CTCF binding strength modulates chromatin architecture through the changes of the local nucleosome repeat length

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

[1] School of Biological 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. Current address: Department of Biological and Medical Sciences, Oxford Brookes University, Headington Campus, Oxford, OX3 0BP, UK

[4] Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA

* Correspondence should be addressed to Vladimir Teif, E-mail: vteif@essex.ac.uk

Running title: CTCF-NRL interplay

Key words: nucleosome repeat length, CTCF, mouse embryonic stem cells, chromatin remodellers

Abstract

Nucleosome repeat length (NRL) defines the average distance between adjacent nucleosomes. When calculated for specific genomic regions, NRL reflects the local nucleosome ordering and characterises its changes during developmental processes. The architectural protein CTCF provides one of the strongest nucleosome positioning signals, setting a decreased NRL for ~20 nucleosomes in its vicinity (thus affecting up to 10% of the mouse genome). We show that upon differentiation of mouse embryonic stem cells (ESCs) to neural progenitor cells and mouse embryonic fibroblasts, a subset of common CTCF sites preserved in all three cell types keeps small NRL despite genome-wide NRL increase. This suggests that differential CTCF binding not only affects 3D genome organisation but also defines genomic regions with conserved nucleosome arrangement. Our analysis revealed that NRL decrease near CTCF is correlated with CTCF affinity for DNA binding. Stronger CTCF binding is linked to increased probability to form chromatin loops and more efficient recruitment of chromatin remodellers. We show that the effect of individual remodellers on decreasing the NRL near CTCF is increasing in the order Brg1≤Chd4<Chd6<Chd1≤Chd2≤EP400≤Chd8<Snf2h.

Introduction

Nucleosomes are positioned along the genome in a non-random way (Baldi, 2019; Lai and Pugh, 2017; Teif and Clarkson, 2019), which is critical for determining the DNA accessibility and genome organisation (Maeshima et al., 2019). A classical parameter characterising the nucleosome spacing is 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 fibers (Bascom et al., 2017; Bass et al., 2019; Nikitina et al., 2017; Risca et al., 2017; Routh et al., 2008).

Ever since the discovery of the nucleosome (Kornberg, 1974; Olins and Olins, 1974) there have been many attempts to compare NRLs of different genomic regions (De Ambrosis et al., 1987; Gottesfeld and Melton, 1978; Lohr et al., 1977) and it has been established that genome-wide NRL changes during cell differentiation (van Holde, 1989; Weintraub, 1978). Recent sequencing-based investigations showed that active regions such as promoters, enhancers and actively transcribed genes usually have shorter NRLs while heterochromatin is characterised by longer NRLs (Baldi et al., 2018; Chereji et al., 2018; Sun et al., 2001; Valouev et al., 2011). Studies performed in Yeast linked NRL changes at transcription start sites (TSS) to a number of specific molecular mechanisms, down to individual chromatin remodellers responsible for increasing/decreasing NRL (Celona et al., 2011; Hennig et al., 2012; Kubik et al., 2019; Mobius et al., 2013; Ocampo et al., 2016; Zhang et al., 2011). However, in higher eukaryotes regulatory regions are very heterogeneous, and although several recent attempts have been made (de Dieuleveult et al., 2016; Giles et al., 2019), it is difficult to come up with a set of definitive remodeller rules determining their effect on NRL. For example, ubiquitous heterogeneity and asymmetry of nucleosome distributions around subsets of different TF binding sites has been noted (Kundaje et al., 2012).

A particularly important nucleosome positioning signal is provided by CTCF, an architectural protein that maintains 3D genome architecture (Merkenschlager and Nora, 2016; Nora et al., 2017; Rao et al., 2017) and can organise up to 20 nucleosomes in its vicinity (Fu et al., 2008) (Fig. 1A). CTCF has hundreds of thousands of potential binding sites in the mouse genome. Usually there are ~30,000-60,000 of CTCF sites bound in a given cell type, which translates to about 1 million of affected nucleosomes (Chen et al., 2012; Shen et al., 2012; Wang et al., 2012; Wiehle et al., 2019).

We previously showed that in mouse embryonic stem cells (ESC), NRL near CTCF is about 10 bp smaller than genome-wide NRL (Teif et al., 2014; Teif et al., 2012). Our analysis demonstrated that purely statistical positioning of nucleosomes near CTCF boundaries would result in a longer NRL than observed experimentally, and the effects of strong nucleosome-positioning DNA sequences, while compatible with the observed NRL, are limited to a small number of CTCF sites (Beshnova et al., 2014). A very recent study has investigated the effect of Snf2 and Brg1 remodellers on NRL in ESCs, suggesting Snf2 as the primary player (Barisic et al., 2019). However, other factors may be at play as well. Thus, the question of what determines the NRL near CTCF remains open, as well as the question of the functional consequences of such small NRLs. Here we will address both these problems in a systematic manner using all available datasets in ESCs.

Results

Location of genomic region with respect to CTCF sites has profound effect on its apparent NRL. Our analysis is based on the “phasogram” type of NRL calculation introduced previously (Teif et al., 2012; Valouev et al., 2011). The idea of this method is to consider all mapped nucleosome reads within the genomic region of interest and calculate the distribution of the distances between nucleosome dyads. This distribution typically shows peaks corresponding to the prevailing 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. To perform bulk calculations of NRLs for many genomic subsets of interest we developed software NRLcalc, which loads the phasograms computed in NucTools and performs linear fitting to calculate NRL (see Methods).

We first noticed that NRL near CTCF depends critically on the distance of the region of NRL calculation to the binding site summit (Fig. 1B). While the phasograms for regions [100, 2000] and [250, 1000], which are both excluding the CTCF site, are quite similar to each other, a region that includes CTCF [-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 from 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 shown by arrows in Fig. 1B. 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 prevailing distance between nucleosomes (Fig. 1C). We thus selected the region [100, 2000] for the following calculations. Below, all NRLs refer to regions [100, 2000] near the summit of TF binding site, unless specified otherwise. 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 (Fig. S1). In the following calculations we used the high-coverage MNase-seq and chemical mapping datasets from (Voong et al., 2016).

In order to investigate the effect of CTCF on NRLs near binding sites of other proteins, we calculated NRLs near binding sites of 18 stemness-related TFs whose binding has been experimentally determined in ESCs using ChIP-seq (Fig. 1D and Fig. S2). The latter analysis revealed that the proximity to CTCF binding sites changes all of these NRLs. When we filtered out TF binding sites that overlap with CTCF, the NRLs for each individual TF increased (Fig. 1D). On the other hand, TF binding sites that overlap with CTCF had significantly smaller NRLs (Fig. S2). Thus, CTCF's effects on NRL are unique, which warrants focusing on CTCF alone for the rest of our study.

The stronger CTCF binds to DNA the smaller is NRL near its binding sites. In order to investigate the effect of CTCF on NRL, we split CTCF sites into 5 quintiles based on the height of their ChIP-seq peaks reported previously (Shen et al., 2012). Comparison of CTCF quintiles in terms of the distribution of nucleosome dyad-to-dyad distances determined by chemical mapping revealed that stronger CTCF binding is associated with smaller NRLs. NRL profiles also changed from one dominant peak in the case of weak CTCF binding to several pronounced peaks in case of the strongest CTCF binding quintile (Fig. 2A and Fig. S3). The calculation of the "classical" NRLs based on MNase-seq data showed a smooth decrease of NRL as the strength of CTCF binding increased (Fig. 2B). We confirmed that this relation is determined by the strength of CTCF binding *per se* by repeating this calculation for all computationally predicted CTCF sites in the mouse genome which were split into quintiles based on the similarity of their motifs to the canonical CTCF motif (Fig. 2B).

