs can inhibit CTCF binding at a subset of sites is consistent with the data in **Figure 3C and 4A**, showing that the subset of CBSs that are not bound by CTCF, are instead bound by other TFs. The enrichment of a particular TF in each cell type is dependent on its transcriptional output, highlighting a mechanism by which cell type-specific transcriptional programs can impact CTCF binding profiles.

SNPs reveal 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 S3A**) 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. For a subset of TFs (**Table S3B**), we then performed allele-specific analysis using ENCODE data from GM12878 to identify SNPs that impact CTCF or TF motifs (binding affinity). In **Figure 7A, 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 7A** is linked in GWAS to colon cancer and cardiometabolic traits.

Conversely, SNPs that affect binding of TFs on one allele can impact binding of CTCF at overlapping sites. In **Figure 7C**, 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. This highlights an example of a TF that binds competitively with CTCF and

Figure 7

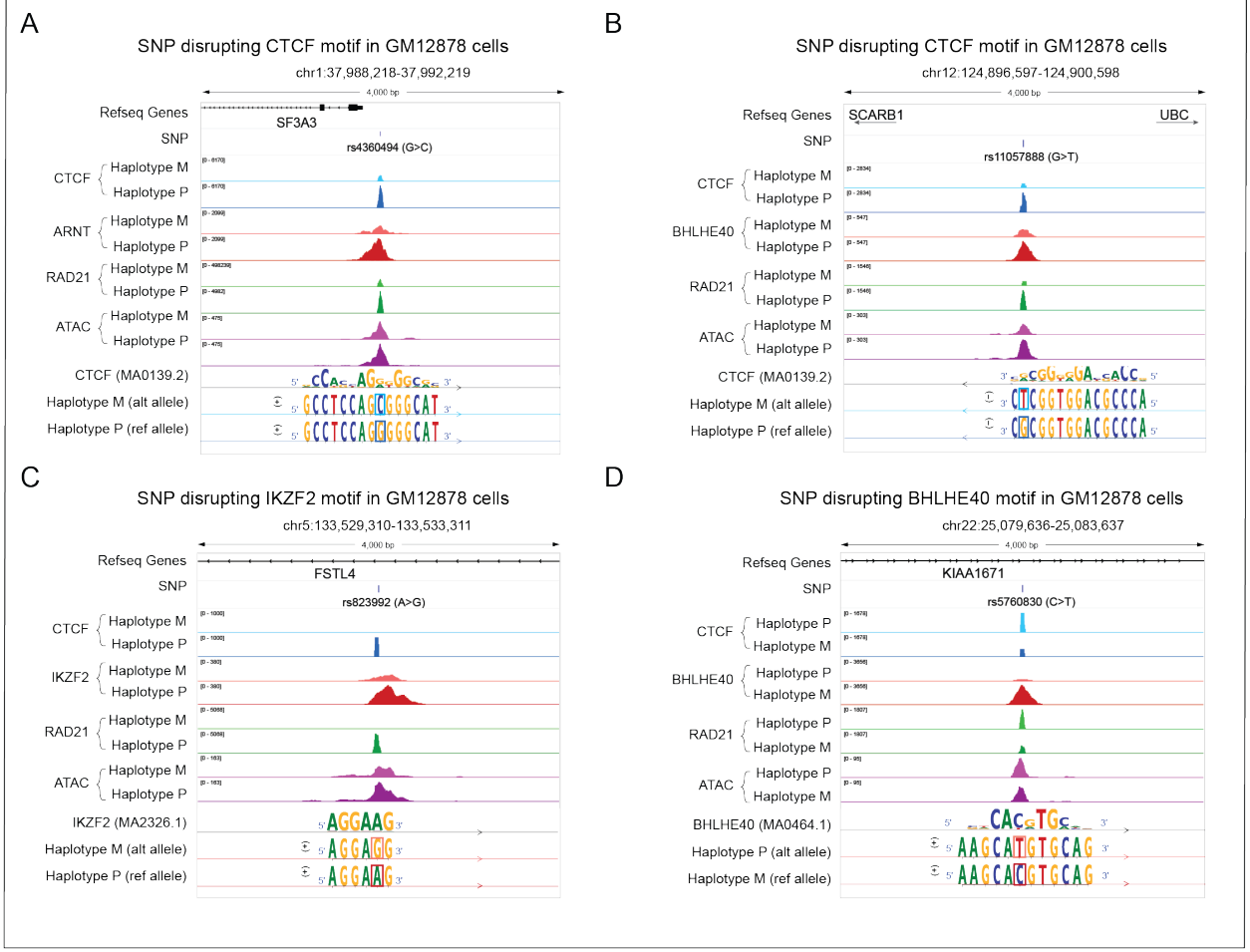


Figure 7. **SNPs reveal both the cooperativity and competitiveness of CTCF-TF cobinding**. **A**: Map showing allele-specific binding (ASB) of CTCF due to a SNP disrupting its motif on the maternal allele in the GM12878 cell line. ARNT binding is also disrupted on the CTCF unbound allele. The loss of CTCF binding is associated with cohesin (RAD21) loss but the chromatin accessibility (ATAC) is similar on both allele suggesting that in this region accessibility is CTCF independent. 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. Of note, the ASB SNP, rs4360494 has been identified as a GWAS SNP associated with colon cancer and cardiometabolic traits. 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.

together with the data **Figure 7B** demonstrates that a single TF can have both a stabilizing and destabilizing effect as shown in **Figure 4A and S5A**.

Discussion

Although it is known that a subset of CTCF binding is cell type specific, the mechanisms underlying state specific control have not been fully elucidated. Here we used WT and a series of mutant CTCF ChIP-seq mESCs data sets previously generated by our lab (*9*), combined with CTCF ChIP-seq from ENCODE data in a wide range of human cell types to (i) characterize accessible versus inaccessible sites, and (ii) determine what factors influence CTCF binding stability and function within these two chromatin environments. Although CTCF is highly conserved in mice and humans, we found different features associated with their respective binding sites. Specifically, mouse cells have proportionately more inaccessible binding sites compared to human cells. In parallel, Up and Downstream CTCF sequences as well as TF motifs are enriched at inaccessible sites in mouse, which we have previously shown generally have weaker CTCF signals and reduced cohesin overlap compared to accessible sites. Since Up and Downstream CTCF sequences and TF motifs stabilize CTCF binding, we speculate that their enrichment at inaccessible sites in mouse, restores the balance of strong versus weak CTCF/cohesin signals in the two species, compensating for the relative reduction of accessible sites in mouse.

Consistent with previous findings (*12*), we show that Upstream (U) motifs contribute to the stability of WT CTCF binding, facilitating its ability to block cohesin on chromatin, while Downstream (D) motifs have a destabilizing effect. Here, we further reveal that when U and D are found together, the stronger stabilizing U motif can overcome the negative impact of the D motif, but only at accessible sites. U and D motifs have enhanced effects on mutant proteins with impaired binding

properties, suggesting they are important for regulating their stability. The presence of U or degenerate U motifs in human cells have a similar stabilizing effect on CTCF, although there are less of them and even fewer D motifs, perhaps because as mentioned above, they are not as important for CBSs found in a predominantly accessible chromatin environment.

CTCF has been shown by us and others to act as a pioneer factor, binding inaccessible as well as accessible chromatin (*9, 23*). Although it is known that chromatin remodelers, such as the NURF component, BPTF can create accessibility at CTCF sites (*10*), it is not known what factors direct remodelers to these locations. To follow up on previously published studies hinting at a role for TFs in cell type specific binding of CTCF in recruiting remodelers, we performed a comprehensive analysis of enrichment of TF motifs in the 35bp region of the CBS in mouse using the WT and nine different CTCF mutant mESC lines (all of which have distinct transcriptional profiles) and 85 ENCODE cell lines. We identified a wide range of motifs corresponding to expressed cell type-specific factors that could perform this function. In human and mouse, the majority of CTCF bound sites are enriched in motifs located within 35 bps of CTCF. Importantly, we demonstrate that CTCF and cohesin signals are strengthened at sites enriched with TF motifs, and by analyzing the effect of a handful of cofactor TFs on CTCF binding profiles, we validated their stabilizing effect. Within a given window, enriched TF motifs show a strong similarity, suggesting redundancy in the regulation of CTCF binding. Furthermore, motifs are degenerate which could allow for differential effects of TFs from the same class or motif cluster. Given the redundancy between factors that can bind, we speculate that at a single CTCF binding site, the presence of distinct TFs in different cell types could fine-tune CTCF stability and function in a lineage or state specific manner.

Use of long read, nano-NOME-seq single molecule analysis allowed us to interrogate the impact of CTCF binding at temporarily unoccupied versus occupied sites within different genomic contexts, which reflects CTCF's residence time. While fluorescent recovery after photobleaching (FRAP) can determine global residence times of CTCF binding (*9*), single molecule analysis enables us to examine changes in the proportion of sites with bound versus temporarily unbound CTCF at distinct subsets of genomic sites, which is not yet feasible by imaging analysis. As expected, we see more unbound molecules at the more dynamically bound inaccessible sites with weak signals, compared to more stable binding at accessible sites with strong signals. We also see the expected changes in the proportion of bound versus temporarily unbound sites at

locations enriched for Up and Downstream CTCF sequences and TF motifs, which we have shown impact CTCF binding stability.

Single molecule analysis reveals that at all CTCF bound CBSs, while nucleosomes are symmetrically positioned around unoccupied sites, they are asymmetrically repositioned around the bound CTCF molecule, such that the N terminal nucleosome is pushed farther away than the C terminal nucleosome. Although weak and less stable CTCF bound sites are associated with more unphased nucleosomes, they retain the asymmetry of repositioned CTCF flanking nucleosomes, however, at inaccessible sites nucleosome spacing is narrower and surrounding nucleosomes are less well phased. These findings highlight the impact of more stable versus dynamic binding, which might be required to achieve nucleosome phasing. Of note, we found that at permanently unbound CBSs in mESCs, CTCF motifs are not masked by nucleosome occupancy but rather by competing TF binding and that nucleosomes are also reorganized by binding of TFs that preclude CTCF binding. However, less phasing is consistent with a shorter average TF residence time compared to CTCF (*24*). Analysis of the binding of a handful of TFs at CBSs devoid of CTCF binding confirm these findings and demonstrate that nucleosomes are symmetrically positioned around their binding sites.

We and others have shown that cohesin binds the N terminal portion of CTCF (*25-28*) and we hypothesized that this interaction could underlie N-terminal weighted asymmetric nucleosome repositioning. However, we found that asymmetry was preserved when we compared nucleosome reorganization at sites with weak CTCF signals, in the presence or absence of cohesin. Nonetheless, there was a striking difference in the organization of the surrounding nucleosomes when cohesin was absent. Thus, while CTCF displaces nucleosomes, cohesin is required for phasing them in an ordered way, suggesting a link between chromatin folding and phasing. Indeed, we found that enrichment of surrounding motifs can also affect the phasing of surrounding nucleosomes with similar positive and negative trends to their effect on CTCF and cohesin stability at differentially accessible sites. However, we showed that stronger CTCF and cohesin signals do not automatically translate into stronger phasing or stronger insulation. Indeed, disruption of phasing by binding of peripheral CTCF ZFs and TFs to surrounding motifs weakens insulation. Thus, CTCF signal strength and nucleosome phasing are better at predicting insulation than CTCF signal strength alone.

Binding of CTCF is affected by DNA methylation (5mC) at only a subset of sites, as discussed previously(*22*). Here we observed that CTCF binding itself alters DNA methylation, rather than the other way around. The 5mC partially reverts to the high baseline levels at CBS proximal nucleosome free upstream and downstream spaces, while the memory of the bound state is retained at surrounding nucleosome positions but only partially retained at the CTCF binding site. In contrast, repositioning alters the 5-hydroxymethyl cytosine status at flanking nucleosomes, indicating a latency in epigenetic transition. Since the residence time of CTCF binding is around ten minutes (*9*), this window of time could be sufficient to induce changes in 5mhC levels but insufficient for induction of changes to the 5mC status.

