Cell Research (2017) :1365-1377.

www.nature.com/cr

ORIGINAL ARTICLE

Molecular mechanism of directional CTCF recognition of a diverse range of genomic sites

Maolu Yin[1, 2], Jiuyu Wang[1], Min Wang[1], Xinmei Li[1, 2], Mo Zhang[3, 4, 5], Qiang Wu[3, 4, 5], Yanli Wang[1, 2, 6]

[1]Key Laboratory of RNA Biology, CAS Center for Excellence in Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China; [2]University of Chinese Academy of Sciences, Beijing 100049, China; [3]Center for Comparative Biomedicine, MOE Key Laboratory of Systems Biomedicine, Institute of Systems Biomedicine, Collaborative Innovative Center of Systems Biomedicine, SCSB, Shanghai Jiao Tong University (SJTU), Shanghai 200240, China; [4]State Key Laboratory of Oncogenes and Related Genes, SJTU Medical School, Shanghai 200240, China; [5]School of Life Sciences and Biotechnology, SJTU, Shanghai 200240, China; [6]Collaborative Innovation Center of Genetics and Development, Shanghai 200438, China

CTCF, a conserved 3D genome architecture protein, determines proper genome-wide chromatin looping interactions through directional binding to specific sequence elements of four modules within numerous CTCF-binding sites (CBSs) by its 11 zinc fingers (ZFs). Here, we report four crystal structures of human CTCF in complex with CBSs of the protocadherin (*Pcdh*) clusters. We show that directional CTCF binding to cognate CBSs of the *Pcdh* enhancers and promoters is achieved through inserting its ZF3, ZFs 4-7, and ZFs 9-11 into the major groove along CBSs, resulting in a sequence-specific recognition of module 4, modules 3 and 2, and module 1, respectively; and ZF8 serves as a spacer element for variable distances between modules 1 and 2. In addition, the base contact with the asymmetric "A" in the central position of modules 2-3, is essential for directional recognition of the CBSs with symmetric core sequences but lacking module 1. Furthermore, CTCF tolerates base changes at specific positions within the degenerated CBS sequences, permitting genome-wide CTCF binding to a diverse range of CBSs. Together, these complex structures provide important insights into the molecular mechanisms for the directionality, diversity, flexibility, dynamics, and conservation of multivalent CTCF binding to its cognate sites across the entire human genome.

Keywords: CTCF; zinc finger protein; 3D genome; topological domains; protein-DNA interaction; higher-order chromatin structure

Cell Research (2017) **27**:1365-1377. doi:10.1038/cr.2017.131; published online 27 October 2017

Introduction

The linear human genome of about 2 m in length is compacted by at least 300 000-fold in a hierarchical manner to fit into the cell nucleus. The higher-order genome organization is intricately related to DNA replication, repair, rearrangement and recombination, RNA processing, and in particular, developmental regulation of gene expression through long-distance chromatin looping interactions between promoters, enhancers, silencers and insulators [1-7].

Correspondence: Yanli Wang[a], Qiang Wu[b]
[a]E-mail: ylwang@ibp.ac.cn
[b]E-mail: qwu123@gmail.com
Received 31 July 2017; revised 9 September 2017; accepted 11 September 2017; published online 27 October 2017

Zinc finger (ZF) is a small well-conserved zinc-binding moiety that is very abundant in the human proteome [8-11]. There are more than 600 or about 3% of genes in human genome that encode ZF proteins [11]. ZFs may interact with diverse partners of proteins, RNAs, and DNAs; however, the most prominent role of ZF proteins is to bind short regions of DNA duplex and to read DNA code for proper expression of genetic information during development and cell division [9-11]. In addition, engineered ZFs have been applied to genome editing [11]. Although great progress has been made for recognition of DNA duplex by ZFs, the molecular mechanisms of specific recognition of genomic element by proteins containing multiple tandem ZFs, such as the CCCTC-binding factor (CTCF), remain poorly understood.

CTCF, a transcription factor containing 11 DNA-binding ZFs, is central for higher-order genome architecture

[5, 6, 12-16]. CTCF is also the prominent insulator-binding protein in mammals, and its insulation activities are intricately related to the location and relative orientation of its binding sites, called CBS [13, 15, 17]. CBSs are enriched at topologically associated domain (TAD) boundaries that function as enhancer-blocking insulators, thus restricting enhancer activity within a particular TAD [15, 17-19]. Interestingly, the boundary CBS pairs between neighboring TADs tend to be organized in a reverse-forward divergent orientation [3, 20, 21]. In contrast, interacting CBS pairs within a TAD tend to be organized in a forward-reverse convergent orientation [1, 3, 20, 22]. Thus, the relative orientation and location of genome-wide CBSs determine the directionality of long-distance DNA looping interactions and higher-order chromatin organization [3, 4, 23, 24].

Large numbers of CBSs spread throughout mammalian genomes, and are found proximal to or within promoters, enhancers, as well as silencers [3, 25-27]. CTCF, together with the associated architectural cohesin complex, determines proper genome-wide chromatin looping interactions by directional and simultaneous binding to numerous CBSs with combinatorial usage of the 11 ZFs of each CTCF (Figure 1A) [3, 4, 13, 14, 23, 28, 29]. Mammalian CBS sequences are highly diverse and comprise four degenerate modules (modules 1-4), with a flexible distance between modules 1 and 2 [3, 13, 20, 27, 28, 30]. In addition, most CBSs lack module 1, containing only the GC-rich modules 2-4, which are often of palindromic nature. Intriguingly, CBSs are recognized by CTCF in

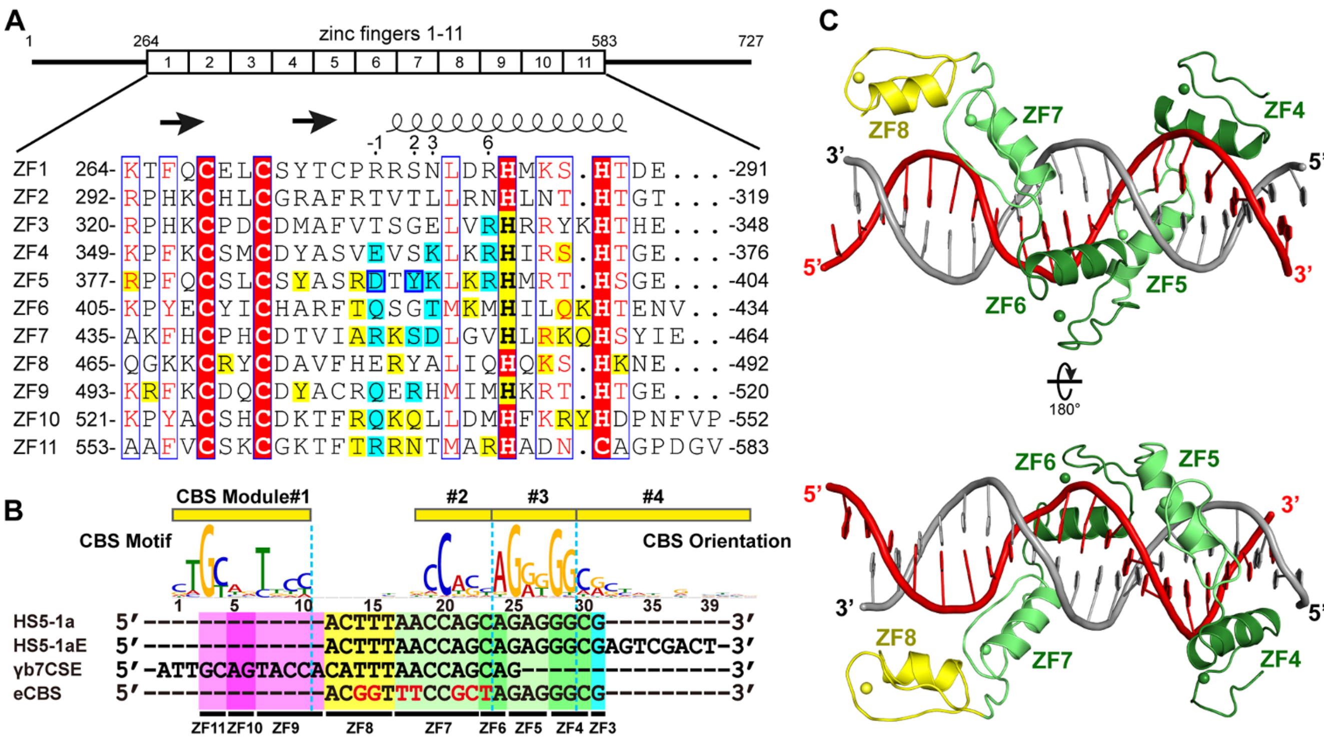


Figure 1 Structure of CTCF ZFs in complex with *Pcdh* enhancer *CBS*. **(A)** Schematic diagram of CTCF. The tandem ZFs 1-11 have been aligned to clearly show their secondary structure and the similarity of their amino acid sequence. The β-strands and α-helix are indicated by arrows and a helix above the alignment, respectively. The amino acids that contact the DNA bases directly or in a water-mediated manner are highlighted by cyan, or the cyan background with a blue square, respectively. The amino acids that contact the phosphate backbones of DNA are highlighted by yellow. The positions of −1, 2, 3, and 6 relative to the start of α-helices are labeled immediately above the alignment. **(B)** The *Pcdh* enhancer and promoter CBS modules and the DNA sequences used for the crystallization. The backgrounds of nucleotides are similar to the color codes of their interacting ZFs as shown in **C** and in all subsequent Figures. The nucleotides within the *Pcdhα* promoter *eCBS* shown in red are diversified nucleotides from those of the *Pcdhα* enhancer *HS5-1a* and of the *Pcdhγ* promoter *γb7*CSE. The 5′ nucleotide “A” of *Pcdhγ* promoter *γb7CSE* was added artificially for technicality of crystallization. **(C)** The crystal structure of ZFs4-8-*HS5-1a* complex. ZFs 4-7 are shown in light green and dark green alternatingly, and ZF8 is shown in yellow. The Crick and Watson DNA strands of the *Pcdhα* enhancer CBS *HS5-1a* are presented in red and grey, respectively. Zn^{2+} is shown as sphere.