Using the same procedure we have also calculated the NRL as a function of the binding strength for all TFs in the mouse genome whose position weight matrices are available in JASPAR2018 (Khan et al., 2018). This analysis revealed that for proteins other that CTCF NRL did not reveal a smooth function of their binding strength (see Fig. S4). Thus, CTCF is a unique protein that shows anticorrelation between the strength of its DNA binding and NRL

Common CTCF sites preserve local nucleosome organisation during ESC differentiation. Then we set to determine the functional consequences of the NRL decrease near CTCF. We investigated the change of NRL near CTCF upon differentiation of ESCs to neural progenitor cells (NPSs) and mouse embryonic fibroblasts (MEFs). We first noted that the stronger CTCF binds to DNA the higher the probability is that this site will remain bound upon differentiation (Fig. 2C). This suggests that the strength of CTCF binding can act as the major factor determining which CTCF sites retain and which are lost upon differentiation (and thus how the 3D structure of the genome will change). In relation to NRL, we showed that NRL near bound CTCF on average increases as the cell differentiates (Fig. 2D and S5). Importantly, common CTCF sites resisted this NRL change, suggesting that CTCF retention upon differentiation at common sites preserves both 3D structure and nucleosome patterns at these loci.

What determines the NRL decrease near CTCF? In order to define the physical mechanisms of NRL decrease near CTCF one has to consider a number of genomic features and molecular factors that potentially can account for the NRL decrease near CTCF:

1) Our previous observations suggested that the strength of CTCF binding is related to the surrounding GC and CpG content (Pavlaki et al., 2018; Wiehle et al., 2019).

Our new calculations performed here show that the strength of CTCF binding is indeed correlated with GC content around CTCF sites (Fig. 3A), as well as the probability that a given site is located in a CpG island (Fig. 3B). Therefore, we one potential hypothesis to check is whether CTCF site location inside vs outside CpG islands has an effect on NRL.

2) Small NRL near CTCF could be simply because CTCF sites are in active regions (promoters or enhancers) which have smaller NRL in comparison with genome-average based on previous studies (Baldi et al., 2018; Valouev et al., 2011). 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 (Fig. 3C).

3) The NRL could depend on whether a given CTCF site forms a boundary of topologically associated domains (TADs) or enhancer-promoter loops. Our analysis using recently published coordinates of TADs and chromatin loops in ESCs (Bonev et al., 2017) showed that there is a positive 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 (Fig. 3C).

4) Nucleosome arrangement could be determined by a specific chromatin remodeller interacting with CTCF. We have processed all available remodeller ChIP-seq datasets in ESCs and plotted the percentage of CTCF sites overlapping with remodeller ChIP-seq peaks (Fig. 3D). 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.

We set to check all four hypotheses formulated above (Fig. 4). CTCF site location inside boundaries of loops or TADs was indeed associated with NRL decrease, which was even more pronounced in CpG islands. We have also derived a systematic rules of remodeller effects on NRL near CTCF, with Brg1 having no detectable effect (based on two independent Brg1 datasets), and Snf2h having the largest effect. The effect of other remodellers is increasing in the order BRG1≤Chd4<Chd6<Chd1≤Chd2≤EP400≤Chd8<Snf2h.

Discussion

We developed a new methodology for quantitative investigations of local NRL changes, and its application revealed a number of interesting observations:

First, we found that NRL critically depends on the distance of the selected genomic region to the summit of the CTCF site. We showed that the CTCF site needs to be 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 (Figure S6), but previous NRL calculations for CTCF-containing regions may need to be re-evaluated.

Second, we found that the NRL decrease near CTCF is correlated with CTCF-DNA binding affinity. This result goes significantly beyond previous observations that stronger CTCF binding is associated with more regular nucleosome ordering near its binding site (Owens et al., 2019; Vainshtein et al., 2017) and may have direct functional implications. Strikingly, the NRL decrease as a function of CTCF binding affinity spans a large interval from 193 bp for weak CTCF-like DNA motifs down to 178 bp for the strongest sites bound in ESCs. None of other DNA-binding proteins showed such behaviour. This uniqueness of CTCF can be explained by the large variability of its binding affinity through different combination of its 11 zinc fingers that allows creating a “CTCF code” (Lobanenkov and Zentner, 2018; Nichols and Corces, 2015).

Third, our calculations showed that the strength of CTCF binding acts as a good predictor of a given CTCF site being preserved upon cell differentiation (which may be used as a foundation for the CTCF code determining its differential binding as the cell progresses along the Waddington-type pathways). Importantly, a subclass of common CTCF sites preserved upon cell differentiation tends to keep a small NRL, while genome-wide NRL increases. 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 (Snyder et al., 2016). 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.

Fourth, we systematised the contributions to NRL decrease determined by each of 8 chromatin remodellers that have been profiled in ESCs (Fig 4B). Our analysis suggests that Snf2h has a major role in this phenomenon, consistent with previous studies of Snf2H knockout in HeLa cells (Wiechens et al., 2016) and ESCs (Barisic et al., 2019). Consistently with the latter study, we found that BRG1 has no detectable effect on NRL near CTCF, although it may be still involved in nucleosome positioning near TAD boundaries (Barutcu et al., 2017). Our investigation also identified Chd8 and EP400 as two novel major players. Previous studies indeed showed that Chd8 physically interacts with CTCF and knockdown of Chd8 abolishes the insulator activity of CTCF sites required for *IGF2* imprinting (Ishihara et al., 2006). Thus, our work revealed a systematic set of remodeller effects on NRL near CTCF and provided the basis for future quantitative investigations of local NRL variations during development.

Materials and Methods

Experimental datasets. Nucleosome positioning and transcription factor binding datasets were obtained from the Gene Expression Omnibus (GEO), Short Read Archive (SRA) and the ENCODE web site as detailed in Table ST1. NRL calculations near CTCF in ESCs were performed using the MNase-seq dataset from (Voong et al., 2016). NRL calculations near 19 stemness-related proteins in ESCs shown in Figure 1D and S1 were performed using the chemical mapping dataset from (Voong et al., 2016). NRL calculations in NPCs and MEFs were based on the MNase-seq datasets from (Teif et al., 2012). MNase-assisted H3 ChIP-seq from (Wiehle et al., 2019) was used for demonstrative purposes in the phasogram calculation in Figure 1C. Coordinates of genomic features and experimental maps of transcription factor and remodeller binding in ESCs were obtained from published sources as detailed in Table S1. The coordinates of loops and TADs described in (Bonev et al., 2017) were provided by the authors in a BED file aligned to the mm10 mouse genome and were 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 2 mismatches. For all other datasets we used processed files with genomic coordinates downloaded from the corresponding database as detailed in 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 center 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 (Vainshtein et al., 2017) were intersected with the corresponding genomic regions of interest using BEDTools (Quinlan, 2014). Average nucleosome occupancy profiles were calculated using NucTols. The phasograms were calculated using NucTools as detailed below.

Binding site prediction. Computationally predicted TF binding sites were determined via scanning the mouse genome with position frequency matrices (PFMs) from the JASPAR2018 database (Khan et al., 2018) using R packages TFBSTools (Tan and Lenhard, 2016) and GenomicRanges (Lawrence et al., 2013). A similarity threshold of 80% was used for all TFs in order to get at least several thousand putative binding sites.

