STRUCTURAL BIOLOGY

The ICF syndrome protein CDCA7 harbors a unique DNA binding domain that recognizes a CpG dyad in the context of a non-B DNA

Swanand Hardikar[1]†, Ren Ren[1]†, Zhengzhou Ying[1]†‡, Jujun Zhou[1]†, John R. Horton[1], Matthew D. Bramble[1], Bin Liu[1,2], Yue Lu[1], Bigang Liu[1], Luis Della Coletta[1], Jianjun Shen[1], Jiameng Dan[1]§, Xing Zhang[1], Xiaodong Cheng[1,2]*, Taiping Chen[1,2]*

***CDCA7*, encoding a protein with a carboxyl-terminal cysteine-rich domain (CRD), is mutated in immunodeficiency, centromeric instability, and facial anomalies (ICF) syndrome, a disease related to hypomethylation of juxtacentromeric satellite DNA. How CDCA7 directs DNA methylation to juxtacentromeric regions is unknown. Here, we show that the CDCA7 CRD adopts a unique zinc-binding structure that recognizes a CpG dyad in a non-B DNA formed by two sequence motifs. CDCA7, but not ICF mutants, preferentially binds the non-B DNA with strand-specific CpG hemi-methylation. The unmethylated sequence motif is highly enriched at centromeres of human chromosomes, whereas the methylated motif is distributed throughout the genome. At S phase, CDCA7, but not ICF mutants, is concentrated in constitutive heterochromatin foci, and the formation of such foci can be inhibited by exogenous hemi-methylated non-B DNA bound by the CRD. Binding of the non-B DNA formed in juxtacentromeric regions during DNA replication provides a mechanism by which CDCA7 controls the specificity of DNA methylation.**

INTRODUCTION

The predominant conformation of DNA in cells is the canonical right-handed B-form double-stranded helix. However, a variety of noncanonical DNA conformations, such as hairpins, cruciforms, left-handed double-helical Z-DNA, triple-stranded H-DNA, G-quadruplexes, and RNA-like four-way junction, have also been recognized (*1*–*8*). Non–B-form structures can affect DNA-dependent processes, including transcription, replication, recombination, and repair, and have been implicated in mutagenesis and genetic instability that are associated with various diseases (*1*, *9*, *10*). Recent evidence suggests that non–B-form DNA is particularly enriched at centromeres in eukaryotes (*11*–*13*), which has led to the hypothesis that non-B DNA structures contribute to centromere specification (*11*). Presumably, some of the biological effects of non-B DNA conformations are mediated by proteins that recognize and/or stabilize them. Numerous non-B DNA binding proteins have been reported (*14*–*16*). However, the molecular mechanisms by which these proteins interact with their corresponding non-B DNA substrates, as well as the functional significance of most such interactions, have not been well characterized.

Immunodeficiency, centromeric instability, and facial anomalies (ICF) syndrome is a rare autosomal recessive disease characterized by immunoglobulin deficiency, facial dysmorphism, intellectual disability, developmental delay, and genomic instability involving the juxtacentromeric regions of chromosomes 1, 9, and 16, (*17*, *18*). These chromosomes have relatively large centromeric and pericentromeric (hereafter peri/centromeric) regions. Patients with ICF usually suffer from recurrent infections in early childhood (*19*–*21*). Recent evidence suggests the involvement of enhanced CD19 activity in immunodeficiency in ICF syndrome (*22*).

A hallmark of ICF syndrome is hypomethylation of specific genomic regions, most notably classical satellite repeats in peri/centromeric regions (*23*). Four ICF-related genes have been identified—*DNMT3B* (DNA methyltransferase 3B), *ZBTB24* (zinc finger–and BTB domain–containing 24), *CDCA7* (cell division cycle–associated 7), and *HELLS* (helicase, lymphoid-specific, also known as lymphoid-specific helicase)—with cases carrying different gene mutations being designated, respectively, as ICF1 (OMIM #242860, *DNMT3B*), ICF2 (OMIM #614069, *ZBTB24*), ICF3 (OMIM #616910, *CDCA7*), ICF4 (OMIM #616911, *HELLS*), and ICFX (unknown) (*24*–*28*). DNMT3B is a de novo DNA methyltransferase that establishes DNA methylation patterns during development (*25*). HELLS, a DNA helicase involved in chromatin remodeling, regulates both de novo and maintenance DNA methylation in an adenosine triphosphatase–dependent manner (*29*–*32*). Recent studies suggest that ZBTB24 and CDCA7 function upstream of HELLS in a molecular pathway that regulates DNA methylation. Specifically, ZBTB24, a C2H2-zinc finger transcription factor, induces *CDCA7* transcription, and CDCA7 recruits HELLS to centromeric satellite repeats, among other regions, to facilitate DNA methylation (*33*–*38*). Thus, CDCA7 plays a key role in determining the specificity of the ZBTB24-CDCA7-HELLS axis in the regulation of DNA methylation.

CDCA7 was identified as a c-Myc–responsive gene (*39*). It is periodically expressed in the cell cycle and reaches the highest level between G_1 and S phase and has been implicated in transcriptional regulation, tumorigenesis, and hematopoietic stem cell emergence (*40*–*45*). However, the fundamental functions of CDCA7 remain poorly understood. In addition to several small functional regions, i.e., a leucine zipper motif, a c-Myc/14-3-3 interaction motif, and a nuclear localization signal, CDCA7 harbors a C-terminal

[1]Department of Epigenetics and Molecular Carcinogenesis, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. [2]Program in Genetics and Epigenetics, The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA.
*Corresponding author. Email: xcheng5@mdanderson.org (X.C.); tchen2@mdanderson.org (T.C.)
†These authors contributed equally to this work.
‡Present address: The Ministry of Education Key Laboratory of Laboratory Medical Diagnostics, College of Laboratory Medicine, Chongqing Medical University, Chongqing 400016, China.
§Present address: State Key Laboratory of Primate Biomedical Research, Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming, Yunnan 650500, China.

cysteine-rich domain (CRD) including four copies of CXXC motif (*28*, *42*). Of note, the identified ICF3 missense mutations in CDCA7 alter three residues in the CRD (*28*), highlighting the functional importance of the domain. Previous studies have shown that ICF3 mutations disrupt the recruitment of the CDCA7-HELLS complex to chromatin (*34*, *37*). Nevertheless, the precise role of the CDCA7 CRD and the chromatin component and feature being recognized are unknown.

In this study, we provide structural and DNA binding data demonstrating that the CDCA7 CRD adopts a unique three-zinc–containing structure that binds a non–B-form DNA in a CpG-specific manner. The three ICF3 missense mutations abolish binding by disrupting, respectively, the interactions with a Zn^{2+}, a guanine of a CpG dinucleotide, and a phosphate group of the DNA backbone. The non-B DNA is formed by two sequence motifs. A 13-mer motif 1 is present throughout the human genome, and an 11-mer motif 2 is highly enriched in centromeric alpha satellite (αSat) repeats. Strand-specific CpG methylation in the non-B DNA exhibits the opposite effects on CDCA7 binding—positively regulated by motif 1 methylation and negatively regulated by motif 2 methylation. We also show that wild-type (WT), but not ICF3-mutant, CDCA7 is concentrated in constitutive heterochromatin foci during the S phase of the cell cycle. The hemi-methylated non-B DNA preferentially bound by the CDCA7 CRD, when introduced in cells, can inhibit CDCA7 foci formation and induce hypomethylation of centromeric satellite DNA. We propose that CDCA7 recruits HELLS and the DNA methylation machinery to centromeric regions by recognizing the non-B DNA formed during DNA replication.

RESULTS

WT CDCA7, but not ICF3 mutants, is concentrated in constitutive heterochromatin foci at S phase

Human (h) and mouse (m) CDCA7 are highly conserved, with ~97% sequence identity in the CRD, where the ICF3 missense mutations are located (Fig. 1A). We previously showed that CDCA7 recruits HELLS to heterochromatin to facilitate DNA methylation of the centromeric minor satellite repeats in mouse embryonic stem cells (mESCs) (*38*). Mutagenesis and coimmunoprecipitation (Co-IP) assays indicated that the leucine zipper motif is required for CDCA7 to interact with HELLS, whereas an ICF3 missense mutation and even deletion of the entire CRD showed no effect on CDCA7-HELLS interaction (fig. S1). Our results were consistent with previous observation that ICF3 missense mutations do not affect the formation of the CDCA7-HELLS complex but prevent the recruitment of the complex to chromatin (*34*, *37*).

To gain insights into the effect of ICF3 mutations on CDCA7 chromatin association, we examined CDCA7 cellular localization. Immunofluorescence (IF) analysis revealed that hemagglutinin (HA)–tagged WT mCDCA7 and the R285H (RH), G305V, and R315H mutants (equivalent to the R274H, G294V, and R304H ICF3 mutants in hCDCA7; Fig. 1A) all localized in the nuclei in transiently transfected NIH3T3 cells (Fig. 1B). WT mCDCA7 was highly enriched in constitutive heterochromatin foci (pattern A)—marked by 4′,6-diamidino-2-phenylindole (DAPI)–bright spots and histone H3 lysine 9 trimethyl (H3K9me3) signal—in a considerable fraction (~30%) of transfected cells, although the majority (~70%) of cells showed no such enrichment (pattern B). In contrast, the ICF3 mutant proteins failed to be concentrated in heterochromatin foci and exhibited only the diffused pattern B (Fig. 1B, right). Similar results were obtained when HA-tagged hCDCA7 proteins were expressed in the human cervical cancer cell line HeLa: concentration of WT hCDCA7, but not the R274H ICF3 mutant, in constitutive heterochromatin foci (fig. S2).