We have shown that enrichment of a wide range of TF motifs stabilize CTCF binding and confirmed this by analyzing the ChIP-seq of a handful of factors at CBSs. The significance of these findings is underscored by our analysis of WT and nine different mutant CTCF expressing mESC cell lines and human tissues from ENCODE, all of which have unique transcriptional programs with individual TFs expressed at different levels. We demonstrate that increases in the expression level of a TF can both increase or decreases the enrichment of its motif at CTCF bound sites, depending on whether the TF has a stabilizing or destabilizing effect. These findings were validated by allele specific analysis which demonstrate that CTCF and TFs have a reciprocal effect on each other's binding such that SNPs affecting either the binding of CTCF or the TF will impact enrichment of the other factor, altering cohesin overlap and accessibility. SNPs in CBSs or TF motifs reveal both the cooperativity and competitiveness of CTCF-TF cobinding. Our finding that cell type-specific TFs can have a negative effect on CTCF binding addresses another long-standing question about why many accessible CBSs remain unbound. Here we show that binding of other factors can compete with CTCF to prevent its occupation. Together, these studies define a mechanism by which transcriptional reprogramming during development and in disease specific states might determine cell type-specific CTCF profiles, linked to local and long-range chromatin organization.

METHODS

Cell line generation and culture

The construction of the vector for cloning transgenic, doxycycline-inducible expression of WT and mutant mouse *Ctcf* cDNA and gene targeting in mouse embryonic stem cells E14Tg2a (karyotype 19, XY; 129/Ola isogenic background) were performed as described in Do *et al., (9)*. Mouse embryonic stem cells and all clones derived from these were cultured under feeder-free conditions

in 0.1% gelatin (Sigma ES-006-B) coated dishes (Falcon, 353003) as described in Do *et al., (9)*. Briefly, the cells were grown in DMEM (Thermo Fisher, 11965-118) supplemented with 15% fetal bovine serum (Thermo Fisher, SH30071.03), 100 U/ml penicillin - 100 µg/ml streptomycin (Sigma, P4458), 1 X GlutaMax supplement (Thermo Fisher, 35050-061), 1 mM sodium pyruvate (Thermo Fisher, 11360-070), 1 X MEM non-essential amino-acids (Thermo Fisher, 11140-50), 50 µM b-mercaptoethanol (Sigma, 38171), 10^4 U/ml leukemia inhibitory factor (Millipore, ESG1107), 3 µM CHIR99021 (Sigma, SML1046) and 1 µM MEK inhibitor PD0325901 (Sigma, PZ0162).

Induction of auxin inducible degradation of CTCF and doxycycline induced expression

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

WGBS and Nano-NOMe-seq sample and library preparation

Nuclei extractraction, GpC methyl-transferase treatment and DNA extraction

After 2 days of IAA and Dox treatment on CTCF mutant cells, 1 million cells were collected and washed with cold PBS at 300g at 4°C. As described in Battaglia *et al*., (*29*), nuclei were extracted using cold lysis buffer (10 mM Tris-HCl pH 7.4, 10 mM NaCl, 0.5 mM spermidine, 1.5 mM MgCl2, 0.1 mM EDTA, 0.25% IGEPAL CA-630) for 3-6 minutes after checking nuclei under a light microscope. The sample was centrifuged for 5 minutes at 800x g at 4°C, washed with 1 ml ice-cold nuclei wash buffer (10 mM Tris-HCl pH 7.4, 50 mM NaCl, 0.5 mM spermidine, 1.5 mM MgCl2, 0.1 mM EDTA), and centrifuged again for 5 minutes at 800x g at 4°C. The nuclei were resuspended in 564 µL 1X GpC MTase reaction buffer (NEB, B0227) and equilibrated to 37°C. This was followed by 6 minutes of enzyme treatment (300 µL 1 M sucrose, 34 µL 10X GpC methyltransferase reaction buffer, 3 µL 32 mM S-Adenosylmethionine (SAM) (NEB, B9003), 100 µL GpC methyltransferase (M.CviPI) (NEB, M0227L)), followed by another 6 minutes of enzyme treatment (50 µL GpC methyltransferase (M.CviPI), 3 µL 32 mM S-Adenosylmethionine (SAM)). An equal volume of pre-warmed 2X stop buffer (20 mM Tris-HCl pH 7.4, 600 mM NaCl, 10 mM EDTA, 1% SDS) was added and mixed well. Proteinase K (200 µg/mL) was added to the GpC methylation reaction and incubated for 4 hours at 56°C, followed by RNase A (100 µg/mL) for 30 minutes at 37°C, followed by an additional round of proteinase K (200 µg/mL) for 30 minutes at 37°C.

HMW DNA was extracted using phenol-chloroform and precipitated by isopropanol. Briefly, an equal volume of phenol-chloroform was added to the samples and mixed vigorously. The samples were centrifuged at 13,000 rpm for 10 minutes at room temperature. The aqueous phase was transferred to a clean tube, and 0.7-1 times the volume of isopropanol was added, mixed well, and incubated at room temperature for 30 minutes. The samples were then centrifuged at 13,000 rpm for 10 minutes at 4°C. The DNA pellet was washed with 1 ml pre-chilled 70% ethanol. After removing all the ethanol, the DNA was dissolved in EB buffer and incubated at 37°C for 30 minutes. The final DNA concentration was quantified using Nanodrop 2000, and molecular weight was measured by genomic DNA screentape (Agilent, 5067-5365).

WGBS-NOMe-seq library preparation

For WGBS-NOMe-seq, the DNA was fragmented prior to bisulfite conversion. Fragmentation was performed via sonication (Diagenode Bioruptor) to produce 350 bp fragments. Bisulfite conversion was performed using the EZ DNA Methylation-Gold Kit (Zymo Research, D5005), and the library was prepared using the xGen TM Methyl-Seq DNA library prep kit (IDT, 10009860) following the manufacturer's protocols. The library was sequenced using 150 bp paired-end (PE) flow cells on a NovaSeqX to achieve a theoretical coverage of 20X per sample.

Nano-NOMe-seq library preparation

For Nano-NOMe-seq, the DNA fragment was sheared to an average size of 20kb with the Covaris G-Tube (#520079) in an Eppendorf 5415D tabletop microcentrifuge following the manufacture's protocol. Sample size was confirmed on genomic screentape with the Agilent Tapestation system. The sheared DNA (1.5ug) was then input into the SQK-LSK114 library prep from Oxford Nanopore to create the Nanopore library. Components were used from the LSK114 kit along with several buffers and enzymes from New England Biolab (NEB) following the LSK114 protocol from Oxford Nanopore. Following library prep, the sample concentration was verified on the Qubit system using the dsDNA High Sensitivity Kit. The concentration and basepair size were used to calculate Femtomolar concentration for loading. The sample was resuspended with library loading beads and loaded at 150-200 femtomolar concentration onto the Promethion R10.4.1 flow cell according to the LSK protocol. Sample were sequenced on the R10.4.1 Promethion flow cell using the Oxford Nanopore Promethion P24 system with a sequencing run time of 72 hrs. Standard settings were applied during the run setup, special settings were applied and fragments less than 1kb rejected and methylation detection was set for both 5mC and 5hmC along with the super accurate basecalling feature turned on.

Datasets and Data Processing

ChIP-seq, RNA-seq and ATAC-seq dataset

The CTCF ChIP-seq, RNA-seq and ATAC-seq data for WT and mutant mESCs were previously generated and processed as described in Do *et al*., (*9*) using the Seq-N-Slide pipeline (*30*).

For human samples, processed data for samples having CTCF ChIP-seq, RNA-seq and ATAC-seq performed on the same biosample were downloaded from ENCODE. The accession numbers of the dataset used in this work are listed in **Table S1A, S1B, S3A and B**. CTCF peaks were scanned for CTCF consensus motif (JASPAR MA0139.2) (*31*) using FIMO (*32*) with a p-value cutoff of $5e10^{-5}$. Only samples with at least 20% of motif containing CTCF peaks were kept for downstream analyses. Accessible CTCF peaks were identified by overlapping CTCF with ATAC peaks using bedtools (*33*).

For the GM12878 cell line, we downloaded an additional set of 145 ENCODE TF ChIP-seq narrow peak files to assess the percentage of CTCF binding sites with TF co-binding (the accession numbers are listed in **Table S3A**. To analyze allele-specific binding the BAM files for CTCF, RAD21, IKZF2, BHLHE40, ARNT and ATAC-seq were also downloaded (the accession numbers are listed in **Table S3B**). The phased genotyping data of GM12878 was retrieved from the Platinium Genomes project (*34*) (accession number PRJEB3246).

To validate the effect of TF co-binding, we downloaded KLF4, OCT4, SOX2, MYC, YY1 and input ChiP-seq data (fastq files) generated in mESCs from GEO (GSE201852 and GSE236376) (*35*). The data were processed using the Seq-N-Slide pipeline (*30*) as described in Do *et al*., (*9*). Briefly, the reads were aligned to mm10 genome with Bowtie2 (*36*) and narrow peaks were called using MACS2 (*37*) in pair-end mode and with the input as control.

The CTCF-oriented heatmaps and profiles were generated using the computeMatrix scale-regions function from deeptools (*38*) using a 15 bp bin and a region body length of 15 bp which corresponds to the length of CTCF core motif.

Hi-C data.

The HIC data were previously generated using Arima HiC kit (A510008) and processed using HiC-Pro and HiCExplorer as described in Do *et al*., (*9*). The HIC matrices were generated at a 10

kb resolution. The insulation scores were computed using GENOVA (*39*) and the CTCF-oriented profiles generated using deeptools (*38*) with a 1 kb bin and a region body length.

WGBS-NOMe-seq and Nano-NOMe-seq processing

WGBS-NOMe-seq fasta files were trimmed using trimGalore (*40*) using "--length 10 -q 30 --paired --three_prime_clip_R1 10 --clip_R2 10" to remove illumina adaptor, exclude low quality reads and the 10 bp on the 3' and 5' end to remove the fill-in nucleotides introduced by the library preparation steps. The data were then aligned against the mm10 reference genome using Bismark aligner (*41*). Duplicated reads were removed using Bismark deduplicate_bismark function and methylation in WCG (ACG and TCG) and HGC (GCA, GCC and GCT) context were called using bismark_methylation_extractor in NOMe-seq mode. Methylation in GCG context is ignored because of ambiguity and CCG to take into account off-target methylation from the GpC methyl-transferase. The first bp on the 3' end of a read was ignored using "--ignore_3prime_r2 1" as it shows a methylation bias. Methylation on positive and negative strands were aggregated since we did not observe substantial hemi-methylation genome-wide and at CTCF peaks. GpC and CpG overlapping the ENCODE blacklisted regions (*42*) were excluded as well as sites with more than 100x coverage.

For nano-NOMe-seq, base and base modification calling were performed on pod5 files using dorado basecaller (*43*) using the "--min-qscore 10 sup,5mC_5hmC" option, followed by alignment using dorado aligner against the mm10 reference genome. Reads with less than 10 MAPQ score were filtered out using samtools (*44*). Methylation in HGC and WCG context was called using dorado pileup using the "--combine-strands" option. Single molecule (SM) methylation was extracted using modkit extract with the "--read-calls-path" option. GpC and CpG overlapping the ENCODE blacklisted regions (*42*) were excluded as well as sites with more than 100x coverage.

CTCF-oriented methylation profiles were generated using deeptools and single molecule heatmaps generated using R. Single molecules with bound or unbound CTCF were defined as molecule containing a methylated or unmethylated GpC in the motif, respectively. No bias was observed in the nucleosome positioning of molecules with and without GpC containing motifs (**Figure S4A**). For KLF4, OCT4, SOX2, MYC and YY1, single molecules with bound and unbound TF were defined as molecule containing a methylated or unmethylated GpC in 10 bp centered by their motif, respectively since most of these TFs do not contain a GC in their consensus motif.