Stratification of TF-DNA binding affinity. In the case of experimentally determined binding sites of CTCF we stratified these into five equally sized quintiles according to the ChIP-seq peak height determined via peak calling performed in the original publication (Shen et al., 2012). In the case of the predicted TF sites, we used the TRAP algorithm (Roider et al., 2007) to predict the affinity of TF for each site. The same operation as described above was performed on these sites, with the sites arranged into quintiles according to the TRAP score.

Phasogram calculation. The “phasograms” representing the histograms of dyad-to-dyad or start-to-start distances were calculated with the NucTools script nucleosome_repeat_length.pl. When paired-end MNase-seq was used, dyad-to-dyad distances were calculated using the center of each read as described previously (Vainshtein et al., 2017). 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 locations, so the DNA fragments span from dyad to dyad.

Automated NRL determination from phasograms. Studying many phasograms proved cumbersome when manually picking the points in a non-automated way. To circumvent this problem, an interactive applet called *NRLcalc* was developed 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. The app allows one to select a smoothing window size to minimise noise in the phasograms. A smoothing window of 20 bp was used in our calculations. The app 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.

Acknowledgements

We thank Boyan Bonev and Giacomo Cavalli for providing the coordinates of chromatin loops and TADs, Stuart Newman for the computer cluster support and Yevhen Vainshtein for the NucTools support and fruitful discussions.

Competing interests

No competing interests declared

Funding

This work was funded by the Wellcome Trust grants 200733/Z/16/Z and 211967/Z/18/Z and by the Frontrunner program of the University of Essex.

Data availability

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

Figure Legends

Figure 1. Setting the methodology for NRL calculation. A) Average nucleosome profile around CTCF binding sites in ESCs. B) Phasograms depicting the normalised frequency of nucleosome dyad-to-dyad distances calculated using NucTools for three different regions near CTCF sites: [100, 2000], [100, 1000] and [-500, 500]. Both [100, 2000] and [100, 1000] patterns oscillate with NRL=174. In the case of the [-500, 500] phasogram additional peaks (indicated by red arrows) appear which correspond to distances between nucleosomes on different sides of CTCF, thus resulting in an apparent NRL<160. C) NRLs calculated from the phasograms shown in panel (B). Region [-500, 500] is characterised by an unrealistically small NRL=155bp which is an artefact of the interference of two waves of distances between nucleosomes located on different sides from CTCF. D) 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. Left: all TF binding sites; right: TF binding sites which do not intersect with CTCF.

Figure 2. CTCF binding strength determines NRL decrease, which has functional implications during differentiation. A) Chemical mapping reveals the fine structure of the NRL distribution for different CTCF quintiles defined based on CTCF binding strength (the height of ChIP-seq CTCF peaks). B) Dependence of NRL on the strength of CTCF binding based on experimental ChIP-seq peaks (black line) and computationally predicted CTCF sites (red line). C) The stronger CTCF binding in ESC the higher the probability that a given CTCF site will be retained upon differentiation to NPCs and MEFs. D) Comparison of NRLs near CTCF during ESC differentiation. Upon differentiation average NRL near CTCF increases (denoted “All”), but common CTCF sites keep the smallest NRL (denoted “Comm”).

Figure 3. What determines the NRL decrease near CTCF? A) CTCF binding sites split into quintiles based on their binding strength are characterised 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) The stronger CTCF binds the higher the

probability that it is co-enriched with different chromatin remodellers indicated on the figure.

Figure 4. The summary of the effects on the value of NRL near CTCF. A) 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 all remodeller peaks; outside of promoters and enhancers; common CTCF sites in ESCs, NPCs and MEFs. The top horizontal dashed line corresponds to the weak CTCF-like motifs from Figure 2D. B) NRLs calculated for CTCF sites that overlap (black) and do not over (red) with chromatin remodeller peaks for eight remodellers experimentally mapped in ESCs. Two Brg1 datasets are denoted as 2009 (Ho et al., 2009) and 2016 (de Dieuleveult et al., 2016). Vertical bars show the standard deviation.

References

Aksoy, I., Jauch, R., Chen, J., Dyla, M., Divakar, U., Bogu, G. K., Teo, R., Leng Ng, C. K., Herath, W., Lili, S., et al. (2013). Oct4 switches partnering from Sox2 to Sox17 to reinterpret the enhancer code and specify endoderm. *EMBO J* **32**, 938-953.

Baldi, S. (2019). Nucleosome positioning and spacing: from genome-wide maps to single arrays. *Essays In Biochemistry*, EBC20180058.

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.

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

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.

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.

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.

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.

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 e524.

Celona, B., Weiner, A., Di Felice, F., Mancuso, F. M., Cesarini, E., Rossi, R. L., Gregory, L., Baban, D., Rossetti, G., Grianti, P., et al. (2011). Substantial histone reduction modulates genomewide nucleosomal occupancy and global transcriptional output. *PLoS Biol* **9**, e1001086.

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.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., et al. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-1117.

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.

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J. and Plath, K. (2017). Cooperative Binding of Transcription Factors Orchestrates Reprogramming. *Cell* **168**, 442-459 e420.

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

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.

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 Genetics* **4**, e1000138.

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. *Epigenetics Chromatin* **12**, 12.

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.

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.

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.

Irizarry, R. A., Wu, H. and Feinberg, A. P. (2009). A species-generalized probabilistic model-based definition of CpG islands. *Mamm Genome* **20**, 674-680.

Ishihara, K., Oshimura, M. and Nakao, M. (2006). CTCF-Dependent Chromatin Insulator Is Linked to Epigenetic Remodeling. *Molecular Cell* **23**, 733-742.

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**, D1284.

Kim, K. Y., Tanaka, Y., Su, J., Cakir, B., Xiang, Y., Patterson, B., Ding, J., Jung, Y. W., Kim, J. H., Hysolli, E., et al. (2018). Uhrf1 regulates active transcriptional marks at bivalent domains in pluripotent stem cells through Setd1a. *Nature communications* **9**, 2583.

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

Krepelova, A., Neri, F., Maldotti, M., Rapelli, S. and Oliviero, S. (2014). Myc and max genome-wide binding sites analysis links the Myc regulatory network with the polycomb and the core pluripotency networks in mouse embryonic stem cells. *PLoS One* **9**, e88933.

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. *bioRxiv*, 592816.

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.

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.

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.

Liu, J., Han, Q., Peng, T., Peng, M., Wei, B., Li, D., Wang, X., Yu, S., Yang, J., Cao, S., et al. (2015). The oncogene c-Jun impedes somatic cell reprogramming. *Nat Cell Biol* **17**, 856-867.

Lizio, M., Harshbarger, J., Shimoji, H., Severin, J., Kasukawa, T., Sahin, S., Abugessaisa, I., Fukuda, S., Hori, F., Ishikawa-Kato, S., et al. (2015). Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* **16**, 22.

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

Local, A., Huang, H., Albuquerque, C. P., Singh, N., Lee, A. Y., Wang, W., Wang, C., Hsia, J. E., Shiau, A. K., Ge, K., et al. (2018). Identification of H3K4me1-associated proteins at mammalian enhancers. *Nat Genet* **50**, 73-82.

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

Maeshima, K., Ide, S. and Babokhov, M. (2019). Dynamic chromatin organization without the 30-nm fiber. *Current Opinion in Cell Biology* **58**, 95-104.