a single-directional fashion [3, 20, 27, 28, 31, 32]. The specific and directional recognition of CBSs by CTCF is quintessential for 3D genome organization in the cell nucleus and mutations of both CTCF and CBSs have been reported in various human diseases [33-35]. However, due to the lack of atomic structures of the CTCF-CBS complex, the precise mechanisms of directional CTCF recognition of diverse CBSs remain unknown.

Three human protocadherin (*Pcdh*) clusters (*Pcdhα*, *β* and *γ*) are organized in a single large locus of about one megabase DNA in the chromosome 5, and the *Pcdh α* and *γ* clusters comprise variable and constant regions (Supplementary information, Figure S1) [36]. The encoded Pcdh proteins play important roles in processes essential for neural circuit assembly in the brain such as individual neuronal identity, isoneuronal self-avoidance and even spacing, and heteroneuronal neurite tiling and co-existence [37, 38]. These *Pcdh* clusters are model genes for investigating 3D genome folding and gene regulation (Supplementary information, Figure S1A) [3, 20]. A repertoire of *Pcdh* promoter CBSs (*CSE* (CTCF-binding conserved sequence element) and *eCBS* (exonic CBS in the *Pcdhα* variable exons)) within variable regions are in the forward orientation, namely from modules 1 to 4 (Supplementary information, Figure S1B-S1F), whereas several CBSs within the enhancer and super-enhancer located downstream of the *Pcdhα* and *γ* clusters, respectively, are in the reverse orientation (from modules 4 to 1; Supplementary information, Figure S1G) [3, 20, 39]. Specific contacts between these convergent forward-reverse CBSs are essential for proper long-distance chromatin interactions between the distal enhancer and its target promoters, activating a set of *Pcdh* promoters in a cell-specific manner in the brain [3, 20].

Here we report the structures of various CTCF ZFs in complex with a set of representative CBSs within enhancers and promoters of the human *Pcdh* gene clusters. Our structural studies reveal that ZFs 4-7 read modules 3 and 2, ZF3 binds module 4, and ZFs 9-11 recognize module 1 in a highly sequence-specific manner; as a result, directionality is imposed on the interaction between CTCF and CBS.

Results

Overall structure of the ZFs 4-8-CBS complex

To understand how CTCF recognizes diverse nucleotide sequences within the modules 2-3 of different *Pcdh* CBSs, and to map which ZFs interact with individual CBS bases, we determined the crystal structure of CTCF ZFs 4-8 in complex with the core sequence of the CBS *HS5-1a* of the *Pcdhα* enhancer (Figure 1B and 1C). The crystal structure of ZFs4-8-*HS5-1a* was solved by single-wave-length anomalous diffraction (SAD) method at 2.0 Å resolution (Supplementary information, Table S1). The 19-bp *Pcdh HS5-1a* DNA duplex used for crystallization consists of modules 2-3 with a one-nucleotide overhang at both 5′ ends (Supplementary information, Figure S2A and S2B). For the purpose of DNA alignment, the nucleotides are counted from 12 to 31 (Figure 1B). The ZFs4-8-*HS5-1a* crystal structure shows that each of ZFs 4-8 is folded into a canonical ββα domain [10, 40], binding a zinc ion sandwiched between the two-stranded antiparallel β-sheet on one side, and an α-helix on the other side. ZFs 4-8 stretch along the entire length of the 19-bp core CBS duplex, mainly interacting with CBS *HS5-1a* on the Crick strand (in red) of the *Pcdh* enhancer (Figure 1C). ZFs 4-8 are anchored to CBS by an extensive network of hydrogen bonds to the phosphate backbone and bases, as well as by electrostatic interactions between ZFs and the phosphate backbone (Supplementary information, Figure S2C and S2D). ZFs 4-7 wrap around the DNA duplex with the N-termini of their α-helices inserted into the major groove, contacting 14 bp of the *HS5-1a* core sequence in total. In contrast, ZF8 (in yellow) runs along one side of *HS5-1a*, forming an extended structure outside the CBS DNA duplex (Figure 1C). Importantly, ZFs bind the enhancer *HS5-1a* spanning modules 2-3 in a one-directional manner, in the order of ZFs 8-4.

ZFs 4-7 recognize CBS modules 3-2 in a single-directional manner

The crystal structure of ZFs4-8-*HS5-1a* complex provides critical insights into the molecular mechanism responsible for CTCF recognition of the conserved CBS core sequences (Figure 1B). Module 2 is mainly recognized by ZFs 6-7. Specifically, ZF7 recognizes nucleotides 20-22 in a sequence-specific manner (Figures 2A and 3A). Nucleotide C20, the most conserved nucleotide of module 2, forms hydrogen bonds with the side-chain of Asp451. Bases of A21 and G22 contact the side-chain of Arg448. ZF6 makes base contacts with C23 and A24 via Thr421 and Gln418, respectively. The human genome contains fully palindromic CBSs with the core sequence of “CCACCAGGTGG”, and A24 is located at the center of this fully palindromic core of modules 2-3. The specific recognition of this central asymmetrical A24 by Gln418 is likely critical for CTCF to discriminate the CBS orientation. The electrophoretic mobility shift assay (EMSA) showed that the substitution of Gln418 by Ala greatly reduces the affinity between ZFs 4-8 and the DNA probe (Supplementary information, Figure S2E). In addition, mutation of A24 to any other three nucleotides reduces the binding affinity with ZFs 4-8 (Supplementary information, Figure S2F). However, the mutation of A24 to T24 has the least effect, consistent with the idea that A24T may switch the directionality of CTCF binding to CBSs with symmetric core sequences and lack of module 1. Finally, the fact that point mutations of A24 to any other three nucleotides are frequently found in cancer genomes [34] of human patients suggests that A24 is critical for CTCF cellular function. Together, these data suggest that A24 is important for directional CTCF binding.

Module 3 is primarily recognized by ZFs 4-5 (Figures 2B and 3A). In particular, ZF5 contacts nucleotides 25-27, and sequence-specifically recognizes G25, A26, and T26′ through Arg396, Lys393, Asp390 and Tyr392 directly or in a water-mediated manner. G25, the most

