CTCF-dependent chromatin boundaries formed by asymmetric nucleosome arrays with decreased linker length

Christopher T. Clarkson[1], Emma A. Deeks[1,2], Ralph Samarista[1,3], Hulkar Mamayusupova[1], Victor B. Zhurkin[4,*] and Vladimir B. Teif [1,*]

[1]School of Life Sciences, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK, [2]Biological Sciences BSc Program, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK, [3]Wellcome Trust Vacation Student and [4]Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA

ABSTRACT

The CCCTC-binding factor (CTCF) organises the genome in 3D through DNA loops and in 1D by setting boundaries isolating different chromatin states, but these processes are not well understood. Here we investigate chromatin boundaries in mouse embryonic stem cells, defined by the regions with decreased Nucleosome Repeat Length (NRL) for ∼20 nucleosomes near CTCF sites, affecting up to 10% of the genome. We found that the nucleosome-depleted region (NDR) near CTCF is asymmetrically located >40 nucleotides 5′-upstream from the centre of CTCF motif. The strength of CTCF binding to DNA and the presence of cohesin is correlated with the decrease of NRL near CTCF, and anti-correlated with the level of asymmetry of the nucleosome array. Individual chromatin remodellers have different contributions, with Snf2h having the strongest effect on the NRL decrease near CTCF and Chd4 playing a major role in the symmetry breaking. Upon differentiation, a subset of preserved, common CTCF sites maintains asymmetric nucleosome pattern and small NRL. The sites which lost CTCF upon differentiation are characterized by nucleosome rearrangement 3′-downstream, with unchanged NDR 5′-upstream of CTCF motifs. Boundaries of topologically associated chromatin domains frequently contain several inward-oriented CTCF motifs whose effects, described above, add up synergistically.

INTRODUCTION

Nucleosomes are positioned along the genome in a non-random way (1–3), which is critical for determining the DNA accessibility and genome organisation (4). A particularly important nucleosome positioning signal is provided by CTCF, an architectural protein that maintains 3D genome architecture (5–7) and can organize up to 20 nucleosomes in its vicinity (8–11) (Figure 1A). Most other TFs do not possess such nucleosome-organizing potential (Supplementary Figure S1). CTCF has ∼100 000 potential binding sites in the mouse genome. Usually there are ∼30 000–60 000 CTCF sites bound in a given cell type, which translates to about 1 million of affected nucleosomes (up to 10% of the mouse genome) (11–14). CTCF is able to act as an insulator between genomic regions with different chromatin states, but how exactly this is achieved is not known. Here, we explore molecular mechanisms of the insulator boundary formation by CTCF through rearrangement of surrounding nucleosome arrays.

One of the ways to characterize genomic nucleosome distribution is through an integral parameter called the nucleosome repeat length (NRL), defined as the average distance between the centres of adjacent nucleosomes. NRL can be defined genome-wide, locally for an individual genomic region or for a set of regions. The local NRL is particularly important, since it reflects different structures of chromatin fibres (15–19). Ever since the discovery of the nucleosome (20,21) there have been many attempts to compare NRLs of different genomic regions (22–24) and it has been established that genome-wide NRL changes during cell differentiation (25,26). Recent sequencing-based investigations showed that active regions such as promoters, enhancers and actively transcribed genes usually have shorter NRLs and heterochromatin is characterized by longer NRLs (27–30). While in Yeast it is possible to link NRL changes to
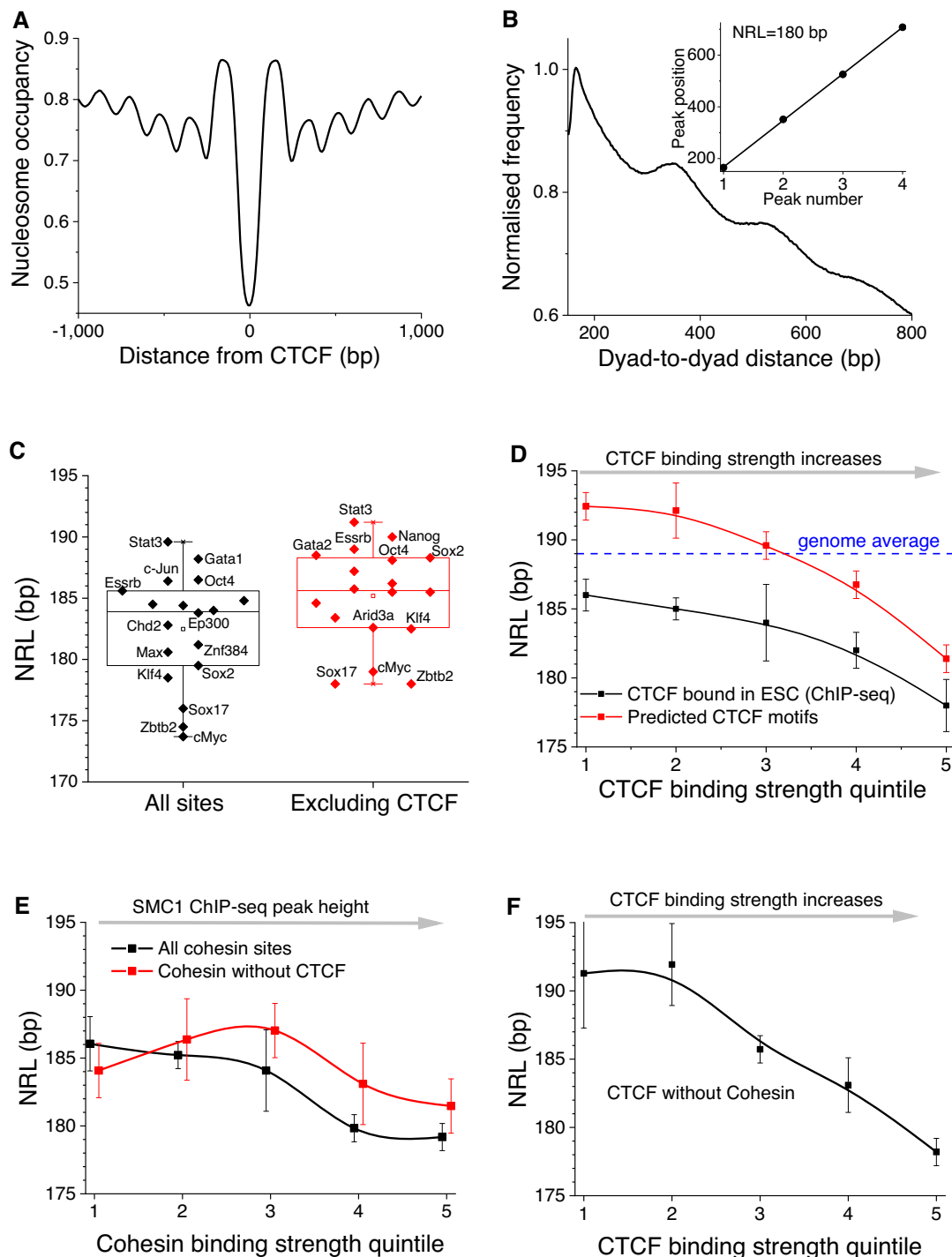
Figure 1. CTCF-dependent decrease of the nucleosome repeat length (NRL). (**A**) Average nucleosome profile based on MNase-seq from Voong *et al.* (42) around CTCF binding sites in ESCs determined by ChIP-seq (12). This profile is calculated without taking into account the directionality of CTCF binding. (**B**) Illustration of the 'phasogram' method of NRL calculation for the region [100, 2000] from the centre of experimental CTCF sites measured in ESCs. The calculation of frequencies of nucleosome dyad-to-dyad distances is followed by the linear regression of the peak positions (insert). (**C**) NRLs calculated near binding sites of 18 stemness-related chromatin proteins in ESCs in the region [100, 2000] from the summit of TF binding ChIP-seq peak, using chemical nucleosome mapping data from Voong *et al.* (42). Left: all TF binding sites; right: TF binding sites which do not intersect with CTCF. Open squares show the average NRL value based on all these TFs. The full list of experimental ChIP-seq datasets used in this calculation is provided in Supplementary Table ST1. (**D**) Dependence of NRL on the strength of CTCF binding based on experimental ChIP-seq peaks from mouse ENCODE (12) stratified into binding strength quintiles by the heights of peaks (black line) and computationally predicted CTCF sites obtained by scanning the mouse genome with TFBStools using >80% similarity for JASPAR matrix MA0139.1 stratified into binding strength quintiles by their TRAP score (red line). (**E**) NRL near bound cohesin, split into 5 quintiles based on the heights of experimental ChIP-seq peaks of the cohesin subunit SMC1 (83), calculated separately for all cohesin sites (black) and cohesin sites that do not contain CTCF motifs (red). (**F**) The same as (D), but only for experimental CTCF peaks that do not overlap with SMC1 peaks. The error bars correspond to the standard deviation of the linear fit across the peaks of the phasogram as explained in panel B.

the action of individual chromatin remodellers (31–35), in mouse or human regulatory regions are very heterogeneous and it is difficult to come up with a set of definitive remodeller rules determining their effect on NRL (36,37).

We previously showed that in mouse embryonic stem cells (ESC), NRL near CTCF is ∼10 bp smaller than genome-wide NRL (10,38). Our analysis demonstrated that purely statistical positioning of nucleosomes near CTCF boundaries would not be enough to explain genome-wide NRL shortening near bound CTCF observed experimentally; also, the effects of strong nucleosome-positioning DNA sequences, while compatible with the observed NRL, are limited to a small number of CTCF sites (39). A very recent study has investigated the effect of Snf2 and Brg1 remodellers on NRL in ESCs, suggesting Snf2 as the primary player (40). However, other factors may be at play as well. Thus, it is still unclear what determines the NRL near CTCF and how different CTCF sites are distinguished from each other e.g. during cell differentiation. Furthermore, recent studies have shown that CTCF can act as a boundary element between different chromatin states (e.g. DNA methylation) linearly spreading along the genome (11,41), but the mechanistic explanation for such a function is not immediately clear from the better established role of CTCF in 3D chromatin looping. Here we address these problems using available experimental datasets in ESCs and their differentiated counterparts.

We show below that the boundaries of nucleosome arrays are encoded in extended DNA regions >200 bp long enclosing individual CTCF motifs. Furthermore, the strength of CTCF binding provides a single 'code' that determines the value of NRL near CTCF, the level of asymmetry of CTCF-dependent nucleosome array boundaries, and eventually serves as a guide for larger-scale chromatin rearrangements during cell differentiation.

MATERIALS AND METHODS

Experimental datasets

Nucleosome positioning, transcription factor and chromatin remodeller binding datasets were obtained from the Gene Expression Omnibus (GEO), Short Read Archive (SRA) and the ENCODE web site as detailed in Supplementary Table ST1. NRL calculations near CTCF in ESCs were performed using the MNase-seq dataset from (42). NRL calculations near 18 stemness-related proteins in ESCs shown in Figure 1C and Supplementary Figure S3 were performed using the chemical mapping dataset from (42). NRL calculations in NPCs and MEFs were based on the MNase-seq datasets from (38). MNase-assisted H3 ChIP-seq from (11) was used for demonstrative purposes in the phasogram calculation in Figure 1B and aggregate profiles in Supplementary Figure S9. A more detailed list of datasets used in each figure is provided in Supplementary Table ST1. Coordinates of genomic features and experimental maps of transcription factor and remodeller binding in ESCs were obtained from published sources as detailed in Supplementary Table ST1. The coordinates of loops described in (43) were kindly provided by the authors in a BED file aligned to the mm10 mouse genome and converted to mm9 using liftOver (UCSC Genome Browser).