Downstream Analyses

Motif logos and Motif spacing enrichment analysis

Motif logos of the aggregated sequences centered on CTCF motif detected by FIMO (*32*) were created using the R package, ggseqlogo (*45*). Significantly enriched spacings of TF motifs in the vicinity of CTCF bound motifs were identified using SpaMo. The motif positional weight matrices (PWM) were downloaded from JASPAR 2024 (CORE vertebrates non-redundant dataset) (*31*) . The PWM for the U and D motifs were derived from Nakahashi *et al., (12)*. The analyses were performed using the 500 bp window centered on each bound CTCF motif stratified by chromatin accessibility. TF motif clustering and trees were performed using the R package, universalmotif (*46*) for each window defined as follow: upstream spacer, U window and further upstream, 0 to 5bp, 6-8 bp and more than 9bp upstream, respectively; downstream spacer, D window and further downstream, 0 to 9 bp, 10-12 bp and more than 13 bp downstream respectively.

Association between TF motif enrichment and TF expression

To ask whether TFs more enriched at accessible sites compared to inaccessible sites were more expressed in WT mESCs and human iPSCs, accessible enriched TFs were defined as TFs with a 2-fold increase in the percentage of motif found within 35 bp of the CTCF motif identified by the SpaMo analysis. A multivariate linear regression using a backward-stepwise selection was then performed to test the association between TF expression level (log2 TPM) and chromatin accessibility controlled for the epistatic effect of the motif orientation and window of enrichment (relative to the CTCF motif). The adjusted expression level by accessibility was estimated by computing the model margins and the p-value by performing the pairwise comparisons of predictive margins using STATA v17.

To test the overall correlation between TF expression and motif enrichment, we performed an aggregated analysis across WT and mutant mESCs or across ENCODE human samples (**Table S1A**). The expression level of TFs and percentage of motif were summed by sample, accessibility, window of enrichment and motif orientation. A multivariate linear regression using a backward-stepwise selection was then performed to test the correlation between the aggregated percentage of enriched motifs and TF expression level (log2 TPM) controlled for the epistatic effect of chromatin accessibility, the motif orientation and window of enrichment (relative to the CTCF motif). The differential effect of these covariates on the TF enrichment was estimated using the adjusted slope (interaction term between the covariate and TF expression) by computing the model margins and the pairwise comparisons of predictive margins using STATA v17.

Bivariate linear regressions between individual TF motif enrichment and expression level were performed using R.

Allele-specific binding analyses

Allele-specific binding (ASB) was performed by separating the ChIP-seq or ATAC-seq reads overlapping phased heterozygous SNPs in the GM12878 cell line using R. Allele-specific BAM and bigwig files were then generated using samtools (*44*) and bamCoverage (*38*). To identify ASB SNPs disrupting the TF motif, allele-specific motif affinity was computed using the R package, atSNP (*47*).

References

1. I. F. Davidson *et al.*, DNA loop extrusion by human cohesin. *Science* **366**, 1338-1345 (2019).
2. G. Fudenberg *et al.*, Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep* **15**, 2038-2049 (2016).
3. B. Dehingia, M. Milewska, M. Janowski, A. Pekowska, CTCF shapes chromatin structure and gene expression in health and disease. *EMBO Rep* **23**, e55146 (2022).
4. C. Guo *et al.*, CTCF-binding elements mediate control of V(D)J recombination. *Nature* **477**, 424-430 (2011).
5. D. Hnisz *et al.*, Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454-1458 (2016).
6. V. Narendra *et al.*, Transcription. CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science* **347**, 1017-1021 (2015).
7. Y. Xiang, X. Zhou, S. L. Hewitt, J. A. Skok, W. T. Garrard, A multifunctional element in the mouse Igkappa locus that specifies repertoire and Ig loci subnuclear location. *J Immunol* **186**, 5356-5366 (2011).
8. A. S. Hansen, CTCF as a boundary factor for cohesin-mediated loop extrusion: evidence for a multi-step mechanism. *Nucleus* **11**, 132-148 (2020).
9. C. Do *et al.*, Binding domain mutations provide insight into CTCF's relationship with chromatin and its ability to act as a chromatin organizer. *Biorxiv.*, (2024).
10. M. Iurlaro *et al.*, Systematic assessment of ISWI subunits shows that NURF creates local accessibility for CTCF. *Nat Genet* **56**, 1203-1212 (2024).

11. K. Essien *et al.*, CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. *Genome Biol* **10**, R131 (2009).
12. H. Nakahashi *et al.*, A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep* **3**, 1678-1689 (2013).
13. R. N. Plasschaert *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).
14. W. Soochit *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).
15. G. Hu *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).
16. I. Jerkovic *et al.*, Genome-Wide Binding of Posterior HOXA/D Transcription Factors Reveals Subgrouping and Association with CTCF. *PLoS Genet* **13**, e1006567 (2017).
17. Y. Liu *et al.*, CTCF coordinates cell fate specification via orchestrating regulatory hubs with pioneer transcription factors. *Cell Rep* **42**, 113259 (2023).
18. C. S. Ross-Innes, G. D. Brown, J. S. Carroll, A co-ordinated interaction between CTCF and ER in breast cancer cells. *BMC Genomics* **12**, 593 (2011).
19. C. Sonmezer *et al.*, Molecular Co-occupancy Identifies Transcription Factor Binding Cooperativity In Vivo. *Mol Cell* **81**, 255-267 e256 (2021).
20. Y. Fu, M. Sinha, C. L. Peterson, Z. Weng, The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet* **4**, e1000138 (2008).
21. T. Whitington, M. C. Frith, J. Johnson, T. L. Bailey, Inferring transcription factor complexes from ChIP-seq data. *Nucleic Acids Res* **39**, e98 (2011).
22. C. Do, J. A. Skok, Factors that determine cell type-specific CTCF binding in health and disease. *Curr Opin Genet Dev* **88**, 102244 (2024).
23. R. T. Pop *et al.*, Identification of mammalian transcription factors that bind to inaccessible chromatin. *Nucleic Acids Res* **51**, 8480-8495 (2023).
24. W. J. de Jonge, H. P. Patel, J. V. W. Meeussen, T. L. Lenstra, Following the tracks: How transcription factor binding dynamics control transcription. *Biophys J* **121**, 1583-1592 (2022).

25. M. Nishana *et al.*, Defining the relative and combined contribution of CTCF and CTCFL to genomic regulation. *Genome Biol* **21**, 108 (2020).
26. E. P. Nora *et al.*, Molecular basis of CTCF binding polarity in genome folding. *Nat Commun* **11**, 5612 (2020).
27. E. M. Pugacheva *et al.*, CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention. *Proc Natl Acad Sci U S A* **117**, 2020-2031 (2020).
28. Y. Li *et al.*, The structural basis for cohesin-CTCF-anchored loops. *Nature* **578**, 472-476 (2020).
29. S. Battaglia *et al.*, Long-range phasing of dynamic, tissue-specific and allele-specific regulatory elements. *Nat Genet* **54**, 1504-1513 (2022).
30. I. Dolgalev, Seq-N-Slide [Computer software]. (2022).
31. I. Rauluseviciute *et al.*, JASPAR 2024: 20th anniversary of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **52**, D174-D182 (2024).
32. T. L. Bailey *et al.*, MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-208 (2009).
33. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842 (2010).
34. M. A. Eberle *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).
35. L. Laugwitz *et al.*, ZSCAN10 deficiency causes a neurodevelopmental disorder with characteristic oto-facial malformations. *Brain* **147**, 2471-2482 (2024).
36. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359 (2012).
37. Y. Zhang *et al.*, Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
38. F. Ramirez, F. Dundar, S. Diehl, B. A. Gruning, T. Manke, deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**, W187-191 (2014).
39. R. H. van der Weide *et al.*, Hi-C analyses with GENOVA: a case study with cohesin variants. *NAR Genom Bioinform* **3**, lqab040 (2021).
40. F. Krueger, TrimGalore. (2021).
41. F. Krueger, S. R. Andrews, Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571-1572 (2011).
42. H. M. Amemiya, A. Kundaje, A. P. Boyle, The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep* **9**, 9354 (2019).

43. nanoporetech, Dorado. https://github.com/nanoporetech/dorado, (2024).
44. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
45. O. Wagih, ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* **33**, 3645-3647 (2017).
46. B. J. Tremblay, universalmotif: An R package for biological motif analysis. *The Journal of Open Source Software* **9(100):7012**, (2024).
47. C. Zuo, S. Shin, S. Keles, atSNP: transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics* **31**, 3353-3355 (2015).

Acknowledgements

This work was supported by 1R35GM122515 (J.S) and NIH P01CA229086 (J.S). GC were supported by a fellowship from the NCC.

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

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. 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.

Supplemental Tables

Table S1A. Accession numbers of the CTCF ChIP-seq, DNAse 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.

Table S2. 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 and mutant mESCs. **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.

Table S3. A. ENCODE accession numbers of the TF ChIP-seq narrow peak files in GM12878 to assess the overall percentage of cobound CTCF and TF sites. **B**. ENCODE accession numbers of the TF ChIP-seq bam files in GM12878 used for the allele-specific binding analyses.

Supplemental Figures

Figure S1

A

U motif at accessible sites				
Samples	**Number of sequences**	**Total number of sequences**	**Percentage (%)**	**A bias (% sequences with A at position 9)**
WT	450	4583	9.82	69.33
K365T	389	3292	11.82	71.21
H455R	52	500	10.4	51.92
Q418R	335	2773	12.08	68.96
R448Q	741	5898	12.56	73.68
R377C	428	4653	9.2	69.39
R377H	379	3508	10.8	67.02
S354F	281	3678	7.64	66.19
R342C	383	3614	10.6	69.71
R339Q	287	2314	12.4	65.16
U motif at inaccessible sites				
WT	1364	5953	22.91	81.45
K365T	1005	3757	26.75	82.49
H455R	53	397	13.35	64.15
Q418R	1870	6071	30.8	84.6
R448Q	1610	5194	31	85.03
R377C	2620	11950	21.92	84.62
R377H	1883	6869	27.41	83.7
S354F	2319	11891	19.5	83.61
R342C	1859	6853	27.13	82.14
R339Q	2021	6782	29.8	82.58
D motif at accessible sites				
WT	122	4583	2.66	87.7
K365T	102	3292	3.1	90.2
H455R	7	500	1.4	85.71
Q418R	97	2773	3.5	88.66
R448Q	233	5898	3.95	87.12
R377C	125	4653	2.69	89.6
R377H	103	3508	2.94	87.38
S354F	86	3678	2.34	90.7
R342C	94	3614	2.6	88.3
R339Q	63	2314	2.72	85.71
D motif at inaccessible sites				
WT	658	5953	11.05	88.3
K365T	456	3757	12.14	89.47
H455R	17	397	4.28	70.59
Q418R	969	6071	15.96	89.16
R448Q	886	5194	17.06	89.39
R377C	1526	11950	12.77	88.4
R377H	937	6869	13.64	88.9
S354F	1336	11891	11.24	89.3
R342C	904	6853	13.19	86.95
R339Q	957	6782	14.11	87.77

Figure S1. A: Proportion and number of sequences with a U or D motif within 35 bp of the CTCF consensus motif at accessible and inaccessible CTCF bound sites in WT and mutant mESCs.