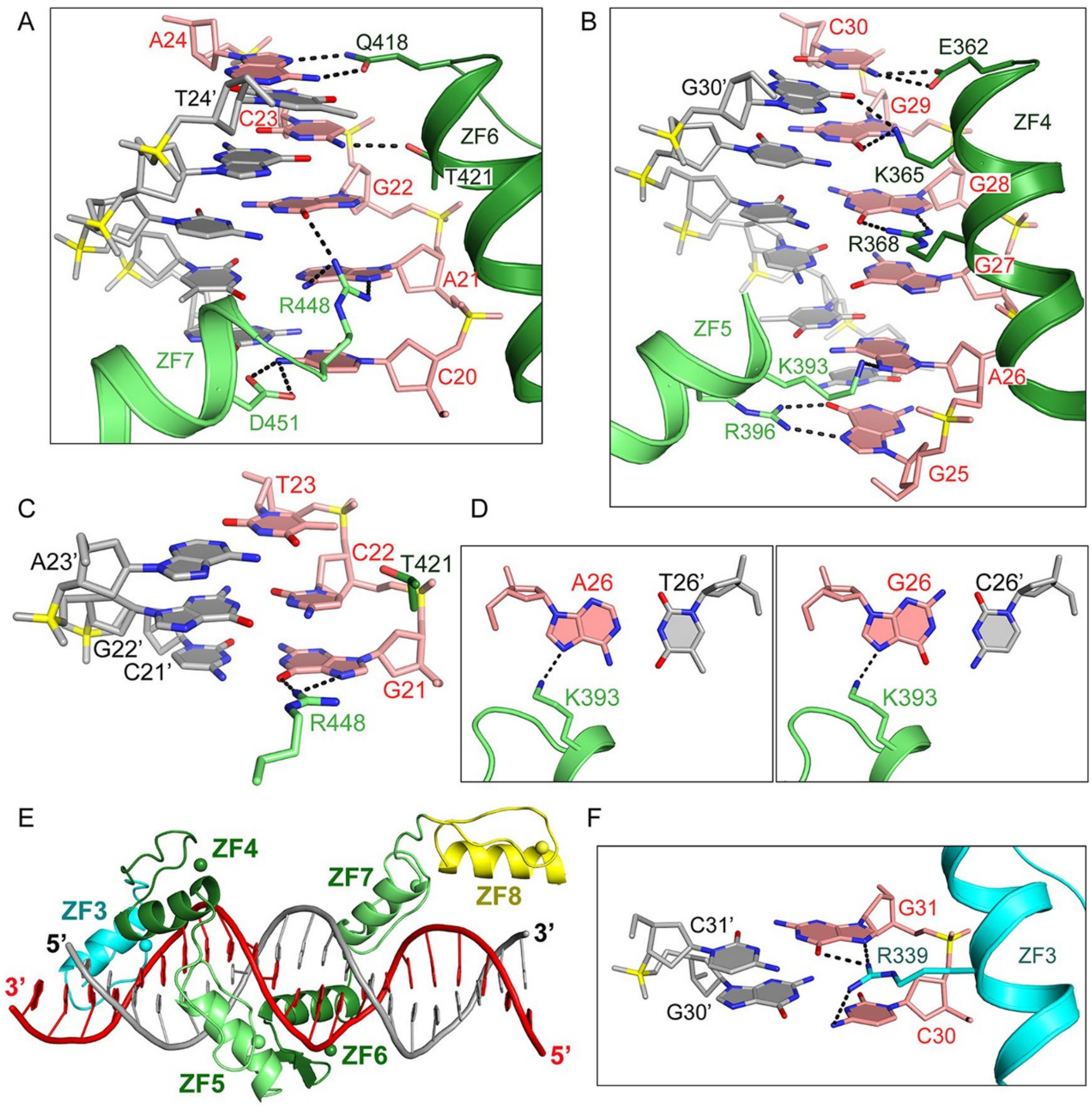


Figure 2 CTCF ZFs 4-7 and ZF3 sequence-specifically recognize *Pcdh* enhancer CBS *HS5-1a* modules 2-3 and module 4, respectively. **(A)** Detailed view of the interactions between the *Pcdhα* enhancer CBS *HS5-1a* (mainly module 2) and ZFs 6-7 in the ZFs4-8-*HS5-1a* complex. The oxygen and nitrogen atoms within the base are shown in red and blue, respectively. The phosphate atoms are colored in yellow, and all other atoms are colored in light red (Watson strand) and grey (Crick strand), respectively. **(B)** Detailed view of the interactions between the CBS module 3 and ZFs 4-5. **(C)** Magnified view of the interactions between nucleotides 21-23 and ZF7 in the ZFs4-8-*eCBS* complex. **(D)** The interaction of A26 with the side-chain of Lys393 in the ZFs4-8-*HS5-1a* complex (left panel), and the interaction of G26 with Lys393 from the modeled structure (right panel). **(E)** Crystal structure of ZFs 3-8 bound to CBS *HS5-1aE*. ZF3 is shown in cyan. Color codes of ZFs 4-8 and DNA are identical to those in Figure 1C. **(F)** The sequence-specific interactions between Arg339 of ZF3 and nucleotides 30-31.

conserved nucleotide within module 3, forms a pair of hydrogen bonds with the side-chain of Arg396. ZF4 recognizes G28-G29-C30 in a sequence-specific fashion. G28, one of the most conserved nucleotide in module 3, is read by Arg368 through base contacts (Figure 2B). The side-chain of Lys365 forms hydrogen bonds with the bases of G29 and G30′. Finally, Glu362 forms hydrogen bonds with the base of C30. As shown in Supplementary information, Figure S2E, mutations of K365A, R368A, or R396A decrease the DNA binding affinity (as assessed by EMSA), validating that these protein-DNA interactions are significant for the association between CTCF and module 3.

Like the canonical ZFs [10, 40], the residues at position –1 (the amino acid immediately preceding the α-helix) within ZFs 4-7 are critical for reading CBS by contacting the bases of CBS *HS5-1a* directly or in a water-mediated fashion (Figure 1A). Similarly, the amino acids at positions 2-3 within ZFs 4-7 and at position 6 of ZFs 3-5 also form base contacts with CBS. However, the residues at positions 2-3 and 6 of ZFs 8-11 are not involved in base recognition, except the Arg508 at position 2 of ZF9, which interacts with G9′ in a base-specific manner (Figures 1A and 3A; discussed below). Interestingly, each of the three most conserved nucleotides, namely C20, G25 and G28, forms two hydrogen bonds with Asp451, Arg396 and Arg368, respectively. These interactions reveal how ZFs 4-7 base-specifically recognize the three highly-conserved nucleotides, and explain why CBS is highly conserved within modules 2-3. In addition, ZF6 reads A24, the single asymmetric central nucleotide within the otherwise palindromic core sequence, enabling ZFs 4-7 to bind modules 2-3 in a single-directional manner.

Structural basis for the diversity of CTCF-CBS binding

To understand the molecular mechanism behind the ability of CTCF to bind a wide range of target CBS sequences, we used the *eCBS* sequence from the *Pcdhα8* promoter region (Figure 1B), alongside which there are 12 highly homologous CBS sites (Supplementary information, Figures S1 and S3A). We solved the structure of ZFs 4-8 in complex with the *Pcdhα8 eCBS* (Supplementary information, Table S1). The overall structure of ZFs 4-8 bound to the *eCBS* sequence of the *Pcdhα8* promoter is similar to that bound to the CBS *HS5-1a* of the *Pcdhα* enhancer (Supplementary information, Figure S3B), suggesting that ZFs 4-8 recognize diverse CBS repertoire of the *Pcdh* promoters in a single-directional manner similar to that of the *Pcdh* enhancers.

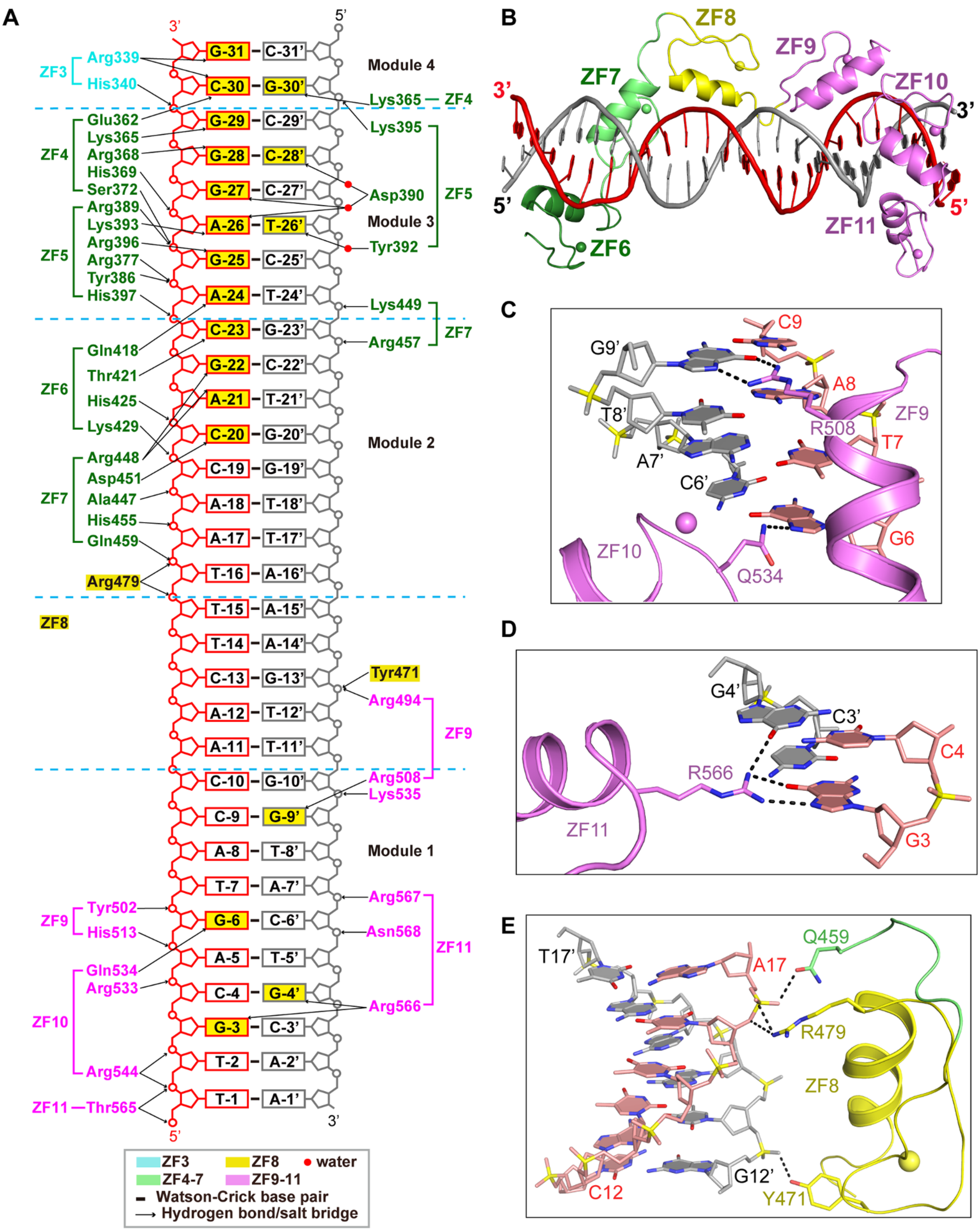
Next, we compared the sequence-specific interactions between ZFs 4-7 and these two CBSs, which have different nucleotides at positions 14-15, 17-18, and 21-23 (highlighted in red in Figure 1B). In the *Pcdhα8* promoter *eCBS*-bound complex, G21 forms two hydrogen bonds with Arg448 (Figure 2C), which is similar to the two hydrogen bonds between A21 and Arg448 observed in the ZFs4-8-*HS5-1a* enhancer complex (Figure 2A). However, exchange of the enhancer CBS *HS5-1a* nucleotides 22-23 to *Pcdhα8* promoter *eCBS* 22-23 abrogates the base-specific interactions with ZFs 6-7 (compare Figure 2C to 2A), suggesting that interactions between nucleotides 22-23 and ZFs 6-7 have no significant effect on the binding of CTCF to CBS. Thus, ZFs 6-7 are able to bind genome-wide CBSs of diverse sequences at positions 21-23.

