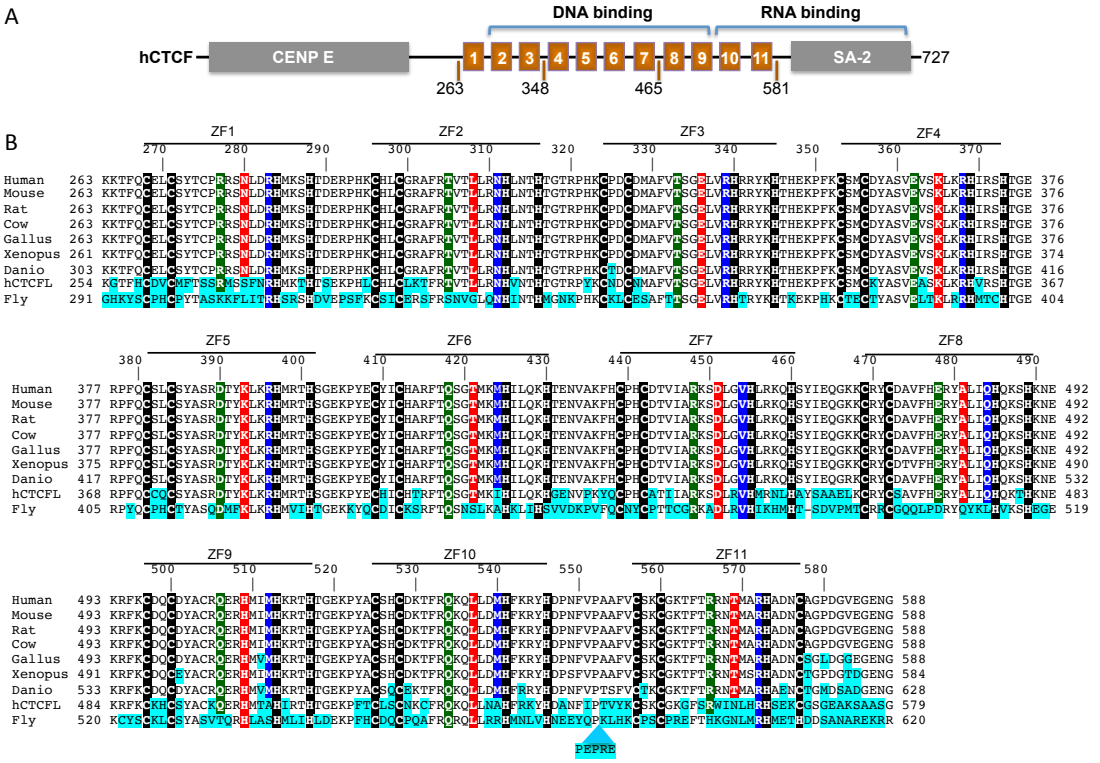
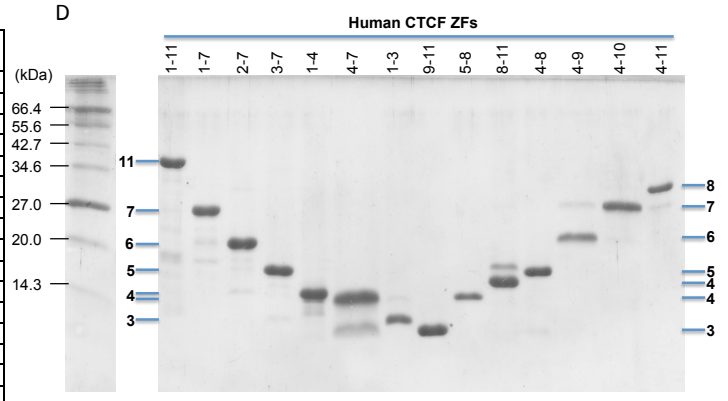


C

Human CTCF ZF	Residues	Construct (pXC #)	Protein yield (mg/L)	Crystals
1-11	263-581	1441	~0.5	
1-7	263-465	1564	~2	-
1-4	273-377	1417	+++ (~25)	
1-3	263-348	1356	+++ (~25)	
2-7	294-465	1565	~2	+
3-11	321-581	1567	~1	
3-9	321-518	1571	0.3	
3-7	321-465	1551	~1	+
4-11	348-581	1568	~2	+
4-10	348-547	1574	~1	+
4-9	348-518	1573	~1.5	
4-8	348-492	1357	~1.5	
4-7	348-465	1202	+++ (~25)	+
	K365T	1518	~1.5	
5-8	377-492	1199	+++ (~25)	+
6-8	405-492	1197	+++ (~25)	+
8-11	464-581	1358	+++ (~25)	
9-11	493-581	1359	+++ (~25)	



