6200–6212 *Nucleic Acids Research, 2016, Vol. 44, No. 13* *Published online 11 April 2016*
doi: 10.1093/nar/gkw249

Defining the multivalent functions of CTCF from chromatin state and three-dimensional chromatin interactions

Yiming Lu[1,†], Guangyu Shan[1,†], Jiguo Xue[1,†], Changsheng Chen[2,*] and Chenggang Zhang[1,*]

[1]Beijing Institute of Radiation Medicine, State Key Laboratory of Proteomics, Cognitive and Mental Health Research Center, Beijing 100850, China and [2]Department of Health Statistics, School of Military Preventive Medicine, Fourth Military Medical University, Xi'an 710032, China

ABSTRACT

CCCTC-binding factor (CTCF) is a multi-functional protein that is assigned various, even contradictory roles in the genome. High-throughput sequencing-based technologies such as ChIP-seq and Hi-C provided us the opportunity to assess the multivalent functions of CTCF in the human genome. The location of CTCF-binding sites with respect to genomic features provides insights into the possible roles of this protein. Here we present the first genome-wide survey and characterization of three important functions of CTCF: enhancer insulator, chromatin barrier and enhancer linker. We developed a novel computational framework to discover the multivalent functions of CTCF based on chromatin state and three-dimensional chromatin architecture. We applied our method to five human cell lines and identified ∼46 000 non-redundant CTCF sites related to the three functions. Disparate effects of these functions on gene expression were found and distinct genomic features of these CTCF sites were characterized in GM12878 cells. Finally, we investigated the cell-type specificities of CTCF sites related to these functions across five cell types. Our study provides new insights into the multivalent functions of CTCF in the human genome.

INTRODUCTION

CCCTC-binding factor (CTCF) is an ubiquitously expressed DNA-binding protein containing a 11 zinc-fingers domain (1). It is widely distributed in the mammalian genomes and is present in various chromatin states, including intergenic, transcribed regions, promoters and enhancers (2,3). CTCF was initially discovered as a negative regulator of the chicken *Myc* gene (4) and now has been found to be involved in many cellular processes.

One important role of CTCF in the genome is to restrict the spread of functional chromatin domains (5,6), namely chromatin barrier. The function of CTCF as chromatin barrier was suggested by studies at some specific loci in the genome. For example, one study showed that mouse transcription factor WT1 can regulate the expression of *Wnt4* gene by modulating the chromatin state of a domain with CTCF-defined boundaries and that silence of CTCF leads to spreading of histone modifications outside the domain and causes aberrant expression of neighboring genes (7). Genome-wide studies also supported the role of CTCF as chromatin barrier. A study utilizing ChIA–PET, which combines ChIP technique with 3C analyses, identified four categories of CTCF-mediated chromatin loops from 1480 CTCF-containing cis-interacting loci in mouse embryonic stem cells. Distinct histone modification signatures representing active or repressive chromatin states were observed to be present between inside and outside of the loops and these profiles were associated with different gene expression patterns. Although the number of identified CTCF-mediated loops is quite small compared to the number of all CTCF-binding sites in the genome, these results still suggest a genome-wide role of CTCF in separating functional domains with distinct chromatin states.

Another important role of CTCF is enhancer insulator, which blocks the communication between enhancer and promoter (8). Many transgenic studies suggested that CTCF could act as an enhancer insulator in a position-dependent manner. For example, activation of a poised CTCF-binding site could down-regulate *Eip75B* gene expression by recruiting CP190 protein and topologically separating an alternative upstream promoter of this gene from its enhancer (9). Observations from some genome-wide analysis also supported the role of CTCF as enhancer insulator. A survey of conserved regulatory motifs in the human genome identified 15 000 CTCF-binding sites separating

[*]To whom correspondence should be addressed. Tel: +86 10 66931590; Fax: +86 10 68169574; Email: zhangcglab@gmail.com
Correspondence may also be addressed to Changsheng Chen. Tel: +86 29 84774853; Fax: +86 29 84774858; Email: chencs@fmmu.edu.cn
[†] These authors contributed equally to the paper first authors.

adjacent genes and these genes show remarkably reduced correlation in gene expression when compared to genes in a similar arrangement but not separated by CTCF-binding sites (10). Another study showed that the correlation between the intensity of H3K4me1 at enhancers and Pol II signal at promoters within each CTCF-demarcated domain was only slightly higher than that of random enhancer-promoter pairs (11). These results supported the role of CTCF as enhancer insulator.

Recent studies showed that CTCF could also help to link distant enhancers to promoters. For example, one study identified >1000 long-range interactions between enhancers and promoters in 1% of the human genome by the ENCODE pilot project using the chromosome conformation capture carbon copy (5C) technique and found that these interactions are often not blocked by sites bound by CTCF (12). Instead, the interacting enhancers are significantly enriched for CTCF-binding sites with DNase I hypersensitive sites and/or active histone modifications such as H3K4me1, H3K4me2 and H3K4me3. Another genome-wide study assessed CTCF-mediated intrachromosomal interactions in mouse ESCs using ChIA-PET and showed that 28% of genes were up-regulated by enhancers-promoters interaction mediated by CTCF, while knockdown of this protein caused down-regulation of these genes (13). These results suggest that one of the main roles of CTCF is to facilitate gene transcription by linking enhancers to promoters, which seems to contradict the idea of enhancer insulator as a predominant role for CTCF.

As a master weaver of the genome (14), the roles of CTCF in the genome are diverse. Besides chromatin barrier, enhancer insulator and enhancer linker, CTCF also plays important roles in the regulation of V(D)J recombination (15), topologically associating domain (TAD) boundary demarcation (16), alternative promoter selection (17) and alternative splicing regulation (18). It is not clear whether the mechanisms underlying its various functions are the same; nevertheless teasing out the common characteristics shared by some functional modes of CTCF could shed light on our understanding of its multivalent roles in the genome. The genome-wide CTCF multivalency was once assessed through diverse usage of its 11 zinc fingers, however, the genomic context of chromatin state surrounding CTCF was overlooked (19). By summarizing the previously reported functions of CTCF, we noticed that its known roles could be grouped into two broad categories: acting as an insulator of adjacent regions with distinct chromatin states or acting as a linker of distal regions with 3D interactions, hinting that the multivalent functions of CTCF could be identified based on the chromatin states of CTCF flanking regions and the 3D interactions between CTCF sites.

In this article, we develop a systemic computational framework for identifying the multivalent functions of CTCF. We aim to address the following questions: (i) how to link chromatin states and genome topology to the multivalent functions of CTCF? (ii) how do the different functions of CTCF affect gene expression? (iii) what are the genomic characteristics of these functions respectively? and (iv) how are the cell-type specificities of the different functions of CTCF? To answer these questions, we apply our method to the genome-wide mapping datasets of chromatin modification and genome topology in five human cell types and identify a set of CTCF sites associated with different functions. The gene expression, sequence motif and genome topology features of these functions are comprehensively investigated in multiple cell types. At last, we assess the cell-type specificities of the multivalent functions of CTCF.

MATERIALS AND METHODS

Data source

The chromatin modification, transcription factor binding site (TFBS), chromatin three-dimensional interacting and gene expression data of five human cell types were obtained from the ENCODE project (20). The five human cell types included B-lymphoblastoid cells (GM12878), cervical carcinoma cells (HeLaS3), human mammary epithelial cells (HMEC), umbilical vein endothelial cells (HUVEC) and erythrocytic leukaemia cells (K562). The ChIP-seq dataset was composed of nine histone modifications, including H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K27ac, H3K27me3, H3K36me3, H3K9me3 and H4K20me1, and broad peaks of these modifications called by Scripture software (21) were directly downloaded from the UCSC ENOCDE website (http://genome.ucsc.edu/ENCODE/). Peaks of DNase I Hypersensitive sites (DHSs) were also downloaded from the ENCODE website. The TFBS dataset was composed of the binding sites of CTCF and Pol2 in five cell types and the binding sites of other 75 TFs in GM12878 cells. The chromatin 3D interacting data of the five cell types was obtained from a study on chromatin looping using the 3D map of human genome at kilobase resolution (22). Specifically, the chromatin loops identified by Rao et al. in the five cell types were downloaded from NCBI GEO (accession number: GSE63525) and were used for 3D genome analysis. The RNA-seq data of two biological duplicates of GM12878 cells was downloaded and mapped to reference human genomes UCSC hg19 (Human Build GRCh37). The annotation data of 23,862 human RefSeq genes, which provides the gene structure information and the positions of the TSS of the genes, was downloaded from the UCSC Genome Browser (23) website (http://www.genome.ucsc.edu/).

Functional annotation of chromatin states

The predicted chromatin states were associated to functional elements using three types of annotated regions: promoters, enhancers and gene transcribed regions. The promoter regions were defined by the 4k regions around the TSS ([$-2k$, $+2k$]) downloaded using the UCSC genome browser. The putative enhancer regions were identified in five human cell types by an AdaBoost model using the genome-wide mapping data of three histone modifications: H3K4me1, H3K4me2 and H3K4me3 as described in the paper (24). The information of gene transcribed regions was also obtained from the UCSC genome browser. We calculated the percentages of each chromatin state overlapping with the three types of annotated regions respectively. Genomic intervals overlapping with multiple types of regions were assigned to a certain type of region based on the priority order: promoter, enhancer and transcribed.