To better understand the binding affinity of ZFs 4-7 to CBS containing diversified nucleotides at positions 26-27, we modeled their structure by replacing nucleotides A26-G27 with G26-T27. Both A26 and G26 contain an N7 atom within their bases, and form one hydrogen bond with Lys393 of ZF5 (Figure 2D), indicating that the 26th nucleotide can either be A or G. The base of nucleotide G27 has no contact with any ZF (Figure 2B), and the substitution by T27 (Figure 1B) or any other nucleotide has therefore no effect on the recognition by CTCF. These data indicate that ZF5 is able to bind numerous genome-wide CBSs containing A/G26 and G/T27.

Together, our analysis shows that ZFs 4-7 tolerate the substitution of nucleotides 21-23 and 26-27, explaining why ZFs 4-7 recognize diverse CBSs found throughout the human genome.

ZF3 recognizes module 4 in a sequence-specific manner

To determine the structural basis for the recognition of module 4, which is located downstream of module 3, we solved the crystal structure of ZFs 2-8 in complex with the extended length of the *Pcdh* enhancer CBS *HS5-1a* (hereafter referred to as *HS5-1aE*; Supplementary information, Figures S1 and S3C). The structures of ZFs 4-8 exhibit high similarity when bound to either the core sequences or the extended length of the enhancer CBS *HS5-1a*. In the ZFs2-8-*HS5-1aE* complex, ZF2 is disordered, whereas the α-helix of ZF3 is positioned in the major groove of module 4 (Figure 2E). This suggests that ZF3 is important for CBS module 4 recognition by CTCF. Furthermore, the side-chain of Arg339 in ZF3 forms hydrogen bonds with the bases of C30 and G31 (Figure 2F). Interestingly, C30 and G31 is located within the "CGCT" box (Supplementary information, Figure S1), the core sequence within CSEs in the promoter region of members of the three mammalian *Pcdh* gene clusters [39]. We noted that the conformity of the *Pcdh* promoter CSE modules 2-3 to the genome-wide type 1 CBSs (Supplementary information, Figure S4) is weaker


A
3'
5'
Arg339
His340
ZF3
G-31 – C-31'
C-30 – G-30'
Module 4
Lys365 — ZF4
Glu362
Lys365
Arg368
His369
Ser372
ZF4
G-29 – C-29'
G-28 – C-28'
G-27 – C-27'
Lys395
Asp390
ZF5
Arg389
Lys393
Arg396
Arg377
Tyr386
His397
ZF5
A-26 – T-26'
G-25 – C-25'
A-24 – T-24'
Module 3
Tyr392
Lys449
Arg457
ZF7
Gln418
Thr421
His425
Lys429
ZF6
C-23 – G-23'
G-22 – C-22'
A-21 – T-21'
C-20 – G-20'
Module 2
Arg448
Asp451
Ala447
His455
Gln459
ZF7
C-19 – G-19'
A-18 – T-18'
A-17 – T-17'
T-16 – A-16'
Arg479
ZF8
T-15 – A-15'
T-14 – A-14'
C-13 – G-13'
A-12 – T-12'
A-11 – T-11'
Tyr471
Arg494
ZF9
C-10 – G-10'
Arg508
Lys535
C-9 – G-9'
A-8 – T-8'
Module 1
T-7 – A-7'
Tyr502
His513
ZF9
G-6 – C-6'
A-5 – T-5'
Arg567
Asn568
ZF11
Gln534
Arg533
ZF10
C-4 – G-4'
G-3 – C-3'
Arg566
Arg544
T-2 – A-2'
T-1 – A-1'
ZF11 — Thr565
5'
3'
ZF3
ZF8
water
ZF4-7
ZF9-11
Watson-Crick base pair
Hydrogen bond/salt bridge
B
ZF8
ZF9
ZF7
ZF10
3'
3'
5'
5'
ZF6
ZF11
C
C9
G9'
A8
T8'
ZF9
R508
T7
A7'
C6'
ZF10
G6
Q534
D
G4'
C3'
C4
R566
ZF11
G3
E
T17'
Q459
A17
R479
ZF8
G12'
C12
Y471

than that of the *Pcdh* enhancer CBS *HS5-1a* (Supplementary information, Figure S1B-S1G). This observation suggests that the interactions between ZF3 and module 4 are able to compensate for the weaker interactions between CTCF and modules 2-3 within diverse CSEs, to ensure the proper binding between CTCF and those CSEs that have suboptimal modules 2-3. In addition, unlike the roughly palindromic sequence of modules 2-3, the sequence of module 4 is asymmetric. Thus, the binding of ZF3 to module 4 further determines the directionality of CTCF binding to numerous oriented CBSs, which is central for proper chromatin looping during 3D genome folding.

ZFs 9-11 recognize module 1

In addition to the core sequence (modules 2-4), ~15% of CBSs contain additional module 1 at the upstream of the core sequence (Supplementary information, Figure S4) [27, 28]. To understand how module 1 is recognized by CTCF, we determined the crystal structure of ZFs 6-11 in complex with a chimeric CBS containing module 1 of CSE of the *Pcdhγb7* promoter and the core sequence of the *Pcdhα HS5-1a* enhancer (hereafter referred to as *γb7CSE*; Figure 3B and Supplementary information, Figure S5A and S5B). Here, nucleotides 14-25 are identical to the corresponding enhancer *HS5-1a* core sequences, and the nucleotides 1-13 are from the *Pcdhγb7* promoter *CSE* bearing CBS module 1 (except for nucleotide "A" at position 8, which replaces the natural genomic "C" purely for crystallization purpose, Figure 1B). It was demonstrated that ZFs 6-11 can actually bind to the chimeric sequence as revealed by EMSA analysis (Supplementary information, Figure S5C). In the complex of ZFs 6-11 bound to the *Pcdhγb7* promoter *CSE*, ZFs 7-8 and its bound DNA (magenta) display a similar structure to the ZFs4-8-*HS5-1a* enhancer complex (green), but ZF6 exhibits subtle differences upon superpositioning ZFs 7-8 (Supplementary information, Figure S5B).

The structure of the ZFs6-11-*γb7CSE* promoter complex presented here shows how CTCF binds to CBS module 1 in a sequence-specific manner. Similar to ZFs 4-6, ZFs 9-11 wrap around the DNA duplex with their α-helix inserted into the major groove of module 1 (Figure 3B). In contrast to ZFs 4-6, in which each ZF recognizes 2-3 bases, each of ZFs 9-11 only reads 1-2 bases (Figure 3A). Arg508 of ZF9 and Gln534 of ZF10 form base contacts with G9′ and G6, respectively (Figure 3C). Arg566 of ZF11 forms two hydrogen bonds with the base of G3, which is the most conserved nucleotide in module 1 (Figures 1B and 3D). In addition, Arg566 also forms a sequence-specific interaction with G4′. Collectively, the structure of ZFs6-11-*γb7CSE* complex explains how ZFs 9-11 recognize CBS module 1, and the interactions between ZFs 9-11 and module 1 critically contribute to the directionality of CTCF binding to CBSs.