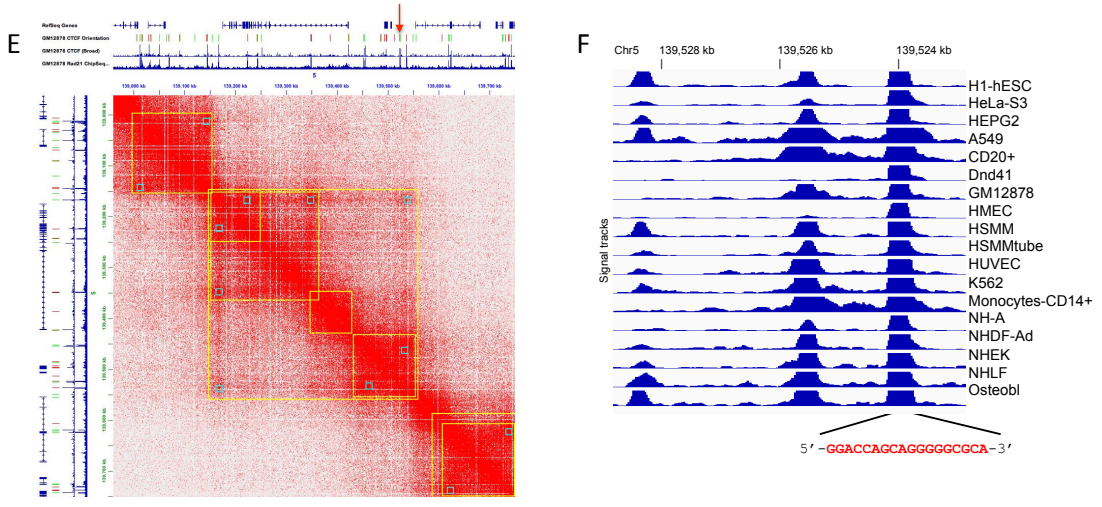


Figure S1. The CTCF family, related to Figure 1

(**A)** Schematic representation of human CTCF. The N-terminal and C-terminal domains of CTCF are known to interact with other proteins, for example, the centromeric protein CENP-E (Xiao et al., 2015) and the cohesin subunit SA2 (Xiao et al., 2011), respectively. In addition, CTCF binds to RNA (Kung et al., 2015; Saldana-Meyer et al., 2014).

(**B**) Sequence alignment of the conserved 11-ZF region from human (NP_006556.1), mouse (NP_851839.1), rat (NP_114012.1), cow (NP_001069216.1), *Gallus* (NP_990663.2), *Xenopus* (NP_001116268.1), *Danio* (NP_001001844.1), human CTCFL/BORIS (NP_542185.2) and fruit fly (NP_648109.1). Numbering of residues in human CTCF is shown above the sequence alignment. The sequence variations are highlighted in cyan. The four Zn-coordinating residues of each finger are highlighted with white letters against black. Side chains from three specific amino acids within each finger that make major groove contacts with the DNA bases of 3-bp triplet are highlighted in blue (for 5' base), red (for central base), and green (for 3' base). The first zinc-coordination His in each finger is assigned reference position 0. Residues prior to this, at positions −1 (blue), −4 (red), and −7 (green) form H-bonds with the exposed edges of the DNA bases in the major groove.

(**B)** The expression constructs for human CTCF created in this study.

(**C)** An 18% SDS-PAGE showing examples of the purified ZF proteins used in this study.

(**E)** The CORE sequence used for DNA binding assays and crystallization (indicated by a red arrow) is located at chromosome 5: 13952372-13952389, at a topologically associating domain (TAD) border. This site is constitutively bound in various cell types. Heat maps depicting observed intra-chromosomal contacts for chromosome 5: 139,000 kb -139,750 kb of in situ MboI Hi-C data in GM12878 cell (Rao et al., 2014). Reference genes, CTCF orientation, GM12878 CTCF chip-seq data (Broad), and Rad21 chip-seq sequence (Stanford) are shown. Peaks are shown in blue squares, and contact domains are shown in yellow squares.