Data pre-processing

For nucleosome positioning, raw sequencing data were aligned to the mouse mm9 genome using Bowtie allowing up to two mismatches. For all other datasets, we used processed files with genomic coordinates downloaded from the corresponding database as detailed in Supplementary Table ST1. Where required, coordinates were converted from mm10 to mm9 since the majority of the datasets were in mm9.

Basic data processing

TF binding-sites were extended from the centre of the site to the region [100, 2000]. In order to find all nucleosomal DNA fragments inside each genomic region of interest, the bed files containing the coordinates of nucleosomes processed using the NucTools pipeline (44) were intersected with the corresponding genomic regions of interest using BedTools (45).

Binding site prediction

Computationally predicted TF binding sites were determined via scanning the mouse mm9 genome with position frequency matrices (PFMs) from the JASPAR2018 database (46) using R packages TFBSTools (47) and GenomicRanges (48). A similarity threshold of 80% was used for all TFs in order to get at least several thousand putative binding sites. In the case of MYC, we used matrix MA0059.1 defined in *Homo sapiens*, since its matrix MA0147.2 defined in *Mus musculus* returned a significantly smaller number of sites. For all other TFs we used default JASPAR matrices provided for *Mus musculus*.

Separation into forward and backward facing CTCF motifs

We used TFBSTools (47) to search on the 5′-3′ prime strand for forward facing CTCF motifs using the JASPAR matrix MA0139.1 and the 3′-5′ strand for motifs that are backwards facing ones. An alternative calculation using RSAT (49) with the same matrix led to similar results.

Calculation of aggregate nucleosome profiles

Aggregate nucleosome profiles were calculated using NucTools with single-base pair resolution (44). The calculation taking into account CTCF motif directionality was done as follows: in the case, if the motif is on the plus strand, the region [−1000, 1000] near CTCF also starts left to right, whereas for the minus strand the position of the region was mirrored with respect to the middle of the CTCF site.

Stratification of TF-DNA binding affinity

In the case of experimentally determined binding sites of CTCF, we stratified 33 880 sites, reported by the mouse ENCODE consortium (12), into five equally sized quintiles according to their ChIP-seq peak height reported in the original publication. In the case of computationally predicted TF sites, we started with 111 480 sites found by

scanning the mouse genome with TFBStools using JAS-PAR matrix MA0139.1 with 80% similarity threshold, and split them into five equal quintiles based on their TRAP score (50). The TRAP score is proportional to the binding probability of CTCF for a given site. In order to calculate the TRAP score, we extended CTCF motifs by 30 nucleotides in both directions and used tRap implementation of the TRAP algorithm in R with default parameters (https://github.com/matthuska/tRap). In the calculations involving CTCF motif directionality (Figures 5–7) we first arranged predicted sites by the TRAP score into quintiles, and after that intersected them with the experimental ChIP-seq peaks of CTCF. Only motifs overlapping with sites that were experimentally detected by ChIP-seq in at least one mouse cell type were retained (including datasets from the mouse ENCODE project (12), GSE27944 (51), GSE96107 (43), GSE114599 (11)). These were further filtered to exclude CTCF sites separated by <1000 bp from annotated transcription start sites (TSSs), which removed about 10% of CTCF sites. TSSs were defined based on the Genomatics Eldorado database (Genomatix GmbH). After these filtering steps we obtained the following numbers of sites in the binding strength quintiles Q1−Q5: 3596 (Q1); 3782 (Q2); 6776 (Q3); 14 776 (Q4); 16 860 (Q5).

Phasogram calculation

The 'phasograms' representing the histograms of dyad-to-dyad or start-to-start distances were calculated with Nuc-Tools. When paired-end MNase-seq was used, dyad-to-dyad distances were calculated using the center of each read as described previously (44). When chemical mapping data was used, this procedure was modified to use the start-to-start distances instead, because in the chemical mapping method the DNA cuts happen at the dyad nucleosome locations. The phasogram was then used for the NRL calculation as explained in Figure 1B. The NRL was defined by the slope of the line connecting the phasogram peaks; this line was determined by linear fitting, taking into account only the phasograms where ANOVA *P*-value for the slope determination is <0.05.

Selection of the location of the region near CTCF for NRL calculations

We noticed that NRL near CTCF depends critically on the distance of the region of NRL calculation to the binding site summit (Supplementary Figure S2). While the phasograms for regions [100, 2000] and [250, 1000] near the summits of the experimental CTCF sites, which both exclude the CTCF site, are quite similar to each other, a region that includes the CTCF peak summit [−500, 500] is characterised by a very different phasogram. However, the latter phasogram is an artefact of the effect of the interference of two 'waves' of distances between nucleosomes: one wave corresponds to the distances between nucleosomes located on the same side of CTCF, and the second wave corresponds to distances between nucleosomes located on different sides from CTCF. The superposition of these two waves results in the appearance of additional peaks (Supplementary Figure S2A). A

linear fit through all the peaks given by the interference of these two waves gives NRL = 155 bp, but this value does not reflect the real prevalent distance between nucleosomes (Supplementary Figure S2B). We thus selected the region [100, 2000] for the following calculations. Below, all NRLs refer to regions [100, 2000] near the summits of TF binding sites, unless specified otherwise. We would like to note that the effect explained above means that some of the previous publications reporting NRL near CTCF may need to be re-evaluated, because the summit of CTCF site needs to be always excluded from the genomic region for robust NRL calculations; otherwise, the apparent NRL is unrealistically small. We checked that this artefact at least does not affect NRL calculations near TSS (Supplementary Figure S2C). Once the region location with respect to the CTCF site is fixed, the phasograms are not significantly affected by the choice of the nucleosome positioning dataset (Supplementary Figure S2D). In the following calculations in ESCs we used the high-coverage MNase-seq and chemical mapping datasets from (42).

Automated NRL determination from phasograms

Studying many phasograms proved cumbersome when manually picking the peak locations in a non-automated way. To circumvent this problem, we developed an interactive applet called *NRLcalc* based on the Shiny R framework (http://shiny.rstudio.com), to allow one to interactively annotate each phasogram such that the NRL could be calculated conveniently. *NRLcalc* allows one to select a smoothing window size to minimize noise in the phasograms. A smoothing window of 20 bp was used in our calculations. The applet also provides the *Next* and *Back* button to allow the user to go through many phasograms, as well as intuitive user interface to load and save data.

Analysis of RNA expression near CTCF

RNA-seq data was downloaded from the GEO GSE98671 (7) and mapped with TopHat (52) to the mm9 genome. The mapped BAM files were converted to BED format with BEDOPS (53). The numbers of RNA reads aligning 1000 bp up- and downstream of CTCF motifs were calculated using BedTools (45), requiring at least 1bp intersection.

TAD analysis

TAD coordinates in ESCs and NPCs reported by Bonev *et al.* (43) for the mm10 genome were converted to mm9 using liftOver. TADs defined as common, lost and gained upon ESC to NPC transition were determined using Bed-Tools' command intersect with parameter -wc. TADs with the rate of overlap between ESCs and NPCs >90% were considered common; those appearing in ESCs and NPCs with an overlap rate <80% were defined as lost and gained correspondingly. The aggregate profiles of CTCF motifs around TAD boundaries were calculated with HOMER (54) at a bin resolution of 5000 bp.

RESULTS

Setup of NRL calculations

Let us base our NRL calculations on the 'phasogram' algorithm introduced previously (27,38). The idea of this method is to consider all mapped nucleosome reads within the genomic region of interest and calculate the distribution of the frequencies of distances between nucleosome dyads. This distribution typically shows peaks corresponding to the prevalent distance between two nearest neighbour nucleosomes followed by the distances between next neighbours. The slope of the line resulting from the linear fit of the positions of the peaks then gives the NRL (Figure 1B). To perform bulk calculations of NRLs for many genomic subsets of interest we developed software *NRLcalc*, which loads the phasograms computed in *NucTools* (44) and performs linear fitting to calculate the NRL (see Materials and Methods).

Each TF is characterised by a unique NRL distribution near its binding sites. For example, we used a recently reported chemical nucleosome mapping dataset (42) to calculate NRLs in the region of up to 2000 bp from the centre of the binding site excluding the central 100 bp (hereafter referred to as region [100, 2000]) for 18 stemness-related TFs whose binding has been experimentally determined in ESCs using ChIP-seq (Figure 1C). This analysis revealed that the proximity to CTCF binding sites unanimously reduced the NRL near these sites. When we filtered out TF binding sites that overlap with CTCF binding sites in ESCs, the NRLs for each individual TF increased (Figure 1C). On the other hand, TF binding sites that overlap with CTCF had significantly smaller NRLs (Supplementary Figure S3).

The strength of CTCF binding correlates with NRL decrease in the adjacent region

To dig deeper into the relationship between CTCF and local chromatin conformation, we split CTCF sites into 5 quintiles of increasing binding strength. Two metrics were used as a means of quantifying CTCF binding strength: (i) Experimentally determined CTCF binding sites in ESCs were split into five quintiles based on the height of the ChIP-seq peaks reported by the mouse ENCODE consortium (12). (ii) Theoretically predicted binding sites defined by scanning the mouse genome using TFBStools (47) with the 19-bp CTCF motif (JASPAR MA0139.1) (46) were split into five quintiles based on their calculated TRAP score that is proportional to the probability of CTCF binding to a given site (50) (see Materials and Methods). In each case, the calculation of the NRL was performed in the region [100, 2000] near CTCF binding sites using MNase-seq data (42). These calculations revealed a smooth decrease of NRL as the strength of CTCF binding increased in the case of both used metrics (Figure 1D). In addition, we used the chemical nucleosome mapping dataset (42) to compare the CTCF quintiles in terms of the distribution of nucleosome dyad-to-dyad distances, which also revealed that stronger CTCF binding is associated with smaller NRLs (Supplementary Figure S4). Thus, the CTCF-dependent NRL decrease is a general, dataset-independent effect. Note that chemical mapping-based NRLs should not be directly compared with MNase-seq ones due to the inherent peculiarities of the chemical mapping experiment that we noticed previously (44); below we will use only MNase-seq and MNase-assisted histone H3 ChIP-seq datasets for nucleosome mapping.

We then asked, whether the same effect on NRL is observed for CTCF's binding partner cohesin. Cohesin is a ring-shaped complex that slides along one or two DNA double helices until it meets CTCF, thus extruding DNA loops (55). Cohesin is able to induce regular nucleosome arrays around it even when not associated with CTCF (Supplementary Figure S5), thus it is interesting whether it has a similar effect on NRL. Cohesin does not have its own DNA sequence preferences, but we can still stratify mapped cohesin locations in terms of the strength of binding using ChIP-seq of cohesin's component SMC1 and sorting its occupancy peaks into quintiles based on their height. Figure 1E shows that, similarly to CTCF, cohesin sites are characterized by the local NRL decrease as cohesin's binding strength increases. However, the effect of cohesin's binding strength on NRL is weaker than that for CTCF, and almost disappears if only the cohesin sites that do not contain CTCF motifs are considered (Figure 1E). On the other hand, the bound CTCFs that do not overlap with bound cohesin in ESCs still display a pronounced effect of CTCF binding strength on NRL (Figure 1F). This effect was also recapitulated for CTCF sites residing at least 10 000 bp outside of annotated TSSs (Supplementary Figure S6), showing that it was not caused by protein coding gene transcription.