ZF8 is a flexible spacer element

To understand the function of ZF8, we further analyzed our CTCF-CBS complex structures. ZF8 is positioned nearly parallel to the axis of the DNA duplex in the region between modules 1 and 2, and stabilizes the DNA located between modules 1 and 2 by contacting the phosphate backbone of nucleotides 12-17 (Figure 3E). Thus, ZF8 connects CTCF zinc fingers that recognize modules 1-2 and functions as a bridge to span the gap between modules 1 and 2 (Supplementary information, Figure S4). In contrast to ZFs 3-7 and ZFs 9-11, which recognize CBS modules in a sequence-specific manner, ZF8 lacks specific contacts with the DNA base, suggesting that ZF8 is not involved in the base-specific recognition of CBS. Thus, ZF8 functions as a bridge-like spacer element that connects modules 1 and 2, but not a reader of CBS bases, explaining the non-conserved gaps in types 2 and 3 genome-wide CBSs (Supplementary information, Figure S4).

Flexibility between modules 1 and 2

In the human genome, the distances between modules 1 and 2 are variable (comparing types 2 and 3 CBSs in Supplementary information, Figure S4), suggesting the flexibility of CTCF binding. To understand the structural basis for CTCF binding to CBSs with flexible length, we compared the orientation of each ZF to its preceding ZF. Similar to the canonical ZFs [10, 40], ZFs 3-6 inserts into the major groove of the CBS duplex, and each ZF interacts with ~3 bp (Figures 2E and 3A). However, ZF7 rotates 35° more than those of ZFs 3-6, and contacts with 6 bp (Figures 3A and 4A). Remarkably, the relative ori-

Figure 3 CTCF ZFs 9-11 sequence-specially recognize *Pcdh* promoter *γb7CSE* module 1, which is located upstream of the core sequence of CBSs. **(A)** Schematic list of intermolecular contacts between CTCF ZFs 3-11 and CBS. Nucleotides involved in the base contacts are highlighted by yellow. The water molecules-mediated base-specific interactions are shown as small red dots. Arrows indicate contacts mediated by hydrogen bonds. **(B)** Overall structure of ZFs6-11-*γb7CSE* complex. ZFs 9-11 are shown in magenta. **(C**, **D)** Zoomed-in view of the base contacts between ZFs 9-11 and the *Pcdh* promoter *γb7CSE* module 1. **(E)** ZF8 contacts with the phosphate backbone of the non-conserved DNA gap, which connects the modules 1 and 2.

entation of ZF8 to ZF7 differs significantly from that of ZFs 4-6; as a result, ZF8 flipped out of the major groove of the CBS duplex (Figure 4A). In addition, the relative orientation of ZF9 to ZF8 is also different from that of ZFs 4-6, enabling ZFs 9-11 to fit again into the major groove of the CBS duplex (Figure 4A and 4B). The inter-finger linker of "TGEKP" is important in determining the angles between consecutive tandem ZFs [10]; however, inspection of linker regions between successive ZFs was unable to reveal the conservation of CTCF ZF linkers or special amino-acid residues in the inter-ZF contacts between ZFs 7 and 8, and between ZFs 8 and 9 (Figure 1A). Our structural studies show that the different conformation of ZFs 7-9, in particular ZF8, allows CTCF

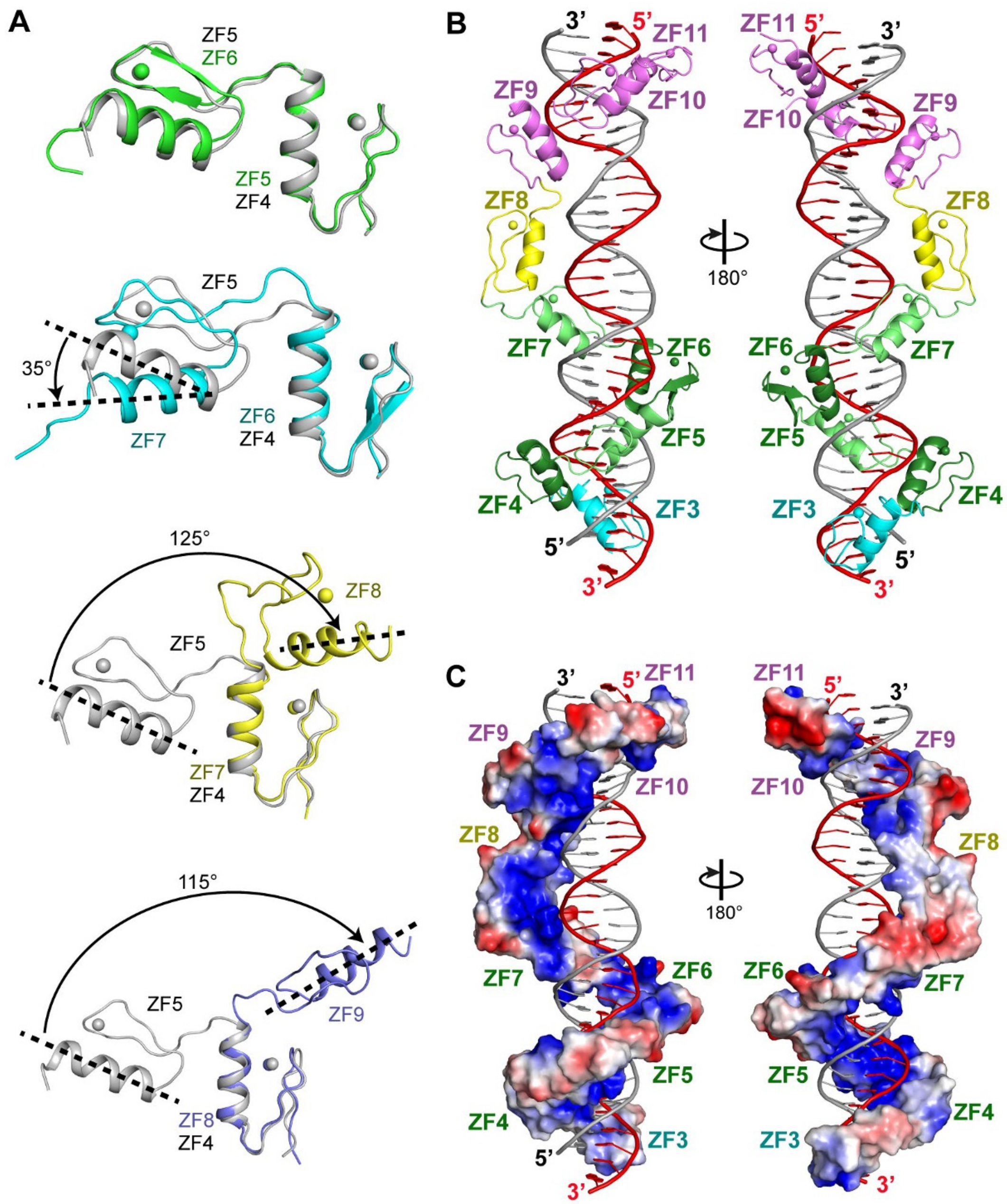


Figure 4 CTCF ZFs 3-11 directionally bind to *Pcdh* CBS modules 4-1. **(A)** Structural comparison between different ZFs. Note that the angles between ZFs 7 and 8, and between ZFs 8 and 9 differ from other canonical consecutive tandem ZF pairs. **(B)** The cartoon view of ZFs3-11-*CBS* complex. **(C)** Corresponding cartoon (DNA) and electrostatically color-coded surface (ZFs 3-11) view of the ZFs3-11-*CBS* complex.

ZF8 to bind the variable gap between CBS modules 1 and 2 in a sequence non-specific manner. In addition, it suggests that the relatively flexible angles between consecutive ZFs of ZFs 7-9 of CTCF endow its flexibility to bind two types of CBSs with either 7 or 8 nucleotides between modules 1 and 2 (Supplementary information, Figure S4).

A model for directional CTCF binding to complete CBS with four modules

To understand how the CTCF binds to the CBS bearing modules 1-4, we modeled ZFs 3-11 in complex with CBS containing modules 1-4 by combining the structural information from the complexes of ZFs2-8-*HS5-1aE* and ZFs6-11-*γb7CSE*. ZFs 3-11 bind CBS in the orientation from module 4 to module 1 (Figure 4B). Consistent with this model, *in vivo* mapping by ChIP-nexus shows that CBS borders occupied by CTCF in the *Pcdh* promoters and enhancers are largely corresponding to modules 1-4 (Supplementary information, Figure S6). ZFs 3-7 form a right-handed helix along the major groove of CBS duplex rotating ~1.5 turns, wrapping and reading the core sequence of modules 2-4 in a single-directional manner. ZFs 9-11 also form a right-handed helix rotating ~2/3 turns and wrapping the duplex of module 1. Finally, ZF8 contacts the spacer between modules 1 and 2.