We also established stable NIH3T3 cell lines expressing HA-tagged mCDCA7 or the RH mutant (Fig. 1C). In agreement with the results of transient transfection (Fig. 1B), IF analysis of the stable cell lines showed that WT mCDCA7 exhibited both pattern A and pattern B, accounting for ~30 and ~70% of the interphase cells, respectively, but the RH mutant exhibited pattern B in all interphase cells (Fig. 1C). During mitosis, both WT and mutant mCDCA7 proteins were present throughout the cells, excluding the chromosomes (Fig. 1C, images on the right). The two stable cell lines (WT and RH mutant) showed no difference in viability and proliferation, and flow cytometry analysis revealed similar cell cycle profiles (Fig. 1D).

The presence of CDCA7 nuclear foci only in a fraction of cells raises the possibility of CDCA7 localization patterns being regulated during the cell cycle. Thus, we first synchronized HA-mCDCA7–expressing NIH3T3 cells at the G_0-G_1 phase by serum starvation [0.5% fetal bovine serum (FBS)] for 48 hours, followed by culturing them in regular medium (containing 10% FBS) for 6, 12, 16, and 20 hours, respectively. As revealed by cell cycle analysis, almost all cells were arrested at G_0-G_1 phase initially (0-hour time point) and remained at G_0-G_1 phase at the 6-hour time point, small fractions reached S (~15%) and G_2-M (~4%) phases at the 12-hour time point, substantially larger fractions were at S (~43%) and G_2-M (~19%) phases at the 16-hour time point, and most cells apparently had entered the next G_1 phase at the 20-hour time point (Fig. 1E). CDCA7 nuclear foci (pattern A cells) were not observed at 0 and 6 hours but appeared in ~17, ~52, and ~29% of the cells at 12, 16, and 20 hours, respectively (Fig. 1, F and G). We conclude that pattern A cells are mostly at S phase, although the localization pattern may persist to G_2 phase in some cells. Together, our results indicate that CDCA7 is enriched in constitutive heterochromatin during DNA replication and that the ICF3 missense mutations in the CRD disrupt such enrichment.

The CDCA7 CRD binds a specific non-B DNA

Our finding suggests that determining how the CRD targets CDCA7 to constitutive heterochromatin is key to understanding the mechanism by which the ZBTB24-CDCA7-HELLS axis specifically regulates methylation of satellite DNA repeats in juxtacentromeric regions (*38*). The CDCA7 CRD has been implicated in DNA binding (*34*). However, electrophoretic mobility shift assay (EMSA) showed that the CDCA7 CRD failed to bind DNA sequences containing the repeating units of several common satellite DNAs in the peri/centromeric regions of human and mouse genomes (fig. S3A).

One possibility is that the CDCA7 CRD recognizes a specific DNA motif or structure that is present in heterochromatic regions of both human and mouse cells. To explore the possibility, we performed systematic evolution of ligands by exponential enrichment (SELEX) (fig. S3B), a technique for identifying single-stranded (ss) "aptamers" recognized by sequence-specific binding proteins (*46*). By screening a library of 30-mer random ssDNA sequences using a recombinant glutathione *S*-transferase (GST) fusion protein comprising mCDCA7 CRD, we identified two highly similar sequences, named Seq-1 and Seq-2 (Fig. 2A). EMSA verified the binding of both sequences by the CRD of mCDCA7 and hCDCA7 and the disruption of binding by ICF3 missense mutations (Fig. 2, B and C, F probe = Seq-1; fig. S4, ss-2 = Seq-2). The binding was highly specific, as the CDCA7 CRD failed to bind the reverse complementary

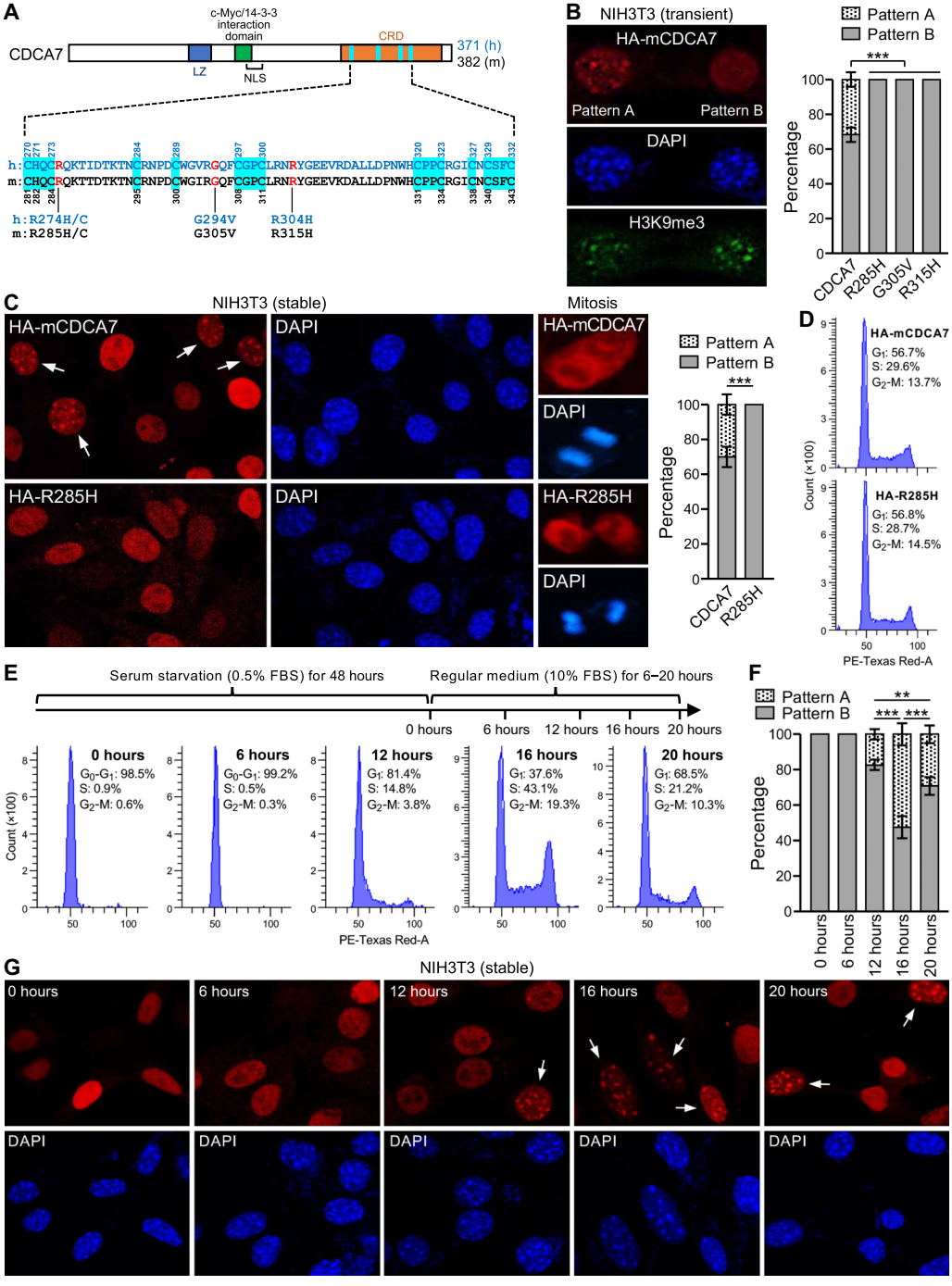


Fig. 1. CDCA7, but not ICF3 mutants, is concentrated in heterochromatic foci at S phase. (**A**) CDCA7 protein with the defined domains: a leucine zipper (LZ) motif, a c-Myc/14-3-3 interaction motif, a nuclear localization signal (NLS), and a CRD that contains four CXXC motifs (highlighted in cyan). The amino acid sequences from the first to fourth CXXC motifs in human (h) and mouse (m) CDCA7 are shown, with the Cys and His residues involved in zinc coordination being highlighted in cyan and numbered. The ICF3 missense mutations are indicated. (**B**) IF data showing that HA-tagged mCDCA7 transiently expressed in NIH3T3 cells exhibits two nuclear localization patterns, pattern A (enrichment in heterochromatin foci) and pattern B (no enrichment in heterochromatin foci), whereas ICF3 mutant proteins (RH, G305V, and R315H) exhibit only pattern B. Shown are representative images (left) and percentages (right) of the two patterns. (**C**) IF data with stable expression of HA-tagged mCDCA7 or the RH mutant in NIH3T3 cells. Shown are representative images (left) and percentages (right) of the two localization patterns. During mitosis, both WT and mutant mCDCA7 proteins are localized throughout the cell, excluding the chromosomes. (**D**) Flow cytometry analysis showing that NIH3T3 cells stably expressing WT or mutant mCDCA7 have similar cell cycle profiles. (**E** to **G**) Cell cycle synchronization experiments using serum starvation, which suggest mCDCA7 enrichment in heterochromatin foci during S phase. Shown are the cell cycle profiles (E), percentages of the two localization patterns (F), and representative images (G) at different time points. The quantification data in (B, C, and F) represent means ± SD from three experiments, with at least 100 interphase cells being counted each time. Statistical analysis was done using one-way analysis of variance (ANOVA). $^{**}P < 0.01$; $^{***}P < 0.001$. Pattern A cells in (C and G) are indicated by arrows.

probe (R probe), a double-stranded DNA (dsDNA) probe (annealed F/R probes) (Fig. 2B, also see fig. S8I below), an RNA probe with the same sequence as F probe, and an RNA/DNA hybrid probe (annealed RNA F probe and DNA R probe) (Fig. 2D). Pull-down and enzyme-linked immunosorbent assay (ELISA) experiments verified that HA-tagged full-length mCDCA7, but not mCDCA7 containing the RH mutant, binds a DNA probe similar to the F probe (fig. S3, C and D).