Figure S1

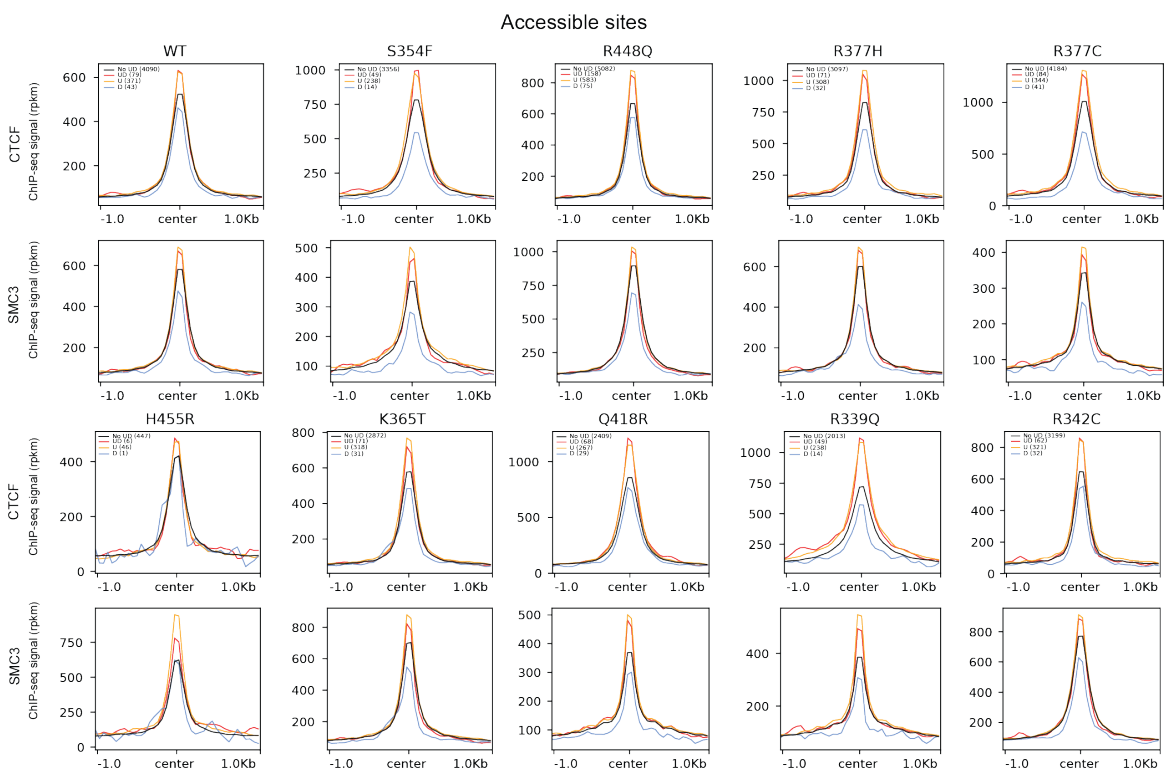
B

U motif at accessible sites							
Sample	Number of sequences	Total number of sequences	Percentage	% sequences with GT at position 9-10)	% sequences with GG at position 9-10)	% sequences with AG at position 9-10)	% sequences with AT at position 9-10)
Placenta Embryonic	240	8436	2.84	32.5	24.58	22.92	17.08
Lower Lobe Of Left Lung Adult	254	10723	2.37	30.31	27.56	23.23	15.35
Right Lobe Of Liver Adult	287	12799	2.24	29.97	23.69	26.48	16.38
Colonic Mucosa Adult	198	7944	2.49	31.82	26.77	22.22	15.15
Astrocyte Adult	244	13115	1.86	31.15	20.49	24.59	20.49
Testis Adult	27	984	2.74	22.22	33.33	25.93	18.52
Ascending Aorta Adult	260	9857	2.64	31.54	25	23.46	16.15
Pancreas Adult	233	8803	2.65	31.33	25.32	21.46	17.6
Stomach Adult	231	8698	2.66	30.3	24.24	25.54	16.02
Spleen Adult	236	8764	2.69	32.2	27.12	21.61	15.68
Gastrocnemius Medialis Adult	219	6857	3.19	33.33	22.83	26.03	14.61
Activated B Cell Adult	176	4396	4	30.68	25	23.86	16.48
Omental Fat Pad Adult	214	5756	3.72	31.78	23.83	24.3	16.82
Mcf 10A Adult	321	17462	1.84	28.97	23.05	26.48	18.07
Heart Left Ventricle Adult	276	12039	2.29	30.07	23.19	27.54	16.67
Left Lung Adult	226	8279	2.73	30.53	29.2	19.47	16.37
Transverse Colon Adult	250	10763	2.32	30	25.2	26	15.2
Calu3 Adult	307	14446	2.13	28.99	24.76	25.08	17.59
Cardiac Muscle Cell Embryonic	310	17202	1.8	28.71	22.9	26.77	18.39
Esophagus Muscularis Mucosa Adult	178	6117	2.91	33.15	24.16	20.22	17.98
Upper Lobe Of Right Lung Adult	234	9971	2.35	31.62	28.63	21.37	14.96
Lower Lobe Of Right Lung Adult	243	10524	2.31	30.86	28.4	21.4	16.05
Ovary Adult	242	8984	2.69	31.82	23.97	24.38	16.53
Thyroid Gland Adult	299	14586	2.05	29.43	24.75	25.75	16.72
Uterus Adult	254	10781	2.36	30.71	23.23	25.59	16.93
Body Of Pancreas Adult	297	14210	2.09	29.29	23.91	27.27	16.16
Myotube Adult	192	11064	1.74	30.21	22.4	21.88	22.4
Hepatocyte Embryonic	311	17821	1.75	29.26	22.83	26.05	18.65
Thoracic Aorta Adult	182	6651	2.74	34.62	23.63	23.08	16.48
Right Lobe Of Liver Child	232	10258	2.26	29.31	28.88	20.69	17.24
Peyer'S Patch Adult	256	10420	2.46	32.42	23.44	24.61	16.41
Caco-2 Adult	277	14036	1.97	30.32	23.47	25.63	16.97
Adrenal Gland Adult	297	13373	2.22	29.97	23.23	25.93	17.17
A673 Child	323	19182	1.68	28.79	23.53	26.63	17.96
Natural Killer Cell Adult	265	9422	2.81	30.57	26.04	24.91	15.09
U motif at inaccessible sites							
Placenta Embryonic	34	1911	1.78	26.47	5.88	35.29	29.41
Lower Lobe Of Left Lung Adult	78	6343	1.23	26.92	11.54	33.33	23.08
Right Lobe Of Liver Adult	41	4387	0.93	21.95	17.07	29.27	26.83
Colonic Mucosa Adult	129	7570	1.7	24.03	19.38	31.78	22.48
Astrocyte Adult	82	4024	2.04	25.61	30.49	28.05	10.98
Testis Adult	298	14371	2.07	31.21	22.15	25.17	17.45
Ascending Aorta Adult	57	4300	1.33	22.81	15.79	33.33	26.32
Pancreas Adult	93	6413	1.45	25.81	15.05	38.71	18.28
Stomach Adult	95	6864	1.38	28.42	18.95	28.42	21.05
Spleen Adult	69	4263	1.62	23.19	7.25	42.03	23.19
Gastrocnemius Medialis Adult	74	5061	1.46	20.27	18.92	27.03	28.38
Activated B Cell Adult	28	1577	1.78	25	10.71	35.71	28.57
Omental Fat Pad Adult	31	1563	1.98	32.26	19.35	25.81	22.58
Mcf 10A Adult	14	1040	1.35	42.86	28.57	14.29	0
Heart Left Ventricle Adult	41	4375	0.94	24.39	17.07	19.51	29.27
Left Lung Adult	105	6784	1.55	27.62	13.33	38.1	19.05
Transverse Colon Adult	84	6517	1.29	27.38	19.05	26.19	23.81
Calu3 Adult	13	1005	1.29	30.77	38.46	7.69	15.38
Cardiac Muscle Cell Embryonic	21	1151	1.82	42.86	28.57	19.05	0
Esophagus Muscularis Mucosa Adult	126	7342	1.72	26.98	21.43	30.95	18.25
Upper Lobe Of Right Lung Adult	104	8577	1.21	23.08	13.46	36.54	23.08
Lower Lobe Of Right Lung Adult	90	7081	1.27	21.11	12.22	38.89	22.22
Ovary Adult	11	1149	0.96	18.18	18.18	18.18	45.45
Thyroid Gland Adult	34	3244	1.05	20.59	17.65	29.41	23.53
Uterus Adult	32	2909	1.1	25	21.88	21.88	31.25
Body Of Pancreas Adult	27	3006	0.9	25.93	11.11	18.52	37.04
Myotube Adult	137	7438	1.84	29.93	23.36	31.39	10.22
Hepatocyte Embryonic	19	1765	1.08	36.84	31.58	21.05	0
Thoracic Aorta Adult	129	8258	1.56	24.03	18.6	31.78	20.93
Right Lobe Of Liver Child	102	9563	1.07	25.49	12.75	38.24	19.61
Peyer'S Patch Adult	75	7767	0.97	20	21.33	30.67	22.67
Caco-2 Adult	55	5548	0.99	20	23.64	30.91	20
Adrenal Gland Adult	28	3344	0.84	28.57	10.71	32.14	25
A673 Child	12	1395	0.86	25	33.33	16.67	0
Natural Killer Cell Adult	52	4166	1.25	26.92	15.38	28.85	26.92

Figure S1. B: Proportion and number of sequences with a U or D motif within 35 bp of the CTCF consensus motif at accessible and inaccessible CTCF bound sites in ENCODE human samples with at least 20% CTCF motif-containing peaks .

Figure S1

C


Accessible sites
WT
S354F
R448Q
R377H
R377C
H455R
K365T
Q418R
R339Q
R342C
CTCF
ChIP-seq signal (rpkm)
SMC3
ChIP-seq signal (rpkm)
-1.0
center
1.0Kb

D

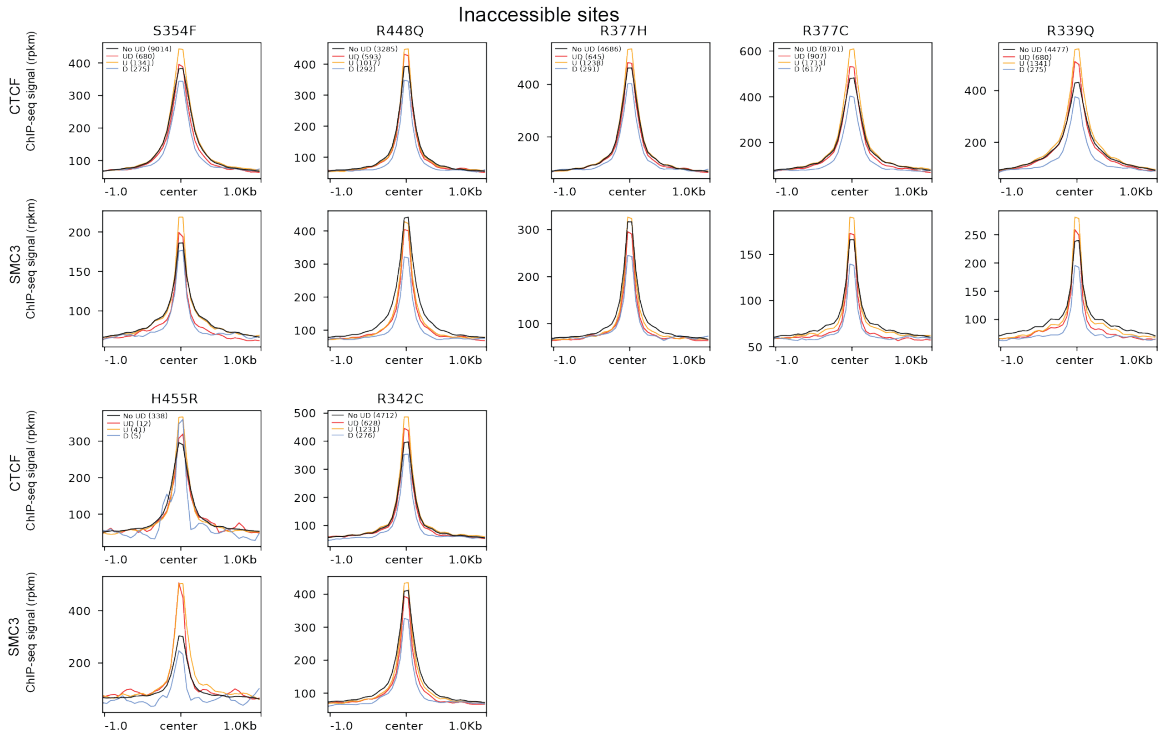

Inaccessible sites
S354F
R448Q
R377H
R377C
R339Q
H455R
R342C
CTCF
ChIP-seq signal (rpkm)
SMC3
ChIP-seq signal (rpkm)
-1.0
center
1.0Kb

Figure S1. C: CTCF and SMC3 binding profiles from ChIP-seq at inaccessible sites in CTCF mutant mESCs at CTCF bound sites containing the CTCF motif alone (no UD), CTCF and the U motif only (U), CTCF and the D motif only (D), or CTCF and both the U and D motifs (UD). The remaining mutants and WT mESC binding profiles at inaccessible sites are shown in **Figure 1D. D:** CTCF and SMC3 binding profiles from ChIP-seq at accessible sites in WT and CTCF mutant mESCs at CTCF bound sites containing the CTCF motif alone (no UD), CTCF and the U motif only (U), CTCF and the D motif only (D), or CTCF and both the U and D motifs (UD).