ZFs 3-11 possess a positively-charged surface facing toward the DNA, stabilizing CBS by interacting with the phosphate backbone (Figure 4C). The base-specific interactions between CTCF and CBS further stabilize the CTCF-CBS complex. Importantly, the base contacts between CTCF and CBS determine the directionality of CTCF binding to oriented CBSs. As discussed above, nucleotide A24 plays an essential role for the orientation of the CBS core sequence. The base-specific contacts between ZF3 and module 4, and between ZFs 9-11 and module 1, further stabilize the CBS binding. Most importantly, they cooperate with the core-binding ZFs 4-7 to determine the directional CTCF binding to CBSs that contain four modules. The observations of directional CTCF binding from our structural studies are consistent with previous biochemical and genetic analyses [3, 27, 28, 31, 32]. In summary, directional CBS binding of CTCF is established through a combination of hydrogen bonds to the asymmetrical central nucleotide “A” within the core sequences of modules 2-3, together with the cooperative reading of modules 1 and 4 by C- and N-terminal ZF clusters, respectively.

Discussion

The architectural balance between genome compaction and chromatin opening in the cell nucleus is essential for developmental gene regulation. The mammalian architectural protein CTCF recognizes numerous insulators in the boundary of topological chromatin domains in each chromosome [13, 15]. In addition, CTCF also binds a large number of genomic sites close to or within promoters as well as enhancers or silencers [3, 13]. For example, CTCF is crucial for allelic and combinatorial *Pcdh* gene expression in individual neurons during brain development through directional and dynamic binding to a repertoire of forward and reverse convergent CBS sites within promoters and enhancers, respectively [3]. In particular, CTCF-mediated specific long-distance chromatin looping interactions between distal enhancers and their cognate promoters determines the promoter choice of combinatorial *Pcdh* gene expression in the brain [20]. However, the molecular mechanisms by which CTCF directionally recognizes convergent oriented forward and reverse CBSs within *Pcdh* promoters and enhancers are not known. In this study, our four crystal structures of CTCF-CBS complexes of *Pcdh* enhancers and promoters reveal that ZF3, ZFs 4-7, and ZFs 9-11 recognize CBS module 4, modules 3-2, and module 1, respectively, thus imposing strictly directional CTCF binding to *Pcdh* CBSs in a cooperative way.

In addition to directionality of CTCF binding to CBSs through specific recognition of the central A24 and cooperation of four modules, our structures of CTCF-CBS complexes have interesting implications for CTCF to recognize extremely diverse CBSs in mammalian genomes. CTCF tolerates a large number of nucleotides within CBSs; however, the three highly conserved CBS nucleotides are each recognized by side-chains of specific CTCF residues through two hydrogen bonds (Figures 2 and 3). These conserved base-specific interactions are also observed in a recent structure of CTCF ZFs 2-9 complexed with an H19 CBS [41]. In this CTCF-H19 complex, however, ZFs 8 and 9 only interact with the DNA phosphate backbone. By contrast, ZF9 forms base-specific interactions in our CTCF-CBS complex structure, suggesting that CTCF may bind various genome-wide CBSs in several distinct manners. Importantly, our structure of CTCF complexed with *Pcdhγb7* promoter CBS reveals the specific recognition of module 1 by ZFs 9-11 and thus explain the variable gap between modules 1 and 2 of the genome-wide types 2 and 3 CBSs (Supplementary information, Figure S4). Although the presence of module 1 may increase the binding affinity between CBSs and CTCF [3, 27, 28], the recognition of CBSs by CTCF *in vivo* is likely quite dynamic. This idea is consistent with the ChIP-nexus data of CTCF binding to CBS *in vivo* as the boundaries not only map to both ends of CBSs but also map around the variable gap of certain *Pcdh* CBSs (Supplementary information, Figure S6). For the physiological function in exquisite 3D genome regulation, CTCF must balance between affinity for cognate CBS and speed for thermodynamic scanning through non-specific DNA sequences. Finally, although there are many multi-finger proteins with multiple tandem ZFs in mammals, the prevalent base-specific recognition mode appears to be only for short stretches of DNA duplex involving modular ensembles of 3-5 consecutive ZFs [10, 11, 40, 42-43].

The DNA bound to the CTCF ZFs adopts a regular B-form, indicating that CTCF does not bend its bound CBS DNA. Although the CBS regions bound by CTCF ZFs are not bent, CTCF may still bend the DNA outside of CBS in collaboration with other protein partners [14, 44]. In our structures, the ZFs bound to CBS are present as monomers. In contrast to previous reports [45], our analysis of CTCF with the C-terminal domain either in the free or DNA-bound states does not show any evidence for CTCF dimerization (Supplementary information, Figure S7). This suggests that CTCF may mediate chromatin looping in a different manner from that proposed earlier although we cannot rule out the possibility of CTCF dimerization through the N-terminal domain.

Methylation of CBSs is known to prevent CTCF binding, which is one of mechanisms of regulating CTCF function in epigenetic processes such as imprinting [46, 47]. On the basis of CTCF-CBS complex, methylation of C20 of the CpG dinucleotide may clash with Asp451 and Arg448 residues of CTCF. It is consistent with the observation that epigenetic methylation of CBS abrogates CTCF binding and abolishes its ability to bridge long-distance chromatin looping interactions between distal enhancers and their target promoters [20]. However, the methylation of C30 likely have little effect on the interaction between CTCF and CBS, as suggested by our structural data.

In conclusion, our structural studies show that the sequence-specific interactions between ZFs and CBSs determine the directionality and conservation of CTCF recognition. In addition, the tolerance for nucleotide substitutions at certain positions of CBSs explains the extreme diversity of mammalian CBSs. Finally, the distinct angles of ZFs 7-8 and of ZFs 8-9 provide the basis for the flexible spacing between CBS modules 1 and 2. Future structural studies of the full-length CTCF protein or its complex with other architectural proteins such as the cohesin complex should provide additional insights into how 3D genome architecture is established, maintained and regulated during developmental control of gene expression.

Materials and Methods

Protein expression and purification

Various truncations, including ZFs 4-8 (residues 349-490), ZFs 2-8 (residues 292-490) and ZFs 6-11 (residues 405-580 with a mutation C504S), were sub-cloned into pET-sumo expression vector and overexpressed in *E. coli* Rosetta (DE3) (Novagen). Cells were harvested and lysed by high-pressure homogenization in lysis buffer containing 20 mM Tris-HCl, pH 7.5, 600 mM NaCl, and 10 mM imidazole. After centrifugation, supernatant was incubated with Ni Sepharose (Qiagen). The bound protein was eluted with lysis buffer containing 200 mM imidazole. The SUMO fusion proteins were cleaved by ULP1 (Ubl-specific protease 1) to remove the His-SUMO tag and were further purified by Heparin column (GE Healthcare). Finally, protein was concentrated to ~10 mg/ml.

Crystallization, data collection and structure determination

Various truncations were incubated with DNA at an equimolar ratio. Then, the complex was further purified on a Superdex 200 16/600 gel filtration column (GE Healthcare) in buffer containing 20 mM Tris-HCl, pH 7.5, 200 mM NaCl, 1 mM $MgCl_2$, 10 µM $ZnSO_4$, 5% glycerol and 1 mM DTT. Samples used for crystallization were measured for final absorbance of ~45 using a Nanodrop spectrophotometer (Thermo Scientific).

Crystals were obtained by mixing 1 µl of CTCF-CBS complex solution and 1 µl of reservoir solution. Crystals of ZFs4-8-*HS5-1a*, ZFs4-8-*eCBS* and ZFs2-8-*HS5-1aE* complex were grown from 0.1 M Bis-Tris, pH 5.6-6.0, 0.2 M NaCl and 17%-24% PEG 3350. Crystals of ZFs6-11-*γb7CSE* complex were grown from 0.1 M Bis-Tris, pH 5.6-6.0, 0.1 M NaCl and 17%-20% PEG 3350. All crystals were cryo-protected using corresponding reservoir buffers with 25% PEG 3350 and flash frozen in liquid nitrogen.

Diffraction data sets of ZFs4-8-*HS5-1a* and ZFs4-8-*eCBS* complexes were collected at beamline BL-19U1 of National Center for Protein Science Shanghai (NCPSS) at Shanghai Synchrotron Radiation Facility (SSRF) [48]. The Zn SAD diffraction data set of ZFs4-8-*HS5-1a,* and the diffraction data sets of ZFs2-8-*HS5-1aE* and ZFs6-11-*γb7CSE* complex were collected at beamlines BL-17U1 at SSRF. All diffraction data sets were processed with HKL2000 [49]. The phases of CTCF ZFs4-8-*HS5-1a* were solved by Zn SAD method using PHENIX Autosol [50]. The structures of ZFs4-8-*eCBS*, ZFs2-8-*HS5-1aE* and ZFs6-11-*γb7CSE* complex were solved by molecular replacement with the structure of ZFs4-8-*Hs5-1a* complex as the model. The model was manually built and adjusted using the program Coot [51]. Iterative cycles of crystallographic refinement were performed using PHENIX. All statistics of data processing and structure refinement of CTCF are summarized in Supplementary information, Table S1. The structure figures were prepared using PyMOL (http://www.pymol.org/).