(**F**) Signal tracks of CTCF ChIP-seq (Broad) in various cell types are shown for chromosome 5 (Consortium, 2012).

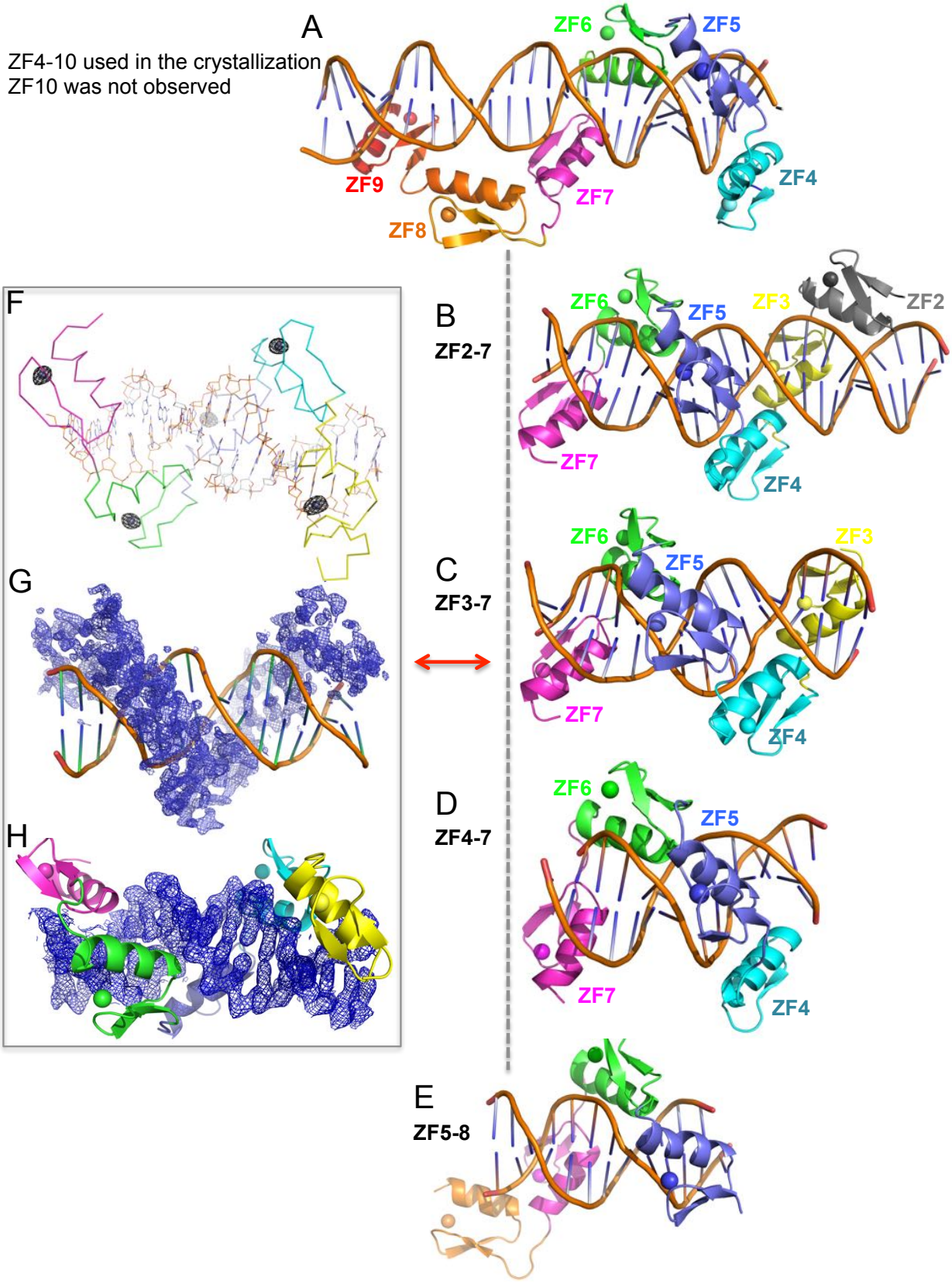

A
ZF4-10 used in the crystallization
ZF10 was not observed
ZF6
ZF5
ZF9
ZF7
ZF4
ZF8
B
ZF2-7
ZF6
ZF5
ZF3
ZF2
ZF7
ZF4
C
ZF3-7
ZF6
ZF5
ZF3
ZF7
ZF4
D
ZF4-7
ZF6
ZF5
ZF7
ZF4
E
ZF5-8
F
G
H