Using the same procedure we have investigated NRL near other chromatin proteins. Firstly, we considered 497 TFs which have position weight matrices in JASPAR2018 (46), and for each of them calculated NRL in the region [100, 2000] from the TF motif as a function of the DNA-binding strength predicted for a given TF. This analysis revealed that for TFs other than CTCF, the NRLs did not reveal a monotonic function of their binding strength (see Figure 2 for examples of TFs relevant to stem cells). We have also performed a similar calculation for chromatin remodellers that have been experimentally profiled in ESCs, asking whether NRL in the region [−1000, 1000] near remodeller depends on the height of the corresponding remodeller peak (Supplementary Figure S7). These calculations did not reveal NRL dependence on the binding strength as in the case of CTCF or cohesin. Thus, CTCF and cohesin are unique proteins whose DNA binding strength is anticorrelated to the NRL value.

The strength of CTCF-DNA binding correlates with GC and CpG content

In order to understand the physical mechanisms of NRL decrease near CTCF, we considered a number of genomic features and molecular factors that could potentially account for the NRL decrease near CTCF (Figure 3). A number of previous observations suggested that the ability of CTCF sites to retain CTCF during cell perturbations is related to the surrounding GC and CpG content (11,56). Our calculations performed here provide more detail on this effect, showing that the strength of CTCF binding is correlated
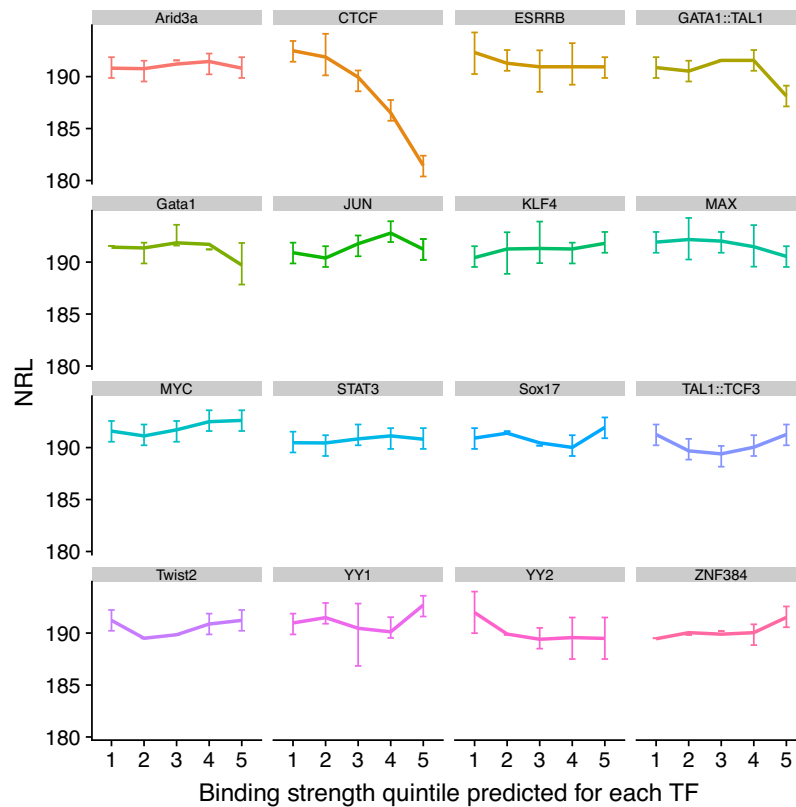
Figure 2. Proteins other than CTCF and cohesin do not show the relationship between DNA-binding strength and NRL near their binding sites. Sixteen representative TFs related to stem cells are shown (similar calculations were performed for 497 TFs listed in JASPAR2018). TF bindings sites used in this analysis were predicted computationally by scanning the mouse genome using TFBStools with the 80% motif similarity cut off and then stratified into five binding strength quintiles based on the TRAP score (see Materials and Methods).

with GC content around CTCF sites (Figure 3A), and that the probability for a given site to be located in a CpG island monotonically increases with the CTCF binding strength (Figure 3B). It is worth noting that the CTCF motif itself is GC-rich, which corresponds to the central peak in Figure 3A, but the effects mentioned above extend to distances >1000 bp from CTCF motif. Furthermore, the CTCF site location inside CpG islands was associated with a significantly decreased NRL in comparison with all CTCF sites (Figure 3D).

The strength of CTCF-DNA binding correlates with the probability of a given site to be inside *cis*-regulatory elements and domain boundaries

Another potential hypothesis is that the small NRL near CTCF could be because CTCF sites are in active regions (promoters, enhancers, etc.) which have a smaller NRL in comparison to genome-average based on previous studies (27,28). Our analysis performed here demonstrated that there is a positive correlation between the strength of CTCF binding and the probability that it is inside a promoter region (Figure 3C). We also used recently published coordinates of topologically associated domains (TADs) and promoter-enhancer loops in ESCs (43) and showed that there is a correlation between the strength of CTCF binding and the probability that it forms a boundary of TADs and even higher correlation for the boundaries of loops (Figure

3C). Furthermore, the NRL near CTCF sites was smaller if these sites were inside borders of loops or TADs, while the NRL value went up if all known regulatory regions were excluded (Figure 3D).

Remodeller-specific effects on NRL near CTCF

Active nucleosome positioning is determined by chromatin remodellers, but the rules of action of individual remodellers are not well defined. In order to clarify remodeller effects on NRL decrease near CTCF, we processed all available remodeller ChIP-seq datasets in ESCs and plotted the percentage of CTCF sites overlapping with remodeller ChIP-seq peaks (Figure 4A). This analysis showed that the stronger CTCF binds the higher the probability that a given CTCF binding site overlaps with remodellers. Particularly large percentage of CTCF sites overlaps with peaks of remodellers Chd4, EP400, Chd8 and BRG1, with Chd4 being the top CTCF-related remodeller. We have also performed similar analysis for three different TFs: CTCFL, Oct4 and c-Jun (Supplementary Figure S8). CTCFL (also known as BORIS), shares a number of sites with CTCF, and unsurprisingly BORIS and CTCF have similar preferences for remodellers. On the other hand, Oct4, which is highly expressed in ESCs, showed a qualitatively similar effect of increasing co-binding with remodellers as its DNA sequence-determined binding strength increases, but the top Oct4-associated remodeller was BRG1 rather than Chd4. As a
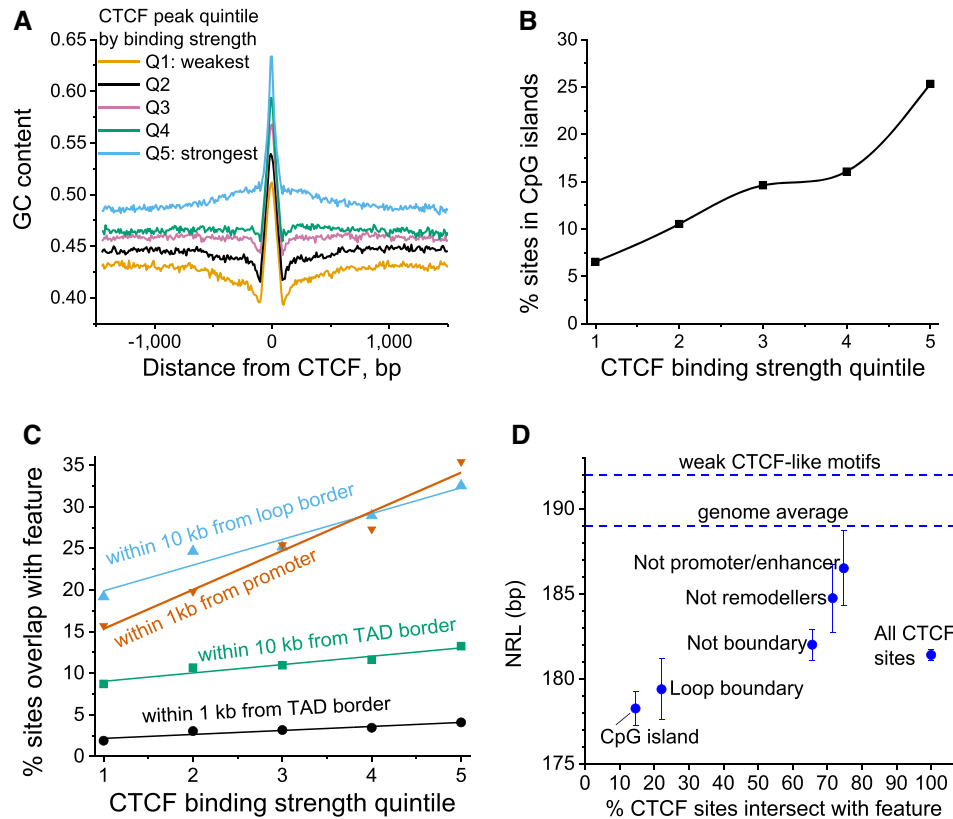
Figure 3. Genetic features correlating with the experimental strength of CTCF binding. (**A**) CTCF binding sites split into quintiles based on their binding strength are characterized by increasing GC content as CTCF binding strength increases. (**B**) The stronger CTCF binding site the higher is the probability that it is located in a CpG island. (**C**) The stronger CTCF binds the higher the probability that it is located in a promoter or forms a boundary of TADs or enhancer-promoter loops. (**D**) NRLs for the following subsets of CTCF sites: all sites bound in ESCs; inside chromatin loop boundary; outside of boundaries of loops and TADs; inside CpG islands; outside of chromatin remodeller peaks; outside of promoters and enhancers. The top horizontal dashed line corresponds to the weak CTCF-like motifs from Figure 1D. Vertical bars show the standard deviation.

negative control, we considered c-Jun, which is not a stem cell TF. As expected, for c-Jun binding sites the percentage of intersection with remodeller peaks did not depend on the predicted strength of c-Jun binding to DNA (Supplementary Figure S8).

Next we set to derive systematic rules of remodeller effects on NRL near CTCF (Figure 4B). By comparing NRLs near CTCF sites overlapping and non-overlapping with each remodeller, we learned that Brg1 has no detectable effect (based on two independent Brg1 datasets), and Snf2h has the strongest effect. The effect of other remodellers on NRL near CTCF is increasing in the order BRG1 ≤ Chd4 < Chd6 < Chd1 ≤ Chd2 ≤ EP400 ≤ Chd8 < Snf2h (Figure 4B).

CTCF motif directionality introduces asymmetry in adjacent nucleosome distribution

All our calculations above were performed without considering the directionality of the CTCF motif. For example, Figure 1A shows a symmetric pattern of nucleosome occupancy around CTCF, which arises due to averaging of different patterns around CTCF motifs in the direction of the plus and minus strand. Now let us always orient the CTCF motif in the same way, left to right (5′ to 3′), and refer to

positions in 5′ direction from the CTCF motif as 'upstream' and 3′ direction as 'downstream'. Using this setup, we calculated aggregate profiles of nucleosome occupancy around CTCF by aligning all regions in 5′ to 3′ direction of the CTCF motif defined by the JASPAR matrix (MA0139.1). In these calculations we considered only CTCF motifs located in ChIP-seq defined peaks in at least one mouse cell type. Furthermore, we excluded CTCF sites that are located inside annotated promoters (see Materials and Methods).