By testing different fragments in the 30-mer Seq-1, we determined that the 4 nucleotides (nt) at the beginning (5′ end) and 1 nt at the end (3′ end) are not required for binding by mCDCA7 CRD (fig. S4, ss-3 to ss-13). The 25-mer CRD-binding sequences of Seq-1 and Seq-2 (Fig. 2E) contain two short dyad symmetries, a pair of tetranucleotides (CGGT and ACCG) in the middle (highlighted in red), and a pair of dinucleotides (GC and GC in Seq-1, TC and GA in Seq-2) at the ends (highlighted in yellow). Two possible non-B DNA structures could be formed: a hairpin by intra-strand base pairing, with two stems (S1 and S2) and two bubbles (B1 and B2) (Fig. 2F) or a more complex structure by inter-strand base pairing, with two symmetric halves (Fig. 2G). Using dsDNA probes formed by annealing single-stranded oligos, we found that both the hairpin (Fig. 2F) and the more complex structure (Fig. 2G) can be bound by mCDCA7 CRD (Fig. 2H). Binding of the non-B DNA was also verified with various other probes (fig. S4, ss-14, ss-15, and ss-18).

Extensive mutagenesis of the non-B DNA revealed the following requirements for CDCA7 binding. First, stem S1 must have at least 2 base pairs (bp) immediately next to bubble B1, and the sequence is not important (fig. S4, ss-6 to ss-13; fig. S6, ss-49 to ss-51, ss-121 to ss-125). Second, the number of nucleotides that form bubble B1 (CCTGT and TTT) cannot be changed (fig. S5, ss-20 to ss-45; fig. S6, ss-52 to ss-54, ss-118 to ss-120), and the TTT sequence is critical (fig. S6, ss-109 to ss-117), but the CCTGT sequence can be CNNGT

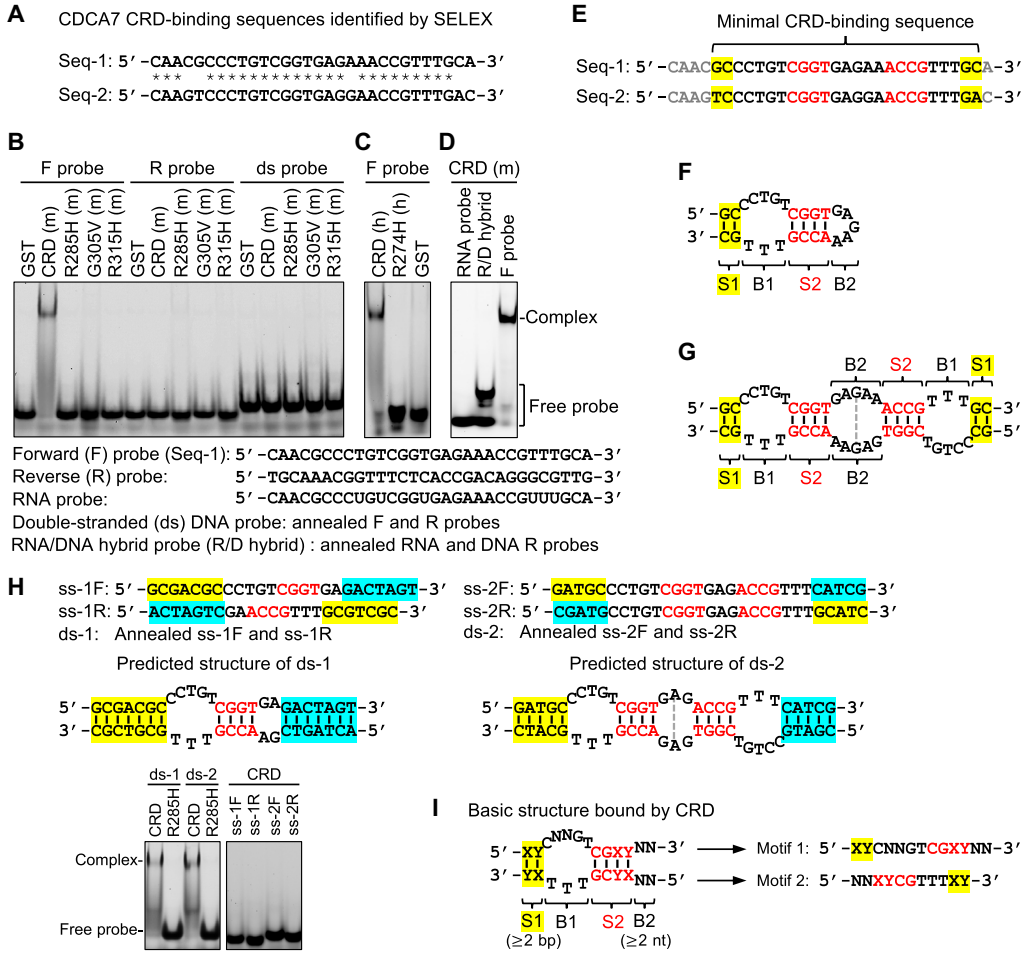


Fig. 2. The CDCA7 CRD binds a specific non-B DNA. (**A**) The two sequences, Seq-1 and Seq-2, identified by SELEX (identical bases indicated by *). (**B** to **D**) EMSA experiments using GST fusion proteins consisting of the CRD of mCDCA7, hCDCA7, or CRD with ICF3 mutations. The probes used for EMSA were shown: ssDNA probes (F probe, identical to Seq-1; R probe, reverse complementary to F probe), RNA probe (sequence identical to that of F probe), dsDNA probe (annealed F and R probes), and RNA/DNA (R/D) hybrid probe (annealed RNA and DNA R probes). (**E**) The minimal CRD-binding sequence determined by mutagenesis (see fig. S4). (**F** and **G**) Possible non-B structures formed by the CRD-binding sequence of Seq-1, through intra-strand (F) or inter-strand (G) base pairing. (**H**) EMSA data with ds probes formed by annealing ss oligos, which showed that the non-B structures shown in (F) and (G) can both be bound by mCDCA7 CRD. Note that the ssDNAs (ss-1F, ss-1R, ss-2F, and ss-2R) used to form ds probes are not bound by the CRD. (**I**) The basic structure bound by the CDCA7 CRD, as defined by mutagenesis results shown in figs. S4 to S6, and the two sequence motifs, motif 1 and motif 2, that form the basic structure. N means any unpaired nucleotide, and X:Y means any base pair.

(N means any nucleotide) (fig. S6, ss-55 to ss-69). Third, stem S2 must be 4 bp (fig. S4, ss-17; fig. S6, ss-70 to ss-81, ss-97 to ss-108), and the C:G and G:C base pairs at positions 1 and 2 are essential, whereas the base pairs at positions 3 and 4 can be formed by any nucleotides (fig. S6, ss-126 to ss-137). Fourth, bubble B2 must have at least 2 nt on each strand (fig. S5, ss-46 to ss-48), but the sequence does not seem to be important (fig. S6, ss-82 to ss-96). Thermal stability analysis of various probes demonstrated that the non-B structure is stable, with the melting temperature (T_m) ranging from 54° to 62° to 75°C, depending on the overall length, in comparison to that of the B-DNA control (73°C) (fig. S7). While DNA binding was examined by EMSA, the binding affinities of some probes were measured by isothermal titration calorimetry (ITC) using the CRD of mCDCD7 or hCDCA7, which generally confirmed the EMSA results (figs. S8 to S10). For example, ITC revealed the importance of the triple-T element for CDCA7 binding, with the middle T being the most critical (fig. S8G).

In summary, the basic structure bound by the CDCA7 CRD is a non–B-form DNA with a stem S1 of at least 2 bp (shown as X:Y and Y:X, meaning any base pairs), a bubble B1 formed by CNNGT on one strand and TTT on the other, and a 4-bp stem S2 containing an essential CpG dyad, followed by a bubble B2 formed by at least two mismatches (Fig. 2I). The two sequence elements that form the basic non-B structure are referred to as motif 1 (5′-XYCNNGTCGXYNN-3′) and motif 2 (5′-NNXYCGTTTXY-3′) with base pairs underlined (Fig. 2I). Notably, these two motifs can exist on two separate DNA strands (Fig. 2H, also see fig. S10 below), raising the possibility of distant motifs coming together to form the non-B DNA structure.

Structural investigations

We used a set of oligonucleotides for co-crystallization trials, with a length varying from 36 to 26 nt by reducing 1 nt at a time from both ends. The 24 nt is a minimal length for mCDCA7 CRD to bind, as further shortened 22 and 20 nt resulted in >20× and >70× reduced binding affinity, respectively (Fig. 3A and fig. S8A). We determined five structures of mCDCA7 CRD in complex with 36 (in two different crystallographic cell dimensions), 34, 32, and 26 nt in the resolution range of 2.6 to 1.6 Å (fig. S11 and table S1). The protein-DNA complexes were all crystallized in space group *C*2, with varied cell dimensions of crystal lattices, containing either one or two protein-ssDNA complexes per crystallographic asymmetric unit. In addition, we determined mCDCA7 or hCDCA7 CRD in complex with a hemi-methylated CpG site in the context of 32 nt (see below). Thus, we describe the structures of complex with the 32 nt determined at 1.9-Å resolution. All 32 nt are clearly resolved in the electron densities.