Figure S2. Summary of structural information of the human CTCF ZF DNA binding domain in complex with DNA derived from this study, related to Figure 2 and Table S1

(**A**) Structure of CTCF ZF4-10 in complex with DNA. The last finger (ZF10) was not visible. Each finger is colored as cyan (ZF4), blue (ZF5), green (ZF6), magenta (ZF7), orange (ZF8) and red (ZF9).

(**B**) Structure of CTCF ZF 2-7 in complex with DNA. ZF2 (in grey) lies in the DNA major groove but does not make base-specific contacts with the DNA sequence used.

(**C-E**) Structures of CTCF ZF3-7 (panel C), ZF4-7 (panel D) and ZF5-8 (panel E) in complex with DNA, respectively. Vertical dashed line indicates all structures shown are aligned with common elements.

(**F**) In the structure of ZF3-7 in complex with DNA, crystallographic phases were calculated by single-wavelength anomalous diffraction (SAD) from five zinc atoms, shown with the anomalous experimental electron density (grey mesh), contoured at 10σ above the mean, for the corresponding zinc atoms.

(**G-H**) The omit electron density (blue mesh), contoured at 2.0σ above the mean, is shown for the protein (panel G) or DNA (panel H).

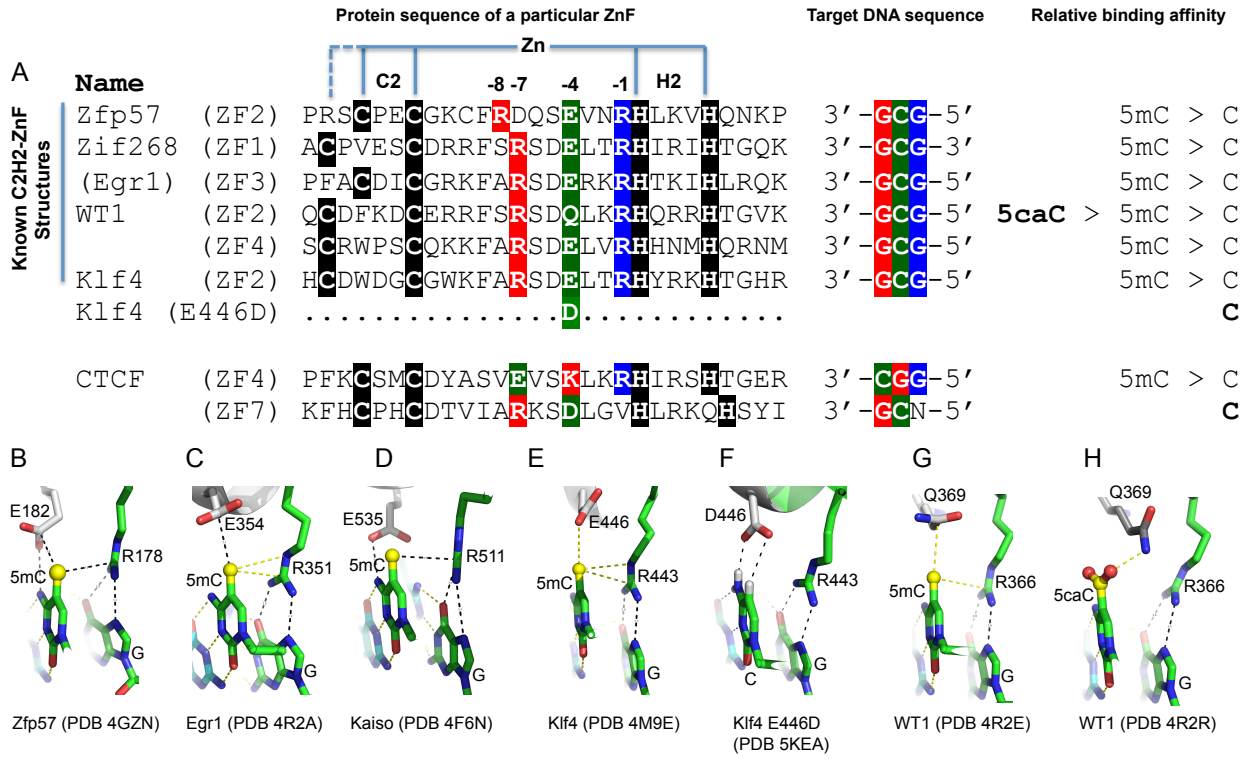


Figure S3. A recognition code for modified cytosine, related to Figure 3

(A) Sequence alignment of representative C2H2 ZF proteins known to interact with methylated (modified) cytosine. The amino acids at positions −1, −4, -7 or −8 and the primary DNA base recognition are highlighted with the same color code (blue, green, and red). In an early report of phage display study of Egr1/Zif268 (Choo and Klug, 1994), an aspartate (D) residue at the -4 position (rather than E in the wild type) shows a distinct preference for binding (unmodified) cytosine, whereas E appears to specify cytosine in wild-type Egr1/Zif268 (Pavletich and Pabo, 1991). This observation led Choo and Klug to comment that "The physical basis for the interaction of aspartate/glutamate and cytosine is not yet clear, since hydrogen-bonding contacts between these groups have yet to be observed in zinc finger cocrystal structures" (Choo and Klug, 1997). In recent studies, we and others showed that Egr1/Zif268 binds methylated DNA (Hashimoto et al., 2014; Zandarashvili et al., 2015).

(B-E) An Arg-Glu pair in Zfp57 (Liu et al., 2012), Egr1 (Hashimoto et al., 2014), Kaiso (Buck-Koehntop et al., 2012), and Klf4 (Liu et al., 2014) recognizes a 5mCpG dinucleotide. In Kaiso, the Arg and Glu residues are from two neighboring fingers (Liu et al., 2013). A negatively charged glutamate prevents the binding of 5caC; both carboxylate groups bear a full -1 charge and thus repel one another electrostatically.

(**F**) In Klf4, substituting Glu446 with aspartate (E446D) resulted in preference for unmodified cytosine (Hashimoto et al., 2016).

(**G-H**) A glutamine allows WT1 bind both 5mC and 5caC (Hashimoto et al., 2014).