Model evaluation

We used a complexity-penalized average silhouette width (PASW) score to evaluate the performances of the PAM clustering models. Silhouette analysis can be used to study the separation distance between the resulting clusters. The silhouette width displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters quantitatively. For each data i, let $a(i)$ be the average dissimilarity of i with all other data within the same cluster and $b(i)$ be the lowest average dissimilarity of i to any other cluster. The silhouette can be defined as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

This measure has a range of [−1, 1]. Silhoette width near +1 indicate that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster. We also limited the model complexity (number of clusters) because too complex models will harm their biological interpretability. As a result, to achieve a balance between model performance and complexity, we refined the ASW to a complexity-penalized ASW (PASW). Let K be the number of clusters, so the PASW could be defined as follows:

$$\text{PASW} = \overline{s(i)} - C \cdot K^2$$

C is the balancing constant so that the value range of the penalizing term equals that of the ASW. In this study, the value range of ASW is ∼0.2, so we set the $C = 0.0005$ ($C{\cdot}K^2$ ranges from 0.002 to 0.2, when K ranges from 2 to 20).

Gene expression measurement

The RNA-seq data of GM12878 was mapped to reference human genome by the TopHat (25) software. Novel transcripts were removed from the analysis. The gene expression levels were reported in FPKM (fragments per kilobase of transcript per million fragments sequenced) by using the Cufflinks (26) software. The FPKM values of a gene in two duplicates were averaged to represent the expression level of this gene. All 23 862 genes were sorted increasingly based on their expression levels and divided into ten classes, each of which contained 10% of the genes. Genes in the first class (bottom 10%) were assigned to an expression level of 1 and genes in the second class were assigned to an expression level of 2, and so on. Genes in the last class were assigned to an expression level of 10.

Sequence motif analysis

For each set of CTCF-binding sites, we used MEME (27) to discover consensus sequence motifs with default parameters. The parallel version of MEME was instructed to report the top five motifs with lengths ranging from 12 to 14 bases (Supplementary Figure S6) in CTCF binding sequences. CTCF binding sites shorter than 12 bases were removed from the analysis.

RESULTS

A computational framework for identifying the multivalent functions of CTCF

To identify the multivalent functions of CTCF, we first predicted the chromatin states of CTCF flanking regions. We adopted an unsupervised Partitioning Around Medoids (PAM) clustering algorithm to identify the chromatin states based on chromatin modifications. The workflow of the PAM method is illustrated in Figure 1. The primary difference between the PAM and the general chromatin segmentation models, such as ChromHMM (28) and Segway (29) is that, instead of predicting chromatin states genome-wide, the PAM model focuses on identifying the chromatin states of the flanking regions surrounding CTCF-binding sites. In this step, we adopted a binarization approach to explicitly model the presence/absence status of each mark in intervals. Specifically, peaks of chromatin marks, including histone modifications and RNA polymerase II occupancy, were called using Scripture (21) software under the assumption of uniform background signal. The presence/absence status of a mark in each flanking region is represented by a binary variable, taking value 1 for at least one peak of the mark presenting in the region and 0 for no peak. A peak of chromatin mark is assigned to a flanking region if the center of the peak presents in the flanking regions or the center of the flanking region presents in the peak. The binarization approach has the advantage that the model does not require a multivariate normal distribution assumption, which is generally violated by the relatively small discrete counts usually found in ChIP-seq datasets, thus enables more robust models to be inferred. We computed the Jaccard distance between each pair of regions based on their combinatorial binary status of marks and separated them into different clusters using the PAM algorithm. We associated these clusters to specific chromatin states according to their chromatin mark profiles and the enrichment of annotated genomic regions (see Methods and Materials).

We next identified the multivalent functions of CTCF based on the chromatin states and genome topology of CTCF flanking regions. We defined two broad functional modes of CTCF: insulator mode and linker mode. The insulator mode of CTCF indicates CTCF locating between two genomic domains with distinct chromatin states, while the linker mode indicates CTCF locating within a genomic domain and interacting with other CTCF via chromatin loops. The insulator mode of CTCF can be identified using the *transition patterns* of chromatin states between two flanking regions, and the linker mode can be identified using the *linkage patterns* of chromatin states between the interacting genomic regions surrounding a pair of CTCF sites mediated by chromatin loops.

Identification of chromatin states of CTCF flanking regions in GM12878 cells

To identify the chromatin states of CTCF flanking regions, we applied the PAM model to a GM12878 epigenetic dataset consisting of the genome-wide occupancy data of nine histone modifications and RNA polymerase II binding sites. We obtained a total of 79 957 CTCF-binding sites

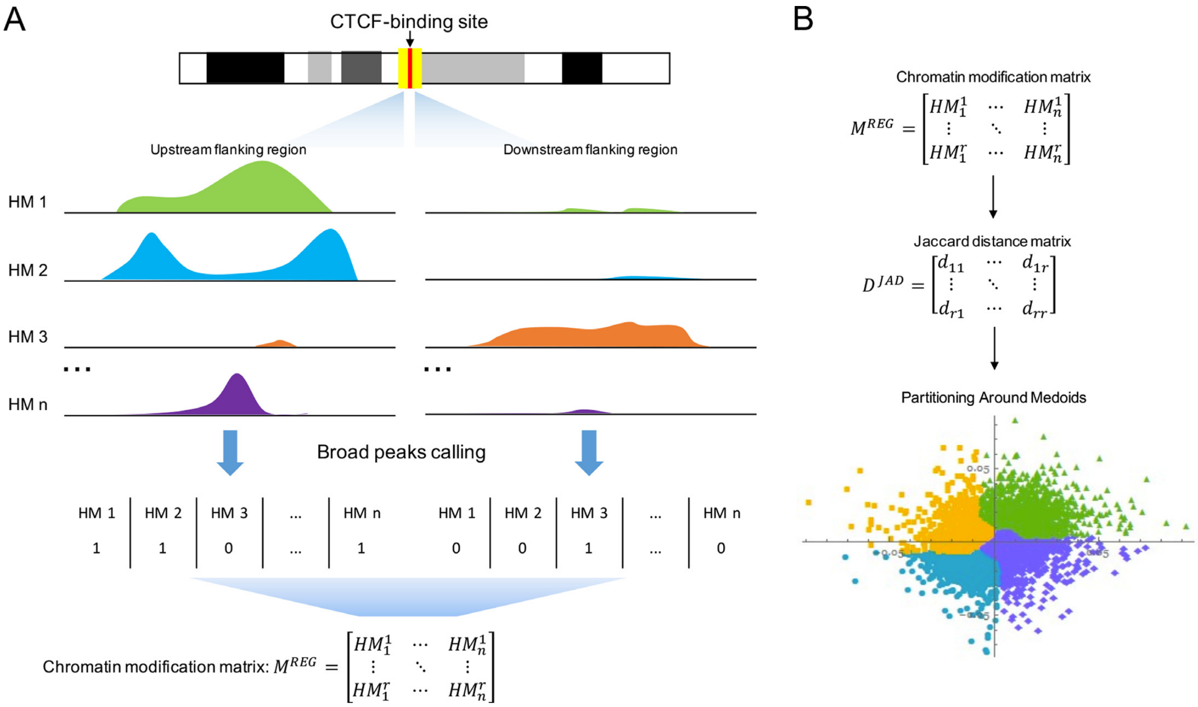
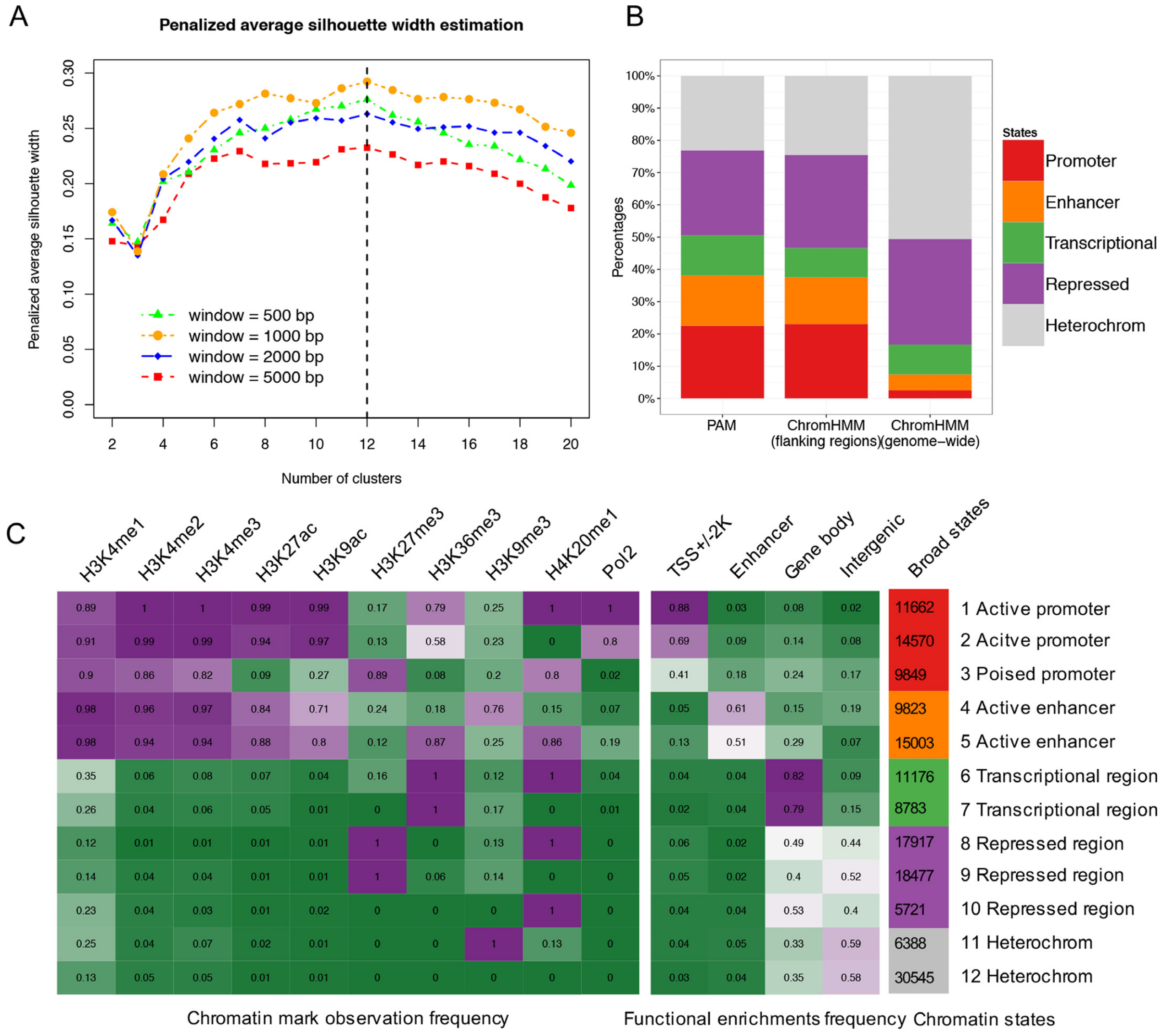