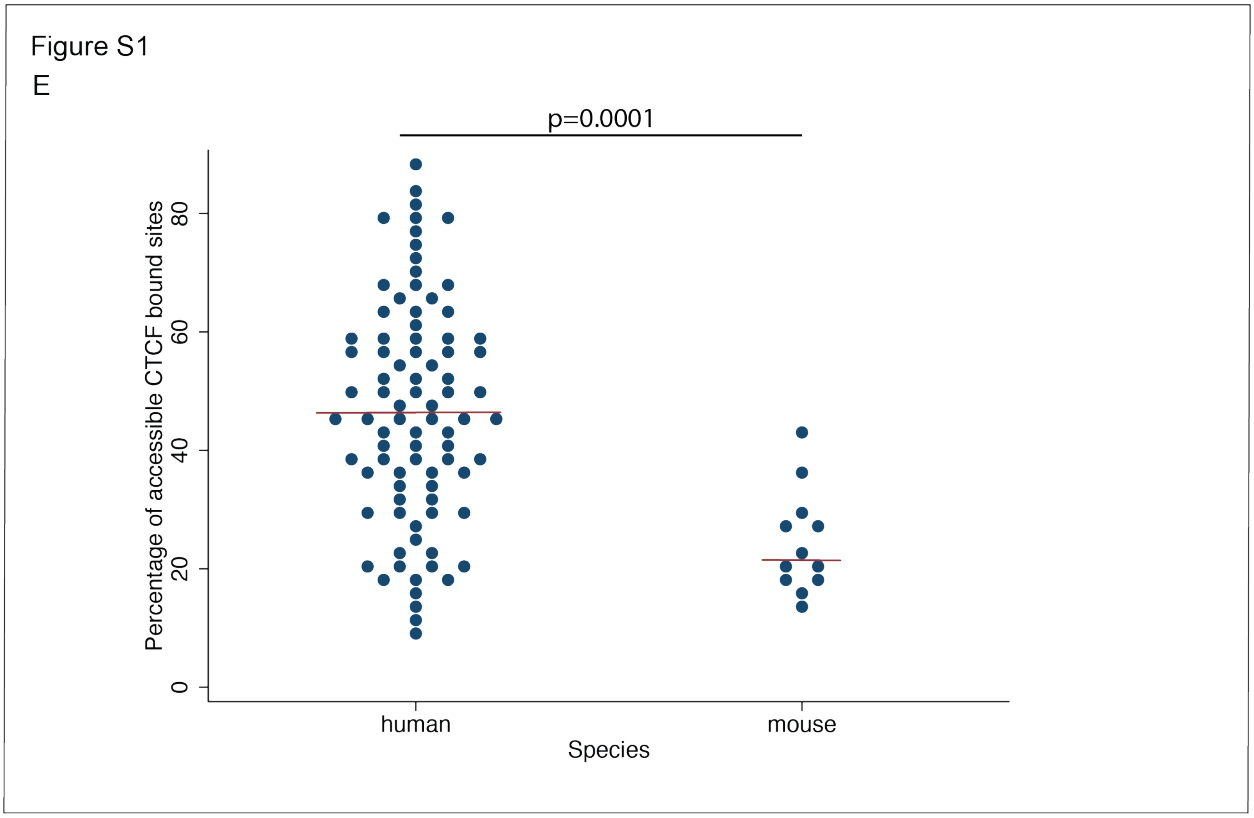


Figure S1. E: Dotplot showing the higher percentage of accessible CTCF bound sites in human versus mouse. The red line represents the median. The p-value has been computed using a t-test. The data related to this table are shown in **Table S1B**.

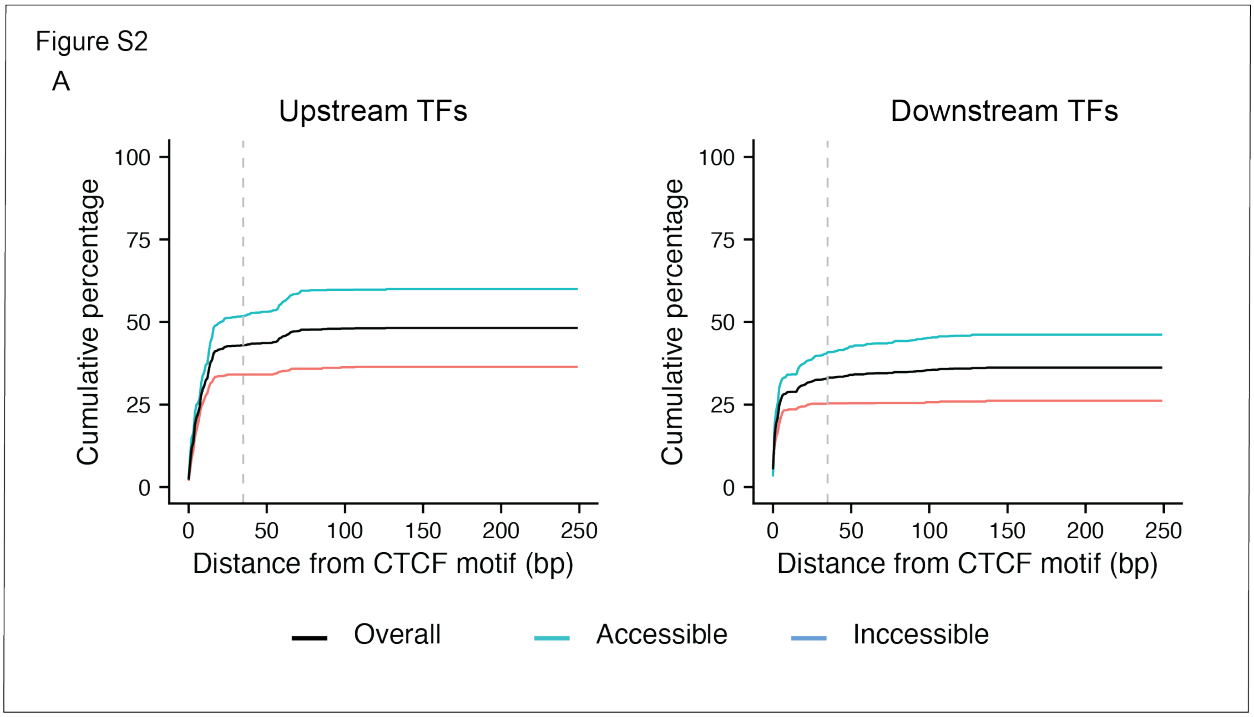


Figure S2. A: Cumulative percentage of CTCF peaks containing a TF motif within 250 bp at accessible and inaccessible sites. The dashed lines correspond to 35 bp upstream or downstream of the CTCF motif.

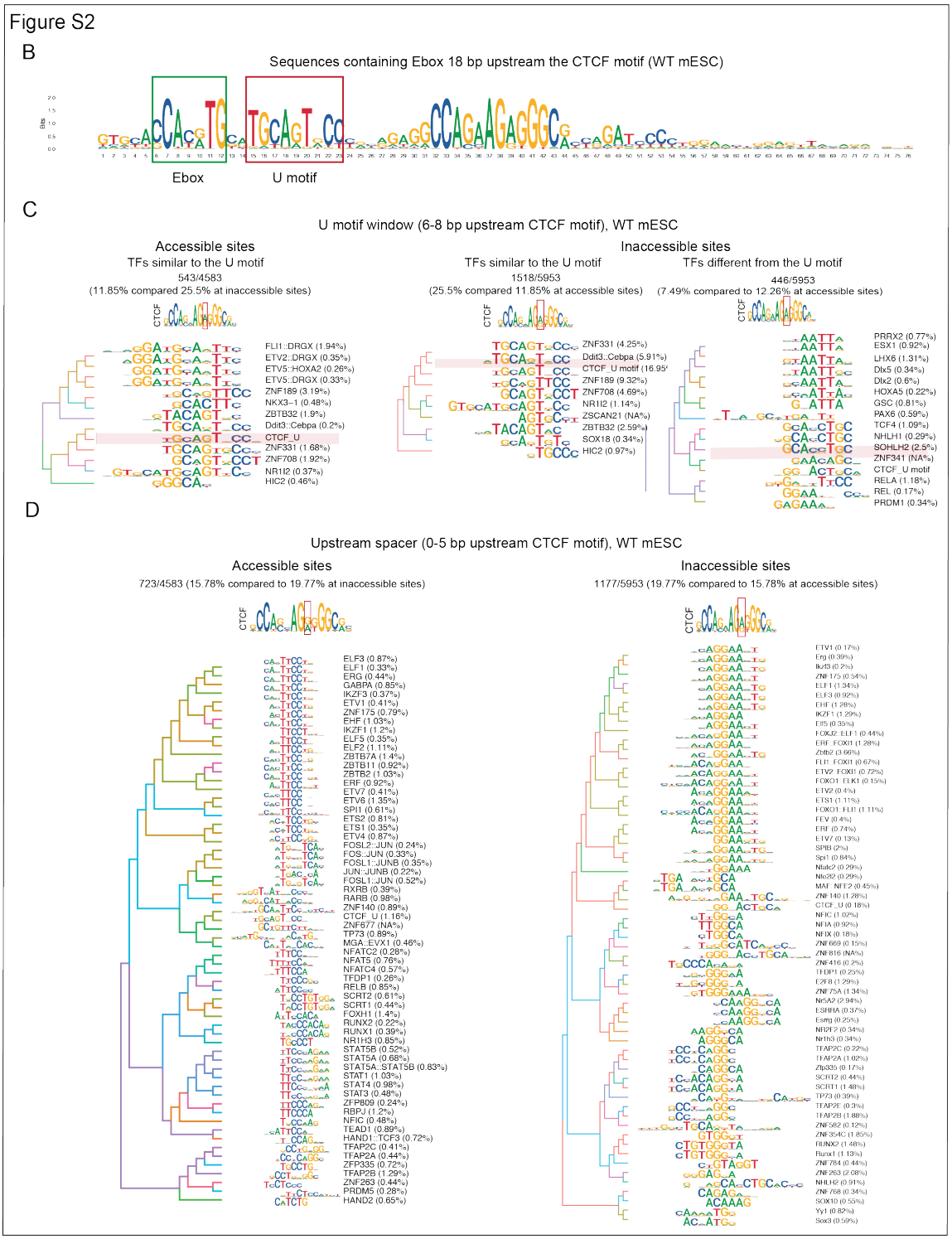


Figure S2. B: DNA logos of the CTCF bound sequences containing an E-box motif 18 bp upstream of the CTCF consensus b showing that the presence of an E-box is highly associated with the presence of the U motif. **C**: Motif logos of the enriched TF motifs found in the U motif window at accessible and inaccessible sites in WT mESCs. The TF motifs different from the U motif at accessible sites are shown in **Figure 2B**. **D**: 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 motifs.