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.

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.

Nichols, M. H. and Corces, V. G. (2015). A CTCF Code for 3D Genome Architecture. *Cell* **162**, 703-705.

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

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.

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.

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

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. *bioRxiv*, 563619.

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.

Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., et al. (2014). RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* **42**, D756-763.

Quinlan, A. R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47**, 11 12 11-34.

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.e324.

Rhee, C., Lee, B.-K., Beck, S., Anjum, A., Cook, K. R., Popowski, M., Tucker, H. O. and Kim, J. (2014). Arid3a is essential to execution of the first cell fate decision via direct embryonic and extraembryonic transcriptional regulation. *Genes & Development* **28**, 2219-2232.

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.

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.

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.

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.

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.

Sun, F.-L., Cuaycong, M. H. and Elgin, S. C. R. (2001). Long-Range Nucleosome Ordering Is Associated with Gene Silencing in <em>Drosophila melanogaster</em> Pericentric Heterochromatin. *Molecular and Cellular Biology* **21**, 2867-2879.

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

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.

Teif, V. B. and Clarkson, C. T. (2019). Nucleosome Positioning. In *Encyclopedia of Bioinformatics and Computational Biology* (ed. S. Ranganathan, M. Gribskov, K. Nakai & C. Schönbach), pp. 308-317. Oxford: Academic Press.

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.

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.

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.

van Holde, K. E. (1989). *Chromatin*. New York: Springer-Verlag.

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 e1515.

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.

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

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 Genetics* **12**, e1005940.

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*.

Xie, L., Torigoe, S. E., Xiao, J., Mai, D. H., Li, L., Davis, F. P., Dong, P., Marie-Nelly, H., Grimm, J., Lavis, L., et al. (2017). A dynamic interplay of enhancer elements regulates Klf4 expression in naive pluripotency. *Genes Dev* **31**, 1795-1808.

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.

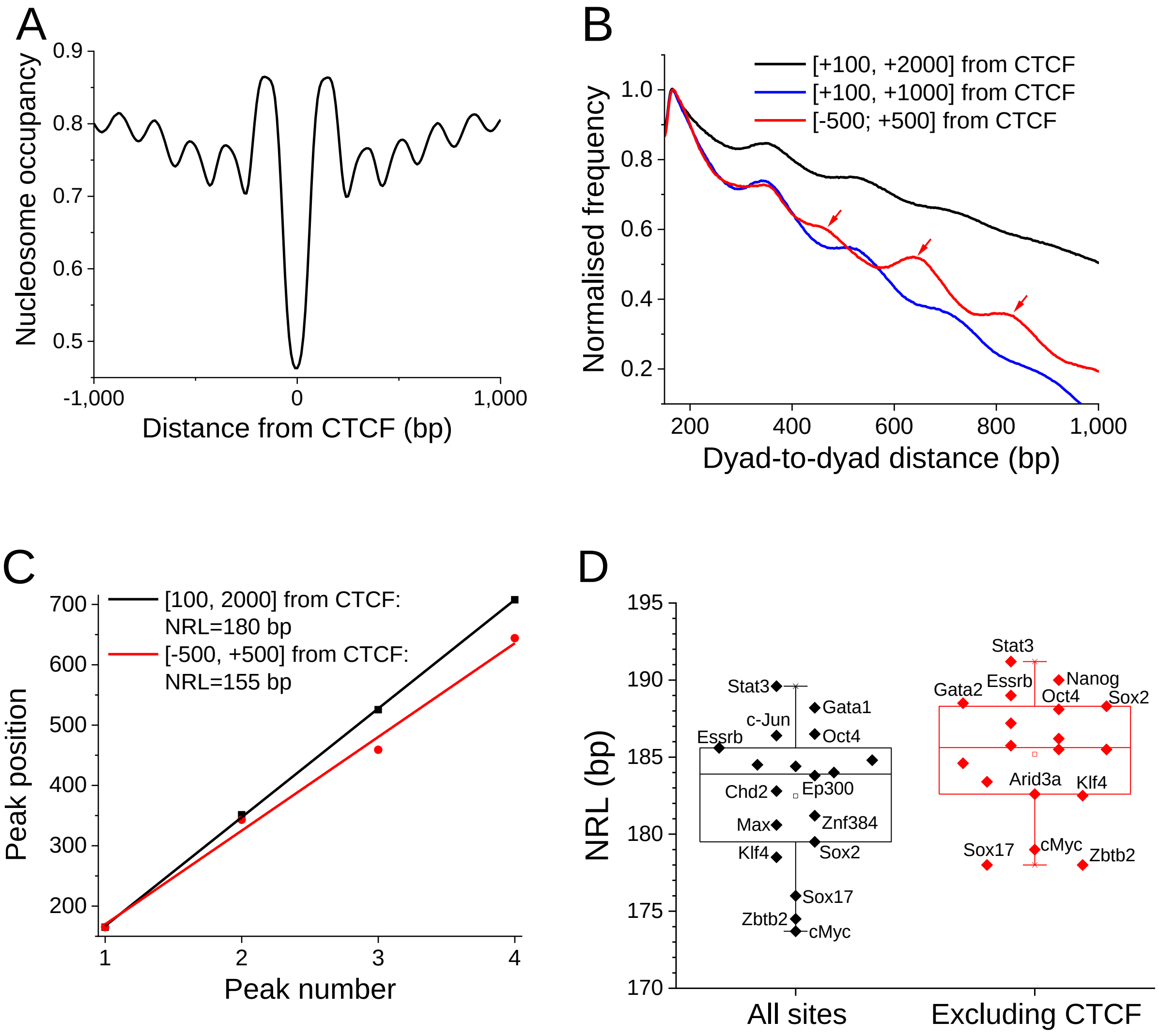

Figure 1. Setting the methodology for NRL calculation. A) Average nucleosome profile around CTCF binding sites in ESCs. B) Phasograms depicting the normalised frequency of nucleosome dyad-to-dyad distances calculated using NucTools for three different regions near CTCF sites: [100, 2000], [100, 1000] and [-500, 500]. Both [100, 2000] and [100, 1000] patterns oscillate with NRL=174. In the case of the [-500, 500] phasogram additional peaks (indicated by red arrows) appear which correspond to distances between nucleosomes on different sides of CTCF, thus resulting in an apparent NRL<160. C) NRLs calculated from the phasograms shown in panel (B). Region [-500, 500] is characterised by an unrealistically small NRL=155bp which is an artefact of the interference of two waves of distances between nucleosomes located on different sides from CTCF. D) 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. Left: all TF binding sites; right: TF binding sites which do not intersect with CTCF.