A
CTCF-binding site
Upstream flanking region
Downstream flanking region
HM 1
HM 2
HM 3
HM n
Broad peaks calling
HM 1 1 | HM 2 1 | HM 3 0 | ... | HM n 1
HM 1 0 | HM 2 0 | HM 3 1 | ... | HM n 0
Chromatin modification matrix: $M^{REG} = \begin{bmatrix} HM_1^1 & \cdots & HM_n^1 \\ \vdots & \ddots & \vdots \\ HM_1^r & \cdots & HM_n^r \end{bmatrix}$
B
Chromatin modification matrix
$M^{REG} = \begin{bmatrix} HM_1^1 & \cdots & HM_n^1 \\ \vdots & \ddots & \vdots \\ HM_1^r & \cdots & HM_n^r \end{bmatrix}$
Jaccard distance matrix
$D^{JAD} = \begin{bmatrix} d_{11} & \cdots & d_{1r} \\ \vdots & \ddots & \vdots \\ d_{r1} & \cdots & d_{rr} \end{bmatrix}$
Partitioning Around Medoids
0.05
-0.05
0.05

Figure 1. The workflow for identifying the chromatin state of CTCF-flanking regions based on the Partitioning Around Medoids clustering algorithm.

in GM12878 cells from the ENCODE project (20). The flanking regions of these CTCF sites were extracted using four different window sizes: 500, 1000, 2000 and 5000 bp, and were then clustered based on the combinatorial presence/absence profiles of the ten chromatin marks by the PAM algorithm. The optimal number of clusters was determined for each window size by optimizing a complexity-penalized average silhouette width (PASW) score (see Methods and Materials section). We screened PAM models with number of clusters ranging from 2 to 20 and found models with 12 clusters gave the highest PASW scores for all four window sizes (Figure 2A). We used a window size of 1000 bp and 12 clusters for further analysis, for these parameters appeared to be able to give the maximized biological interpretability with minimized model complexity. Annotating the clustered regions surrounding CTCF sites using known transcription start sites (TSSs), putative enhancers and transcribed gene regions revealed five broad classes of chromatin states, which were associated to promoters, enhancers, transcribed, repressed and heterochromatin regions, respectively (Figure 2C and Supplementary Figure S1). In addition to the 12-cluster model, we also tested our method by training a 10-cluster and a 15-cluster model, from which five broad chromatin states were identified respectively (Supplementary Figure S2). We found the vast majority of the regions assigned the same broad state overlapped across the 10-, 12- and 15-cluster model (Supplementary Figure S3), demonstrating our clustering method is highly robust to the number of clusters.

We compared our results with a genome-wide chromatin state map predicted by ChromHMM—a method for identifying chromatin states on a genome-wide scale—using the same chromatin marks and an interval of 1000 bp. After ChromHMM learned and evaluated a set of models ranging from 11 to 15 states, we focused on a 12-state model that provided best resolution to resolve biologically meaningful chromatin patterns (Supplementary Figure S4). These 12 chromatin states were assigned to five broad classes as described above. We assigned chromatin states to CTCF flanking regions by overlapping them with the predicted states by ChromHMM. We found the ChromHMM and PAM models gave very similar predictions on chromatin states of CTCF flanking regions, as shown in Figure 2B. Specifically, the PAM model predicted 22.6% promoter, 15.5% enhancer, 12.5% transcribed, 30.3% repressed and 19.1% heterochromatin states and ChromHMM predicted 22.6% promoter, 14.2% enhancer, 9.0% transcribed, 33.2% repressed and 21.0% heterochromatin states. Nevertheless, we used the PAM model for the further analysis of chromatin state transition between CTCF flanking regions, because most CTCF sites embedded within ChromHMM intervals and it is hard for us to determine the specific states of the two flanking regions.

As expected, the states enriched in the flanking regions of CTCF were distinct from the genome-wide chromatin state map predicted by ChromHMM, which consisted of 2.5% promoter, 5.0% enhancer, 9.2% transcribed, 32.7% repressed and 50.6% heterochromatin states (Figure 2B). The promoter state was most highly enriched around CTCF sites (log odds ratio = 3.18 for both PAM and ChromHMM model; *P*-value = 2.88E−41, Chi-squared test) and enhancer was also highly enriched at CTCF sites (log odds ratio = 1.65/1.52 for PAM/ChromHMM; *P*-value = 1.76E−14/4.94E−12, Chi-squared test). In contrast, heterochromatin was highly depleted at CTCF sites (log odds ratio = −1.41/−1.27 for PAM/ChromHMM; *P*-



H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	H3K27me3	H3K36me3	H3K9me3	H4K20me1	Pol2	TSS+/-2K	Enhancer	Gene body	Intergenic	Broad states	
0.89	1	1	0.99	0.99	0.17	0.79	0.25	1	1	0.88	0.03	0.08	0.02	11662	1 Active promoter
0.91	0.99	0.99	0.94	0.97	0.13	0.58	0.23	0	0.8	0.69	0.09	0.14	0.08	14570	2 Acitve promoter
0.9	0.86	0.82	0.09	0.27	0.89	0.08	0.2	0.8	0.02	0.41	0.18	0.24	0.17	9849	3 Poised promoter
0.98	0.96	0.97	0.84	0.71	0.24	0.18	0.76	0.15	0.07	0.05	0.61	0.15	0.19	9823	4 Active enhancer
0.98	0.94	0.94	0.88	0.8	0.12	0.87	0.25	0.86	0.19	0.13	0.51	0.29	0.07	15003	5 Active enhancer
0.35	0.06	0.08	0.07	0.04	0.16	1	0.12	1	0.04	0.04	0.04	0.82	0.09	11176	6 Transcriptional region
0.26	0.04	0.06	0.05	0.01	0	1	0.17	0	0.01	0.02	0.04	0.79	0.15	8783	7 Transcriptional region
0.12	0.01	0.01	0.01	0.01	1	0	0.13	1	0	0.06	0.02	0.49	0.44	17917	8 Repressed region
0.14	0.04	0.04	0.01	0.01	1	0.06	0.14	0	0	0.05	0.02	0.4	0.52	18477	9 Repressed region
0.23	0.04	0.03	0.01	0.02	0	0	0	1	0	0.04	0.04	0.53	0.4	5721	10 Repressed region
0.25	0.04	0.07	0.02	0.01	0	0	1	0.13	0	0.04	0.05	0.33	0.59	6388	11 Heterochrom
0.13	0.05	0.05	0.01	0.01	0	0	0	0	0	0.03	0.04	0.35	0.58	30545	12 Heterochrom

Chromatin mark observation frequency
Functional enrichments frequency
Chromatin states

Figure 2. Identification of chromatin states of CTCF flanking regions in GM12878 cells. (**A**) The curves of penalized average silhouette width for models with four window sizes: 500, 1000, 2000, 5000 bp and number of clusters ranging from 2 to 20. (**B**) The percentages of five broad types of chromatin states at CTCF-flanking regions and/or whole genome identified by PAM and ChromHMM. (**C**) Observation frequency of 10 chromatin marks and enrichment frequency of four functional elements associated with 12 chromatin states and five broad states identified by the PAM method.

value = 3.83E−49/4.48E−43, Chi-squared test). The transcribed and repressed states are not significantly enriched or depleted at CTCF sites compared to the genome-wide map, the log odds ratios for the two states were -0.04 and 0.02 (*P*-value = 0.938 and 0.936, Chi-squared test) for ChromHMM and 0.44 and −0.11 for PAM (*P*-value = 0.0214 and 0.207, Chi-squared test).