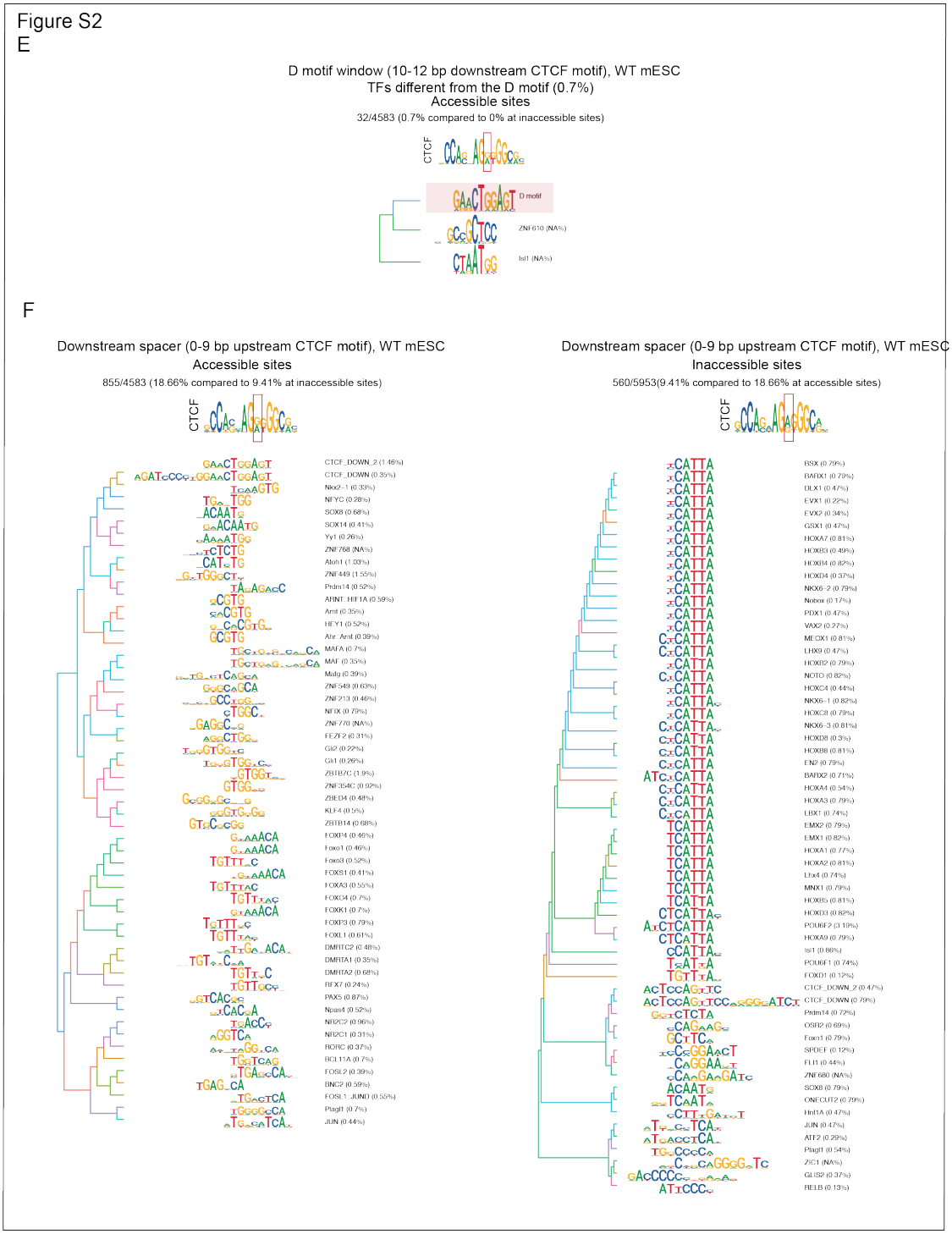


Figure S2. E: Motif logos of the enriched TF motifs found in the D window at accessible sites in WT mESCs. No motifs were significantly enriched at inaccessible sites. **F**: Motif logos of the enriched TF motifs found in the downstream spacer window at accessible and inaccessible sites in WT mESCs.

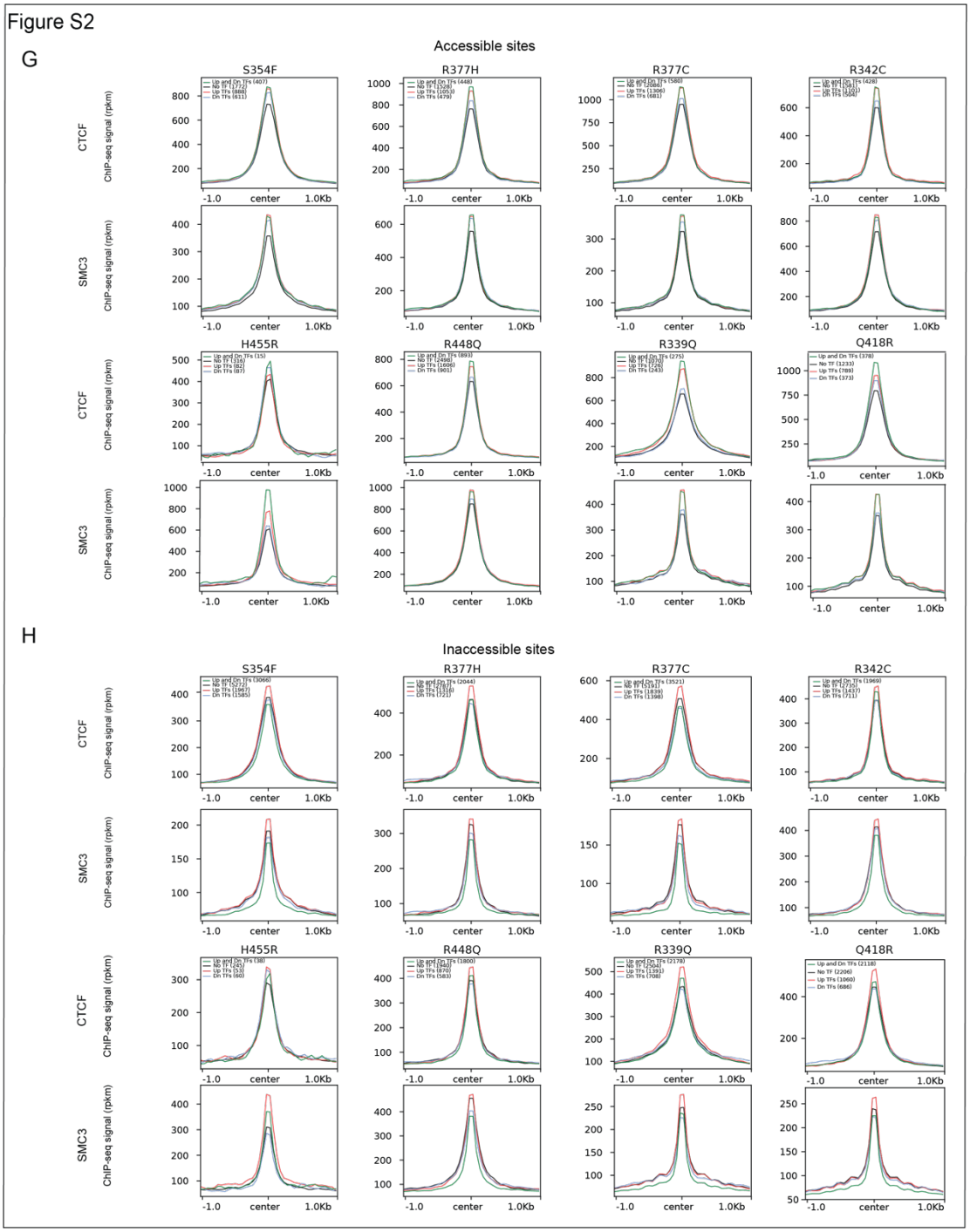


Figure S2. G: CTCF and SMC3 binding profiles from ChIP-seq at accessible sites in CTCF mutants mESCs at CBSs containing at least one enriched TF motif upstream, downstream or both upstream and downstream compared to CBSs without any TF motifs in the surrounding 35 bp region. The remaining mutant and WT mESC binding profiles at accessible sites are shown in **Figure 2C. H:** CTCF and SMC3 binding profiles from ChIP-seq at inaccessible sites in WT mESC and CTCF mutants at CBSs containing at least one enriched TF motif upstream, downstream or both upstream and downstream, compared to CTCF binding sites without any TF motifs within 35 bp.

Figure S3

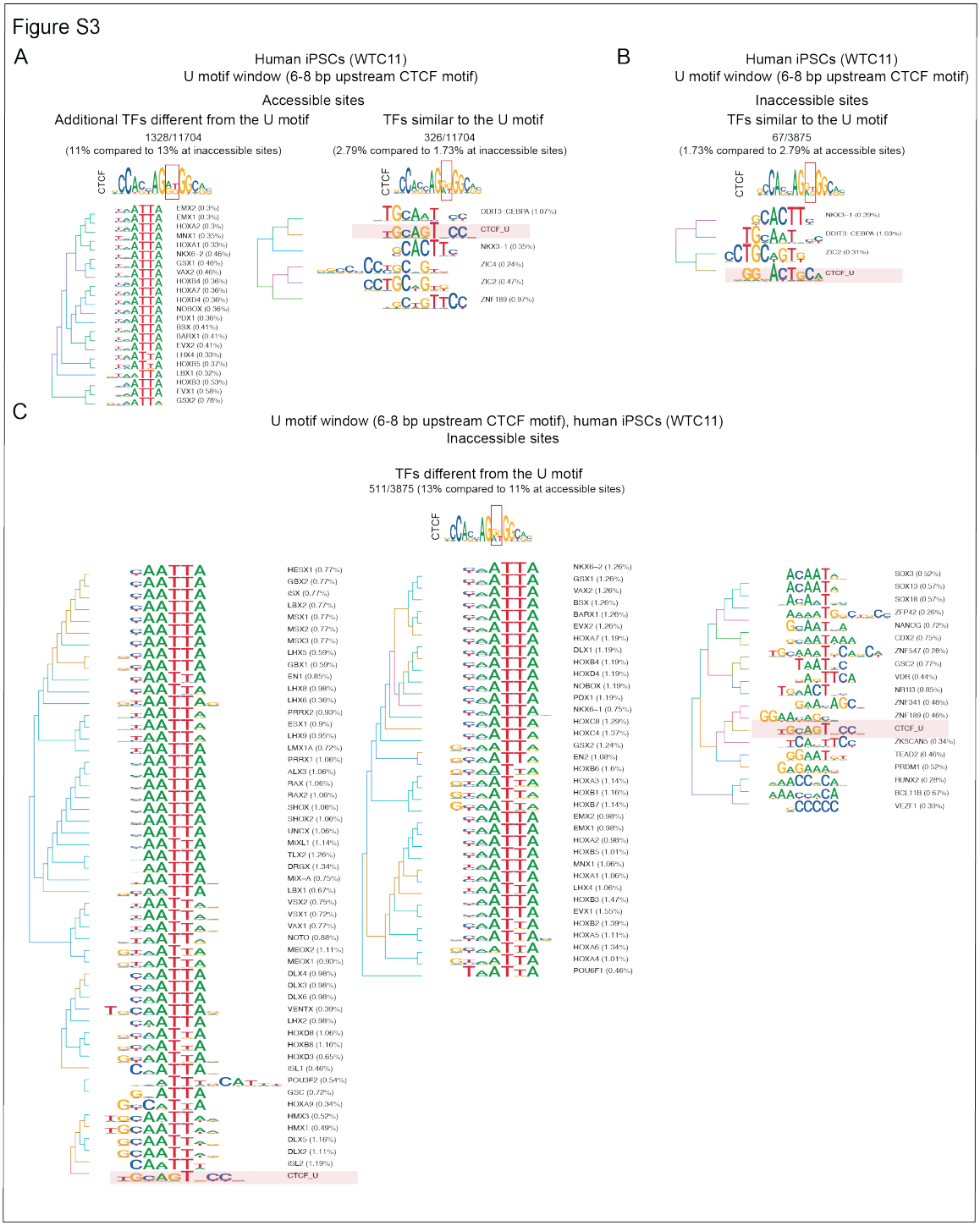

A
Human iPSCs (WTC11)
U motif window (6-8 bp upstream CTCF motif)
Accessible sites
Additional TFs different from the U motif
1328/11704
(11% compared to 13% at inaccessible sites)
TFs similar to the U motif
326/11704
(2.79% compared to 1.73% at inaccessible sites)
B
Human iPSCs (WTC11)
U motif window (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)
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)
CTCF