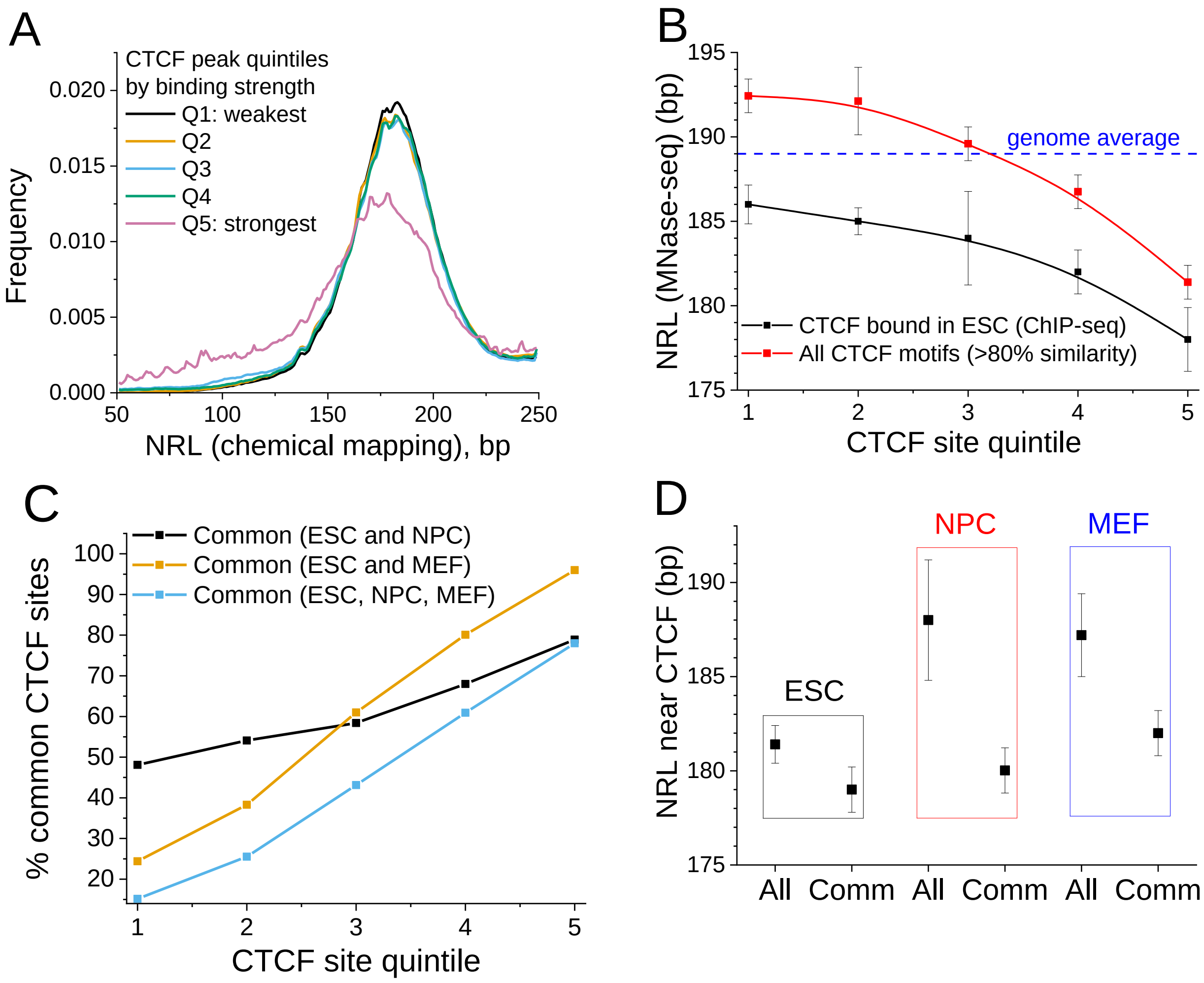

Figure 2. CTCF binding strength determines NRL decrease, which has functional implications during differentiation. A) Chemical mapping reveals the fine structure of the NRL distribution for different CTCF quintiles defined based on CTCF binding strength (the height of ChIP-seq CTCF peaks). B) Dependence of NRL on the strength of CTCF binding based on experimental ChIP-seq peaks (black line) and computationally predicted CTCF sites (red line). C) The stronger CTCF binding in ESC the higher the probability that a given CTCF site will be retained upon differentiation to NPCs and MEFs. D) Comparison of NRLs near CTCF during ESC differentiation. Upon differentiation average NRL near CTCF increases (denoted "All"), but common CTCF sites keep the smallest NRL (denoted "Comm").

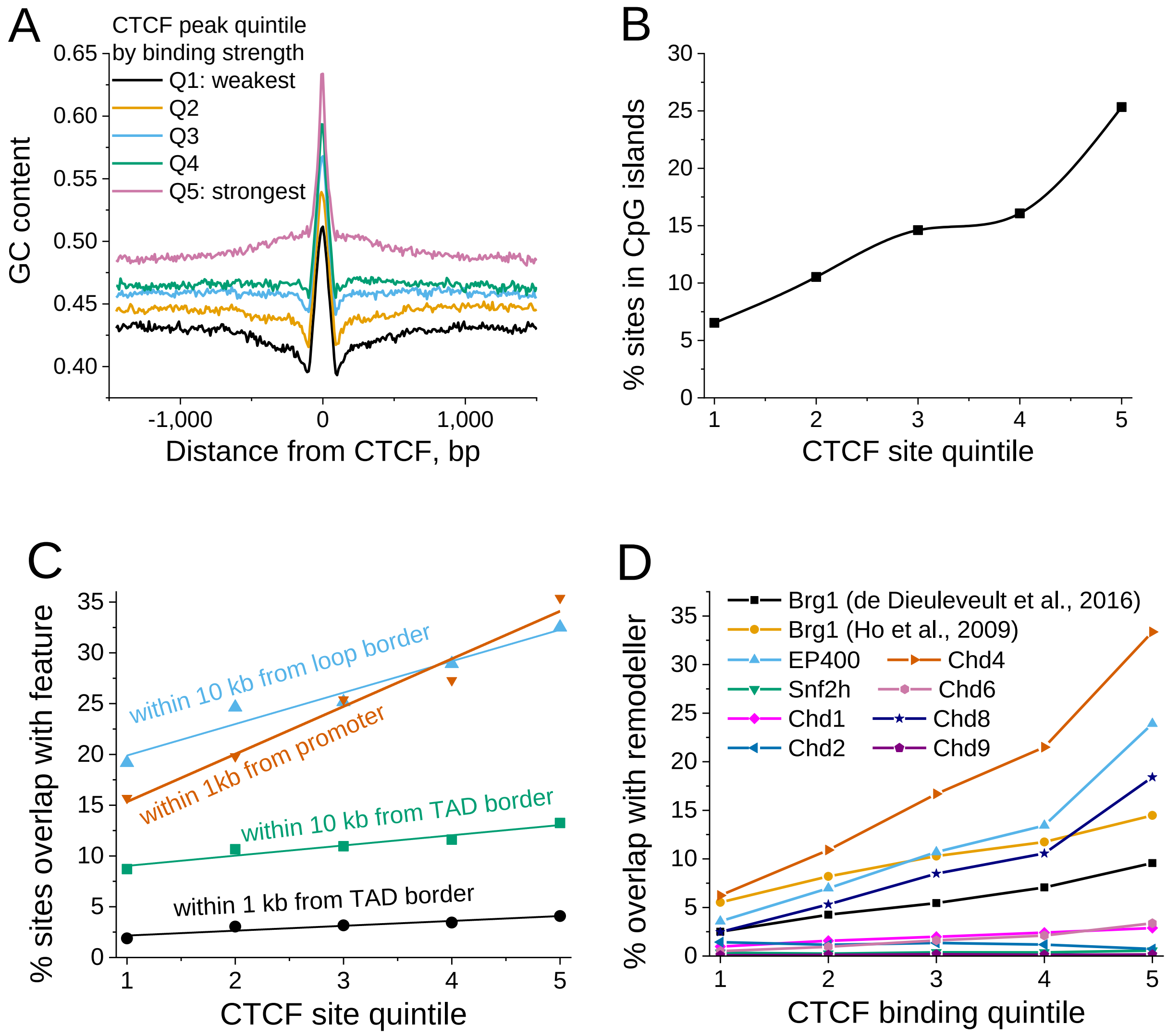

Figure 3. What determines the NRL decrease near CTCF? A) CTCF binding sites split into quintiles based on their binding strength are characterised by increasing GC content as CTCF binding strength increases. B) The stronger CTCF binding site the higher is the probability that it is located in a CpG island. 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) The stronger CTCF binds the higher the probability that it is co-enriched with different chromatin remodellers indicated on the figure.