Figure 5A shows the aggregate profiles of MNase-seq nucleosome occupancy (42) around CTCF in ESCs taking into account the motif directionality. Here, the wave-like pattern of the nucleosome occupancy around CTCF sites reveals strong asymmetry. To the best of our knowledge this is the first report of such a pronounced nucleosome asymmetry around CTCF motifs. Counterintuitively, the weaker the CTCF binding, the stronger is the asymmetry. Such an asymmetry is similar to what is usually observed near promoters, except that we have excluded from this calculation CTCF sites that overlap with promoters. We have also confirmed this effect using MNase-assisted H3 ChIP-seq dataset (Figure S9) and plotted the occupancy of RNA Pol II around CTCF (Figure 5B). Pol II occupancy shows CTCF-dependent enrichment, which increases with the increase of CTCF binding strength. Weak CTCF sites which
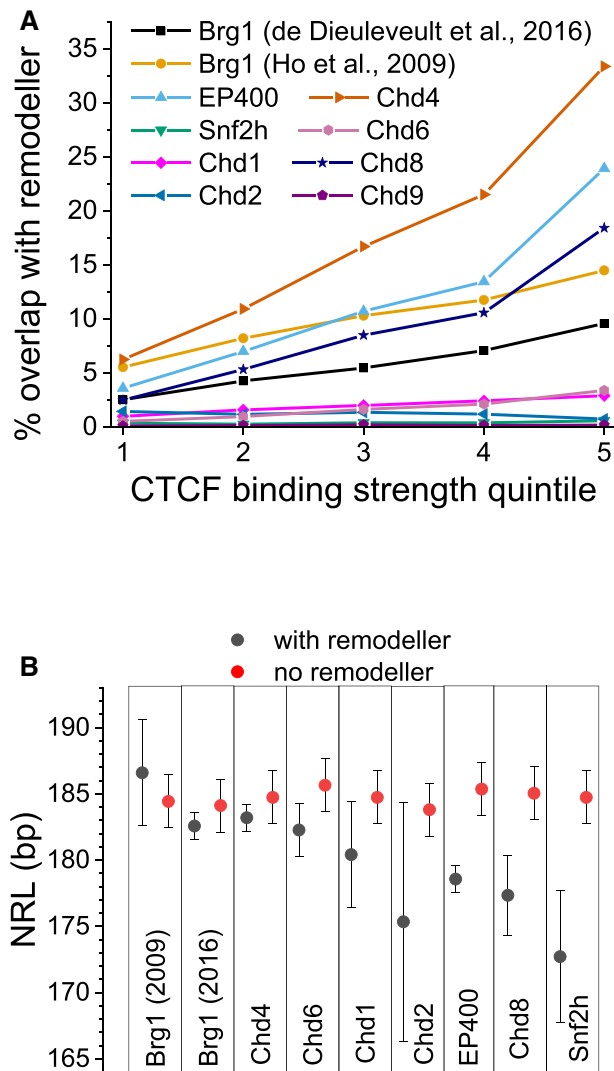
Figure 4. Effects of different chromatin remodellers on the value of NRL near CTCF. (**A**) The stronger CTCF binds the higher is the probability that it is co-enriched with different chromatin remodellers indicated on the figure. The enrichment was defined as the ratio of CTCF sites overlapping with ChIP-seq peaks of a given remodeller to the total number of CTCF sites in a given quintile. (**B**) NRLs calculated near CTCF sites that overlap (black) and do not overlap (red) with ChIP-seq peaks of eight chromatin remodellers experimentally mapped in ESCs. Remodeller names are indicated on the figure. Two Brg1 datasets reported in 2009 and 2016 are taken from separate publications, (84) and (36) respectively.

have the strongest asymmetry are devoid of Pol II. Thus, the asymmetry of nucleosome occupancy near CTCF is not due to Pol II-dependent transcription.

The most striking feature of the asymmetric nucleosome profiles near CTCF is that the deepest point of the nucleosome-depleted region is shifted about 41 bp 'upstream' in 5′ direction from the centre of the CTCF motif. This is different from what is usually assumed based on symmetric profiles such as in Figure 1A. Interestingly, the first strong nucleosome peak at 105 bp 'downstream' in 3′ direction from CTCF appears similarly for all CTCF site quin-

tiles, whereas the next peak at 165 bp 'downstream' in 3′ direction from CTCF is extremely sensitive to the CTCF binding strength. There are also several other nucleosome occupancy peaks that display strong sensitivity to the CTCF binding strength.

The CTCF-dependent peak of nucleosome occupancy 3′-downstream of CTCF can be attributed to Chd4

In order to determine the structural origin of the nucleosome occupancy peak at 165 bp from the CTCF, motif we calculated aggregate profiles of all chromatin remodellers using their ChIP-seq binding datasets in ESCs (Supplementary Figure S10). Interestingly, Figure S10 shows that the remodellers position themselves between nucleosomes. Chd4 is the only remodeller characterized by a CTCF-dependent peak at position +165 bp (Figure 5C). The peak of Chd4 at this location is quite pronounced, which is consistent with Chd4 being the top CTCF-associated remodeller (Figure 4A). Thus, Chd4 plays an important role in establishing the asymmetry of nucleosome positioning, while it does not affect the NRL value *per se* (Figure 4B). On the other hand, another remodeller Snf2h affects the value of NRL and the regularity of the nucleosome near CTCF (see Supplementary Figure S11, plotted using the recent Snf2h knockout data (40)).

CTCF creates asymmetric nucleosome arrays; cohesin symmetrises them

Next we investigated the interplay between CTCF and cohesin in relation to the asymmetry of nucleosome arrays. Cohesin's subunits Rad21 and SMC1 bind quite symmetrically with respect to the CTCF motif (Supplementary Figure S12) and they have a dramatic effect on the symmetry of nucleosome arrays around CTCF (Figure 5D). Our calculations showed that for all CTCF binding strength quintiles, CTCFs which are not co-bound with cohesin create asymmetric and less regular nucleosome arrays, whereas CTCFs co-bound with cohesin create more symmetric and more regular arrays of nucleosomes (Figure 5D).

The value of NRL in the region 3′-downstream of the CTCF motif linearly depends on the CTCF binding strength

The effect of CTCF motif directionality introduces a significant correction to the NRL dependence on the CTCF binding strength that we found above (Figure 5E and F). When performing NRL calculations separately for the region [100, 2000] 3′-downstream and region [−2000, −100] 5′-upstream from the centre of the CTCF motif, we noticed that the most regular behaviour is observed 3′-downstream, where the effect can be described by a linear dependence (Figure 5F). We also checked whether the appearance of the nucleosome occupancy peak 165 bp downstream of CTCF is the main determinant of the NRL decrease. The recalculation of the NRL in the interval [300, 2000] 3′-downstream from CTCF showed that while the NRL decrease is less steep, it still follows the same trend (Supplementary Figure S13).
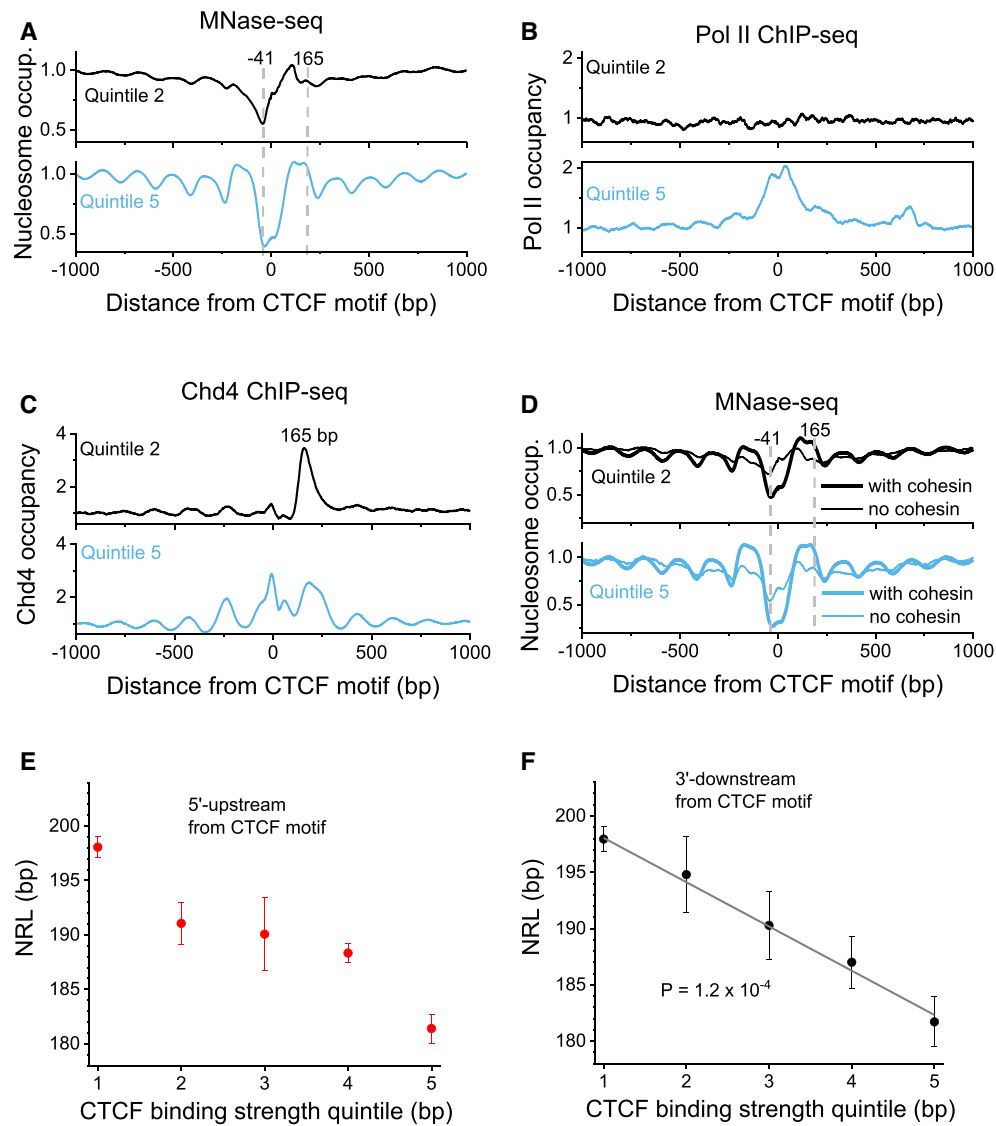
Figure 5. Combined effects of CTCF motif directionality and binding strength on nucleosome positioning. (**A**) Aggregate nucleosome profiles based on MNase-seq (42) around CTCF motifs outside promoters which coincide with experimentally verified binding sites in at least one mouse cell types, taking into account the DNA strand directionality. The strong peak at 105 bp from the centre of CTCF motif appears for all CTCF quintiles. On the other hand, the nucleosome peak at position 165 is sensitive to the strength of CTCF binding and increases as the strength of CTCF binding increases from weak binding at quintile 2 to strong binding at quintile 5. (**B**) CTCF binding outside of promoters is associated with CTCF-dependent Pol II enrichment. In the weakest CTCF quintile there is no Pol II enrichment, so the promoter-like nucleosome occupancy near CTCF is not due to Pol II. (**C**) The binding of Chd4 (and not any other experimentally profiled remodeller) shows a CTCF dependent peak at 165 bp, coinciding with the nucleosome occupancy peak. (**D**) Nucleosome positioning based on MNase-seq, as in panel A, but CTCF sites are split into those that overlap with the cohesin subunit SMC1 (thick line) and do not overlap with SMC1 (thin line). E and F) NRL as a function of CTCF binding strength quintile corrected for the CTCF motif directionality. (**E**) NRL calculated in the region [−2000, 100] in 5′ direction ('upstream') of the centre of CTCF motif. (**F**) NRL calculated in the region [100, 2000] in 3′ direction ('downstream') of the centre of CTCF motif. In the latter case NRL dependence of CTCF binding strength can be fitted as a straight line (*t*-test $P = 1.2 \times 10^{-4}$).