Figure S3. A: Motif logos of the enriched TF motifs found in the U motif window at accessible sites in human iPSCs (WCT11)**.** The remaining motifs different from the U motif are shown in **Figure 2B. 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)**.** The motifs different from the U motif are shown in **Figure 2B. 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)**.**

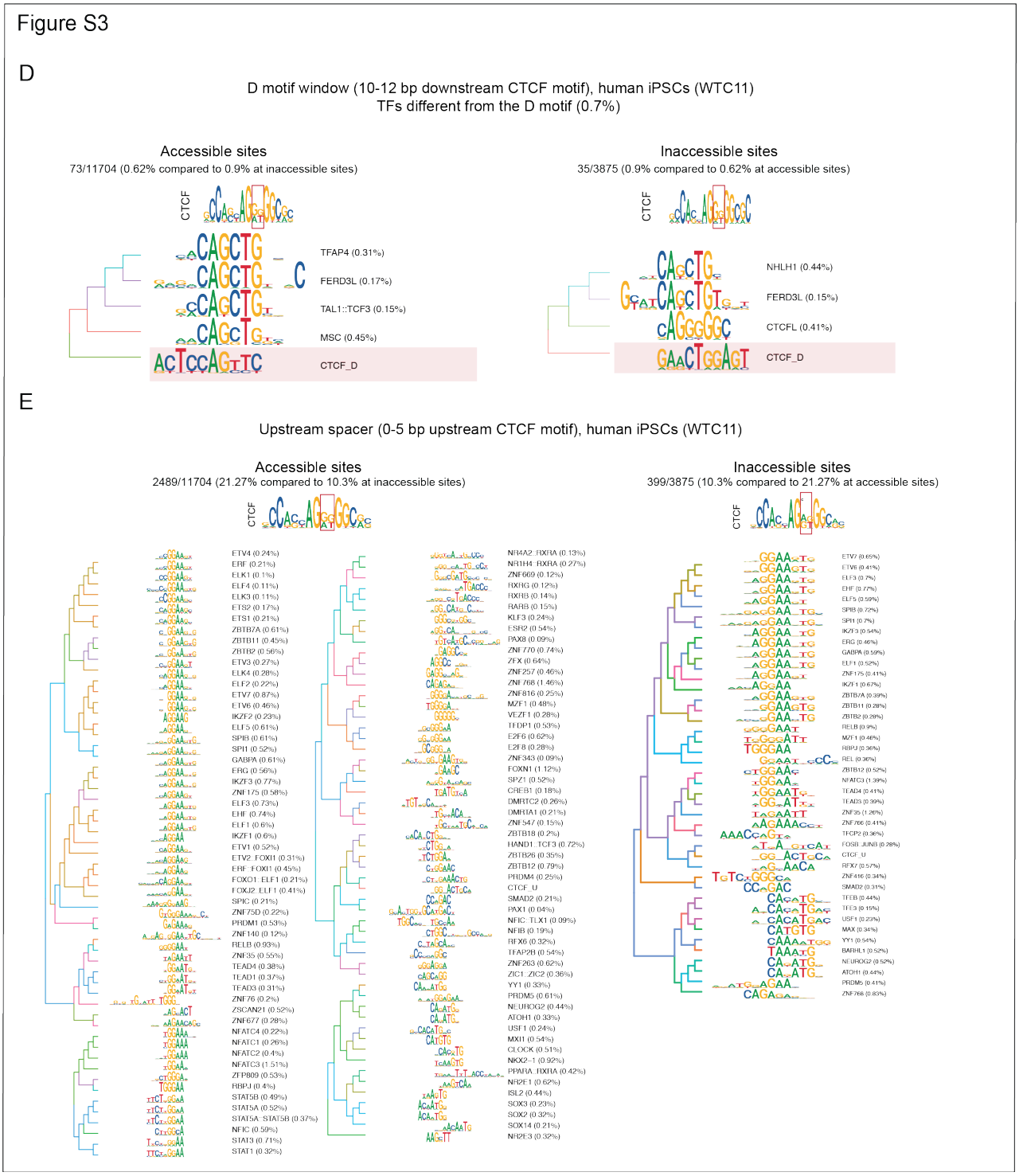
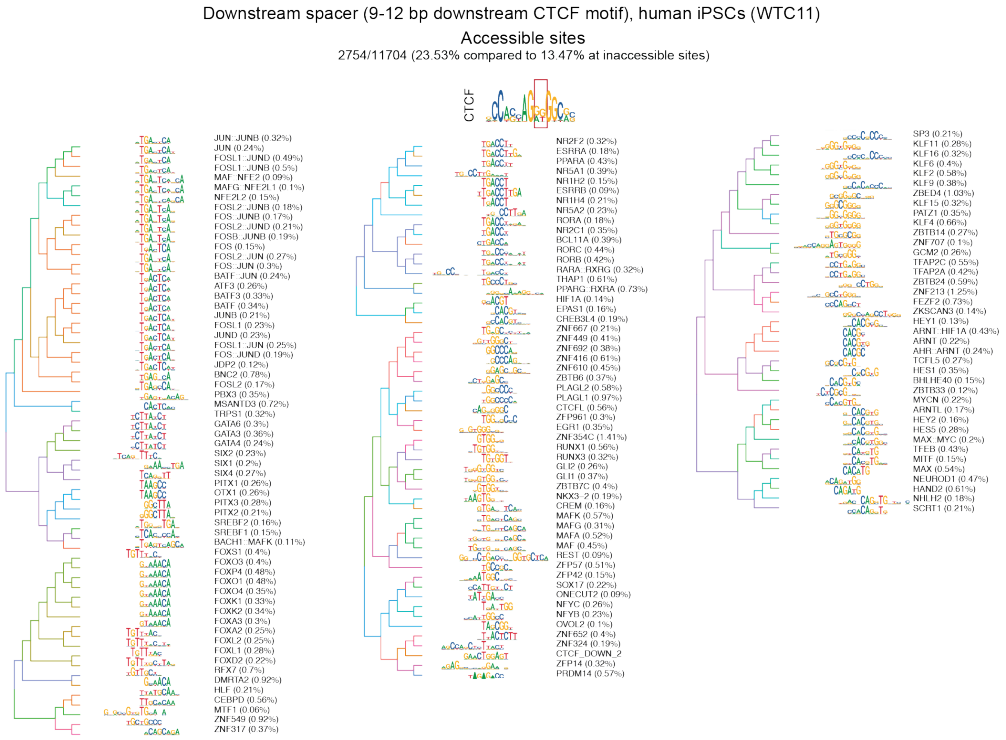


Figure S3. D: Motif logos of the enriched TF motifs found in the D motif window at accessible and inaccessible sites in human iPSCs (WCT11). **D**: Motif logos of the enriched TF motifs found in the upstream spacer at accessible and inaccessible sites in human iPSCs (WCT11)**.** This window is enriched in ETS motifs.

Figure S3

F



G

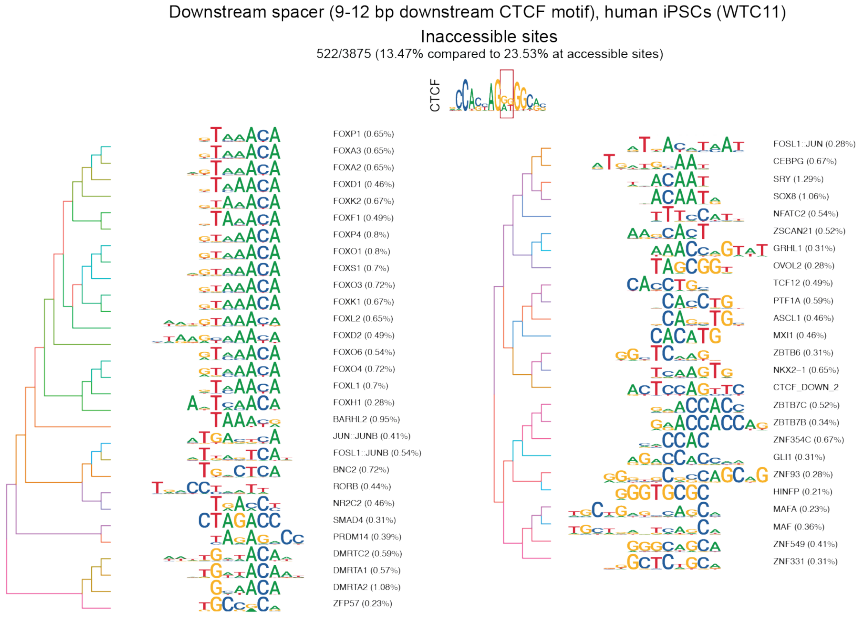


Figure S3. F: Motif logos of the enriched TF motifs found in the downstream spacer at accessible and inaccessible (**G**) sites in human iPSCs (WCT11)**.**

Figure S4

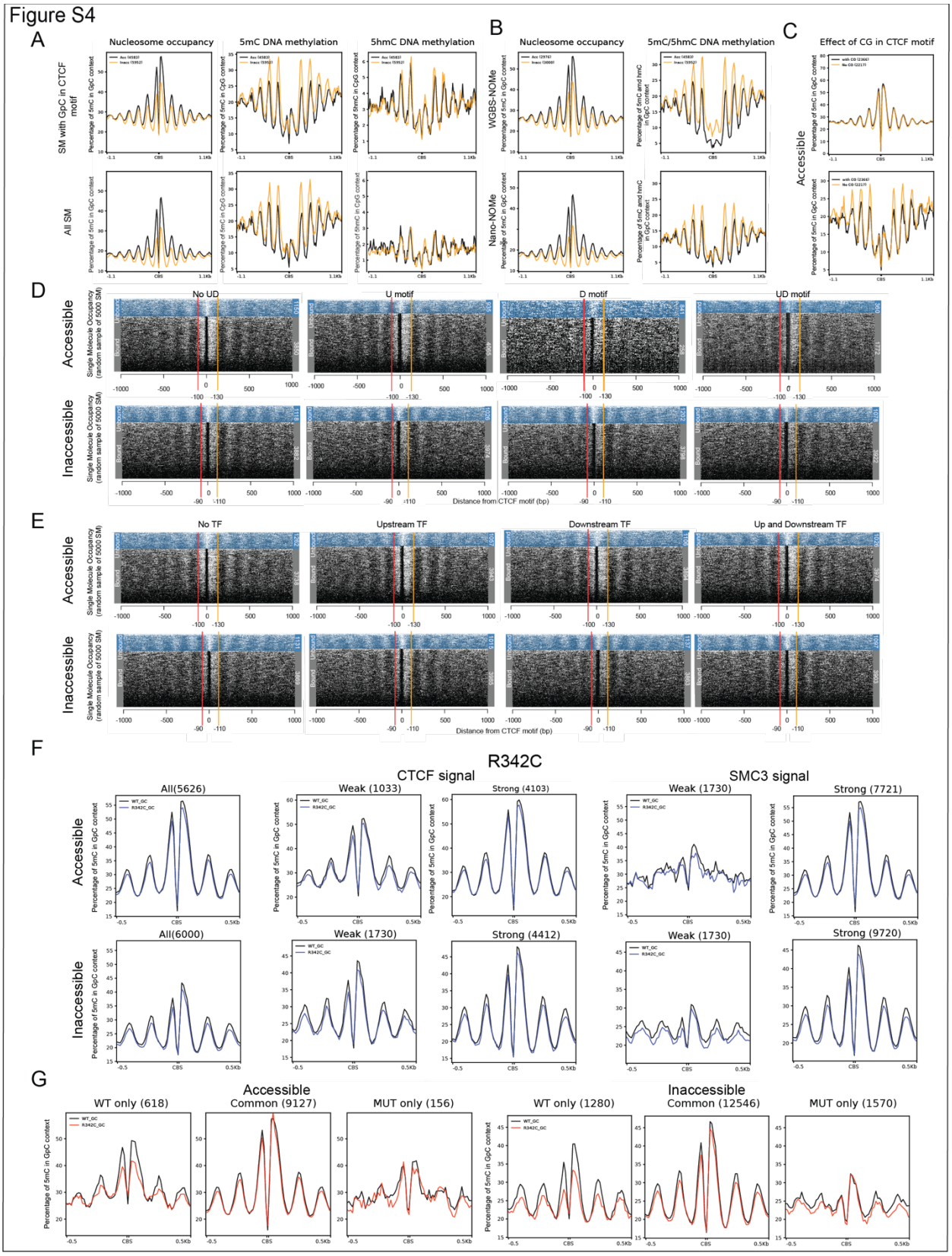

A
Nucleosome occupancy
5mC DNA methylation
5hmC DNA methylation
SM with GpC in CTCF motif
All SM
B
Nucleosome occupancy
5mC/5hmC DNA methylation
WGBS-NOMe
Nano-NOMe
C
Effect of CG in CTCF motif
Accessible
D
No UD
U motif
D motif
UD motif
Accessible
Inaccessible
Single Molecule Occupancy (random sample of 5000 SM)
Distance from CTCF motif (bp)
E
No TF
Upstream TF
Downstream TF
Up and Downstream TF
Accessible
Inaccessible
Distance from CTCF motif (bp)
F
R342C
CTCF signal
SMC3 signal
All(5626)
Weak (1033)
Strong (4103)
Weak (1730)
Strong (7721)
Accessible
All(6000)
Weak (1730)
Strong (4412)
Weak (1730)
Strong (9720)
Inaccessible
G
Accessible
WT only (618)
Common (9127)
MUT only (156)
Inaccessible
WT only (1280)
Common (12546)
MUT only (1570)
Percentage of 5mC in GpC context