Genome-wide identification of the insulator mode of CTCF in GM12878 cells

To assess the transition patterns of chromatin states between the two flanking regions of CTCF sites, we constructed a symmetric transition matrix with the number of transition events between each two of the five broad chromatin states (Figure 3A). We found about 83.9% (67,044) of the CTCF sites had two flanking regions of the same broad class of chromatin states and about 16.1% (12 913) sites had different broad classes of chromatin states at their flanking regions. To identify the transition patterns highly enriched at CTCF binding sites, we randomly permuted the flanking regions surrounding CTCF sites and sampled 12,913 permuted transition events. We found three types of transition patterns were significantly enriched at CTCF sites: Promoter/Enhancer, Promoter/Repressed and Enhancer/Transcribed (*P*-value < 2.2E−16, Chi-squared test; Figure 3B). We also compared the transition patterns at CTCF sites with the genome-wide transition patterns generated by ChromHMM and found that the transition patterns at CTCF sites were also distinct from those of the whole genome (Supplementary Figure S5A).

The frequent transitions between promoter and enhancer surrounding CTCF sites were unexpected, for enhancers generally serve as TSS-distal regulators. Considering the number of Promoter/Enhancer transitions took up 19.9% (5530/27 842) of all enhancer- and promoter-related CTCF sites, we attributed these transitions mainly to the close re-

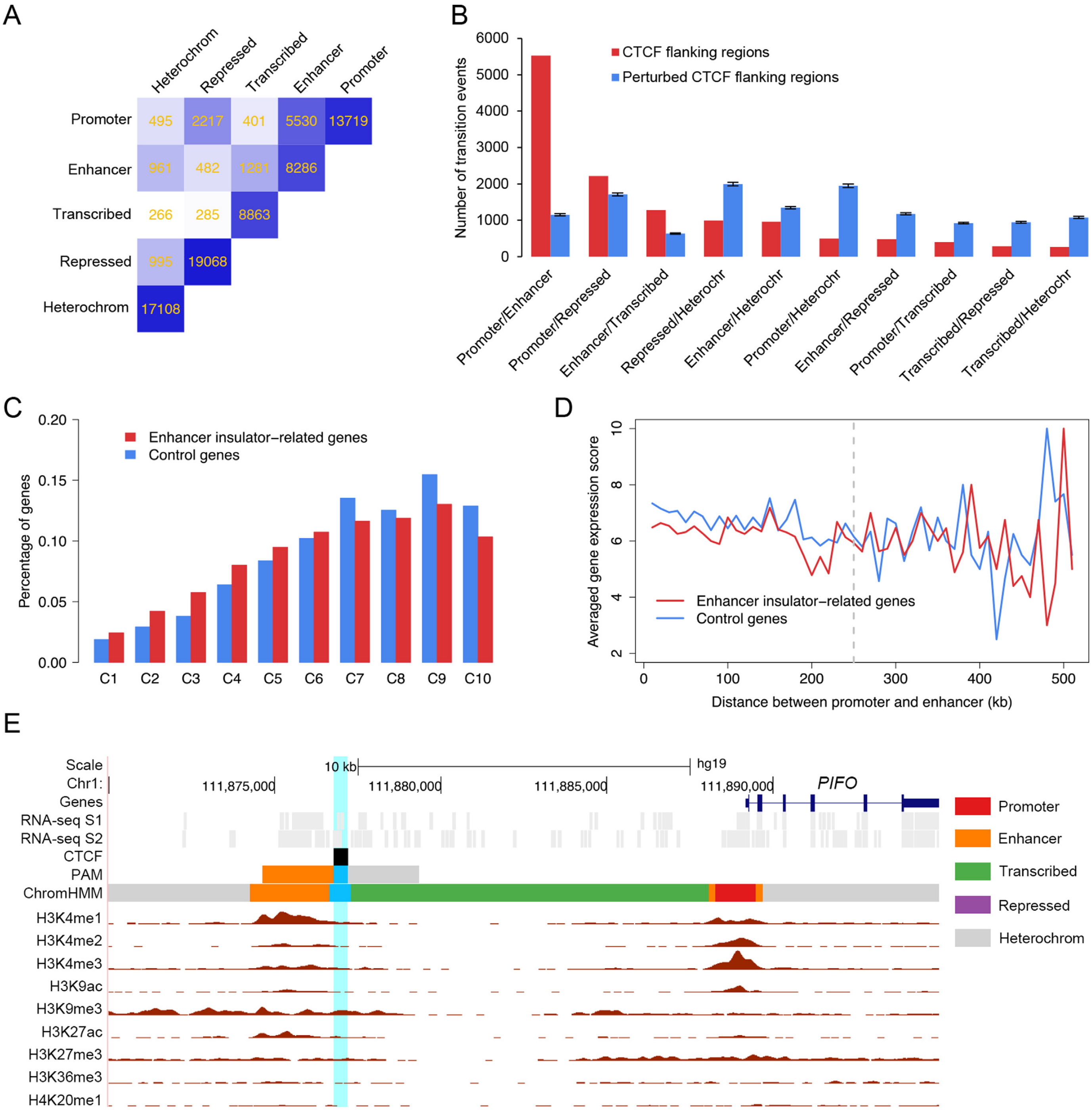


Figure 3. Identification of the enhancer insulator CTCF sites in GM12878 cells. (**A**) The transition matrix among five broad types of chromatin states surrounding CTCF sites. The number in each cell of the matrix represents the number of transition events between the two corresponding states. (**B**) The comparison of number of transitions between CTCF flanking regions and the perturbed CTCF flanking regions. Error bar represents the standard deviation of event count from 10 times of random perturbations. (**C**) The distributions of the enhancer insulator-related genes and the control genes across 10 gene clusters. Gene were sorted by increasing expression level from C1 to C10. (**D**) The curves of average gene expression score against the distance between promoter and enhancer. (**E**) An example of enhancer insulator CTCF site and its closest gene *PIFO*, which has an active promoter but is very weakly transcribed. From top to bottom, the display tracks are genomic position, gene structure, RNA-seq signals with two replicates, broad chromatin states predicted by PAM and ChromHMM, ChIP-seq signals of nine histone modifications, respectively. All these signals are specific to GM12878 cells (not include genomic position and gene structure).

semblance of chromatin signatures between promoter and enhancer (Figure 2C), which might cause a higher confusion rate between the two states. To verify this idea, we investigated the role of these CTCF sites in gene transcription. We constructed both a CTCF-positive gene set containing 4,031 genes whose promoters were involved in the Promoter/Enhancer transitions and a CTCF-negative gene set containing 4168 genes whose promoters are adjacent to enhancers but not separated by CTCF. We then evenly divided all human genes (23 862 genes) into ten classes according to their expression levels in GM12878 and examined their distributional differences between the two gene sets. We found genes in CTCF-positive set were significantly up-regulated compared to those in CTCF-negative set (P-value = 2.46E−12, Kendall's correlation test; Supplementary Figure S5B). In addition, we also identified factors that colocalized with these CTCF sites in GM12878. We examined the percentages of these sites overlapping with the binding sites of 75 transcription factors (TFs) and found POLR2A and RUNX3 were the most frequently colocalized factors (Supplementary Figure S5C). POLR2A is the largest RNA polymerase II subunit and RUNX3 is a member of the runt domain-containing family of TF binding to enhancers and promoters (30). These results indicated that CTCF associated with Promoter/Enhancer transitions might act as transcriptional activators.

Next, we described two functions of CTCF that are associated with the insulator mode of CTCF.

CTCF as an enhancer insulator. To investigate the potential role of CTCF as an enhancer insulator, we considered three types of transitions: Enhancer/Transcribed, Enhancer/Repressed and Enhancer/Heterochromatin, consisting of 2724 events (Supplementary Table S1). Enhancers and promoters associated with these transitions were separated by CTCF and at least one other chromatin domain, including transcribed, repressed or heterochromatin regions. CTCF sites distal from both enhancers and promoters were excluded from this analysis to avoid false positive Enhance-Promoter pairs. To investigate the regulatory role of CTCF as enhancer insulators, we constructed both an enhancer insulator-related gene set and a control gene set. The enhancer insulator-related set was constituted by genes whose promoters lie on the opposite side of the enhancers in the transitions so that promoters and enhancers were separated by both CTCF and at least one transcribed/repressed/heterochromatin domains, while the control set was constituted by genes whose promoters are located in the same side of the enhancers in the transitions (adjacent promoters and enhancers were removed) so that they were separated only by the transcribed/repressed/heterochromatin domain(s). We examined the distributional difference of those two gene sets among ten gene classes as described above and found that genes in the enhancer insulator set were significantly down-regulated compared to those in the control set (P-value = 1.58E−7, Kendall's correlation test, Figure 3C), suggesting these CTCF sites act as functional enhancer insulators. To further examine whether the distance between enhancers and promoters affects their roles in gene expression, we sorted genes in the two sets based on the distances between promoters and enhancers and calculated their averaged expression scores in each 10 kb window from 0 up to 500 kb, respectively. Interestingly, we found the enhancer insulator-related genes were ubiquitously down-regulated within a distance of 250 kb, beyond this distance there was no significant difference between the two gene sets (Figure 3D). This distance is similar to the loop size ∼200 kb at which transition of histone modification was observed, as reported previously (13), suggesting the enhancer insulator role of CTCF may be related to chromatin looping. Figure 3E shows an example of enhancer insulator CTCF site and its closest gene *PIFO*, which has an active promoter but is very weakly transcribed.