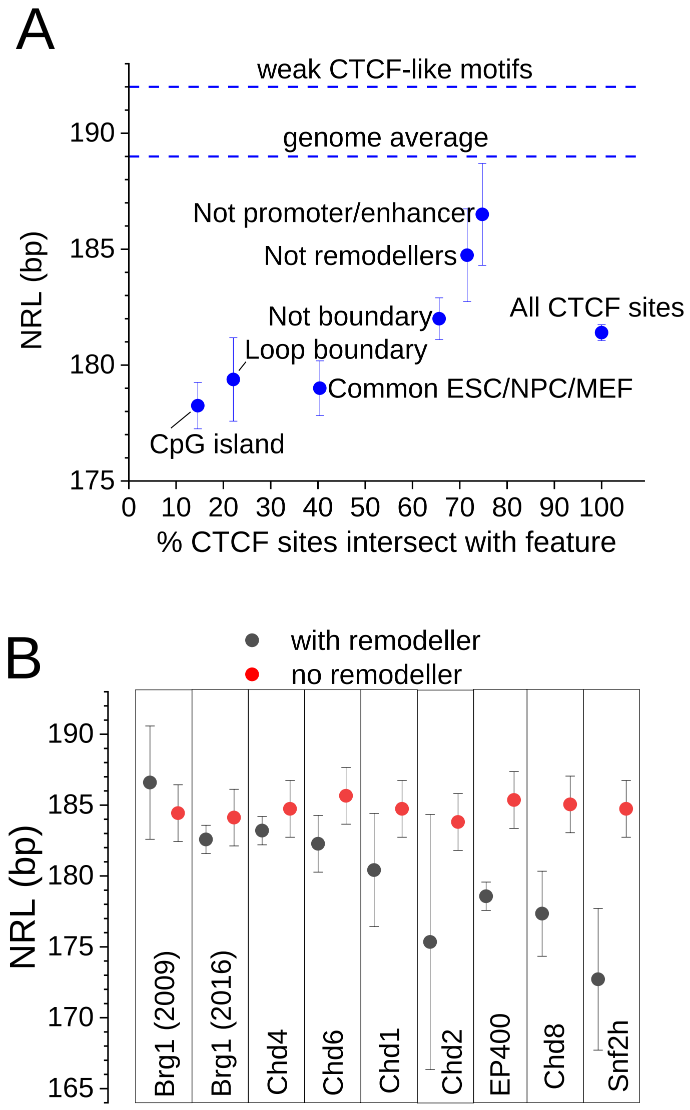

Figure 4. The summary of the effects on the value of NRL near CTCF. A) NRLs for the following subsets of CTCF sites: all sites bound in ESCs; inside chromatin loop boundary; outside of boundaries of loops and TADs; inside CpG islands; outside all remodeller peaks; outside of promoters and enhancers; common CTCF sites in ESCs, NPCs and MEFs. The top horizontal dashed line corresponds to the weak CTCF-like motifs from Figure 2D. B) NRLs calculated for CTCF sites that overlap (black) and do not over (red) with chromatin remodeller peaks for eight remodellers experimentally mapped in ESCs. Two Brg1 datasets are denoted as 2009 (Ho et al., 2009) and 2016 (de Dieuleveult et al., 2016). Vertical bars show the standard deviation.

Supplemental Materials

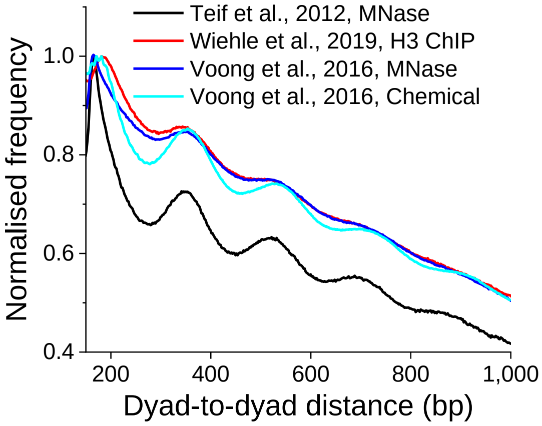

Figure S1. Phasograms calculated for region [100, 2000] near CTCF site for two different MNase-seq datasest from (Teif et al., 2012) and (Voong et al., 2016), MNase-assisted H3 ChIP-seq from (Wiehle et al., 2019) and chemical mapping from (Voong et al., 2016).

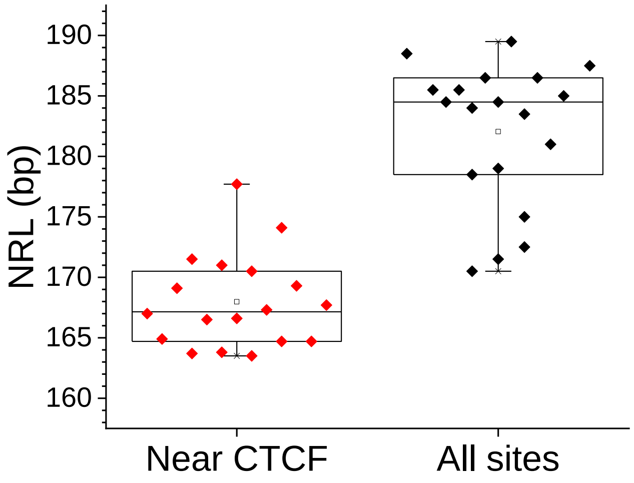

Figure S2. NRLs calculated near binding sites of 18 stemness-related chromatin proteins in ESCs in the region [250, 1000] from the TF sites. The same TF datasets as in Fig. 1D are used. Left: TF binding sites in the vicinity of CTCF; right: all TF binding sites irrespective of their distance from CTCT.