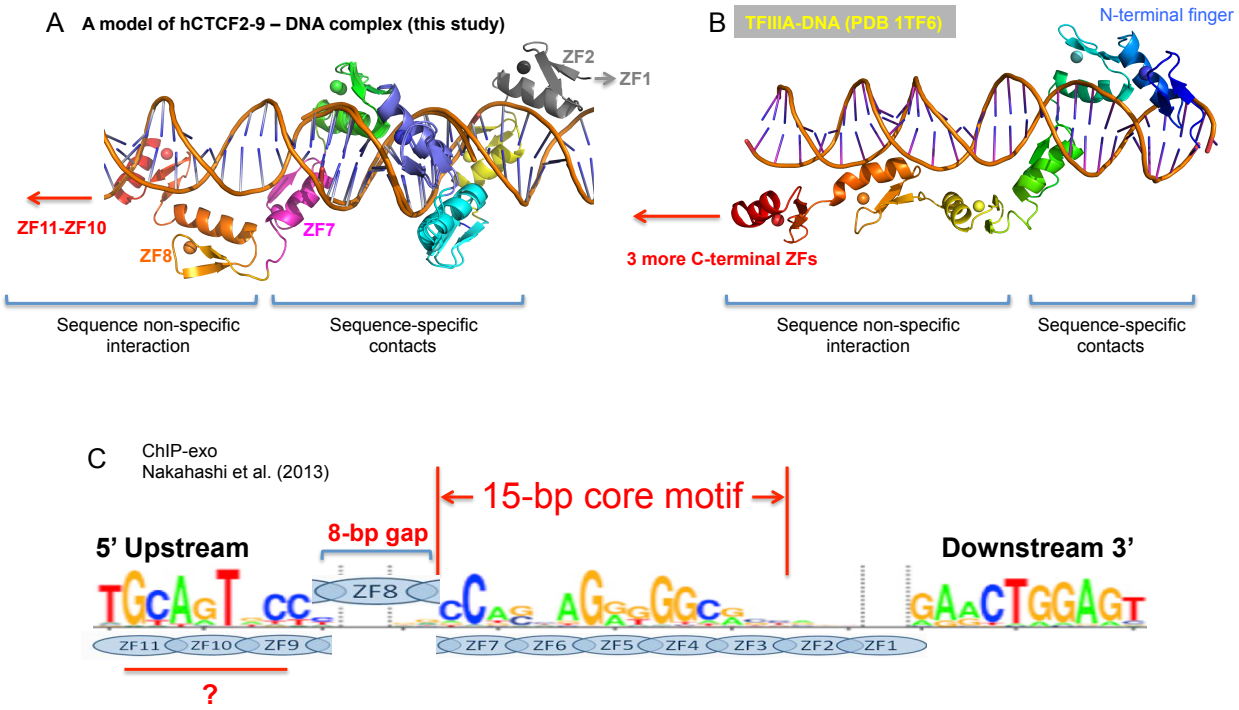


Figure S4. CTCF ZF8-9 spans across DNA phosphate backbone, related to Figure 4

(**A-B**) Comparison of our model of CTCF ZF2-9 to an earlier example of transcription factor TFIIIA (Nolte et al., 1998) suggests that the C-terminal fingers could span the greater length of the DNA duplexes without additional base specific interactions.

(C) In a study by Nakahashi et al., in addition to the 15-bp CORE sequence, a second 5' upstream motif found at approximately 15% of CTCF-binding sites (Nakahashi et al., 2013). However, we were not able to confirm the binding of the C-terminal ZF9-11 to DNA for the number of sequences tested (data not shown). In addition, our model does not explain how the 3' downstream DNA motif destabilizes CTCF occupancy (Nakahashi et al., 2013). One possible explanation is that other DNA binding protein(s) in the ChIP-exo experiments occupy the 5' upstream and/or 3' downstream sequences.

Table S1. Summary of X-ray data collection from SERCAT 22-ID beamline and refinement statistics (*), related to Figures 2 and S2