Cell culture

HEK293T cells were cultured in DMEM (Hyclone) supplemented with 10% (v/v) FBS (Gibco) and 1% penicillin-streptomycin. Cells were grown in a humidified incubator at 37 °C with 5% CO_2.

ChIP-nexus

ChIP-nexus experiments were performed according to the published protocol [52] with modifications. Briefly, HEK293T cells (~2

$\times$ 10^7) were crosslinked with 1% (v/v) formaldehyde for 10 min at room temperature (20-25 °C), and crosslinking was quenched by 125 mM glycine (final concentration). Crosslinked cells were lysed and washed twice using the ChIP buffer (10 mM Tris-HCl, pH 7.5, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.1% SDS, 1× protease inhibitors, 0.15 M NaCl) to obtain nuclei. Nuclear samples were sonicated using Bioruptor Sonicator to fragmentize DNA to sizes ranging from 100 to 10 000 bp. Chromatins were then immunoprecipitated with 5-10 µl CTCF antibody (Millipore, 07-729) by slow rotation at 4 °C overnight. The mixture was then incubated with 50 µl Protein G Magnetic beads (Life Technologies, 10004D) for another 3 h at 4 °C. The magnetic beads with enriched chromatin were then sequentially washed with the washing buffer A (10 mM TE, 0.1% Triton X-100), washing buffer B (150 mM NaCl, 20 mM Tris-HCl, pH 8.0, 5 mM EDTA, 5.2% sucrose, 1.0% Triton X-100, 0.2% SDS), washing buffer C (250 mM NaCl, 5 mM Tris-HCl, pH 8.0, 25 mM HEPES, 0.5% Triton X-100, 0.05% sodium deoxycholate, 0.5 mM EDTA), washing buffer D (250 mM LiCl, 0.5% NP-40, 10 mM Tris-HCl, pH 8.0, 0.5% sodium deoxycholate, 10 mM EDTA), and the Tris buffer (10 mM Tris-HCl, pH 7.5).

The chromatin ends were repaired with 2 µl end-repair enzyme mixture (NEB, E6050S) in 50 µl of 1× buffer for 30 min at 20 °C and then washed again with the buffers A-D. The chromatin ends were dA-tailed by adding 3 µl of Klenow exo- (NEB, M0212S) and 10 µl of 1 mM dATP (NEB, N0440S) in 50 µl of 1× NEB buffer 2 at 37 °C for 30 min. The Nexus adaptors (5′ phosphate GATCG GAAGA GCACA CGTCT GGATC CACGA CGCTC TTCC, 5′ phosphate TCAGA GTCGA GATCG GAAGA GCGTC GTGGA TCCAG ACGTG TGCTC TTCCG ATCT) were added by incubating with 50 µl of 1× Ligase mix (NEB, M0367S) for 1 h at 25 °C. The ends were filled with 1 µl of Klenow exo- and then with 1.5 µl of T4 DNA polymerase (NEB, M0203S) to generate blunt-ended DNAs. Beads-DNA samples were then treated with 20 U Lambda Exonuclease (NEB, M0262S) in the Exonuclease buffer (0.2% Triton X-100 and 1% DMSO in 100 µl of 1× Lambda buffer) by incubation at 37 °C for 1 h on thermomixer at 1 000 rpm. Samples were then digested by 75 U RecJf exonuclease (NEB, M0264S) in 100 µl of 1× NEB buffer 2 containing 0.2% Triton X-100 and 1% DMSO at 37 °C for 1 h on thermomixer at 1 000 rpm, and then washed three times with the RIPA buffer (50 mM HEPES, pH 7.5, 1 mM EDTA, 0.7% sodium deoxycholate, 1% NP-40, 0.5 M LiCl). To elute DNA, samples were incubated with 200 µl elution buffer (50 mM Tris-HCl, pH 8.0, 10 mM EDTA, 1% SDS) at 65 °C for 30 min on thermomixer at 1 000 rpm. The samples were reverse-crosslinked at 65 °C overnight, and then incubated with 2 µl RNase A (Thermo, EN0531) at 37 °C for 2 h, and then with 8 µl proteinase K (NEB, P8107S) at 55 °C for 4 h. The ssDNAs were extracted with 400 µl of phenol:chloroform:isopentanol, and then precipitated with ethanol and sodium acetate. The precipitated DNAs were resuspended with 10 µl nuclease-free water.

The ssDNA was incubated at 95 °C for 5 min and then chilled on ice immediately to denature single-strand DNA. The ssDNA was circularized with 75 U ssDNA Ligase (Epicentre, CL4111K), 0.05 mM ATP, 2.5 mM $MnCl_2$ in 15 µl of 1× CircLigase buffer at 60 °C for 1 h. The samples were incubated with 0.2 µM cut-oligo (GAAGA GCGTC GTGGA TCCAG ACGTG) in 50 µl of 1× FastDigest buffer (Thermo, FD0054) in thermocycler with the annealing program: 95 °C for 1 min, slowly cool down at 1% ramp to 25 °C for 1 min, hold at 25 °C for 30 min. The samples were incubated with 3 µl FastDigest *Bam*HI at 37 °C for 30 min to linearize DNA. DNA was precipitated using ethanol and sodium acetate, and resuspended with 20 µl water. To amplify samples by PCR, 25 µl of Q5 polymerase (NEB, M0543S), 2.5 µl of 10 mM Illumina universal primer and index primer were added, and the PCR program was as follows: 98 °C for 30 s to denature DNA, followed by 16 cycles of 98 °C for 10 s and 65 °C for 75 s, with a final extension at 65 °C for 5 min. PCR products of sizes ranging from 150 to 300 bp were purified from 2% agarose gel using QIAquick gel extraction kit (Qiagen, 28604). Purified DNA samples were sequenced on an Illumina HiSeq platform.

Bioinformatics

About 20 million single-end reads were aligned to hg19 using Bowtie2 [53] after removing barcodes. MACS [54] was used for finding peaks and MACE was used for measuring CTCF occupancy border (footprint). CTCF motif was discovered by MEME suite [55], and heatmap was generated by a custom R code.

Static light scattering

Various truncations, including ZF4-622 (residues 349-622), ZF4-658 (residues 349-658) and ZF4-727 (residues 349-727), were purified as above and were incubated with DNA (*HS5-1aR*, as shown in Supplementary information, Figure S7) in buffer containing 20 mM Tris-HCl, pH 7.5, 100 mM NaCl, 1 mM $MgCl_2$, 10 µM $ZnSO_4$, 1 mM DTT, and 5% glycerol. The complexes were further purified by superdex 200 10/300 GL (GE Healthcare). Size exclusion chromatography column (WYATT, product# WTC-030S5) connected to MALS detector (DAWN HELEOSII) was used to determine the molar mass of our prepared CTCF-DNA complexes.

Electrophoretic mobility shift assay

Each of the purified proteins (1 µM) was incubated for 15 min at room temperature with 1 µM, 3 µM and 5 µM of duplex oligonucleotide labeled with γ-^{32}P in binding buffer (25 mM Tris-HCl, pH 7.5, 50 mM NaCl, 1 mM $MgCl_2$, 5% glycerol, 10 µM $ZnSO_4$, 1 mM DTT). The binding complex was electrophoresed on 6% non-denaturing polyacrylamide gels in 1× Tris-glycine buffer (25 mM Tris-HCl, 192 mM glycine) for 1.5 h. The ^{32}P-labeled DNA was then detected by the image plate (Fuji film BAS-SR2040). Images were obtained using the Typhoon FLA 7000 IP System (GE Healthcare).

Accession numbers

The atomic coordinates of the CTCF ZF4-8-*HS5-1a*, ZF4-8-*eCBS*, ZF2-8-*HS5-1E*, and ZF6-11-*γb7CSE* complexes have been deposited in the Protein Data Bank with accession numbers 5YEG, 5YEH, 5YEF, and 5YEL. The ChIP-nexus data have been deposited in the NCBI GEO (Gene Expression Omnibus) database with accession number GSE103651.

Acknowledgments

We thank the staffs from BL18U1 and BL19U1 beamlines of National Center for Protein Science Shanghai (NCPSS) at Shanghai Synchrotron Radiation Facility (SSRF), and staffs of BL-17U1 beamline at SSRF for assistance with data collection. We thank Dr Lingling Chen (Institute of Biochemistry and Cell Biology, Shanghai) for providing the human CTCF construct. We thank Hongjie Zhang at the IBP radioactive isotope laboratory for the guidance in handling radiolabeled chemicals. We also thank Dr Torsten Juelich for critical reading of the manuscript and linguistic assistance. This work was supported by the Ministry of Science and Technology of China (2017YFA0504203 and 2014CB910102), the National Natural Science Foundation of China (31630015, 91440201, 31571335, 31400640, 31630039, 91640118 and 31470820), the Science and Technology Commission of Shanghai Municipality (14JC1403601), and the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB08010203).