The asymmetric nucleosome depletion 5′-upstream of CTCF/CTCFL motifs is encoded in DNA repeats and may be linked to their transcription

Next we calculated the average nucleotide distribution around CTCF sites used above taking into account the orientation of CTCF motifs. This revealed an unexpected nucleotide pattern in the extended region near CTCF (Figure 6A). The nucleosome depletion in the region around −41 bp upstream of CTCF is associated with a decrease

of GC content. This is consistent with previous observations that high AT-content and in particular poly(dA:dT)-tracts have strong nucleosome-excluding properties (57). It is worth noting that the CTCF motif used in our calculations is just 19 bp, but the length of the highly structured area near CTCF is more than 200 bp. This means that the CTCF motif is frequently encountered as part of a much larger DNA sequence organisation, some type of sequence repeats that are primarily responsible for the establishment of the asymmetric boundaries around CTCF.
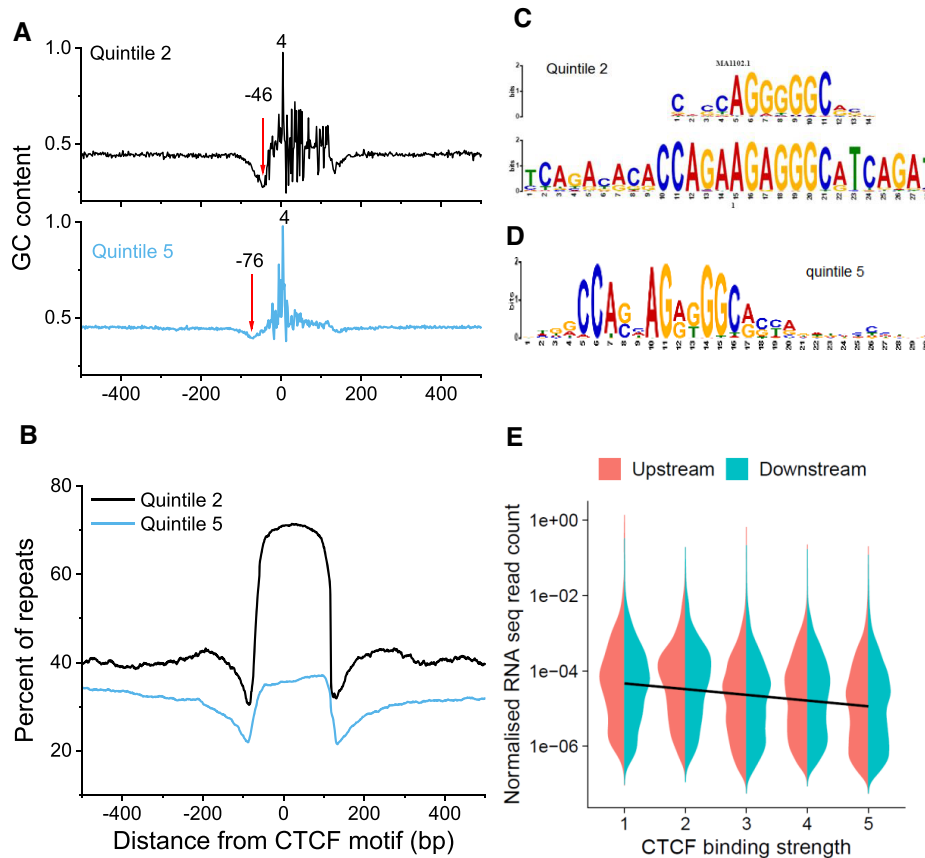
Figure 6. Effects of the nucleotide content around CTCF sites. (**A**) Average GC content around CTCF motifs for CTCF binding strength quintiles 2 and 5. (**B**) The percentage of repeats determined by the USCS Genome Browser's Repeat Masker as a function of the distance from the middle of CTCF motifs. (**C**) The sequence of the consensus motif in quintile 2 with the smallest *P*-value. The best TF match for the quintile 2 consensus motif is CTCFL (Boris) (JASPAR MA1102.1). (**D**) The sequence of the consensus motif in quintile 5. The quintile 5 consensus sequence contains the classical CTCF motif (JASPAR MA0139.1). (**E**) Violin plot showing the numbers of RNA reads expressed from the regions [−1000; 0] and [0; 1000], respectively upstream (red) and downstream (blue) of CTCF binding sites, as a function of CTCF binding strength. The straight line is a linear fit through all the points, showing a general decrease of the number of RNA reads as CTCF binding strength increases ($P = 1.2e−11$). The linear fits performed separately across 'downstream' or 'upstream' regions are not distinguishable.

Indeed, 50% of the CTCF motifs used in our calculations in Figures 5 and 6 overlapped with repeats defined by the UCSC Genome Browser repeat masker. Furthermore, the percentage of repeats given by the repeat masker shows a very structured profile with an extended region (>200 bp) near CTCF strongly enriched with repeats (Figure 6B).

Another interesting finding shown in Figure 6C and D is that when we subjected each binding strength quintile to a separate *de novo* motif discovery, the strongest quintile 5 was associated with the classical CTCF motif (JASPAR MA0139.1), whereas a weak quintile 2 was associated with CTCFL (BORIS) defined by the JASPAR matrix MA1102.1. To the best of our knowledge this is the first indication that CTCF and CTCFL may have different effects on nucleosomal organisation (Figure 5A).

We have also checked whether the nucleosome depletion 5′-upstream of CTCF is related to transposon transcription. Using coordinates of ChIP-seq peaks of RNA Pol III determined previously in ESCs (58), we found that 33% of co-localisations of TFIIIC and Pol III and 17% of co-localisations of SINE repeats and Pol III overlapped with our CTCF motifs. Thus, not only the DNA repeats are re-

sponsible for the AT-rich region 5′-upstream of CTCF, but also their Pol III-dependent transcription may be linked to the asymmetric nucleosome depletion pattern.

CTCF binding directly affects expression of adjacent RNA

In order to investigate quantitatively the effect of CTCF on RNA expression, we plotted the normalized amount of total RNA reads within [−1000, 1000] from CTCF as a function of CTCF binding strength (Figure 6E). It showed that the strong CTCF binding correlates with the weaker expression of neighbouring RNA ($P = 1.2e−11$). There was no significant asymmetry in RNA expression up- or downstream of the CTCF motif.

Nucleosome-depleted boundaries 5′-upstream of CTCF motif are preserved even if binding CTCF is lost during cell differentiation

Next we compared nucleosome positioning around CTCF motifs upon differentiation of ESCs to neural progenitor cells (NPCs), as well as in the differentiated mouse embryonic fibroblasts (MEFs) using MNase-seq data from (38)
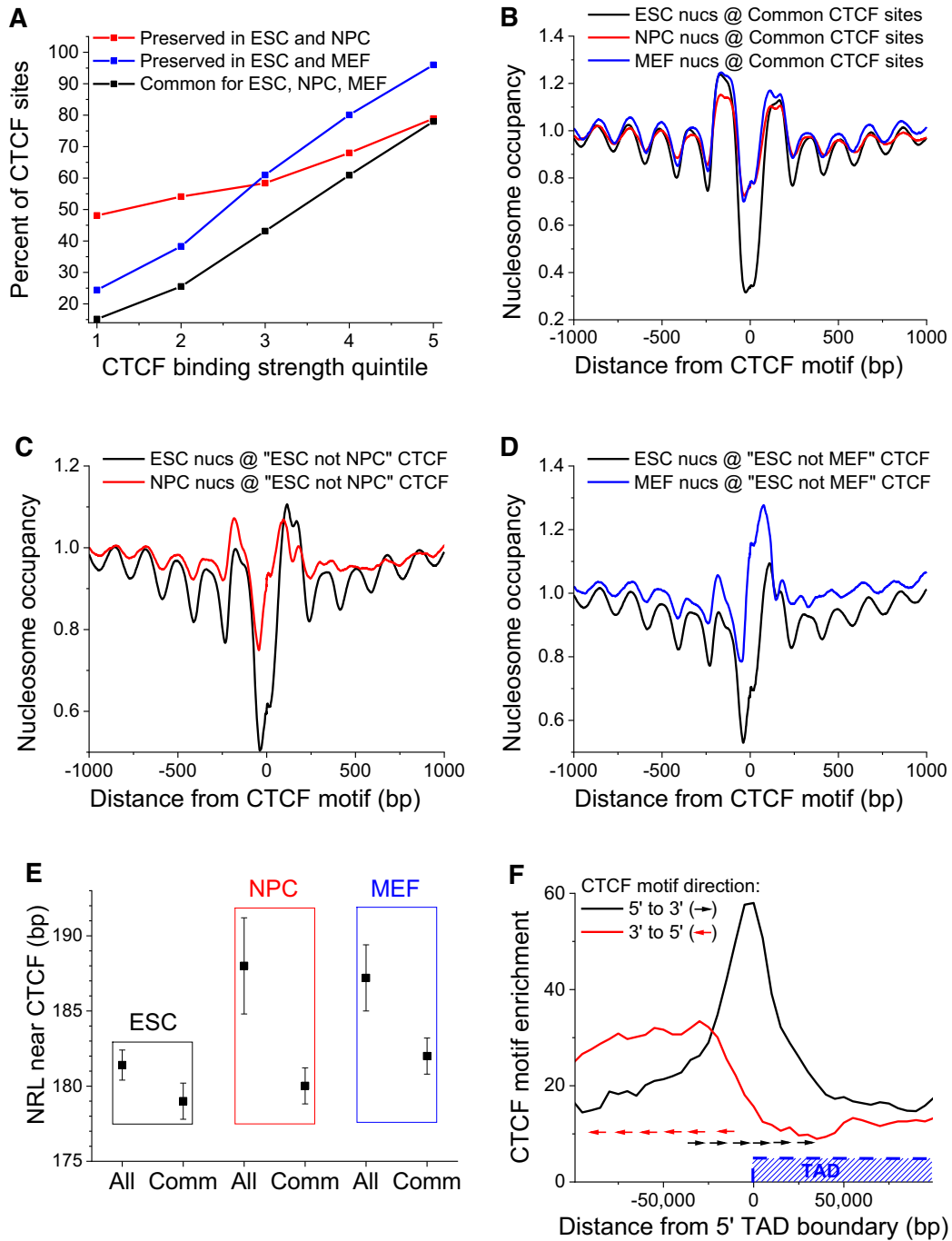
Figure 7. Effects of asymmetric CTCF-dependent boundaries in stem cell differentiation. (**A**) The fraction of CTCF sites preserved upon differentiation of ESCs to NPCs and MEFs as a function of CTCF binding strength. CTCF sites preserves in all these three cell types are termed 'common'. (**B**) Nucleosome occupancy in ESCs (black), NPCs (red) and MEFs (blue) around CTCF sites common between ESC, NPC and MEF, calculated taking into account CTCF motif directionality. (**C**) Nucleosome occupancy around 'ESC not MEF' sites that are present in ESCs (black line) but lost in MEFs (red line) taking into account CTCF motif directionality. (**D**) Nucleosome occupancy around 'ESC not NPC' sites that are present in ESCs (black line) but lost in NPCs (red line) taking into account CTCF motif directionality. Note that in differentiated cells a nucleosome is being positioned to cover the 'lost' CTCF sites, but nucleosome depletion on the left of CTCF is still preserved. (**E**) NRLs in region [100, 2000] from CTCF's experimental binding site summit calculated without taking into account the motif directionality. Upon differentiation average NRL near CTCF increases (denoted 'All'), but common CTCF sites keep the smallest NRL (denoted 'Comm'). (**F**) Enrichment of the strongest CTCF motifs (5th quintile) near 5'-boundaries of TADs in ESC, calculated separately for CTCF motifs oriented 5'-to-3' (black) and 3'-to-5' (red). The TAD is located to the right from the 5'-boundary. The arrows show an example of CTCF motif distribution for an individual region.