CTCF as a chromatin barrier. The chromatin barrier role of CTCF is associated with six types of state transition events, including Promoter/Repressed, Enhancer/Repressed, Transcribed/Repressed, Promoter/Heterochromatin, Enhancer/ Heterochromatin and Transcribed/Heterochromatin, consisting of 4706 events (Supplementary Table S2). To investigate the regulatory role of CTCF as chromatin barrier, we constructed both a chromatin barrier-related gene set who lie on the opposite sides of the repressed or heterochromatin domain in the transitions and a control gene set who lie on the same side of the transitions. We examined the distributional difference of those two gene sets among the ten gene classes and found that genes in the chromatin barrier set were significantly up-regulated compared to those in the control set (P-value < 2.2E−16, Kendall's correlation test, Figure 4A), suggesting these CTCF sites act as chromatin barriers. We investigated the distance effect of CTCF sites as chromatin barriers and found that the chromatin barrier-related genes were ubiquitously up-regulated within a distance of 250 kb and no significant difference was observed beyond this distance, which closely resembled the distance effect of enhancer insulator (Figure 4B). Figure 4C shows an example of chromatin barrier CTCF site separating two adjacent genes *TDRD10* and *UBE2Q1*, which locate at different chromatin states and display distinct expression levels.

Genome-wide identification of the linker mode of CTCF in GM12878 cells

We assessed the linkage pattern of chromatin states between the long-range interacting CTCF sites identified by Hi-C data in GM12878. To avoid confusion, we used the 67 044 non-transition CTCF sites whose flanking regions were of the same chromatin state to generate a symmetric linkage matrix by calculating the numbers of linkage events between two interacting CTCF sites. Among all possible linkage patterns, we found five types of linkages are significantly enriched at long-range interacting CTCF sites mediated by chromatin loops compared to randomly shuffled CTCF pairs (Figure 5A). They included linkages between Repressed/Repressed, Promoter/Enhancer, Transcribed/Repressed, Enhancer/Enhancer and Enhancer/Transcribed. We next described one function of CTCF that is associated with the linker mode.

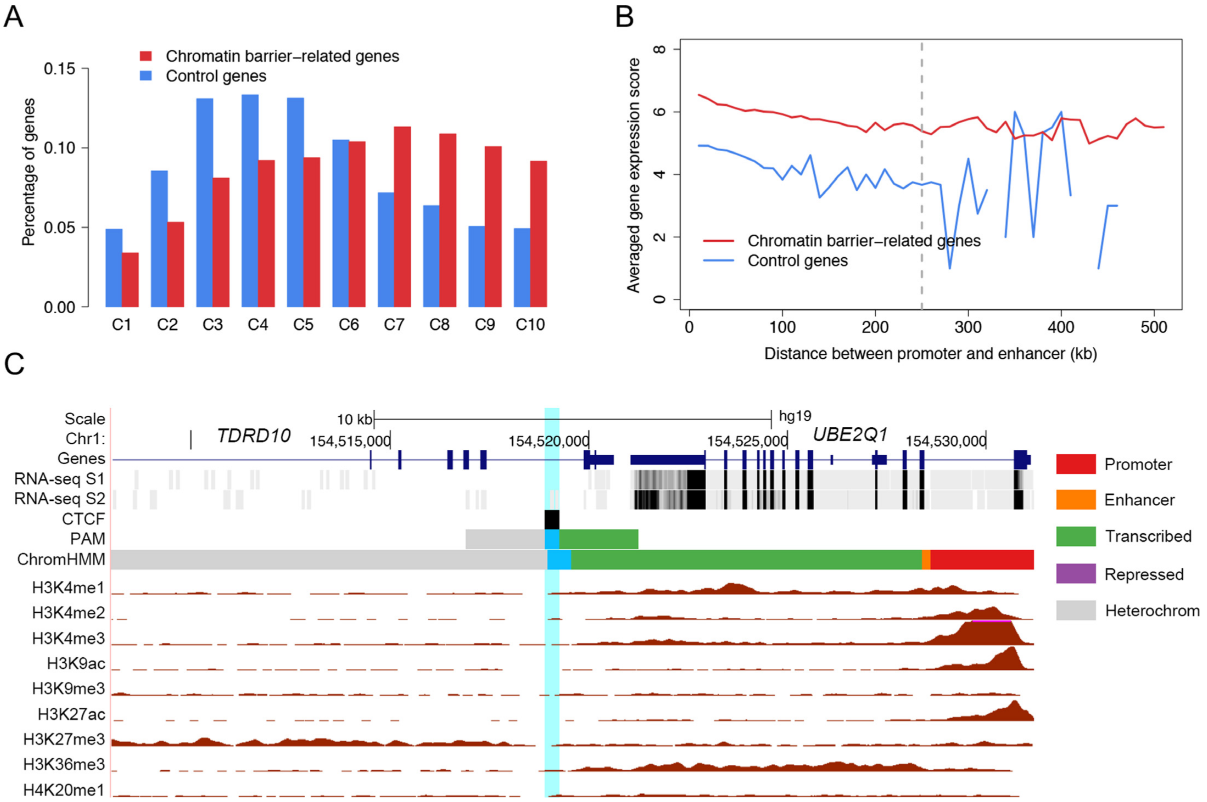


Figure 4. Identification of the chromatin barrier CTCF sites in GM12878 cells. (**A**) The distributions of the chromatin barrier-related genes and the control genes across 10 gene clusters. Gene were sorted by increasing expression level from C1 to C10. (**B**) The curves of average gene expression score against the distance between promoter and repressed or heterochromatin regions. (**C**) An example of chromatin barrier CTCF site separating two adjacent genes *TDRD10* and *UBE2Q1*, which locate at different chromatin states and display distinct expression levels. From top to bottom, the display tracks are genomic position, gene structure, RNA-seq signals with two replicates, broad chromatin states predicted by PAM and ChromHMM, ChIP-seq signals of nine histone modifications, respectively.

CTCF as an enhancer linker. The enhancer linker function of CTCF is associated with the Promoter/Enhancer linkage, containing 981 events (Supplementary Table S3). To investigate the regulatory role of CTCF as enhancer linker, we constructed both an enhancer linker gene set whose promoters were linked to distal enhancers in these linkage events and a non-enhancer linker gene set whose promoters were linked to non-enhancer chromatin domains. We examined the distributional difference of those two gene sets among the 10 gene classes and found that genes in the enhancer linker set were significantly up-regulated compared to those in non-enhancer linker set (*P*-value = 9.82E−6, Kendall's correlation test, Figure 5B), suggesting these CTCF sites could link distal enhancers to promoters through chromatin loops. Figure 5C shows an example of enhancer linker CTCF site and the linked gene *APITD1*, which has an active promoter and is actively transcribed. The bimodal distributions of H3K4me2 and H3K4me3 around TSS indicated the nucleosome occupancy by TFs.

Characterization of the three distinct functional modes of CTCF

We assessed several important genomic features of the different CTCF functional sites. These features included binding motifs, colocalized factors, chromatin accessibility and binding orientation.

Sequence motif enrichment analysis in 10,000 sites randomly sampled from 79,957 global CTCF sites showed that the typical consensus DNA sequence for CTCF-binding sites (22,31), written as 5′-CCACNAGGTGGCAG-3′ (denoted CM1), was the most enriched motif. We also found that the top five enriched motifs in enhancer insulator and chromatin barrier CTCF sites closely resembled those of global sites (Supplementary Figure S6) and CM1 was also the most enriched motif in enhancer insulator (1029 forward and 1460 reverse; *E*-value = 1.8E−932 and 6.0E−1205) and chromatin barrier CTCF sites (2276 forward and 1833 reverse; *E*-value = 7.5E−1747 and 2.9E−1535). Interestingly, we found another motif, written as 5′-CCCCTCCCCCTCCC-3′ (denoted CM2), was the most enriched motif (324 forward and 384 reverse; *E*-value = 6.7E−109 and 5.7E−126) in enhancer linker CTCF sites instead of CM1. This motif is the first discovered binding motif of CTCF (4).