Figure S4: CTCF binding displaces flanking nucleosome with an N-terminal bias in WT and R342C mutant mESCs. A, B and **C**: profiles showing similar nucleosome occupancy phasing and DNA methylation for GpC-containing CBSs compared to non GpC-containing CBSs (**A**), nano-NOMe-seq compared to WGBS-NOMe-seq (**B**) and CpG-containing CBSs compared to non CpG-containing CBSs from NOMe-seq data (**C**), although the baseline level of methylation varies. **D**: Heatmaps of 5000 randomly sampled bound and unbound molecules at accessible and inaccessible CTCF peaks in the absence or presence of the U and/or D motifs in WT mESCs. For the less prevalent accessible sites with D motifs, all the molecules (1099) are shown. The profiles are shown in **Figure 3E**. **E**, Heatmaps of 5000 randomly sampled bound and unbound molecules at accessible and inaccessible CTCF peaks in the absence or presence of the upstream and/or downstream TF motifs in WT mESCs. The profiles are shown in **Figure 3F. F**: Profiles from left to right show the nucleosome occupancy (5mC in HGC context) of bound molecules from WGBS-NOMe-seq in R342C mutant mESCs at all accessible and inaccessible CTCF peaks, weak or strong CTCF peaks and CTCF peaks without or with a strong SMC3 overlap. **G**: Profiles showing the nucleosome occupancy (5mC in HGC context) of bound molecules from WGBS-NOMe-seq in R342C mutant mESCs at accessible and inaccessible WT only, common and mutant only CTCF peaks. All the profiles in this figure are oriented on the C to N terminal CTCF core motif.

Figure S5

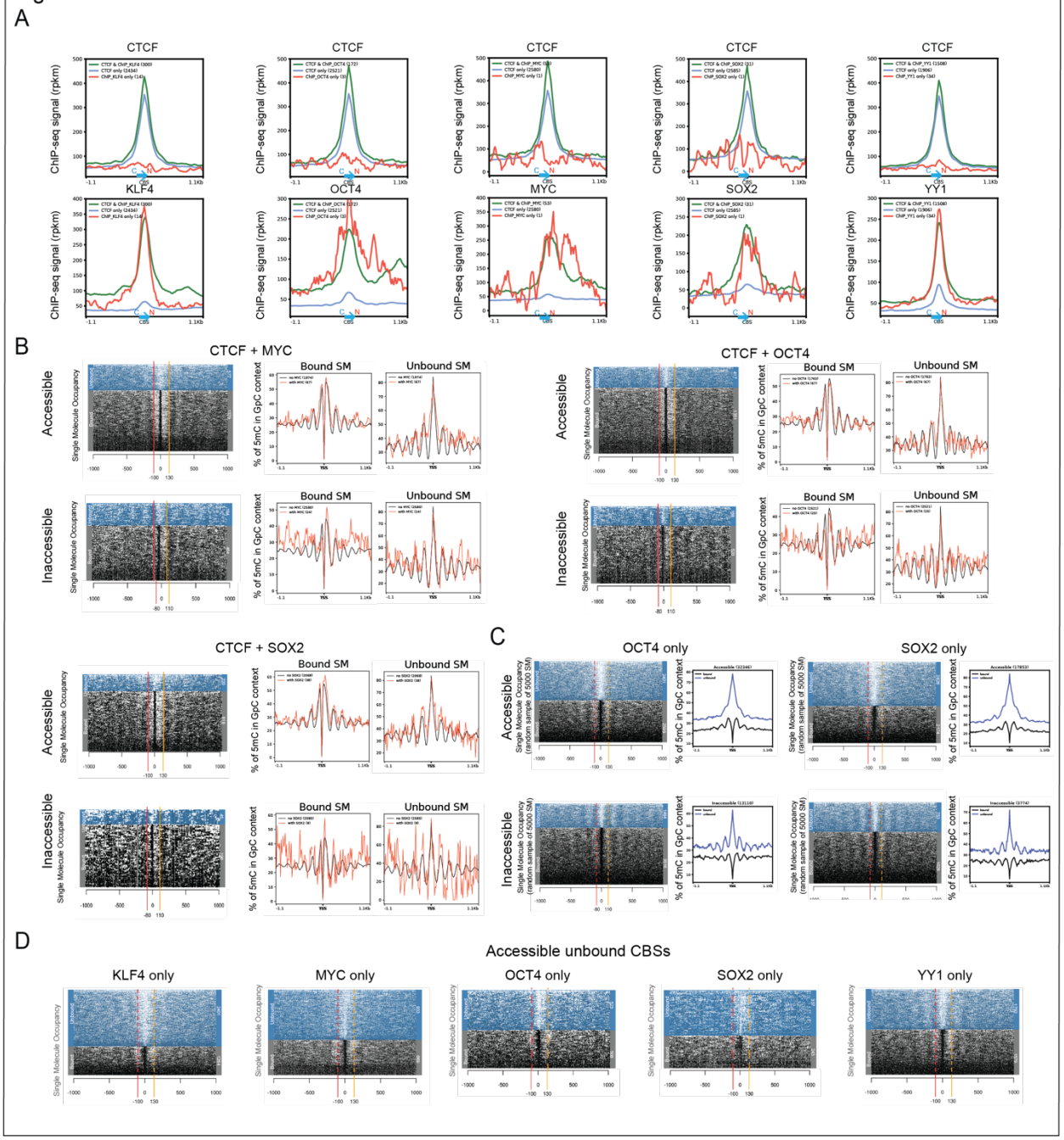

A
CTCF
KLF4
OCT4
MYC
SOX2
YY1
ChIP-seq signal (rpkm)
B
CTCF + MYC
CTCF + OCT4
CTCF + SOX2
Accessible
Inaccessible
Single Molecule Occupancy
% of 5mC in GpC context
Bound SM
Unbound SM
C
OCT4 only
SOX2 only
D
Accessible unbound CBSs
KLF4 only
MYC only
OCT4 only
SOX2 only
YY1 only

Figure S5: Individual TF strengthen CTCF signal and disrupt CTCF-dependent nucleosome phasing. A: 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 sites where they are co-bound (green) compared to CTCF only (blue) and TF only sites (red) at inaccessible regions. 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. The profiles and heatmaps at accessible sites are shown in **Figure 4A**. **B**: Single molecule heatmaps of the nucleosome positioning centered on CBSs oriented on the positive strand at sites co-bound by CTCF and MYC, OCT4 or SOX2 for bound (black) and unbound (blue) molecules. The profiles compare the nucleosome position at co-bound (red) versus CTCF only bound sites (without spamo enriched TF motifs, or MYC, OCT4 or SOX2 binding detected by ChIP-seq) (black) for bound and unbound SM. The effect of KLF4 and YY1 are shown in **Figure 4B. C:** Single molecule heatmaps and profiles demonstrate the CTCF-independent effect of each TF on nucleosome positioning. Each heatmap is centered on the OCT4, or SOX2 motif (oriented on the positive strand) at sites with no bound CTCF or CTCF motif. The data for KLF4, MYC and YY1 are shown in **Figure 4C**. **D**: Single molecule heatmaps of the nucleosome positioning centered on the KLF4, MYC, OCT4, SOX2 or YY1 motif (oriented on positive strand) at bound sites without CTCF binding but with a CTCF motif.

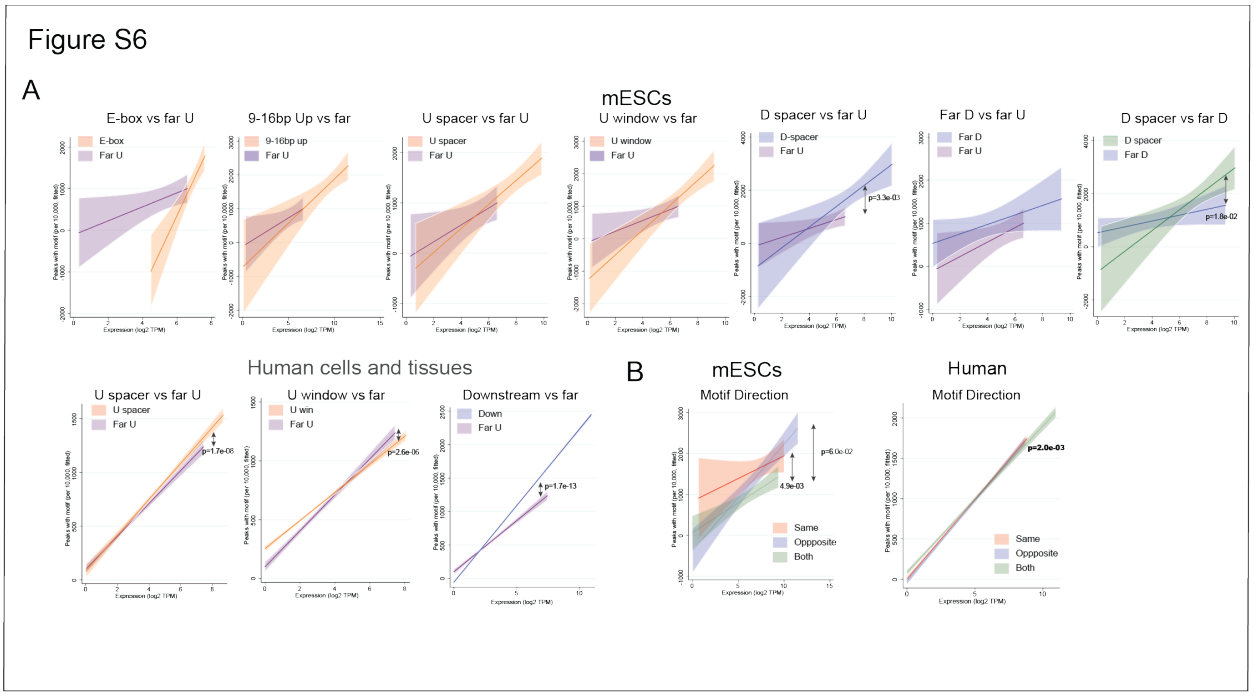


Figure S6: Independent effect of the TF motif window and of their motif orientation relative to CTCF motif on the correlation between TF motif enrichment and expression level. A: Fitted correlation between TF expression and the percentage of CTCF sites found with a TF motif showing the differential effects of the windows in mESCs and human iPSCs. **B:** Fitted correlation between TF expression and the percentage of CTCF sites found with a TF motif showing the differential effects of the TF motif orientation in mESCs and human iPSCs.