Non–B-form DNA

Instead of forming a hairpin structure, the two annealed single-stranded oligos couple to each other and form a non-B DNA conformation with two symmetric halves, and each half is bound by one CDCA7 CRD (Fig. 3B). This implies that in the crystal, the protein domain and ssDNA are in equal molar ratio. As expected, the non-B DNA can be divided into four sections: a 6-bp stem S1, a bubble formed by a 5-nt bulged loop of the top strand and a 3-nt triple T of the bottom strand, a 4-bp stem S2, and two purine mismatches. The 6-bp stem S1 was coaxially stacked with the neighboring DNA molecule, forming a pseudo-continuous duplex between the two DNA molecules throughout the crystal lattice (Fig. 3C). We numbered the DNA sequence as 1 to 13 for the top strand (motif 1) according to the basic CRD-binding structure (Fig. 2I) and used subscribed T_1, T_2, and T_3 for the triple-T element of the bottom strand (motif 2) (Fig. 3B).

The axes of the two stems, the 6-bp stem S1 and the 4-bp stem S2, are in an L-shaped ~90° sharp turn (Fig. 3B). The base pairs in the two stem regions obey the Watson-Crick hydrogen bonding (H-bond) patterns (Fig. 3, D, E, and N to Q). The sharp turn of the helix is mediated by the 5-nt bulged loop. The bulge contains an intra-strand C3:G6 base pair (Fig. 3F), which perfectly stacks with the last base pair of stem S1 (Fig. 3G). For the 2 nt between C3 and G6, C4 protrudes from the bulge (Fig. 3H) and T5 stacks on the other side of the C3:G6 base pair (Fig. 3I). DNA binding assays revealed that the C3:G6 base pair cannot be changed to the three other base pairs (fig. S5, ss-138 to ss-140; fig. S9D), but the 2 nt (C4 and T5) between C3 and G6 can be substituted with any nucleotides (fig. S6, ss-58 to ss-63).

The last nucleotide of the bulge, T7, is surrounded by two thymine residues of the triple-T element of the opposite strand: stacking with T_1 (Fig. 3J) and making a single H-bond with T_2 (Fig. 3K). The T7 and T_2 mismatched bases stack with the first base pair of stem S2 (Fig. 3L). The last thymine residue, T_3 of the triple-T element, locates in the minor groove side of stem S2 and spans the first two base pairs of the CGCT tetranucleotide sequence (Fig. 3M). A single-nucleotide substitution at positions of T_1, T_2, or T_3 to cytosine (T > C) led to a substantial reduction in binding affinity by factors of 11.5×, 58×, or 11×, respectively (fig. S8G). This reduction can be attributed to the alteration in H-bond potential, which occurs because of the switch in H-bond donor and acceptor roles along the Watson-Crick edge between the N3 and O4 atoms of thymine and the N3 and N4 atoms of cytosine (fig. S8G).

After the 4-bp stem S2 (Fig. 3, N to Q), the first purine mismatch at position 12 forms a noncanonical base pair via two hydrogen bonds (Fig. 3R). Like A:T base pair having two H-bonds, the G:A mismatch has the same thermal stability (*47*). The second purine mismatch has their separate ways, with one (A13) staying stacked with the neighboring bases and the other (G13) flipping out and stacking with the side chain of Trp^{301} (Fig. 3S). Subsequently, the inter-strand sugar-sugar distance decreased to 3.8 Å from that of ~10.5 Å in the stem B-DNA. In addition, divalent metal ions (Mg^{2+} used in the crystallization) bind between the phosphate groups of close apposition of DNA strands as well as bridge between two unpaired bases (fig. S12, A to C). The C2′ atoms of deoxyribose rings, particularly those in the non-base paired regions (B1 and B2), are in van der Waals contacts with other bases or protein side chains (fig. S12, D to F), suggesting that relief from the steric repulsion of the ribose 2′-OH group in RNA can allow non-B DNA to fold on its own and/or interact with CDCA7.

Three-zinc–containing DNA binding domain

Unlike any other DNA binding domains, the CDCA7 CRD adopts a "cross-braced" topology of three Zn^{2+}-coordinating residues by 11 cysteines and one histidine (Fig. 4, A and B; the Zn^{2+}-coordinating residues are highlighted and labeled in Fig. 1A). The first zinc ion (Zn1) is coordinated by two CXXC motifs, $C_{281}HQC_{284}$ and $C_{308}GPC_{311}$. His^{282}, immediately following Cys^{281}, points to the opposite direction and forms a second set of zinc coordination (Zn2) with $C_{338}X$-$C_{340}XXC_{343}$. The third zinc ion (Zn3) is organized by $C_{295}XXXXC_{300}$ and $C_{331}XXC_{334}$. The three sets of zinc coordination residues are interconnected: His^{282} connects Zn1 and Zn2, the three-residue linker

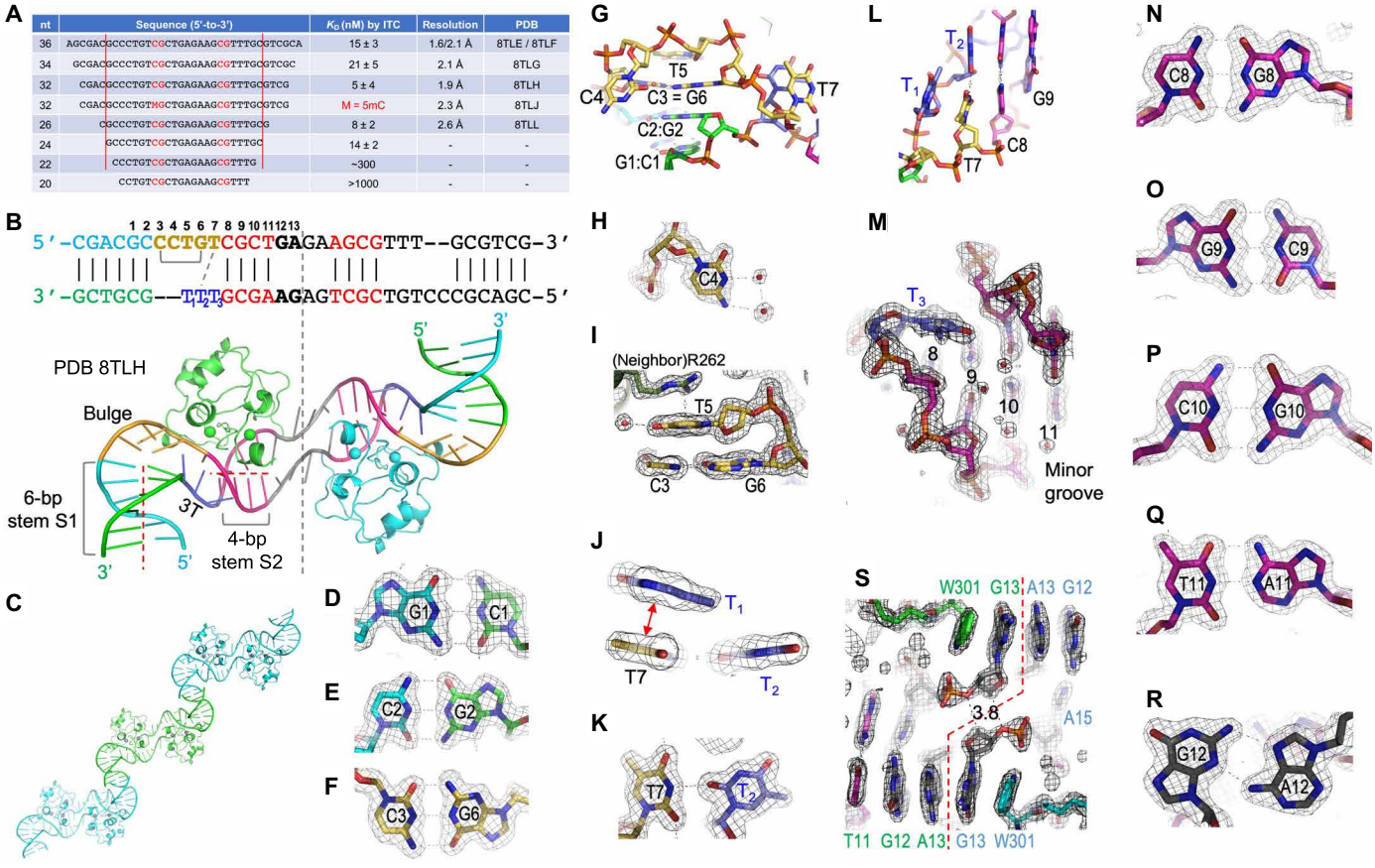


Fig. 3. The CDCA7-bound DNA adopts a non-B-form conformation. (**A**) Summary of DNA sequences used for ITC binding measurements and structural information [Protein Data Bank (PDB) accession number and corresponding resolutions]. (**B**) The 32-nt oligos used for co-crystallization and the nucleotide numbering above the sequence. The subscribed T_1, T_2, and T_3 are used for the triple-T element of the bottom strand. The 6-bp stem S1, the 4-bp stem S2, bulged region, and 3T are labeled and colored coded. The two dashed red lines indicate the axes of the two stems, which are in a ~ 90° turn. (**C**) Two neighboring CDCA7 CRD-DNA complexes form a pseudo-continuous duplex between the two DNA stems. (**D**) G1:C1 base pair. (**E**) C2:G2 base pair. (**F** and **G**) An intra-strand C3 and G6 base pair (F), which stacks with the last base pair of stem S1 (G). (**H**) C4 protrudes from the bulge and is unstacked. (**I**) T5 stacks with C3:G6 on one side and Arg^{262} of neighboring molecule on the other side. (**J** and **K**) T7 stacks with T_1 and H-bonds with T_2. (**L**) T7 and T_2 mismatch stacks with the first C:G base pair of stem S2. (**M**) T_3 is located in the minor groove of CpG dinucleotide at positions 8 and 9. (**N** to **Q**) The four base pairs of stem S2. (**R**) A purine mismatch at position 12. (**S**) Trp^{301} stacks with an extrahelical G13. The dashed red line indicates the two symmetric halves of the crystalized complexes. Two pairs of deoxyribose C4′ atom of G13 and the O3′ atom of G13 from the opposite strand are in close contact. The DNA omit electron density was contoured at 3σ above the mean.