DNA-binding factor colocalization analyses were performed in different CTCF sets based on the binding sites of 75 TFs in GM12878. We found the profiles of colocalization percentages of the top 10 factors in the enhancer insulator and chromatin barrier CTCF sets closely resembled that of global CTCF sites. The cohesion subunits RAD21

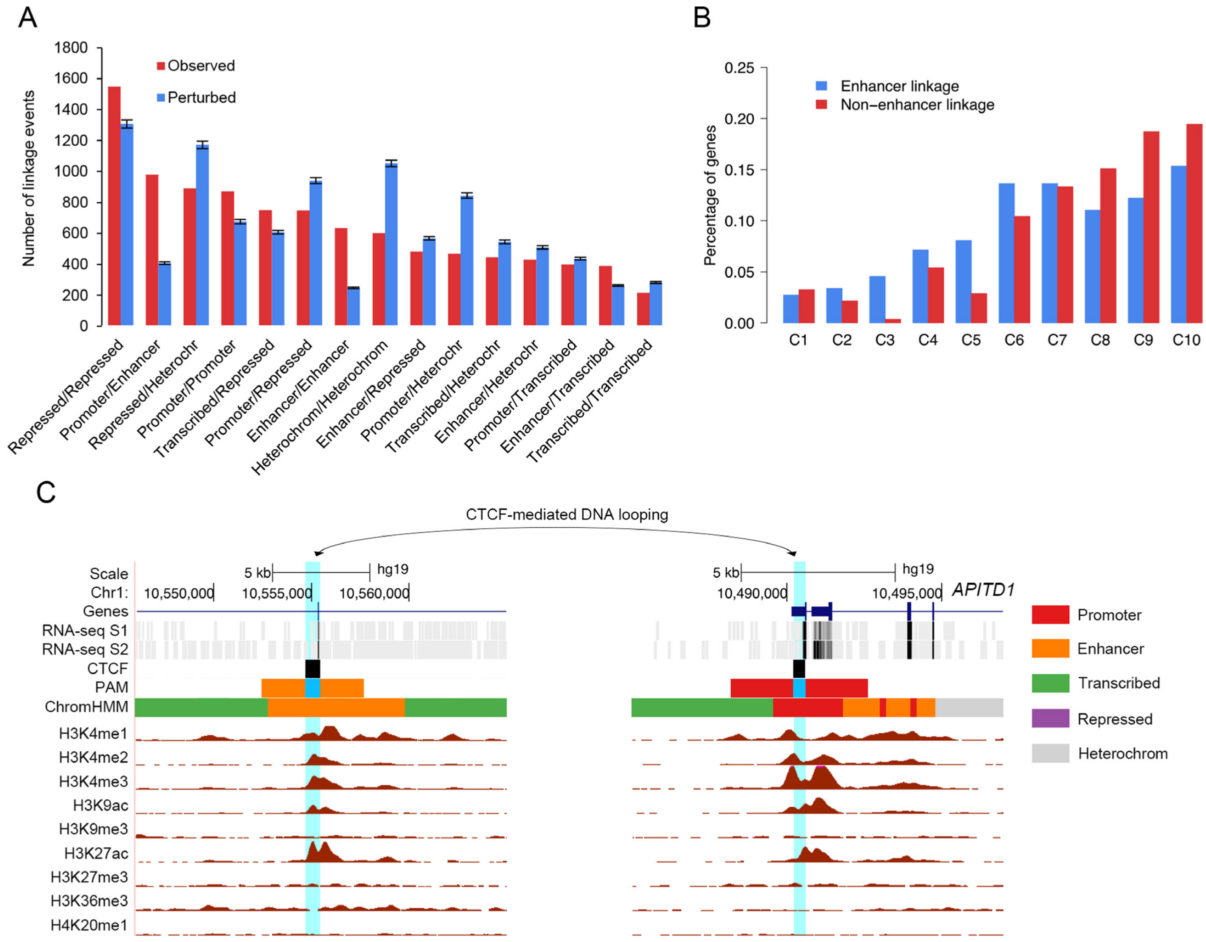


Figure 5. Identification of the enhancer linker CTCF sites in GM12878 cells. (**A**) The comparison of number of linkages between CTCF flanking regions and the perturbed CTCF flanking regions. Error bar represents the standard deviation of event count from 10 times of random perturbations. (**B**) The distributions of the enhancer linker-related genes and the control genes across 10 gene clusters. Gene were sorted by increasing expression level from C1 to C10. (**C**) An example of enhancer linker CTCF site and the linked gene *APITD1*, which has an active promoter and is actively transcribed. The bimodal distributions of H3K4me2 and H3K4me3 around TSS indicated the nucleosome occupancy by TFs. From top to bottom, the display tracks are genomic position, gene structure, RNA-seq signals of two replicates, broad chromatin states predicted by PAM and ChromHMM, ChIP-seq signals of nine histone modifications, respectively.

and SMC3 were the top two colocalized factors in these sites (Figure 6A). In contrast, we found the top 10 colocalized factors in enhancer linker sites were distinct from the global ones (Figure 6B) and POLR2A and RUNX3 were the top two colocalized factors in enhancer linker CTCF sites.

We used the mapping data of DNase I hypersensitivity sites (DHSs) to assess the chromatin accessibility of different CTCF functional sites. We found the DHS coverage varied greatly among enhancer insulator, chromatin barrier, enhancer linker and global CTCF sites (Figure 6C). Enhancer linker CTCF sites exhibited the highest chromatin accessibility among different CTCF sets and 61.5% of these sites overlapped with DHSs, distinct from the 46.4% of global set (P-value < 2.2E−16, Wilcoxon rank sum test). Chromatin barrier CTCF sites also showed relatively high accessibility, with a 55.0% overlapping rate with DHSs (P-value < 2.2E−16, Wilcoxon rank sum test). Interestingly, enhancer insulator CTCF sites showed the lowest chromatin accessibility among four CTCF sets, with a 42.7% overlapping rate, significantly lower than that of global set (P-value = 4.0E−9, Wilcoxon rank sum test). This result supported the idea that enhancer insulator blocks the enhancer-promoter interactions by reducing the local chromatin accessibility (32).

Considering the importance of CTCF orientation in chromatin looping (33), we investigated the orientation of different CTCF functional sites. We first assessed the orientation of the global CTCF sites linked by chromatin loops based on the typical CTCF-binding motif CM1. Among four possible orientations of a pair of CTCF sites on the same chromosome: (i) forward-reverse, (ii) forward-forward, (iii) reverse-reverse, and (iv) reverse-forward, the counts of convergent (forward-reverse) and divergent (reverse-forward) orientation of 11 393 CM1-containing CTCF pairs were 8704–174 (50-fold enrichment, well consistent with the 51-fold reported by Rao *et al.*) (Figure 6D, Supplementary Figure S7) (22). In enhancer insulator CTCF sites, the counts were 971–12 (81-fold enrichment, convergent versus divergent); in chromatin barrier, 1666–24 (69-fold); in enhancer linker, 381–18 (21-fold). We

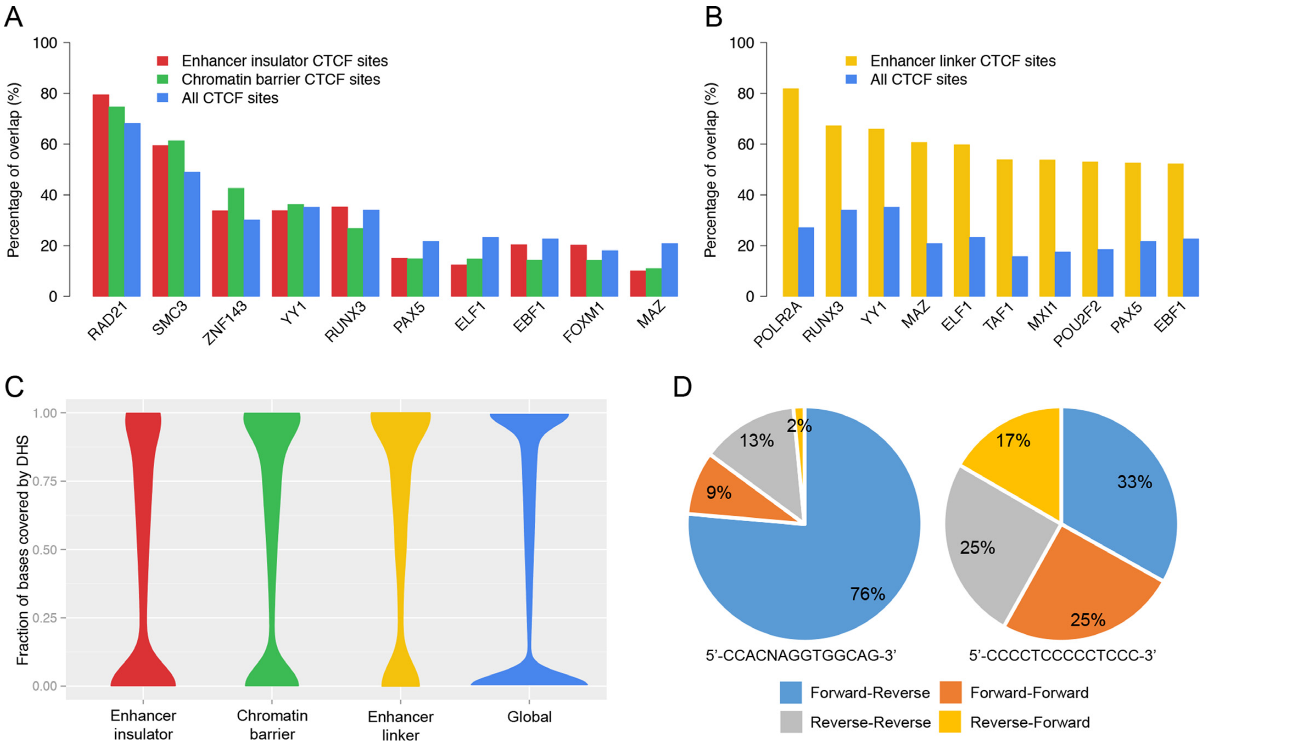


Figure 6. Characterization of the three functional roles of CTCF sites in GM12878 cells. (**A**) The top 10 colocalized TFs associated to both enhancer insulator and chromatin barrier CTCF sites using all CTCF sites as background. (**B**) The top 10 colocalized TFs associated to enhancer linker CTCF sites using all CTCF sites as background. (**C**) Violin plots of the DHS coverage of enhancer insulator, chromatin barrier, enhancer linker and global CTCF sites. (**D**) Percentages of four orientations of the 11 393 CM1-containing CTCF-pairs (left, 5′-CCACNAGGTGGCAG-3′) and 2846 CM2-containing pairs (right, 5′-CCCCTCCCCCTCCC-3′).