Author Contributions

MY and XL expressed, purified, and grew crystals of CTCF ZFs-*CBS* complexes. MW carried out all cloning experiments. MZ performed ChIP-nexus. MY and JW collected X-ray diffraction data. YW and MY designed experiments and solved all crystal structures. YW and QW analyzed data and wrote the manuscript (YW contributed to Figures 1-4 and Supplementary information, Figures S2, S3, S5 and S7; QW contributed to Supplementary information, Figures S1, S4 and S6). YW supervised and coordinated this project.

Competing Financial Interests

The authors declare no competing financial interests.

References

1 Rao SS, Huntley MH, Durand NC, *et al*. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 2014; **159**:1665-1680.

2 Levine M, Cattoglio C, Tjian R. Looping back to leap forward: transcription enters a new era. *Cell* 2014; **157**:13-25.

3 Guo Y, Xu Q, Canzio D, *et al.* CRISPR inversion of CTCF sites alters genome topology and enhancer/promoter function. *Cell* 2015; **162**:900-910.

4 Tang Z, Luo OJ, Li X, *et al.* CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. *Cell* 2015; **163**:1611-1627.

5 Dekker J, Mirny L. The 3D genome as moderator of chromosomal communication. *Cell* 2016; **164**:1110-1121.

6 Merkenschlager M, Nora EP. CTCF and cohesin in genome folding and transcriptional gene regulation. *Annu Rev Genomics Hum Genet* 2016; **17**:17-43.

7 Narendra V, Bulajic M, Dekker J, Mazzoni EO, Reinberg D. CTCF-mediated topological boundaries during development foster appropriate gene regulation. *Genes Dev* 2016; **30**:2657-2662.

8 Miller J, McLachlan AD, Klug A. Repetitive zinc-binding domains in the protein transcription factor IIIA from *Xenopus* oocytes. *EMBO J* 1985; **4**:1609-1614.

9 Berg JM. Zinc fingers and other metal-binding domains. Elements for interactions between macromolecules. *J Biol Chem* 1990; **265**:6513-6516.

10 Wolfe SA, Nekludova L, Pabo CO. DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct* 2000; **29**:183-212.

11 Klug A. The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annu Rev Biochem* 2010; **79**:213-231.

12 Lobanenkov VV, Nicolas RH, Adler VV, *et al.* 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* 1990; **5**:1743-1753.

13 Ong CT, Corces VG. CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet* 2014; **15**:234-246.

14 Nichols MH, Corces VG. A CTCF code for 3D genome architecture. *Cell* 2015; **162**:703-705.

15 Huang H, Wu Q. CRISPR double cutting through the labyrinthine architecture of 3D genomes. *J Genet Genomics* 2016; **43**:273-288.

16 Nora EP, Goloborodko A, Valton AL, *et al.* Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. *Cell* 2017; **169**:930-944.

17 Ghirlando R, Felsenfeld G. CTCF: making the right connections. *Genes Dev* 2016; **30**:881-891.

18 Dixon JR, Gorkin DU, Ren B. Chromatin domains: the unit of chromosome organization. *Mol Cell* 2016; **62**:668-680.

19 Hnisz D, Day DS, Young RA. Insulated neighborhoods: structural and functional units of mammalian gene control. *Cell* 2016; **167**:1188-1200.

20 Guo Y, Monahan K, Wu H, *et al.* CTCF/cohesin-mediated DNA looping is required for protocadherin α promoter choice. *Proc Natl Acad Sci USA* 2012; **109**:21081-21086.

21 Gomez-Marin C, Tena JJ, Acemel RD, *et al.* Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. *Proc Natl Acad Sci USA* 2015; **112**:7542-7547.

22 Vietri Rudan M, Barrington C, Henderson S, *et al.* Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep* 2015; **10**:1297-1309.

23 de Wit E, Vos ES, Holwerda SJ, *et al.* CTCF binding polarity determines chromatin looping. *Mol Cell* 2015; **60**:676-684.

24 Sanborn AL, Rao SS, Huang SC, *et al.* Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci USA* 2015; **112**:E6456-E6465.

25 Kim TH, Abdullaev ZK, Smith AD, *et al*. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. *Cell* 2007; **128**:1231-1245.

26 Wang H, Maurano MT, Qu H, *et al.* Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res* 2012; **22**:1680-1688.

27 Schmidt D, Schwalie PC, Wilson MD, *et al.* Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* 2012; **148**:335-348.

28 Nakahashi H, Kwon KR, Resch W, *et al.* A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep* 2013; **3**:1678-1689.

29 Hansen AS, Pustova I, Cattoglio C, Tjian R, Darzacq X. CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *Elife* 2017; **6**:e25776.

30 Rhee HS, Pugh BF. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* 2011; **147**:1408-1419.
31 Quitschke WW, Taheny MJ, Fochtmann LJ, Vostrov AA. Differential effect of zinc finger deletions on the binding of CTCF to the promoter of the amyloid precursor protein gene. *Nucleic Acids Res* 2000; **28**:3370-3378.
32 Renda M, Baglivo I, Burgess-Beusse B, *et al.* Critical DNA binding interactions of the insulator protein CTCF: a small number of zinc fingers mediate strong binding, and a single finger-DNA interaction controls binding at imprinted loci. *J Biol Chem* 2007; **282**:33336-33345.
33 Kemp CJ, Moore JM, Moser R, *et al.* CTCF haploinsufficiency destabilizes DNA methylation and predisposes to cancer. *Cell Rep* 2014; **7**:1020-1029.
34 Katainen R, Dave K, Pitkanen E, *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat Genet* 2015; **47**:818-821.
35 Ji X, Dadon DB, Powell BE, *et al.* 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell* 2016; **18**:262-275.
36 Wu Q, Maniatis T. A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell* 1999; **97**:779-790.
37 Zipursky SL, Sanes JR. Chemoaffinity revisited: Dscams, protocadherins, and neural circuit assembly. *Cell* 2010; **143**:343-353.
38 Chen WV, Nwakezw CL, Denny CA, *et al.* Pcdhαc2 is required for axonal tiling and assembly of serotonergic circuitries in mice. *Science* 2017; **356**:406-411.
39 Wu Q, Zhang T, Cheng JF, *et al.* Comparative DNA sequence analysis of mouse and human protocadherin gene clusters. *Genome Res* 2001; **11**:389-404.
40 Nolte RT, Conlin RM, Harrison SC, Brown RS. Differing roles for zinc fingers in DNA recognition: structure of a six-finger transcription factor IIIA complex. *Proc Natl Acad Sci USA* 1998; **95**:2938-2943.
41 Hashimoto H, Wang D, Horton JR, *et al*. Structural basis for the versatile and methylation-dependent binding of CTCF to DNA. *Mol Cell* 2017; **66**:711-720.
42 Pavletich NP, Pabo CO. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* 1991; **252**:809-817.
43 Bedard M, Roy V, Montagne M, Lavigne P. Structural insights into c-Myc-interacting zinc finger protein-1 (Miz-1) delineate domains required for DNA scanning and sequence-specific binding. *J Biol Chem* 2017; **292**:3323-3340.
44 MacPherson MJ, Sadowski PD. The CTCF insulator protein forms an unusual DNA structure. *BMC Mol Biol* 2010; **11**:101.
45 Yusufzai TM, Tagami H, Nakatani Y, Felsenfeld G. CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol Cell* 2004; **13**:291-298.
46 Bell AC, Felsenfeld G. Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature* 2000; **405**:482-485.
47 Hark AT, Schoenherr CJ, Katz DJ, *et al*. CTCF mediates methylation-sensitive enhancer-blocking activity at the *H19/Igf2* locus. *Nature* 2000; **405**:486-489.
48 Wang QS, Yu F, Huang S, *et al*. The macromolecular crystallography beamline of SSRF. *Nucl Sci Tech* 2015; **26**:12-17.
49 Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol* 1997; **276**:307-326.
50 Adams PD, Grosse-kunstleve RW, Hung LW, *et al.* PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr D Biol Crystallogr* 2002; **58**:1948-1954.
51 Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* 2010; **66**:486-501.
52 He Q, Johnston J, Zeitlinger J. ChIP-nexus enables improved detection of *in vivo* transcription factor binding footprints. *Nat Biotechnol* 2015; **33**:395-401.
53 Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; **9**:357-359.
54 Zhang Y, Liu T, Meyer CA, *et al*. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 2008; **9**:R137.
55 Bailey TL, Boden M, Buske FA, *et al*. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 2009; **37**:W202-W208.

(**Supplementary information** is linked to the online version of the paper on the *Cell Research* website.)