CTCF		ZF2-7	ZF3-7	ZF3-7 (**M**=5mC)	ZF4-7	ZF5-8	ZF5-8	ZF6-8	ZF4-10	ZF4-11
DNA	(5'-3')	GCC**A**GC**A**GGGGGCGCTAGTGAGG	GCC**A**GC**A**GGGGGCGCTA	GCC**A**GC**A**GGGGG**M**GCTA	C**A**GC**A**GGGGGCGC	GTGCC**A**GC**A**GGGG	TTGCC**A**GC**A**GGGG	GTTGCC**G**C**G**TG	CAGTGCCCACAGAGGCC**A**GC**A**GGGGGCG	
	(3'-5')	CGG**T**CG**T**CCCCCGCGATCACTCC	CGG**T**CG**T**CCCCCGCGAT	CGG**T**CG**T**CCCCCG**M**GAT	GG**T**CG**T**CCCCCGC	CCACGG**T**CG**T**CCC	CAACGG**T**CG**T**CCC	CAACGG**C**GC**A**C	GTCACGGGTGTCTCCGG**T**CG**T**CCCCCGC	
PDB		5T0U	5KKQ	5T00	5K5H	5K5I	5K5J	5K5L	5UND	-
Wavelength (Å)		1.27046	1.27046	1.0	1.27046	1.27046	1.0000	1.2829	1.0	1.0
Space Group		*P*2	*P*1	*P*1	$P4_12_12$	$P6_5$	$P4_12_12$	$P2_12_12$	$P2_1$	$P2_1$
Unit cell (Å)		101.6, 41.0, 106.7	41.0, 44.9, 86.7	41.0, 44.9, 86.8	52.8, 52.8, 166.5	45.1, 45.1, 260.6	58.9, 58.9, 172.3	126.8, 52.4, 69.1	75.1, 73.8, 93.3	75.5, 74.3, 92.7
α, β, γ (°)		90.0, 92.2, 90.0	98.4, 92.4, 94.8	98.3, 92.4, 94.8	90, 90, 90	90, 90, 120	90, 90, 90	90, 90, 90	90.0, 91.4, 90.0	90.0, 92.7, 90.0
Resolution (Å)		53.30-3.20 (3.28-3.20)	28.94-1.74 (1.80-1.74)	41.96-2.19 (2.27-2.19)	50.32-3.11 (3.19-3.11)	23.26-2.15 (2.23-2.15)	27.87-2.29 (2.37-2.29)	46.7-3.13 (3.21-3.13)	34.60-2.55 (2.64-2.55)	34.50-3.69 (3.82-3.69)
[a] R_{merge}		0.163 (0.350)	0.075 (0.549)	0.117 (0.733)	0.088 (0.593)	0.077 (0.709)	0.048 (0.726)	0.114 (0.392)	0.044 (0.458)	0.057 (0.329)
[b] <I/σI>		8.2 (2.3)	12.4 (1.6)	9.33 (2.43)	17.56 (4.3)	27.7 (2.2)	27.3 (2.7)	14.9 (3.6)	14.9 (1.4)	12.9 (3.3)
Completeness (%)		99.3 (94.3)	92.6 (64.1)	98.7 (91.6)	84.1 (47.5)	99.3 (97.2)	98.7 (100.0)	98.7 (91.1)	99.8 (98.7)	100 (100)
Redundancy		5.1 (4.7)	3.6 (1.8)	7.4 (7.0)	36.6 (22.0)	12.7 (4.1)	6.1 (6.2)	7.5 (4.8)	6.3 (5.4)	7.2 (7.2)
CC 1/2, CC		(/ 0.985)	(0.637 / 0.882)	(/ 0.960)	(0.990 / 0.996)	(0.647 / 0.886)	(0.901 / 0.974)	(0.980 / 0.995)	(0.590 / 0.862)	(0.898 / 0.973)
Reflections (observed)		76,745	204,410	230,608	151,106	207,071	87,503	63,398	209,009	79,994