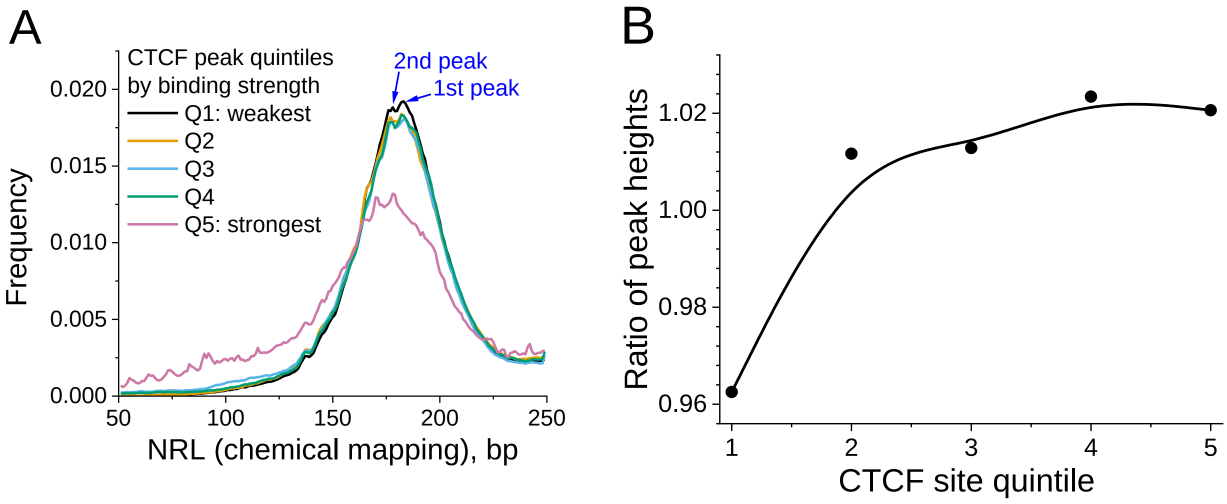

Figure S3. A) A histogram of nucleosome dyad-to-dyad distances determine dusing chemical mapping for different CTCF quintiles defined based on CTCF binding strength determined by the height of ChIP-seq CTCF peaks (the same as in Figure 2A in the main text). B) The ratio between heights of 2nd peak and 1st peak of the distribution of lengths of chemical mapping-based dyad-to-dyad distances shown in Fig. 1C as a function of the CTCF site quintile based on the heights of CTCF ChIP-seq peaks.

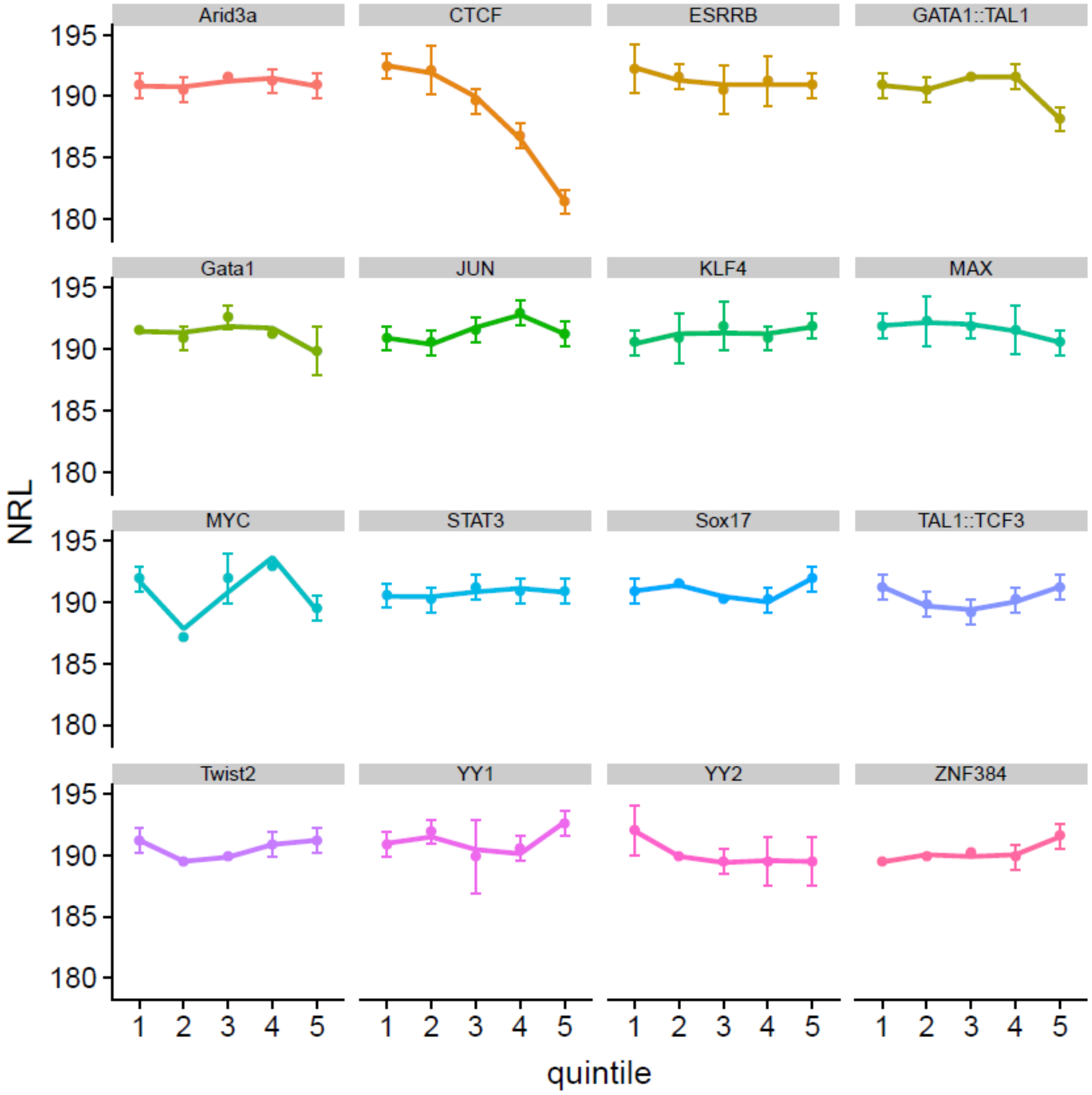

Figure S4. Proteins other than CTCF do not show the relationship between DNA-binding strength and NRL near their binding sites. 16 TFs related to stem cells are shown. TFBS used in this analysis were predicted computationally.