and CTCF ChIP-seq data from (12,43) (Figure 7A). Notably, stronger CTCF binding to DNA increases the probability that a given site will remain bound upon differentiation. This suggests that the sequence-dependent strength of CTCF binding can act as the 'CTCF code', determining which CTCF sites are retained and lost upon differentiation (and thus how the 3D structure of the genome will change). Our further analysis revealed that common CTCF sites that are present in all three states are characterized by quite minor asymmetry of nucleosome organisation (Figure 7B). On the other hand, CTCF sites that are lost upon ESC differentiation to NPCs and MEFs have more profound asymmetry of the nucleosome pattern around them (Figure 7C and D). Upon differentiation both in NPCs and MEFs, the array of nucleosomes 3′-downstream of the CTCF motif is shifted to cover the CTCF site. It is worth noting that nucleosome positioning in this region is only partly CTCF-dependent. For example, inside the [−100, 100] region around CTCF, the percentage of nucleosomes covering the CTCF motifs that lost CTCF upon differentiation changes from 47% to 60% upon ESC to MEF transition, and from 42% to 54% upon ESC differentiation to NPC. Interestingly, the nucleosome-depleted region 5′-upstream of CTCF still remains open upon differentiation. The latter effect was also confirmed for the case of CTCF sites that are not bound by CTCF in ESCs and become bound in MEFs (Supplementary Figure S15).

Common CTCF sites preserve local nucleosome organisation during ESC differentiation

Then, we set to determine the functional consequences of the NRL decrease near CTCF. NRL near bound CTCF on average increases as the cells differentiate from ESCs to NPCs or MEFs (Figure 7E and Supplementary Figure S16). However, common CTCF sites resist this NRL change, suggesting that CTCF retention at common sites upon differentiation preserves both 3D structure and nucleosome patterns at these loci. As we have established previously (Figure 5F), the effect of the active CTCF-dependent NRL decrease is mostly pronounced in the region 3′-downstream of CTCF motifs. The NRL increase near CTCF upon cell differentiation is also mostly in the 3′-downstream region (Supplementary Figure S17).

Directed CTCF motifs mark TAD boundaries

Our previous calculations were performed at the level of boundaries formed by single CTCF motifs. However, in some cases chromatin boundaries are created by cumulative action of several CTCF motifs located not far from each other. In particular, our calculations showed that CTCF motifs oriented toward the inner part of TAD are centred at the TAD boundaries, whereas the outward-looking CTCF motifs are enriched at the outer side of the boundaries (Figure 7F). TADs that were lost upon differentiation demonstrate a smaller enrichment of CTCF motifs near them (Supplementary Figure S18), which suggests that CTCF motifs at functionally important chromatin boundaries may act additively. Thus, the effects of individual CTCF motifs described above can be summed up at a region of up to several kb, to act synergistically at the boundaries between large chromatin domains.

DISCUSSION

We developed a new *NRLcalc* methodology to investigate nucleosome rearrangement and NRL changes near TF binding motifs distinguished by their orientation and binding strength. The application of this method to CTCF and cohesin binding sites revealed a number of new effects (Figure 8):

Firstly, we found that contrary to previous assumptions, the nucleosome arrangement near CTCF motifs is asymmetric and to a large degree hard-wired in the sequence of the DNA region >200 bp long including the CTCF motif (Figures 5A and 6A). The asymmetry in this case is not just a consequence of heterogeneity of nucleosome distributions around subsets of sites (59), but is a generic feature across all CTCF sites. The nucleosome-depleted region, which was previously believed to coincide with the CTCF binding site (10,39), is actually shifted 5′-upstream of CTCF motif (Figure 5). This nucleosome depletion is associated with AT-rich DNA sequence repeats which may disfavour nucleosome formation (57) and introduce bending of the double helix near CTCF (60,61). The effect of CTCF is modulated by its binding partner cohesin, which symmetrises the nucleosome arrays when it co-binds with CTCF (Figure 5D).

The asymmetric nucleosome-depleted regions near CTCF resemble the pattern observed near TSS, and the corresponding effect of NRL decrease as the gene activity increases (Supplementary Figure S14). Importantly, this effect is observed even for CTCF sites that are separated by more than 10,000 bp from the nearest annotated TSS (Supplementary Figure S6). Thus, the effects reported here are not directly related to gene transcription by Pol II. However, they may be linked to transcription of transposons such as Pol III-dependent SINE repeats. Several publications suggested an important role of transposons in the evolution of CTCF sites (62–66), and also it is known that mouse SINE B2 repeats can act as insulators (domain boundaries) *per se* (67). In addition, our data suggests that CTCF may play active role in transposon functioning as transcribed units separating nucleosome arrays. This is in line with recent reports about transcribed transposons associated with CTCF sites (68). Interestingly, previous publications reported that TFIIIC binds to RNA Pol III at tRNA genes and acts as a barrier against the spreading of heterochromatin (69) – this barrier function can be now re-interpreted in light of our results on the association of CTCF with Pol III as well as Pol II outside of gene promoters (Figure 5B). The importance of repetitive DNA sequences in the formation of chromatin boundaries near CTCF is further strengthened by the possibility of non-consensus TF binding in these regions (70). Unexpectedly, the effect of CTCF on the expression of RNA from adjacent locations is short-range repression, which becomes stronger as CTCF binding increases (Figure 6E).

We also showed that the asymmetry of the nucleosome signatures depends on the DNA-defined strength of CTCF binding and may be in addition determined by the CTCF/BORIS competition, because 'weak' CTCF binding
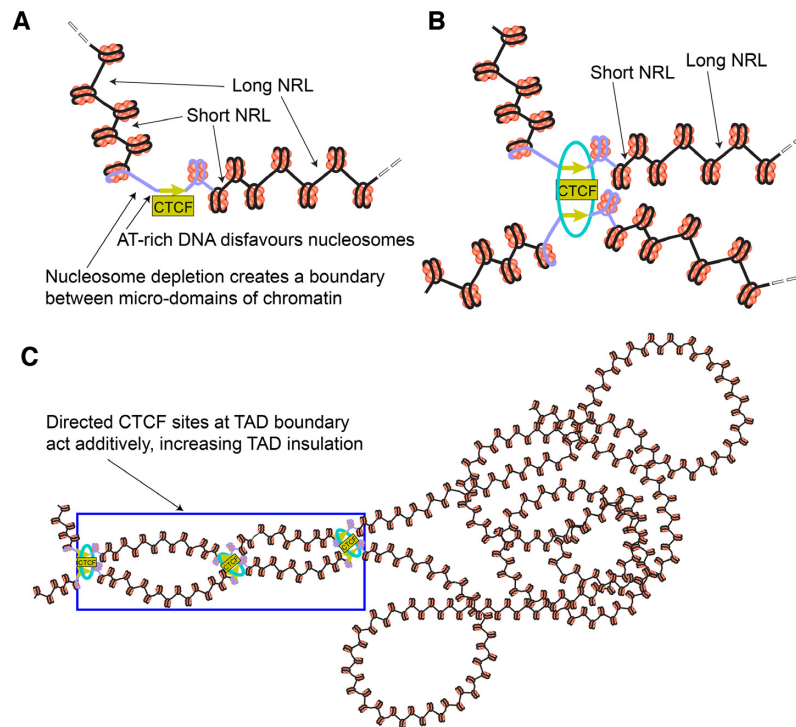
Figure 8. Schematic illustration of the effect of CTCF binding strength and motif orientation on the nucleosome arrangement in a single genomic region (**A**), at the base of a loop (**B**), and as part of a chromatin boundary containing several CTCF motifs (**C**). An extended DNA region including CTCF motif is enriched with repetitive sequences that define the mechanical properties of this region as a chromatin boundary (shown in violet colour)—see Figures 5A, 6A, D and Supplementary Figure S4. The region 5′-upstream of CTCF motif contains AT-rich sequences that disfavour nucleosome formation and may account for DNA bending in the complex with CTCF. Such regions can be due to DNA repeats such as SINEs, some of which are transcribed by Pol III that interact with CTCF. In analogy to the coding gene transcription the region 5′-upstream of the CTCF motif is depleted of the '-1' nucleosome. In the region 3′-downstream of CTCF motif chromatin remodellers including Chd4 and Snf2h determine the regularity of the nucleosome array. The nucleosomes located close to CTCF are separated by shorter linkers and nucleosomes further away from CTCF are separated by longer linkers, reaching the genome-average linker length at distances where CTCF effects disappear (corresponding to NRL change from ∼180 bp near strong CTCF sites to ∼190 bp genome-average, see Figure 3D). The cohesin ring is represented by the cyan ellipse. In the chromatin boundary containing several CTCF motifs, the effects described above for individual CTCFs may add up to increase chromatin domain insulation through the construction of special nucleosome array packing at the boundary, physically preventing interactions between adjacent TADs.

sites are enriched with the CTCFL recognition motif (Figure 6). BORIS has been previously proposed to interfere with CTCF binding (71), and our results further substantiate its role in the 'CTCF code' (43) that defines differential CTCF/BORIS binding.

Secondly, we found that the NRL decrease near CTCF is correlated with CTCF-DNA binding affinity (Figures 1D, F and 5F). This result goes significantly beyond previous observations that the CTCF binding strength is related to a more regular nucleosome ordering near its binding site (44,72) and may have direct functional implications. Strikingly, the variation of NRL as a function of CTCF binding affinity can be as large as ∼20 bp (the difference between NRL near the weakest CTCF-like motifs and the strongest CTCF-bound sites). Cohesin has a similar effect, but it is pronounced only when cohesin co-binds with CTCF. None of other DNA-binding proteins showed such behaviour (Figure 2). This uniqueness of CTCF can be explained by the large variability of its binding affinity through different combinations of its 11 zinc fingers that allows creating a 'CTCF code' (61,71,73). The effect of the NRL dependence on CTCF binding strength is most profound 3′-downstream of CTCF motifs, where it can be approximated

by a linear function (Figure 5F). This strong nucleosome patterning downstream but not upstream of CTCF is comparable to that of transcription start sites (TSSs) of protein-coding genes. In analogy, this effect could provide an additional argument that this may be linked to the transcription of non-coding repeats enclosing CTCF including Pol III-dependent SINEs.