between Cys^{334} and Cys^{338} links Zn3 to Zn2, and Gly^{305} is part of the linker between Cys^{300} (for Zn3) and Cys^{308} (for Zn1). Gly^{305}, an ICF3 mutated residue, sits right next to Zn3, with an interatomic distance of 4.5 Å (Fig. 4C). The substitution of Gly^{305} with valine (G305V) would result in repulsion and disruption of Zn3 binding. The cross-braced three-zinc coordination folds the CDCA7 CRD into a globular domain with five short helices and two short strands (Fig. 4D). The unique zinc coordination by CDCA7 differs from the three-zinc-ion–coordinating histone-binding ADD domain in ATRX and DNMT3 family members (fig. S13).

DNA base–specific recognition

The triple-T element and the tetranucleotide stem S2 provide most of the functionally important interactions with CDCA7. We made the following observations. First, the N3 atom of T_1 along the Watson-Crick edge makes an H-bond with Ser^{268}, which in turn interacts with the phosphate group between T5 and G6 of the opposite strand (Fig. 4E). Second, Gln^{286} spans two neighboring stacking bases, T_2 and the Gua of the C:G base pair at position 8 (Fig. 4F). Like T_1, the Gln^{286}-mediated interaction with T_2 is via the N3 atom of the Watson-Crick edge (Fig. 4G), and $T_2 > C$ substitution reduced the binding by a factor of 58× (fig. S8G). Third, the CpG dinucleotides of stem S2 have the most extensive interactions in both the major and minor grooves. On the major groove side, the C:G base pair at position 8 has Gln^{286} interaction with the O6 atom of Gua and the main chain carbonyl oxygen atom of Cys^{284} interaction with the N4 atom of Cyt (Fig. 4H). We note that the side chain of Gln^{286} has saturated potential of its ability to form H-bonds: its amide nitrogen atom and carbonyl oxygen atom each having two H-bonds. On the minor groove side, T_3 of the triple-T element provides an H-bond with the N2 atom of Gua (G8). Fourth, the G:C base pair at position 9 has Arg^{285} on the major groove side, forming the bident H-bonds with the Gua (via the guanidinium group) and an H-bond with the O4 atom of Cyt (via the main chain carbonyl oxygen) (Fig. 4I). The Arg-Gua interaction is common in specific Gua recognition, but it is unique to have an Arg residue (involving both side-chain and

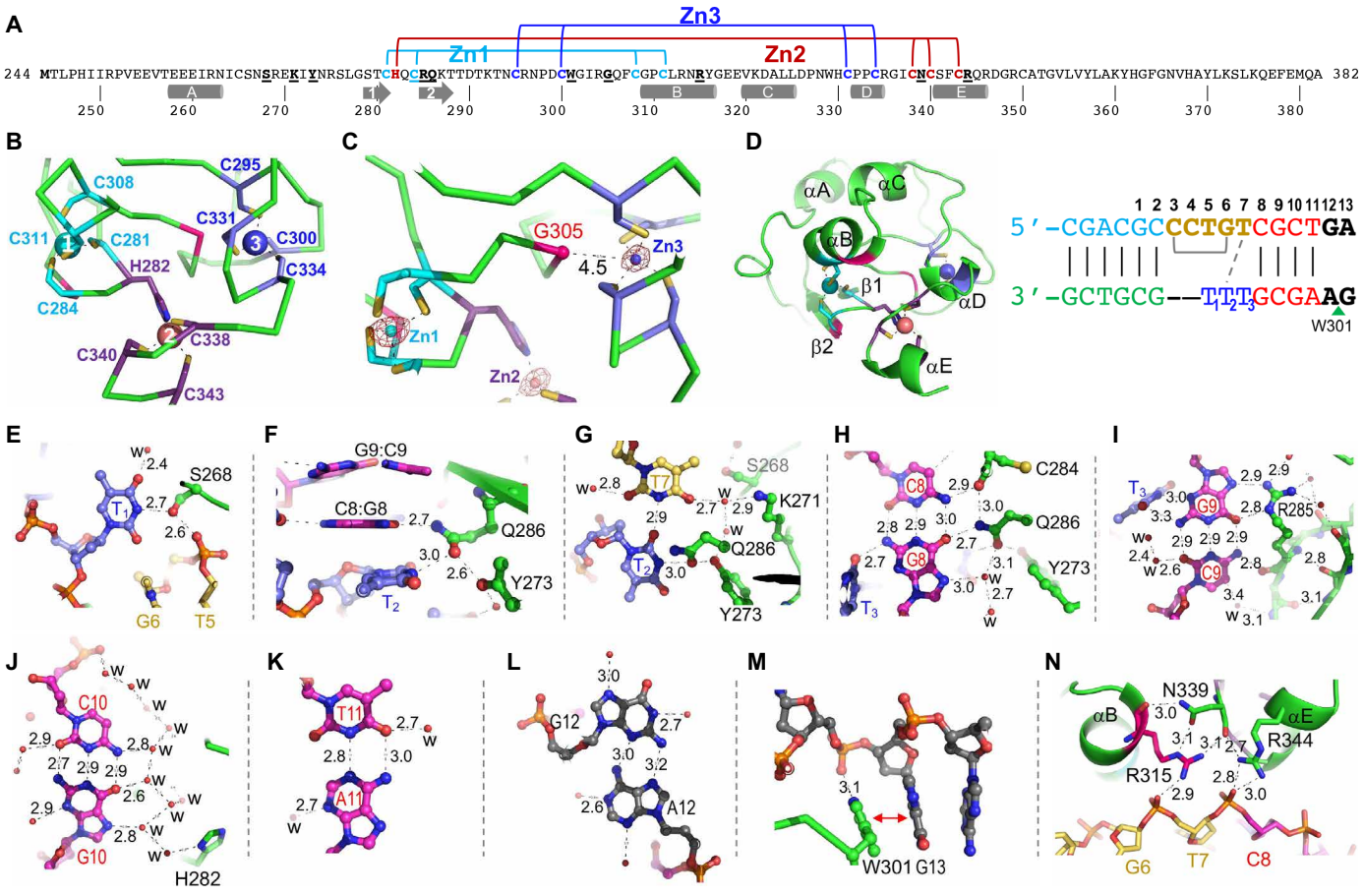


Fig. 4. mCDCA7-DNA interactions. (**A**) Cross-brace topology of three Zn binding. Bold residues involved in chelating Zn and DNA binding are underlined. (**B**) Three zinc atoms are coordinated by 11 cysteines and one histidine. (**C**) Gly^{305}, an ICF3 mutated residue, sits right next to Zn3 and links Zn1 and Zn3. The Zn omit electron density is contoured at 5σ above the mean. (**D**) A globular CRD with five short helices and two short strands. The residues after helix E are disordered. For convenience, the DNA sequence and its numbering are shown again. (**E**) Ser^{268} interacts with T_1. (**F**) Gln^{286} bridges with T_2 and G8. (**G**) Gln^{286} interacts with T_2. (**H**) The main chain carbonyl oxygen of Cys^{284} and Gln^{286} interact with C:G base pair at DNA position 8. (**I**) Arg^{285} interacts with G:C base pair at DNA position 9 via both side chain and main chain atoms. T_3 provides two additional H-bonds with G9 at the DNA minor groove. (**J**) Water-mediated interactions with C:G base pair at position 10. (**K**) Water-mediated interactions with T:A base pair at position 11. (**L**) A purine mismatch at DNA position 12. (**M**) Trp^{301} stacks with G13 and forms an H-bond with the phosphate group. (**N**) Arg^{315} and Arg^{344} interact with two neighboring DNA phosphate groups.

main-chain atoms) interaction with both bases of a G:C base pair (fig. S12, G to I). The ICF3 mutation of Arg^{285}-to-His or -Cys (RH/C) would certainly disrupt the Gua recognition. On the minor groove side, T_3 provides two additional H-bonds with the N2 and N3 atoms of Gua (G9). Emphasizing these interactions is the fact that the next two base pairs of stem S2, C:G at position 10 and T:A at position 11, as well as the purine mismatch at position 12, do not conduct direct interactions with CDCA7, instead forming extensive water-mediated interactions (particularly C:G at position 10) (Fig. 4, J to L). Last, Trp^{310} forms an aromatic stack interaction with the extra helical Gua at position 13 as well as an H-bond with the associated phosphate group (Fig. 4M). As shown by the DNA binding assays, an intra-strand C3:G6 base pair of the bulged loop cannot be altered by the three other base pairs. We observed a water-mediated interaction between G6 and Arg^{269}, which also stacks with T_1 (fig. S12D).