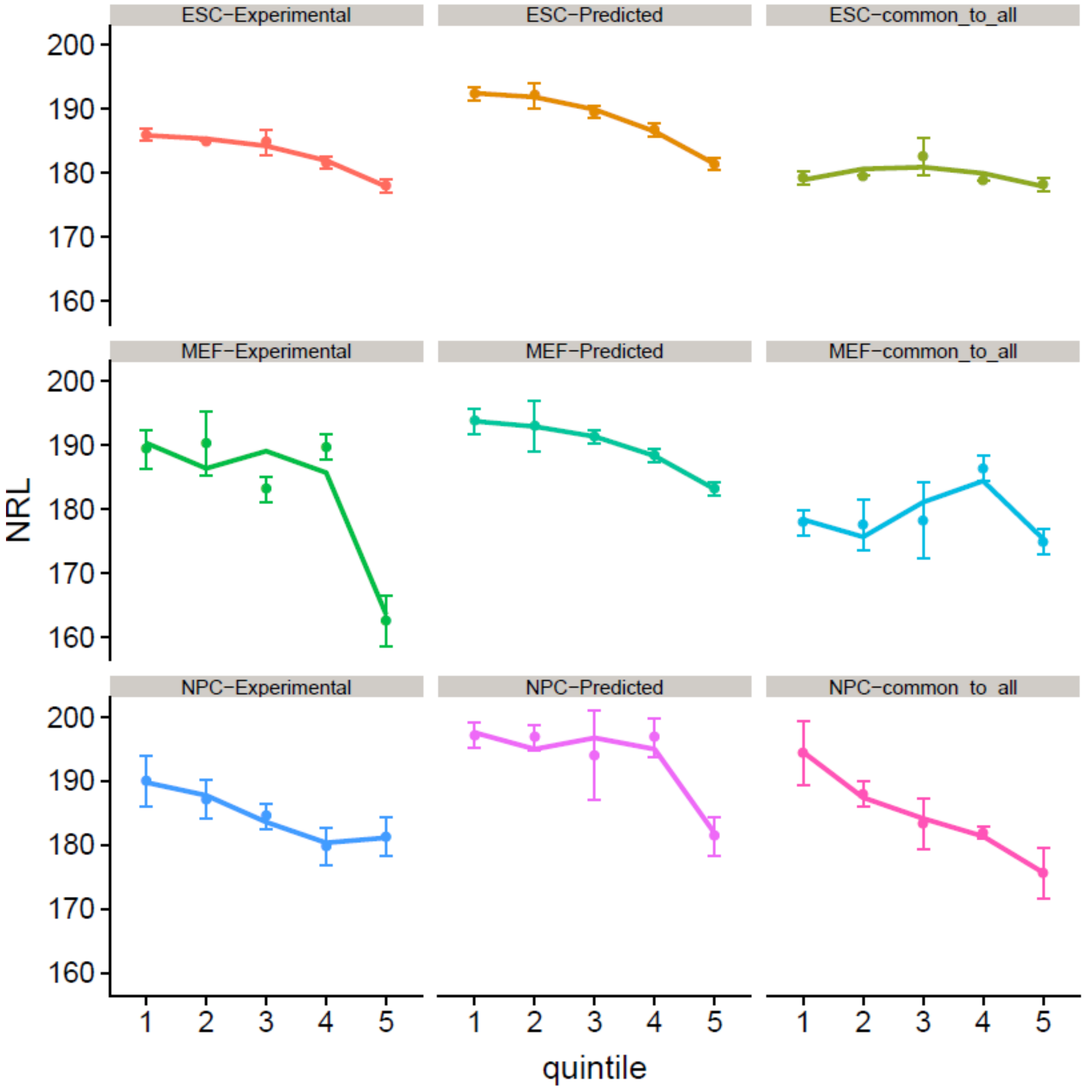

Figure S5. Effect of ESC differentiation on NRL. NRL values are calculated as a function of the CTCF site quintile. Top row: NRLs calculated based on the MNase-seq dataset in ESCs from Teif et al., 2012 for all experimental CTCF sites in ESCs determined in Shen et al, 2012 (left), all computationally predicted CTCF sites (middle) and common CTCF sites that have been determined experimentally in each of ESCs, NPCs and MEFs (right). Middle row: NRLs calculated based on the MNase-seq dataset in MEFs from Teif et al., 2012 for all experimental CTCF sites in MEF determined in Shen et al, 2012 (left), all computationally predicted CTCF sites (middle) and common CTCF sites that have been determined experimentally in each of ESCs, NPCs and MEFs (right). Bottom row: NRLs calculated based on the MNase-seq dataset in NPCs from Teif et al., 2012 for all experimental CTCF sites in NPCs determined in Bonev et al., 2017 (left), all computationally predicted CTCF sites (middle) and common CTCF sites that have been determined experimentally in each of ESCs, NPCs and MEFs (right).

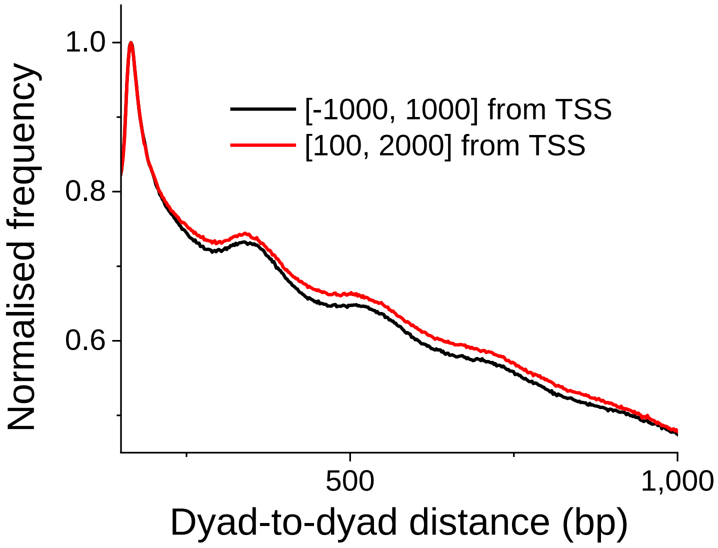

Figure S6. Comparison of the phasograms showing the normalised frequency of nucleosome dyad-to-dyad distances for the regions [-1000, 1000] and [100, 2000] from TSS. The NRLs calculated based on these phasograms are not significantly different (172+/-1 bp vs. 168+/-3 bp respectively).

Supplemental Table ST1. Summary of experimental datasets used in this study

Name	Accession #	Reference
Nanog in ESC	GSM3123484	(Kim et al., 2018)
Oct4 in ESC	GSM2417142	(Chronis et al., 2017)
Sox2 in ESC	GSM2417143	(Chronis et al., 2017)
Klf4 in ESC	GSM2417144	(Chronis et al., 2017)
cMyc in ESC	GSM2417145	(Chronis et al., 2017)
Esrrb in ESC	GSM2561449	(Xie et al., 2017)
GATA1 in ESC	GSM453997	(Chen et al., 2008)
Ep300 in ESC	ENCSR000CCD GSM918750	Mouse ENCODE
Tal1 in ESC	ENCSR000DIN GSM923579	Mouse ENCODE
Zbtb2 in ESC	GSM2716083	(Karemaker and Vermeulen, 2018)
ZNF384 in ESC	ENCSR000ERV GSM1003807	Mouse ENCODE
CTCF in ESC and MEF	ENCSR000CCB GSM918748	(Shen et al., 2012)
STAT3 in ESC	GSM288353	(Chen et al., 2008)
GATA2 in ESC	ENCSR000DIE GSM923587	Mouse ENCODE
cJun in ESC	GSM1587320	(Liu et al., 2015)
Max in ESC	GSM1171650	(Krepelova et al., 2014)
Arid3a in ESC	GSM1370509	(Rhee et al., 2014)
Sox17 in ESC	GSM1059856	(Aksoy et al., 2013)
CpG islands	Obtained from authors' web site	(Irizarry et al., 2009)
BRG1 in ESC	GSM359413	(Ho et al., 2009)
Snf2h in ESC	GSE80049	(Local et al., 2018)
Nucleosome Chemical mapping in ESC	GSE82127	(Voong et al., 2016)
MNase-seq in ESC (Voong et al., 2016)	GSM2183911	(Voong et al., 2016)
MNase-seq (Teif et al, 2012)	GSE40896	(Teif et al., 2012)
MNase-H3-ChIP-seq	GSE114599	(Wiehle et al., 2019)
BRG1 in ESC	GSE64825	(de Dieuleveult et al., 2016)
Chd1 in ESC	GSE64825	(de Dieuleveult et al., 2016)
Chd2 in ESC	GSE64825	(de Dieuleveult et al., 2016)
Chd4 in ESC	GSE64825	(de Dieuleveult et al., 2016)
Chd6 in ESC	GSE64825	(de Dieuleveult et al., 2016)

Chd8 in ESC	GSE64825	(de Dieuleveult et al., 2016)
Chd9 in ESC	GSE64825	(de Dieuleveult et al., 2016)
EP400 in ESC	GSE64825	(de Dieuleveult et al., 2016)
Chromatin loops and TADs in ESC	BED files provided by the authors	(Bonev et al., 2017)
Promoters	RefSeq	(Pruitt et al., 2014)
Enhancers	http://fantom.gsc.riken.jp (all permissive enhancers)	(Lizio et al., 2015)