Thus, our data suggests that the nucleosome arrangement near CTCF is defined by an active, remodeller-dependent process. Therefore, we analysed the contributions to this process by each of 8 chromatin remodellers that have been experimentally profiled in ESCs (Figure 4). We found that Snf2h has a major role in NRL decrease near CTCF, consistent with previous studies of Snf2H knockout in HeLa cells (74) and ESCs (40). In accord with the latter study, we observed that BRG1 has no detectable effect on NRL near CTCF, although it may be still involved in nucleosome positioning near TAD boundaries (75). Our investigation also identified Chd8 and EP400 as regulators of NRL near CTCF (Figure 4B, Supplementary Figure S10). These findings are consistent with the previous investigations that showed that Chd8 physically interacts with CTCF and knockdown of Chd8 abolishes the insulator activity of

CTCF sites required for IGF2 imprinting (76). One can hypothesise that this kind of insulator activity of CTCF is related to the boundary created by the nucleosome-free region 5′-upstream of the CTCF motif reported here, which may physically prevent the spreading of DNA methylation and other epigenetic modifications. Interestingly, our analysis revealed that the main chromatin remodeller responsible for the asymmetry of the nucleosome array near CTCF is Chd4. We show that Chd4 is both the top CTCF-associated remodeller (Figure 4A) and the sole remodeller responsible for the CTCF-dependent nucleosome occupancy peak 3′-downstream of the CTCF motif (Figure 5C). This finding may be important in the context of recent studies indicating that Chd4 is increasing the nucleosome density at regulatory regions (77).

The third major finding of this work concerns the effects of CTCF motif directionality and binding strength on nucleosome rearrangement during cell differentiation. Our calculations showed that the binding affinity is a good predictor for a given CTCF site being preserved upon cell differentiation (Figure 7A). This may be used as a foundation for the 'CTCF code' determining its differential binding as the cell progresses along the Waddington-type pathways. A specific subclass of common CTCF sites preserved upon cell differentiation tends to keep a small NRL, while the average NRL near all CTCF sites increases due to the active nucleosome repositioning 3′-downstream of CTCF motifs (Figure 7). A previous study reported a related distinction of common versus non-common CTCF sites based on the distance between the two nucleosomes downstream and upstream of CTCF (78). The preservation of NRL for common CTCF sites may give rise to a new effect where differential CTCF binding defines extended regions which do not change (or change minimally) their nucleosome positioning. Unexpectedly, the nucleosome-depleted region 5′-upstream of the CTCF motif remains even after CTCF depletion from a given site during differentiation. These nucleosome-depleted regions can have important functional roles, including the preservation of chromatin states while CTCF-dependent loops are dynamic and frequently break and reform throughout the cell cycle (79). For example, if the spreading of some chemical modifications of DNA or histones along the genomic coordinate requires enzymes cooperatively binding to the adjacent nucleosomes, then the consistent lack of a nucleosome at a given location can stop the propagation of the 'epigenetic wave'.

Finally, our finding of the asymmetry of CTCF-dependent chromatin boundaries at the scale of several nucleosomes may provide the missing mechanistic explanation for the asymmetry of chromatin boundaries at the scale of hundreds to thousands of nucleosomes reported recently (80,81). As we showed, TAD boundaries often contain several directed CTCF motifs (Figure 7F, Supplementary Figure S18). One can speculate that in this case the effects of individual CTCF sites accumulate, leading to the formation of a specific, asymmetric and 3D-structured nucleosome organisation at TAD boundary (schematically represented in Figure 8C). Such additivity of individual CTCF motifs could explain previous observations where the removal of part of the DNA sequence responsible for the boundary does not lead to the complete loss of TAD insulation (82).

In general, the asymmetric nucleosome organisation near CTCF reported here can be particularly interesting in light of the ongoing debate on the functional roles of chromatin boundaries in gene regulation.

DATA AVAILABILITY

Our software is available at https://github.com/chrisclarkson/NRLcalc

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Teif,V.B. and Clarkson,C.T. (2019) In: Ranganathan,S, Gribskov,M, Nakai,K and Schönbach,C (eds). *Encyclopedia of Bioinformatics and Computational Biology*. Academic Press, Oxford, pp. 308–317.
2. Baldi,S. (2019) Nucleosome positioning and spacing: from genome-wide maps to single arrays. *Essays Biochem.*, **63**, 5–14.
3. Lai,W.K.M. and Pugh,B.F. (2017) Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nat. Rev. Mol. Cell Biol.*, **18**, 548–562.
4. Maeshima,K., Ide,S. and Babokhov,M. (2019) Dynamic chromatin organization without the 30-nm fiber. *Curr. Opin. Cell Biol.*, **58**, 95–104.
5. Merkenschlager,M. and Nora,E.P. (2016) CTCF and cohesin in genome folding and transcriptional gene regulation. *Annu. Rev. Genomics Hum. Genet.*, **17**, 17–43.
6. Rao,S.S.P., Huang,S.-C., Glenn St Hilaire,B., Engreitz,J.M., Perez,E.M., Kieffer-Kwon,K.-R., Sanborn,A.L., Johnstone,S.E., Bascom,G.D., Bochkov,I.D. *et al.* (2017) Cohesin loss eliminates all loop domains. *Cell*, **171**, 305–320.
7. Nora,E.P., Goloborodko,A., Valton,A.L., Gibcus,J.H., Uebersohn,A., Abdennur,N., Dekker,J., Mirny,L.A. and Bruneau,B.G. (2017) Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell*, **169**, 930–944.
8. Fu,Y., Sinha,M., Peterson,C.L. and Weng,Z. (2008) The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLos Genet.*, **4**, e1000138.
9. Kelly,T.K., Liu,Y., Lay,F.D., Liang,G., Berman,B.P. and Jones,P.A. (2012) Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.*, **22**, 2497–2506.

10. Teif,V.B., Beshnova,D.A., Vainshtein,Y., Marth,C., Mallm,J.P., Höfer,T. and Rippe,K. (2014) Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development. *Genome Res.*, **24**, 1285–1295.

11. Wiehle,L., Thorn,G.J., Raddatz,G., Clarkson,C.T., Rippe,K., Lyko,F., Breiling,A. and Teif,V.B. (2019) DNA (de)methylation in embryonic stem cells controls CTCF-dependent chromatin boundaries. *Genome Res.*, **29**, 750–761.

12. Shen,Y., Yue,F., McCleary,D.F., Ye,Z., Edsall,L., Kuan,S., Wagner,U., Dixon,J., Lee,L., Lobanenkov,V.V. *et al.* (2012) A map of the cis-regulatory sequences in the mouse genome. *Nature*, **488**, 116–120.

13. Wang,H., Maurano,M.T., Qu,H., Varley,K.E., Gertz,J., Pauli,F., Lee,K., Canfield,T., Weaver,M., Sandstrom,R. *et al.* (2012) Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.*, **22**, 1680–1688.

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

15. Routh,A., Sandin,S. and Rhodes,D. (2008) Nucleosome repeat length and linker histone stoichiometry determine chromatin fiber structure. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 8872–8877.

16. Bass,M.V., Nikitina,T., Norouzi,D., Zhurkin,V.B. and Grigoryev,S.A. (2019) Nucleosome spacing periodically modulates nucleosome chain folding and DNA topology in circular nucleosome arrays. *J. Biol. Chem.*, **294**, 4233–4246.

17. Bascom,G.D., Kim,T. and Schlick,T. (2017) Kilobase pair chromatin fiber contacts promoted by living-system-like DNA linker length distributions and nucleosome depletion. *J. Phys. Chem. B*, **121**, 3882–3894.

18. Risca,V.I., Denny,S.K., Straight,A.F. and Greenleaf,W.J. (2017) Variable chromatin structure revealed by in situ spatially correlated DNA cleavage mapping. *Nature*, **541**, 237–241.

19. Nikitina,T., Norouzi,D., Grigoryev,S.A. and Zhurkin,V.B. (2017) DNA topology in chromatin is defined by nucleosome spacing. *Sci. Adv.*, **3**, e1700957.

20. Olins,A.L. and Olins,D.E. (1974) Spheroid chromatin units (v bodies). *Science*, **183**, 330–332.

21. Kornberg,R.D. (1974) Chromatin structure: a repeating unit of histones and DNA. *Science*, **184**, 868–871.

22. Lohr,D., Tatchell,K. and Van Holde,K.E. (1977) On the occurrence of nucleosome phasing in chromatin. *Cell*, **12**, 829–836.

23. Gottesfeld,J.M. and Melton,D.A. (1978) The length of nucleosome-associated DNA is the same in both transcribed and nontranscribed regions of chromatin. *Nature*, **273**, 317–319.

24. De Ambrosis,A., Ferrari,N., Bonassi,S. and Vidali,G. (1987) Nucleosomal repeat length in active and inactive genes. *FEBS Lett.*, **225**, 120–122.

25. Weintraub,H. (1978) The nucleosome repeat length increases during erythropoiesis in the chick. *Nucleic Acids Res.*, **5**, 1179–1188.

26. van Holde,K.E. (1989) *Chromatin*. Springer-Verlag, NY.

27. Valouev,A., Johnson,S.M., Boyd,S.D., Smith,C.L., Fire,A.Z. and Sidow,A. (2011) Determinants of nucleosome organization in primary human cells. *Nature*, **474**, 516–520.

28. Baldi,S., Krebs,S., Blum,H. and Becker,P.B. (2018) Genome-wide measurement of local nucleosome array regularity and spacing by nanopore sequencing. *Nat. Struct. Mol. Biol.*, **25**, 894–901.

29. Chereji,R.V., Ramachandran,S., Bryson,T.D. and Henikoff,S. (2018) Precise genome-wide mapping of single nucleosomes and linkers in vivo. *Genome Biol.*, **19**, 19.

30. Sun,F.-L., Cuaycong,M.H. and Elgin,S.C.R. (2001) Long-Range nucleosome ordering is associated with gene silencing in *Drosophila melanogaster* pericentric heterochromatin. *Mol. Cell. Biol.*, **21**, 2867–2879.

31. Zhang,Z., Wippo,C.J., Wal,M., Ward,E., Korber,P. and Pugh,B.F. (2011) A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science*, **332**, 977–980.

32. Hennig,B.P., Bendrin,K., Zhou,Y. and Fischer,T. (2012) Chd1 chromatin remodelers maintain nucleosome organization and repress cryptic transcription. *EMBO Rep.*, **13**, 997–1003.

33. Kubik,S., Challal,D., Bruzzone,M.J., Dreos,R., Mattarocci,S., Bucher,P., Libri,D. and Shore,D. (2019) Opposing chromatin remodelers control transcription initiation frequency and start site selection. *Nat. Struct. Mol. Biol.*, **26**, 744–754.

34. Ocampo,J., Chereji,R.V., Eriksson,P.R. and Clark,D.J. (2016) The ISW1 and CHD1 ATP-dependent chromatin remodelers compete to set nucleosome spacing in vivo. *Nucleic Acids Res.*, **44**, 4625–4635.

35. Mobius,W., Osberg,B., Tsankov,A.M., Rando,O.J. and Gerland,U. (2013) Toward a unified physical model of nucleosome patterns flanking transcription start sites. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 5719–5724.

36. de Dieuleveult,M., Yen,K., Hmitou,I., Depaux,A., Boussouar,F., Bou Dargham,D., Jounier,S., Humbertclaude,H., Ribierre,F., Baulard,C. *et al.* (2016) Genome-wide nucleosome specificity and function of chromatin remodellers in ES cells. *Nature*, **530**, 113–116.