In addition to the base-specific interactions, CDCA7 contacts 10 phosphate groups, including Arg^{315} interaction with the phosphate group between G6 and T7 and Arg^{344} interaction with the phosphate group between T7 and C8 (Fig. 4N). The two long side-chain conformations of arginine residues are further stabilized by Asn^{339} (Fig. 4N), enhancing the Arg-DNA phosphate contacts. The ICF3 mutation of R315H, the substitution of Arg^{315} by a shorter His side chain, would be disruptive for DNA binding. In summary, CDCA7 residues associated with Zn1 binding, Cys^{284}-Arg^{285}-Qln^{286}, Arg^{315} between Zn1 and Zn2 binding, and Zn3-associated Arg^{344} provide the most functionally important interactions in recognizing the DNA bases of the CpG dinucleotides as well as DNA phosphate backbone.

One of the two motifs that form the non-B DNA is highly enriched at the centromeres of human chromosomes

The basic non-B DNA structure bound by the CDCA7 CRD is formed by two sequence motifs: a 13-mer motif 1 (5′-XYCNNGTCGX YNN-3′) and an 11-mer motif 2 (5′-NNXYCGTTTXY-3′) (Fig. 2I). The two motifs that form the structure do not necessarily need to be continuous or adjacent to each other and, indeed, can exist in two separate DNA strands (Fig. 2H and fig. S10). Thus, we searched the

two motifs in the complete human genome assembly T2T-CHM13v2.0 (*48*, *49*), including the recently assembled Y chromosome (*50*). Motif 1 was distributed throughout the genome, with modest enrichment toward one or both ends of some chromosomes. Motif 2, notably, showed a prominent peak on each chromosome at the centromere, as evidenced by centromere protein A (CENP-A) occupation (Fig. 5 and fig. S14). Human centromeres are defined by αSat, an AT-rich repeat family composed of ~171-bp monomers, which can occur either as large arrays of higher-order repeats (HORs) or stretches of divergent monomers (*49*, *51*). Annotations of satellite repeats (*49*, *50*) confirmed that the centromeric peaks were present in αSat, mostly in active HORs (Fig. 5 and fig. S14), where kinetochore proteins bind (*52*, *53*). The CGTTT sequence of motif 2, which provides most of the functionally important interactions with CDCA7 (Fig. 4), is present in many αSat sequences deposited in GenBank. Motif 1 was depleted at motif 2 peaks at centromeres (Fig. 5 and fig. S14), perhaps due to the highly biased sequence of the αSat.

In addition to dramatic centromeric enrichment, broad peaks of motif 2 were observed in the pericentromeric regions of chr 1, 15, and, less abundantly, chr 16 (Fig. 5, A, C, and D). They were present in HSat2 (chr 1 and 16) and HSat3 (chr 15), respectively (Fig. 5, A, C, and D). The broad motif 2 peaks on chr 1 and 15, like centromeric peaks, were accompanied by motif 1 depletion (Fig. 5, A and C), whereas motifs 1 and 2 were both slightly enriched in the pericentromeric region of chr 16 (Fig. 5D). A notable exception is HSat3

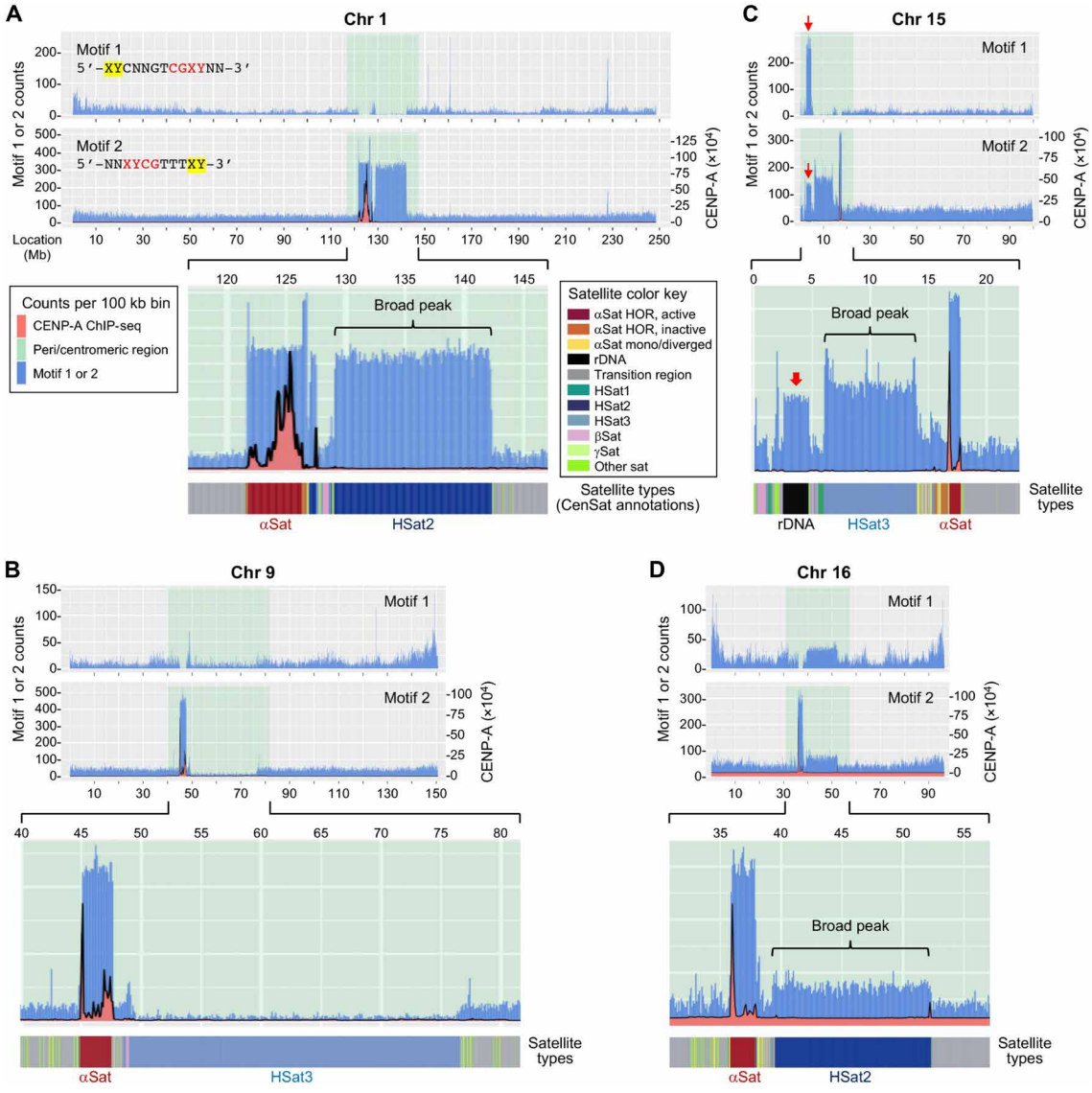


Fig. 5. Motif 2 of the non-B DNA is highly enriched at centromeres of human chromosomes. Motifs 1 and 2 that form the non-B DNA (see Fig. 2I) were searched in the T2T_CHM13v2.0 human genome assembly. Shown are the results of chr 1 (**A**), chr 9 (**B**), chr 15 (**C**), and chr 16 (**D**). The results of other chromosomes are shown in fig. S14. Top: chromosome-wide views of motifs 1 and 2 counts (blue). Highlighted in light green are the peri/centromeric regions. *X* axis, chromosome locations (Mb); *y* axis on the left, motif 1 or 2 counts; *y* axis on the right, CENP-A ChIP signal (orange, $\times 10^4$). Middle: 5× expansion of the peri/centromeric regions of the motif 2 panel (the *y* axis is not proportionally magnified). Bottom: CenSat annotations of satellite types in peri/centromeric regions.

in the pericentromeric region of chr 9, which is frequently hypomethylated in ICF syndrome (*23*), showed depletion of motif 2 (Fig. 5B).

The five acrocentric chromosomes (chr 13, 14, 15, 21, and 22) formed a unique group in that each had a major peak of both motifs 1 and 2—with motif 1 being more abundant—at the same location in the pericentromeric region, which coincides with ribosomal DNA (rDNA) repeats (Fig. 5C and fig. S14, K, L, Q, and R, indicated by red arrows). Chr 14 and 22 each had an additional peak of motifs 1 and 2 at the very end of the p arm, present in satellite DNA annotated as "other satellite" (fig. S14, L and R, indicated by blue arrows). Analysis of the colocalized motifs 1 and 2 peaks on the acrocentric chromosomes revealed considerable numbers of paired motifs (table S2), suggesting abundant opportunities for neighboring motifs 1 and 2 to form the non-B DNA structure in these regions.

On chr Y, both motifs 1 and 2 showed similar distribution patterns—one major peak at the very end of the p arm (indicated by blue arrows in fig. S14T), two major peaks in the pericentromeric region (indicated by red arrows), and many peaks in the large heterochromatic q12 region—with motif 2 being consistently more abundant than motif 1. Satellite annotations revealed that they were present in gamma satellite (γSat) (p-arm end), other satellite (pericentromeric region), and HSat3 (q12 region), respectively (fig. S14T).