Reflections (unique)		14,911 (1,018)	57,383 (4,006)	31,064 (2,903)	5,709 (219)	16,250 (1579)	14,353 (1419)	8,481 (755)	33,435 (3,301)	11,181 (1,102)
Phasing		MR-SAD	Zn-SAD	MR	Zn-SAD	Zn-SAD	MR	Zn-SAD	MR	MR
Bijvoet pairs		-	53,575	-	2,639	15,716	-	6,745		
FOM		-	0.9	-	0.33	0.33	-	0.35		
Refinement										
Resolution (Å)		3.20	1.74	2.19	3.11	2.17	2.29	3.12	2.55	Refinement
No. Reflections		14,798	57,333	30,987	3,385	15,437	14,269	8,420	33,362	Not
[c] R_{work}/ [d] R_{free}		0.230 / 0.269	0.170 / 0.199	0.221 / 0.245	0.237 / 0.264	0.219/0.245	0.208/0.254	0.231/0.270	0.211 / 0.237	Continued
No. Atoms										
	Protein	2765	2395	2339	906	934	938	1471	2510	
	DNA	1874	1453	1386	527	527	550	856	2102	
	Zn	12	10	10	4	5	4	7	12	
	Solvent	12	328	97	6	16	38	-	107	
B-factors (Å^2)										
	Protein	81.5	41.7	70.8	89.5	69.9	64.1	75.3	92.0	
	DNA	82.0	38.3	61.5	55.3	58.8	64.8	98.7	130.9	
	Zn	84.2	37.5	60.4	73.7	76.5	57.0	88.5	88.1	
	Solvent	43.5	41.5	44.9	10.1	59.3	55.8	-	67.8	
R.m.s. deviations										
	Bond length (Å)	0.005	0.02	0.004	0.004	0.004	0.004	0.005	0.004	
	Bond angles (°)	0.5	1.5	0.6	0.5	0.4	0.5	0.8	0.4	
All atom clashscore		1.4	3.7	2.2	10.3	0.8	0.7	3.7	2.3	
Ramachandran (%)										
	Favored	95.2	98.6	98.0	98.2	99.1	98.2	95.2	97.4	
	Allowed	4.8	1.4	2.0	1.8	0.9	1.8	4.8	2.6	
C_β deviation		0	0	0	0	0	0	0	0	

* Values in parenthesis correspond to highest resolution shell; [a] $R_{merge} = \Sigma \mid I - \langle I \rangle \mid / \Sigma I$, where I is the observed intensity and <I> is the averaged intensity from multiple observations; [b] <I/σI> = averaged ratio of the intensity (I) to the error of the intensity (σI); [c] $R_{work} = \Sigma \mid Fobs - Fcal \mid / \Sigma \mid Fobs \mid$, where Fobs and Fcal are the observed and calculated structure factors, respectively; [d] R_{free} was calculated using a randomly chosen subset (5%) of the reflections not used in refinement.