37. Giles,K.A., Gould,C.M., Du,Q., Skvortsova,K., Song,J.Z., Maddugoda,M.P., Achinger-Kawecka,J., Stirzaker,C., Clark,S.J. and Taberlay,P.C. (2019) Integrated epigenomic analysis stratifies chromatin remodellers into distinct functional groups. *Epigenet. Chromatin*, **12**, 12.

38. Teif,V.B., Vainshtein,Y., Caudron-Herger,M., Mallm,J.P., Marth,C., Höfer,T. and Rippe,K. (2012) Genome-wide nucleosome positioning during embryonic stem cell development. *Nat. Struct. Mol. Biol.*, **19**, 1185–1192.

39. Beshnova,D.A., Cherstvy,A.G., Vainshtein,Y. and Teif,V.B. (2014) Regulation of the nucleosome repeat length in vivo by the DNA sequence, protein concentrations and long-range interactions. *PLoS Comput. Biol.*, **10**, e1003698.

40. Barisic,D., Stadler,M.B., Iurlaro,M. and Schübeler,D. (2019) Mammalian ISWI and SWI/SNF selectively mediate binding of distinct transcription factors. *Nature*, **569**, 136–140.

41. Jenkinson,G., Pujadas,E., Goutsias,J. and Feinberg,A.P. (2017) Potential energy landscapes identify the information-theoretic nature of the epigenome. *Nat. Genet.*, **49**, 719–729.

42. Voong,L.N., Xi,L., Sebeson,A.C., Xiong,B., Wang,J.P. and Wang,X. (2016) Insights into nucleosome organization in mouse embryonic stem cells through chemical mapping. *Cell*, **167**, 1555–1570.

43. Bonev,B., Mendelson Cohen,N., Szabo,Q., Fritsch,L., Papadopoulos,G.L., Lubling,Y., Xu,X., Lv,X., Hugnot,J.P., Tanay,A. *et al.* (2017) Multiscale 3D genome rewiring during mouse neural development. *Cell*, **171**, 557–572.

44. Vainshtein,Y., Rippe,K. and Teif,V.B. (2017) NucTools: analysis of chromatin feature occupancy profiles from high-throughput sequencing data. *BMC Genomics*, **18**, 158.

45. Quinlan,A.R. (2014) BEDTools: The swiss-army tool for genome feature analysis. *Curr Protoc Bioinformatics*, **47**, 11.12.1–11.12.34.

46. Khan,A., Fornes,O., Stigliani,A., Gheorghe,M., Castro-Mondragon,J.A., van der Lee,R., Bessy,A., Cheneby,J., Kulkarni,S.R., Tan,G. *et al.* (2018) JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.*, **46**, D260–D266.

47. Tan,G. and Lenhard,B. (2016) TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics*, **32**, 1555–1556.

48. Lawrence,M., Huber,W., Pages,H., Aboyoun,P., Carlson,M., Gentleman,R., Morgan,M.T. and Carey,V.J. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.

49. Castro-Mondragon,J.A., Jaeger,S., Thieffry,D., Thomas-Chollier,M. and van Helden,J. (2017) RSAT matrix-clustering: dynamic exploration and redundancy reduction of transcription factor binding motif collections. *Nucleic Acids Res.*, **45**, e119.

50. Roider,H.G., Kanhere,A., Manke,T. and Vingron,M. (2007) Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, **23**, 134–141.

51. Martin,D., Pantoja,C., Fernandez Minan,A., Valdes-Quezada,C., Molto,E., Matesanz,F., Bogdanovic,O., de la Calle-Mustienes,E., Dominguez,O., Taher,L. *et al.* (2011) Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes. *Nat. Struct. Mol. Biol.*, **18**, 708–714.

52. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.

53. Neph,S., Reynolds,A.P., Kuehn,M.S. and Stamatoyannopoulos,J.A. (2016) Operating on genomic ranges using BEDOPS. *Methods Mol. Biol.*, **1418**, 267–281.

54. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.

55. Fudenberg,G., Imakaev,M., Lu,C., Goloborodko,A., Abdennur,N. and Mirny,L.A. (2016) Formation of chromosomal domains by loop extrusion. *Cell Rep.*, **15**, 2038–2049.

56. Pavlaki,I., Docquier,F., Chernukhin,I., Kita,G., Gretton,S., Clarkson,C.T., Teif,V.B. and Klenova,E. (2018) Poly(ADP-ribosyl)ation associated changes in CTCF-chromatin binding and gene expression in breast cells. *Biochim. Biophys. Acta Gene Regul. Mech.*, **1861**, 718–730.

57. Segal,E. and Widom,J. (2009) Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.*, **19**, 65–71.

58. Carriere,L., Graziani,S., Alibert,O., Ghavi-Helm,Y., Boussouar,F., Humbertclaude,H., Jounier,S., Aude,J.C., Keime,C., Murvai,J. *et al.* (2012) Genomic binding of Pol III transcription machinery and relationship with TFIIS transcription factor distribution in mouse embryonic stem cells. *Nucleic Acids Res.*, **40**, 270–283.

59. Kundaje,A., Kyriazopoulou-Panagiotopoulou,S., Libbrecht,M., Smith,C.L., Raha,D., Winters,E.E., Johnson,S.M., Snyder,M., Batzoglou,S. and Sidow,A. (2012) Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Res.*, **22**, 1735–1747.

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

61. Nichols,M.H. and Corces,V.G. (2015) A CTCF code for 3D genome architecture. *Cell*, **162**, 703–705.

62. Zhang,Y., Li,T., Preissl,S., Amaral,M.L., Grinstein,J.D., Farah,E.N., Destici,E., Qiu,Y., Hu,R., Lee,A.Y. *et al.* (2019) Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells . *Nature Genetics*, **51**, 1380–1388.

63. Choudhary,M.N., Friedman,R.Z., Wang,J.T., Jang,H.S., Zhuo,X. and Wang,T. (2018) Co-opted transposons help perpetuate conserved higher-order chromosomal structures. BioRxiv doi: https://doi.org/10.1101/485342, 05 December 2018, preprint: not peer reviewed.

64. Kentepozidou,E., Aitken,S.J., Feig,C., Stefflova,K., Ibarra-Soria,X., Odom,D.T., Roller,M. and Flicek,P. (2019) Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. BioRxiv doi: https://doi.org/10.1101/668855, 12 June 2019, preprint: not peer reviewed.

65. Schmidt,D., Schwalie,P.C., Wilson,M.D., Ballester,B., Goncalves,A., Kutter,C., Brown,G.D., Marshall,A., Flicek,P. and Odom,D.T. (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell*, **148**, 335–348.

66. Bourque,G., Leong,B., Vega,V.B., Chen,X., Lee,Y.L., Srinivasan,K.G., Chew,J.L., Ruan,Y., Wei,C.L., Ng,H.H. *et al.* (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.*, **18**, 1752–1762.

67. Lunyak,V.V., Prefontaine,G.G., Nunez,E., Cramer,T., Ju,B.G., Ohgi,K.A., Hutt,K., Roy,R., Garcia-Diaz,A., Zhu,X. *et al.* (2007) Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science*, **317**, 248–251.

68. Zhang,Y., Li,T., Preissl,S., Amaral,M.L., Grinstein,J.D., Farah,E.N., Destici,E., Qiu,Y., Hu,R., Lee,A.Y. *et al.* (2019) Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat. Genet.*, **51**, 1380–1388.

69. Simms,T.A., Dugas,S.L., Gremillion,J.C., Ibos,M.E., Dandurand,M.N., Toliver,T.T., Edwards,D.J. and Donze,D. (2008) TFIIIC binding sites function as both heterochromatin barriers and chromatin insulators in Saccharomyces cerevisiae. *Eukaryot. Cell.*, **7**, 2078–2086.

70. Afek,A., Schipper,J.L., Horton,J., Gordan,R. and Lukatsky,D.B. (2014) Protein-DNA binding in the absence of specific base-pair recognition. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 17140–17145.

71. Lobanenkov,V.V. and Zentner,G.E. (2018) Discovering a binary CTCF code with a little help from BORIS. *Nucleus*, **9**, 33–41.

72. Owens,N., Papadopoulou,T., Festuccia,N., Tachtsidi,A., Gonzalez,I., Dubois,A., Vandormael-Pournin,S., Nora,E.P., Bruneau,B.G., Cohen-Tannoudji,M. *et al.* (2019) CTCF confers local nucleosome resiliency after DNA replication and during mitosis. *eLife*, **8**, e47898.

73. Fang,R., Wang,C., Skogerbo,G. and Zhang,Z. (2015) Functional diversity of CTCFs is encoded in their binding motifs. *BMC Genomics*, **16**, 649.

74. Wiechens,N., Singh,V., Gkikopoulos,T., Schofield,P., Rocha,S. and Owen-Hughes,T. (2016) The chromatin remodelling enzymes SNF2H and SNF2L position nucleosomes adjacent to CTCF and other transcription factors. *PLos Genet.*, **12**, e1005940.

75. Barutcu,A.R., Lian,J.B., Stein,J.L., Stein,G.S. and Imbalzano,A.N. (2017) The connection between BRG1, CTCF and topoisomerases at TAD boundaries. *Nucleus*, **8**, 150–155.

76. Ishihara,K., Oshimura,M. and Nakao,M. (2006) CTCF-Dependent chromatin insulator is linked to epigenetic remodeling. *Mol. Cell*, **23**, 733–742.

77. Bornelov,S., Reynolds,N., Xenophontos,M., Gharbi,S., Johnstone,E., Floyd,R., Ralser,M., Signolet,J., Loos,R., Dietmann,S. *et al.* (2018) The nucleosome remodeling and deacetylation complex modulates chromatin structure at sites of active transcription to fine-tune gene expression. *Mol. Cell*, **71**, 56–72.

78. Snyder,M.W., Kircher,M., Hill,A.J., Daza,R.M. and Shendure,J. (2016) Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell*, **164**, 57–68.

79. Hansen,A.S., Pustova,I., Cattoglio,C., Tjian,R. and Darzacq,X. (2017) CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *Elife*, **6**, e25776.

80. Barrington,C., Georgopoulou,D., Pezic,D., Varsally,W., Herrero,J. and Hadjur,S. (2019) Enhancer accessibility and CTCF occupancy underlie asymmetric TAD architecture and cell type specific genome topology. *Nat. Commun.*, **10**, 2908.

81. Nanni,L., Wang,C., Manders,F., Groh,L., Haro,P., Oldenkamp,R., Ceri,S. and Logie,C. (2019) The CTCF anatomy of topologically associating domains. BioRxiv doi: https://doi.org/10.1101/746610, 28 August 2019, preprint: not peer reviewed.

82. Barutcu,A.R., Maass,P.G., Lewandowski,J.P., Weiner,C.L. and Rinn,J.L. (2018) A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. *Nature Communications*, **9**, 1444.

83. Sun,F., Chronis,C., Kronenberg,M., Chen,X.F., Su,T., Lay,F.D., Plath,K., Kurdistani,S.K. and Carey,M.F. (2019) Promoter-enhancer communication occurs primarily within insulated neighborhoods. *Mol. Cell*, **73**, 250–263.

84. Ho,L., Jothi,R., Ronan,J.L., Cui,K., Zhao,K. and Crabtree,G.R. (2009) An embryonic stem cell chromatin remodeling complex, esBAF, is an essential component of the core pluripotency transcriptional network. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 5187–5191.