Examination of the 4-bp stem S2 in peri/centromeric regions revealed chromosome-specific sequence preferences. For example, chr 14 and 16 strongly prefer CGCT and CGAT, respectively, in motif 1 (table S3), and most chromosomes prefer CCCG, TCCG, and TTCG in motif 2 (table S4). We note that the CCCGTTT sequence of motif 2 is identical to part of the 17-bp CENP-B box (CTTCGTTGGAAACGGGA) in αSat (*54*).

Strand-specific CpG methylation in the non-B DNA has opposite effects on CDCA7 binding

Stem S2 of the non-B DNA contains a CpG dyad that is essential for CDCA7 binding (Figs. 2I and 4H and I). Thus, we assessed the effects of CpG methylation. By annealing two single-stranded oligos with or without methylated cytosine (depicted as M), we generated probes that were unmethylated, fully methylated (on both strands), or hemi-methylated on either the forward strand (motif 1–containing Hemi-F) or reverse strand (motif 2–containing Hemi-R) (Fig. 6A). As revealed by EMSA, the mCDCA7 CRD failed to bind the fully methylated and Hemi-R probes but showed stronger binding to the Hemi-F probe, compared to the unmethylated probe (Fig. 6A).

To confirm the effects of methylation, we performed ITC using mCDCA7 CRD. The binding affinity of Hemi-F [dissociation constant (K_D) 1.6 to 2 nM] was higher than that of the unmethylated probe (K_D 9 nM) at the condition of 150 mM NaCl, whereas the binding affinity of Hemi-R (>0.5 μM) decreased by >300-fold compared to the Hemi-F probe, and the fully methylated probe failed to be bound (Fig. 6B). To further verify the results, we measured the binding affinities under increased ionic strengths (150 to 450 mM NaCl). While the binding affinities of the unmethylated probe decreased with increases in ionic strengths (K_D went from 9 nM at 150 mM NaCl to 170 nM at 450 mM NaCl), the binding affinities of the Hemi-F probe remained unchanged (K_D 2 to 7 nM under the concentrations tested) (Fig. 6C and fig. S8E).

Next, we determined the structures of the CRD of mCDCA7 and hCDCA7, respectively, in complex with hemi-methylated DNA in the context of 32 nt (table S1). Structural data indicated that the methyl group of 5-methylcytosine (5mC) at base pair position 8 (Hemi-F) is accommodated by forming a van der Waals contact and a weak C─H•••O type H-bond (3.5 Å) with the main-chain carbonyl oxygen of Gln^{283} (Fig. 6D). In addition, the methyl group of 5mC makes a van der Waals contact with Arg^{285}, which interacts with the neighboring Gua of the same DNA strand (Fig. 6E). We also modeled 5mC onto the cytosine of the opposite strand at the next G:C base pair (Fig. 6F). The methyl group is placed as close as 2.3 Å to the main-chain carbonyl oxygen of Arg^{285}, which is likely to cause steric obstruction with Arg^{285} in this conformation, perhaps explaining the substantial inhibition of binding to cytosine methylation at base pair position 9 (Hemi-R).

The C-terminal 36 residues of CDCA7 contributes to high affinity DNA binding

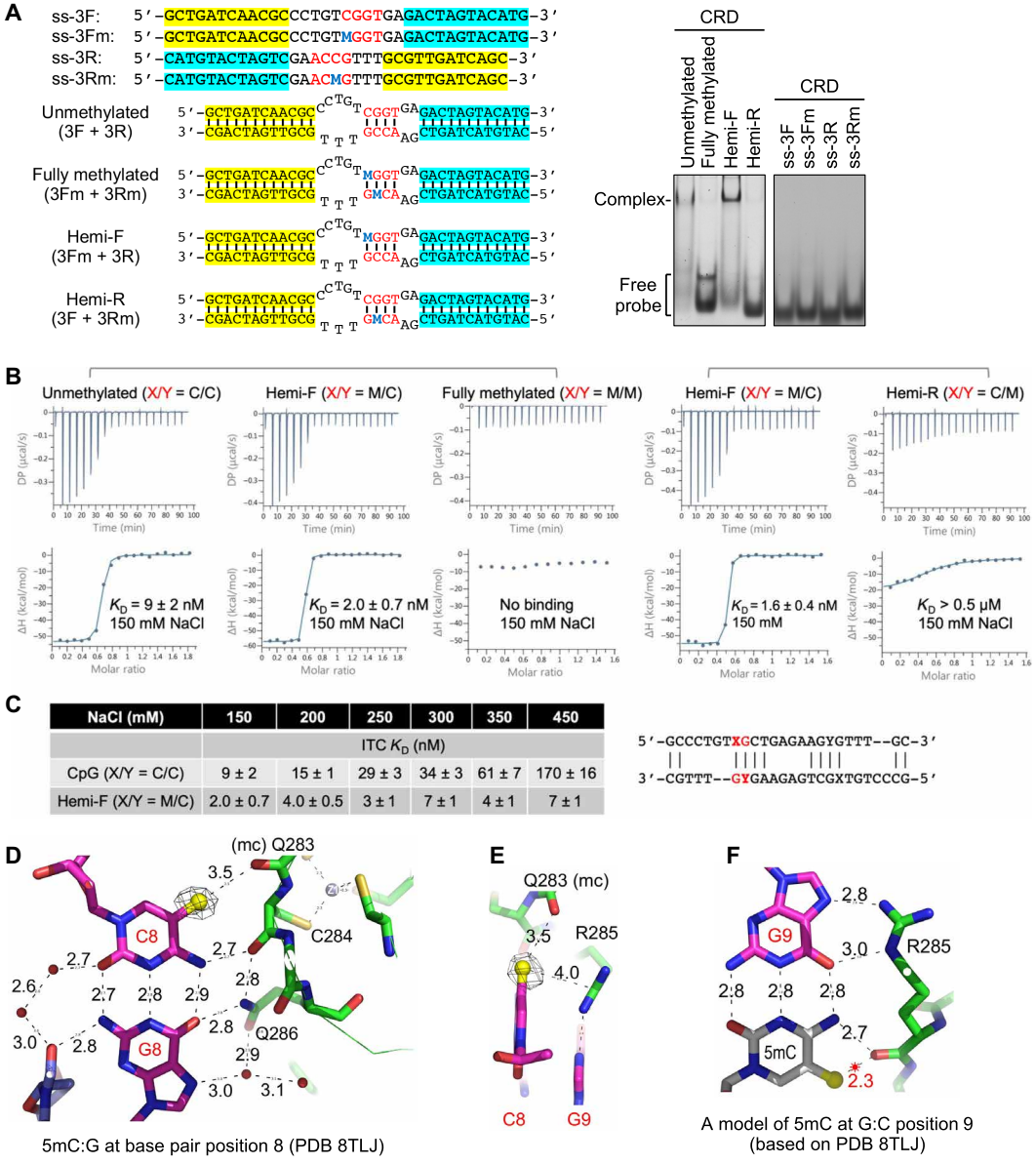
In our study, we observed that the structure extends only up to Arg^{346} in mCDCA7 and Arg^{335} in hCDCA7 despite our intention to express the complete C terminus (mouse residues 241 to 382; human residues 235 to 371). The absence of the final 36 residues at the C terminus could be due to structural disorder or degradation during the purification of the recombinant proteins. Notably, we observed two distinct protein bands during the purification process (fig. S15A). Initially, our focus was on the shorter band, hCDCA7(S). However, by implementing a modified protein purification protocol and using protease inhibitors, we successfully isolated the longer form of hCDCA7, which we have designated as hCDCA7(L).

We repeated the binding assays using EMSA with a non-B DNA composed of two separate strands (fig. S15B). Under conditions of 20 nM DNA probe and 150 mM NaCl, we noted strong binding of hCDCA7(L) with the Hemi-F and unmethylated DNA probes, followed by Hemi-R and fully methylated DNA. This observation aligns with our findings using the shorter form of hCDCA6(S) (Fig. 6B).

We further investigated the binding of fully complementary dsDNA by hCDCA7(L). As anticipated, there was no detectable binding to either unmethylated or fully methylated DNA probes. However, unexpectedly, hCDCA7(L) exhibited binding to the hemi-methylated top strand and a much weaker binding to the hemi-methylated bottom strand (fig. S15C), although the binding affinities were significantly lower than those to the corresponding non-B DNA probes (fig. S15, compare B and C). These results suggest that hCDCA7(L) has a higher affinity compared to the short form and that its binding is dependent on both the DNA structure (non-B versus B-DNA) and CpG methylation status. Specifically, the order of binding affinity for hCDCA7(L) is as follows: non-B hemi-F > non-B DNA > non-B hemi-R > B-DNA hemi F > B-DNA hemi-R.

Hemi-F, but not Hemi-R, inhibits CDCA7 foci formation and induces hypomethylation of centromeric satellite DNA

Collectively, our findings suggest that the CDCA7 CRD would preferentially bind a specific non-B DNA with stem S2 being hemi-methylated on the forward strand (motif 1), which perhaps contributes to CDCA7 enrichment in constitutive heterochromatin. We asked whether exogenous Hemi-F would affect CDCA7 foci formation. Different amounts of the Hemi-F, Hemi-R, or unmethylated probe (Fig. 6A) were transfected into NIH3T3 cells stably expressing HA-mCDCA7 (Fig. 1C). IF analysis at 48 hours after transfection revealed that Hemi-F, but not Hemi-R or unmethylated probe, inhibited the enrichment of mCDCA7 in heterochromatin foci (pattern A) in a dose-dependent manner (Fig. 7, A and B).