also investigated the CTCF orientation based on the motif CM2. We found the tendency of orientation was rather mild for CM2 (Figure 6D, Supplementary Figure S7). The counts of convergent and divergent orientation were 943–473 (2-fold enrichment) in 2846 CM2-containing CTCF pairs. In enhancer insulator, the counts were 100-40 (2.5-fold); in chromatin barrier, 177–85 (2-fold); in enhancer linker, 141–90 (1.5-fold). We were also interested to find that CTCF sites containing CM2 were not as frequently linked by chromatin loops as CM1-containing sites. Only 1384 sites were linked by chromatin loops in 10 000 randomly sampled CM2-containing CTCF sites, significantly less than the 5610 sites in 10 000 random CM1-containing sites (*P*-value < 2.2E−16, Chi-squared test). These results suggested that motif CM2 might not directly associated to the looping capacity of CTCF.

Identification of the multivalent functions of CTCF in multiple human cell types

Comparing the multivalent functions of CTCF across diverse cell types can contribute to understanding differences in regulatory mechanisms of CTCF between cell types. We applied our method to other four human cell types to identify the multivalent functions of CTCF. The four cell types include HeLaS3, HUVEC, HMEC and K562, of which both chromatin modification profiling and 3D chromatin interacting data were obtained from the ENCODE project (see Methods and Materials section). The PAM clustering was performed on the four cell types separately as we did on GM12878. The optimal number of clusters in HeLaS3, HUVEC, HMEC and K562 were 12, 12, 13 and 12 (Supplementary Figure S8). We identified 16 601, 15 340, 11 194 and 14 110 chromatin state transition events and 9725, 3273, 7197 and 17 461 chromatin state linkage events at the CTCF sites of HeLaS3, HMEC, HUVEC and K562, respectively (Figure 7A,C). Among the transition events, we identified 4188, 3219, 4595 and 4373 enhancer insulator CTCF sites and 6602, 5214, 4384 and 4840 chromatin barrier CTCF sites in the four cell types, respectively (Figure 7B and Supplementary Tables S1 and S2). Among the linkage events, we identified 979, 399, 320 and 1197 enhancer linker events (pairs of CTCF sites) in the four cell types, respectively (Figure 7D and Supplementary Table S3).

We assessed the regulatory roles of the different modes of CTCF in HeLaS3, HUVEC and K562. HMEC was excluded from this analysis because standard RNA-seq data for this cell is unavailable. Similar or even more significant trends of gene expression were observed for the different functional roles of CTCF in the three examined cell lines (Supplementary Figures S9–S11). The distributional difference of enhancer insulator-related gene set and the corresponding control gene set among ten gene classes were 2.19E−7, 8.44E−15 and 2.43E−12 (Kendall's correlation tests) in HeLaS3, HUVEC and K562, respectively. The distributional difference of chromatin barrier-related gene set and the corresponding control gene set were all < 2.2E−16 (Kendall's correlation tests) in three cell types. The distribu-

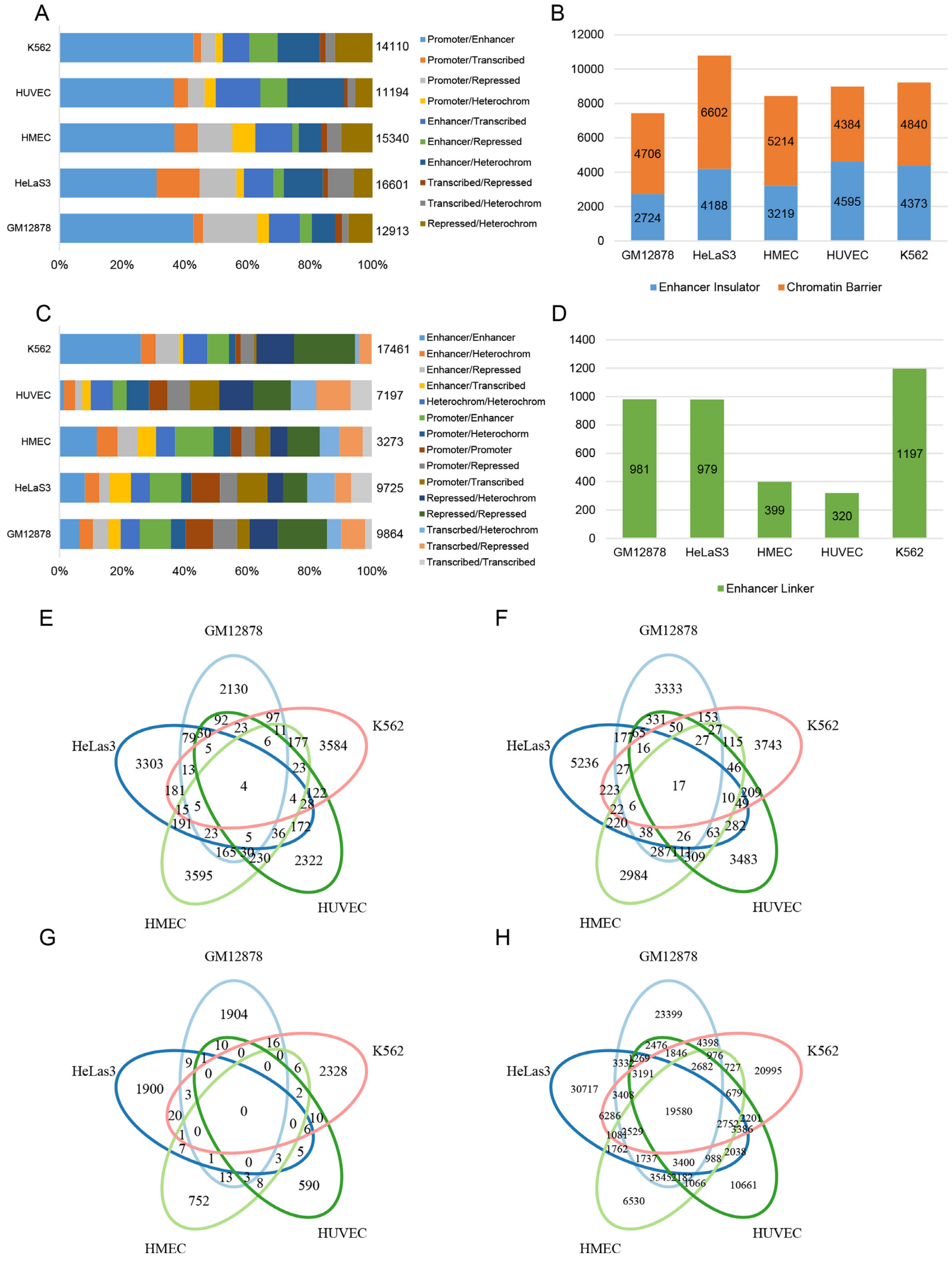


Figure 7. Identification of the three functions of CTCF in five human cell types. (**A**) Percentages of 10 types of chromatin state transitions in five cell types. The total numbers of transitions were noted at the right side of the bars. (**B**) The numbers of CTCF sites identified to be associated to the enhancer insulator and chromatin barrier functions in five cell types. (**C**) Percentages of 15 types of chromatin state linkages in five cell types. The total numbers of linkages were noted at the right side of the bars. (**D**) The numbers of CTCF sites identified to be associated to the enhancer linker function in five cell types. Venn plots of CTCF sites related to (**E**) enhancer insulator, (**F**) chromatin barrier, (**G**) enhancer linker and H) global non-redundant CTCF sites.

tional difference of enhancer linker-related gene set and the corresponding control gene set were all 4.44E−6, 1.67E−6 and 1.16E−5 (Kendall's correlation tests), respectively.

To examine the cell type specificity of the CTCF sites acting as enhancer insulator, chromatin barrier and enhancer linker, we investigated whether these sites were shared across different cell types. Two CTCF sites related to the same function and located within a distance of 1kb in two cell types were treated as a shared CTCF site by two cell types. We found that CTCF sites associated to the three functions showed high cell type specificities. Among 16 701, 21 685 and 7598 non-redundant CTCF sites associated with enhancer insulator, chromatin barrier and enhancer linker, only 261 (1.56%), 600 (2.77%) and 20 (0.263%) sites were shared by three or more cell types and 4 (0.02%), 17 (0.08%) and 0 (0%) were shared by five cell types, respectively (Figure 7E–G). In contrast, 51 686 of the 171 819 (30.1%) global non-redundant CTCF sites were shared by three or more cell types and 19 580 (11.4%) were shared by five cell types (Figure 7H). We also performed Gene Ontology (GO) enrichment analyses using DAVID (34) in the different CTCF functional sites in the five cell types. We found most enriched GO terms were associated with cell type-specific functions in enhancer insulator and enhancer linker-related genes (Supplementary Tables S4 and S5). For instance in enhancer insulator, these include lymphocyte differentiation (GM12878-specific, *P*-value = 2.24E−4), angiogenesis (HUVEC, *P*-value = 6.32E−5); in enhancer linker, B-cell apoptosis (GM12878, *P*-value = 1.2E−4), tube development (HUVEC, *P*-value = 1.52E−5). In chromatin barrier-related genes, more general cell functions were enriched, including cell morphogenesis (GM12878, HMEC, HUVEC and K562, *P*-value = 7.16E−6, 3.6E−4, 2.98E−4 and 5.48E−3, respectively) and neuron differentiation (GM12878, HMEC, HUVEC, *P*-value = 1.69E−6, 4.08E−5 and 7.31E−6, respectively, Supplementary Table S6). These results indicated that enhancer insulator and enhancer linker CTCF sites are more cell type-specific regulators compared to chromatin barrier sites, highlighting their roles in the specific gene expression program of each cell type.

DISCUSSION

We present here the first genome-wide survey and characterization of the multivalent functions of CTCF in the human genome. We defined two functional modes of CTCF: the insulator mode and the linker mode, and identified three important regulatory functions related to them: enhancer insulator, chromatin barrier and enhancer linker, using the chromatin states of CTCF flanking regions and 3D chromatin interactions. We applied our method to the epigenomic datasets of five human cell types and identified a large set of CTCF sites associated with these three functions. On average, there are ∼10 000 CTCF binding sites associated with these three functions in one cell line, demonstrating the multivalent functions of CTCF are ubiquitous in the human genome. We also found the consensus binding motifs and colocalized TFs of CTCF related to the two functional modes were significantly different, suggesting the mechanisms of its multivalent regulatory roles may not be the same. Besides, we found the cell type specificities of the CTCF sites related to these functions were extremely high, indicating their important regulatory roles in cell type specific gene expression.

Although we have identified ∼46 000 sites related to the three important regulatory functions of CTCF in five cell lines, these sites only accounts for a portion (45 984/171 819, 26.8%) of all CTCF binding sites identified by ChIP-seq. A major reason is that the roles of CTCF in the human genome are complex and not limited to the three identified functions. For example, besides linkages between promoter and enhancer, other types of linkages were also significantly enriched at CTCF sites, including promoter and promoter, transcribed and repressed, enhancer and enhancer and enhancer and transcribed. These linkages may be associated to other functions of CTCF, such as alternative promoter selection (17) and alternative splicing regulation (18). However, further information on the multiple promoter usage and exon selection is needed to validate these ideas. We think that our identification of the three functions of CTCF represents the tip of the iceberg, and more studies on the multivalent functions of CTCF will soon emerge.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Basic Research Project (973 program) [2012CB518200] (http://www.973.gov.cn/); General Program [31401141, 81573251] of the Natural Science Foundation of China (www.nsfc.gov.cn); State Key Laboratory of Proteomics of China [SKLP-Y201303, SKLP-O201104, SKLP- K201004] (www.bprc.ac.cn); Special Key Programs for Science and Technology of China [2012ZX09102301-016]. Funding for open access charge: National Basic Research Project (973 program) [2012CB518200] (http://www.973.gov.cn/); General Program [31401141, 81573251] of the Natural Science Foundation of China (www.nsfc.gov.cn); State Key Laboratory of Proteomics of China [SKLP-Y201303, SKLP-O201104, SKLP- K201004] (www.bprc.ac.cn); Special Key Programs for Science and Technology of China [2012ZX09102301-016].

Conflict of interest statement. None declared.

REFERENCE

1. Ohlsson,R., Renkawitz,R. and Lobanenkov,V. (2001) CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet.: TIG*, **17**, 520–527.
2. 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.
3. 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.
4. Lobanenkov,V.V., Nicolas,R.H., Adler,V.V., Paterson,H., Klenova,E.M., Polotskaja,A.V. and Goodwin,G.H. (1990) A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene*, **5**, 1743–1753.

5. Kellum,R. and Schedl,P. (1991) A position-effect assay for boundaries of higher order chromosomal domains. *Cell*, **64**, 941–950.
6. Cuddapah,S., Jothi,R., Schones,D.E., Roh,T.Y., Cui,K. and Zhao,K. (2009) Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.*, **19**, 24–32.
7. Essafi,A., Webb,A., Berry,R.L., Slight,J., Burn,S.F., Spraggon,L., Velecela,V., Martinez-Estrada,O.M., Wiltshire,J.H., Roberts,S.G. *et al.* (2011) A wt1-controlled chromatin switching mechanism underpins tissue-specific wnt4 activation and repression. *Dev. Cell*, **21**, 559–574.
8. Bell,A.C., West,A.G. and Felsenfeld,G. (1999) The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*, **98**, 387–396.
9. Wood,A.M., Van Bortle,K., Ramos,E., Takenaka,N., Rohrbaugh,M., Jones,B.C., Jones,K.C. and Corces,V.G. (2011) Regulation of chromatin organization and inducible gene expression by a Drosophila insulator. *Mol. Cell*, **44**, 29–38.
10. Xie,X., Mikkelsen,T.S., Gnirke,A., Lindblad-Toh,K., Kellis,M. and Lander,E.S. (2007) Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 7145–7150.
11. 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.
12. Sanyal,A., Lajoie,B.R., Jain,G. and Dekker,J. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.
13. Handoko,L., Xu,H., Li,G., Ngan,C.Y., Chew,E., Schnapp,M., Lee,C.W., Ye,C., Ping,J.L., Mulawadi,F. *et al.* (2011) CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.*, **43**, 630–638.
14. Phillips,J.E. and Corces,V.G. (2009) CTCF: master weaver of the genome. *Cell*, **137**, 1194–1211.
15. Chaumeil,J. and Skok,J.A. (2012) The role of CTCF in regulating V(D)J recombination. *Curr. Opin. Immunol.*, **24**, 153–159.
16. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
17. Guo,Y., Monahan,K., Wu,H., Gertz,J., Varley,K.E., Li,W., Myers,R.M., Maniatis,T. and Wu,Q. (2012) CTCF/cohesin-mediated DNA looping is required for protocadherin alpha promoter choice. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 21081–21086.
18. Shukla,S., Kavak,E., Gregory,M., Imashimizu,M., Shutinoski,B., Kashlev,M., Oberdoerffer,P., Sandberg,R. and Oberdoerffer,S. (2011) CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, **479**, 74–79.
19. Nakahashi,H., Kwon,K.R., Resch,W., Vian,L., Dose,M., Stavreva,D., Hakim,O., Pruett,N., Nelson,S., Yamane,A. *et al.* (2013) A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep.*, **3**, 1678–1689.
20. Consortium,E.P. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
21. Guttman,M., Garber,M., Levin,J.Z., Donaghey,J., Robinson,J., Adiconis,X., Fan,L., Koziol,M.J., Gnirke,A., Nusbaum,C. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, **28**, 503–510.
22. Rao,S.S., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
23. Rosenbloom,K.R., Armstrong,J., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.
24. Lu,Y., Qu,W., Shan,G. and Zhang,C. (2015) DELTA: a distal enhancer locating tool based on AdaBoost algorithm and shape features of chromatin modifications. *PLoS One*, **10**, e0130622.
25. Trapnell,C., Pachter,L. and Salzberg,S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
26. Roberts,A., Pimentel,H., Trapnell,C. and Pachter,L. (2011) Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, **27**, 2325–2329.
27. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
28. Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.
29. Hoffman,M.M., Buske,O.J., Wang,J., Weng,Z., Bilmes,J.A. and Noble,W.S. (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.
30. Levanon,D., Eisenstein,M. and Groner,Y. (1998) Site-directed mutagenesis supports a three-dimensional model of the runt domain. *J. Mol. Biol.*, **277**, 509–512.
31. Kim,T.H., Abdullaev,Z.K., Smith,A.D., Ching,K.A., Loukinov,D.I., Green,R.D., Zhang,M.Q., Lobanenkov,V.V. and Ren,B. (2007) Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell*, **128**, 1231–1245.
32. Ong,C.T. and Corces,V.G. (2011) Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.*, **12**, 283–293.
33. Guo,Y., Xu,Q., Canzio,D., Shou,J., Li,J., Gorkin,D.U., Jung,I., Wu,H., Zhai,Y., Tang,Y. *et al.* (2015) CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell*, **162**, 900–910.
34. Huang,D.W., Sherman,B.T., Tan,Q., Kir,J., Liu,D., Bryant,D., Guo,Y., Stephens,R., Baseler,M.W., Lane,H.C. *et al.* (2007) DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.*, **35**, W169–W175.