NaCl (mM)	150	200	250	300	350	450
	ITC K_D (nM)					
CpG (X/Y = C/C)	9 ± 2	15 ± 1	29 ± 3	34 ± 3	61 ± 7	170 ± 16
Hemi-F (X/Y = M/C)	2.0 ± 0.7	4.0 ± 0.5	3 ± 1	7 ± 1	4 ± 1	7 ± 1

Fig. 6. Strand-specific CpG methylation of stem S2 shows the opposite effects on CDCA7 binding. (**A**) EMSA data showing that mCDCA7 CRD–mediated binding of the non-B DNA is enhanced by methylation of the forward strand of stem S2 (hemi-F probe) but inhibited by methylation of the reverse strand (hemi-R probe) or both strands (fully methylated probe). Methylated cytosine (M) is shown in blue. Note that the ss oligos, before annealing, failed to be bound (right). (**B**) ITC assays with similar probes, which confirm that mCDCA7 CRD binds the hemi-F probe with higher affinity (K_D = 1.6 to 2.0 nM) compared to that of the unmethylated probe (K_D = 9 nM) but binds hemi-R weakly and does not bind the fully methylated probe. The mean and error estimate of K_D was derived from individual curve fitting. (**C**) ITC assays with increased salt concentrations, which show that the binding affinity of the unmethylated probe decreases with increasing concentrations of NaCl, whereas the binding affinity of the hemi-F probe is not affected by as high as 450 mM NaCl. (**D**) Structure of mCDCA7 in complex with methylated cytosine at position 8 (hemi-F). The methyl group is recognized by the main chain carbonyl oxygen of Gln[283]. (**E**) A methyl-Arg-Gua triad involving the neighboring 5mC and Gua of the same DNA strand. (**F**) A model of cytosine methylation at position 9 (hemi-R) resulting in repulsion with Arg[285].

We previously showed that disruption of *Cdca7* in mESCs results in substantial hypomethylation of the minor satellite repeats in centromeric regions (*38*) (Fig. 7C). Thus, we assessed the effects of Hemi-F and Hemi-R on DNA methylation by repeatedly transfecting the probes (once every other day) into WT mESCs. As shown in Fig. 7D, Hemi-F, but not Hemi-R, induced detectable loss of methylation at the minor satellite repeats after two times of transfection (day 4) and more obvious hypomethylation after three times of transfection (day 6). These results support the idea that CRD-mediated recognition of hemi-methylated non-B DNA formed at centromeres contributes to the specificity of CDCA7 in regulating DNA methylation.

HA-mCDCA7 ChIP-seq analysis reveals enrichment of motif 1–containing reads among unmapped reads

To identify genomic sequences that CDCA7 binds to, we performed chromatin immunoprecipitation sequencing (ChIP-seq). Facing the

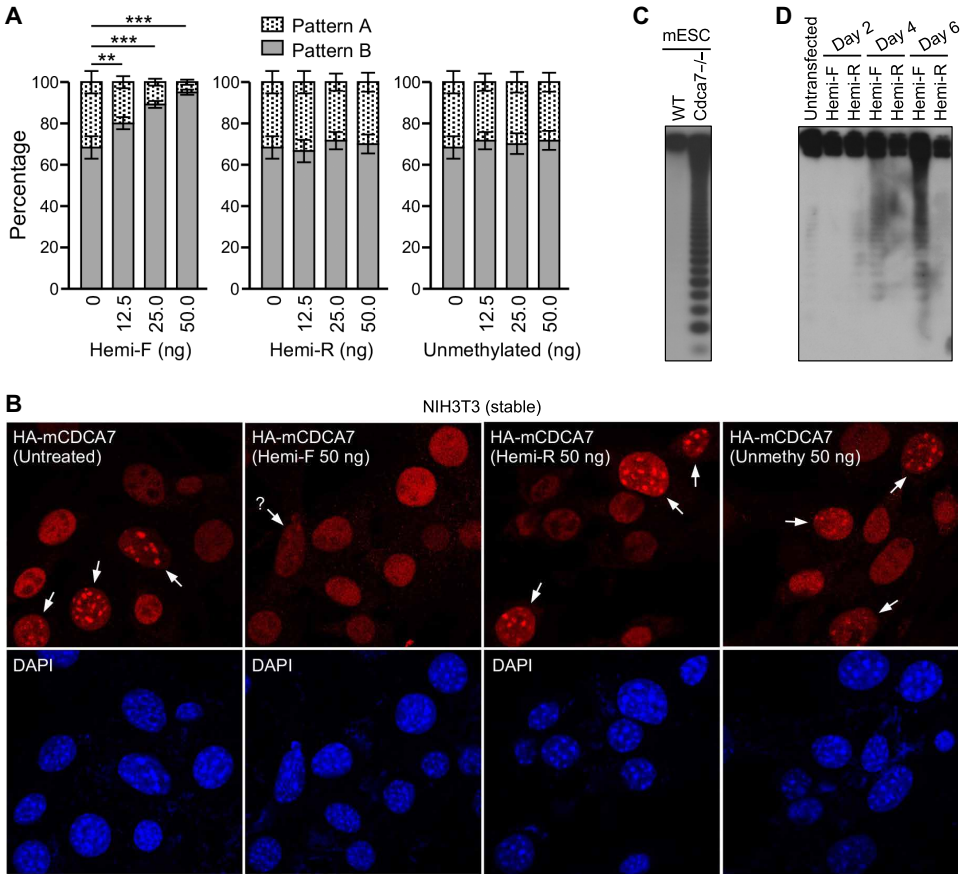


Fig. 7. Exogenous Hemi-F inhibits CDCA7 foci formation and induces hypomethylation of minor satellite repeats in murine cells. (**A** and **B**) IF data showing that transfection of Hemi-F, but not Hemi-R or unmethylated probe, in NIH3T3 cells stably expressing HA-mCDCA7 results in reduced numbers of cells with CDCA7 heterochromatin foci (pattern A). IF was performed 48 hours after transfection of the probes. Shown are the percentages of the two localization patterns (means ± SD from three experiments) (A) and representative images (B). Statistical analysis was done using one-way ANOVA. $^{**}P < 0.01$; $^{***}P < 0.001$. (**C**) Southern blotting analysis of genomic DNA from WT and *Cdca7*$^{-/-}$ mESCs after digestion with the methylation-sensitive restriction enzyme *Hpa*II, which shows substantial hypomethylation of minor satellite repeats in *Cdca7*-deficient cells (as evidenced by the smear on the gel, which indicates digestion of minor satellite DNA into smaller fragments). (**D**) Southern blot showing that repeated transfections of Hemi-F, but not Hemi-R, in WT mESCs induce loss of methylation at minor satellite repeats.

challenge of not having a ChIP-grade antibody for CDCA7, we opted to introduce HA-tagged WT mCDCA7 or the RH mutant into *Cdca7*$^{-/-}$ mESCs (*38*) (the decision to use murine cells was made before the release of the completed human genome assembly). To avoid the complications of overexpression, we selected stable clones (two per genotype: WT-4 and WT-9, RH-5 and RH-10) that exhibited expression levels of mCDCA7 comparable to those of endogenous mCDCA7 in WT mESCs (fig. S16A) for ChIP analysis using an HA antibody. This method has been successfully applied in our previous work to uncover a specific ZBTB24-binding motif, establishing *CDCA7* as a direct target gene of ZBTB24 (*36*).

For each sample, we generated approximately 35 to 60 million reads. Model-based analysis of ChIP-seq (MACS2) (*55*) revealed either no peaks or only a small number of peaks, a result that was somewhat unexpected but consistent with the notion that CDCA7 predominantly binds to sequences within peri/centromeric satellite repeats. The ChIP-seq libraries were sequenced in a 50-bp single-read run. Many reads containing CDCA7-binding sites were too short to be mapped to unique locations in repetitive regions. Significant fractions of reads with motif 1 (CNNGTCGXY) or motif 2 (XYCGTTT) were either unmapped or multimapped (fig. S16B). We noticed that motif 1–containing reads were enriched among unmapped reads in WT, but not RH, ChIP samples, compared to their input (fig. S16B, red arrows). Analyzing reads with individual motif 1s revealed that most of the 256 (4N × 4N × 4X × 4Y) instances were enriched in WT ChIP samples. However, a few high-count motif 1 sequences (e.g., CAAGTCGTC, CAAGTCGTA, and CACGTCGTA) showed no enrichment or high percentages of reads in input samples, skewing the data (fig. S17). On average, motif 1–containing reads were enriched ~12.6-fold among unmapped reads (Fig. 8A). Grouping motif 1s by different XY combinations on stem S2 showed substantial enrichment, ~7 to 24-fold, among unmapped reads in all groups (Fig. 8B). The enrichment was specific, as the RH mutant showed reduced binding, and reads with "motif 1 control sequences" showed no enrichment (Fig. 8, A and B). These results suggest that mCDCA7 binds motif 1 in murine cells.

It is unclear why motif 2–containing reads were not enriched in the ChIP-seq dataset (Fig. 8A and fig. S16B). The high percentages of motif 2–containing reads among multimapped and unmapped reads in input samples (fig. S16B) might have partially "masked" the

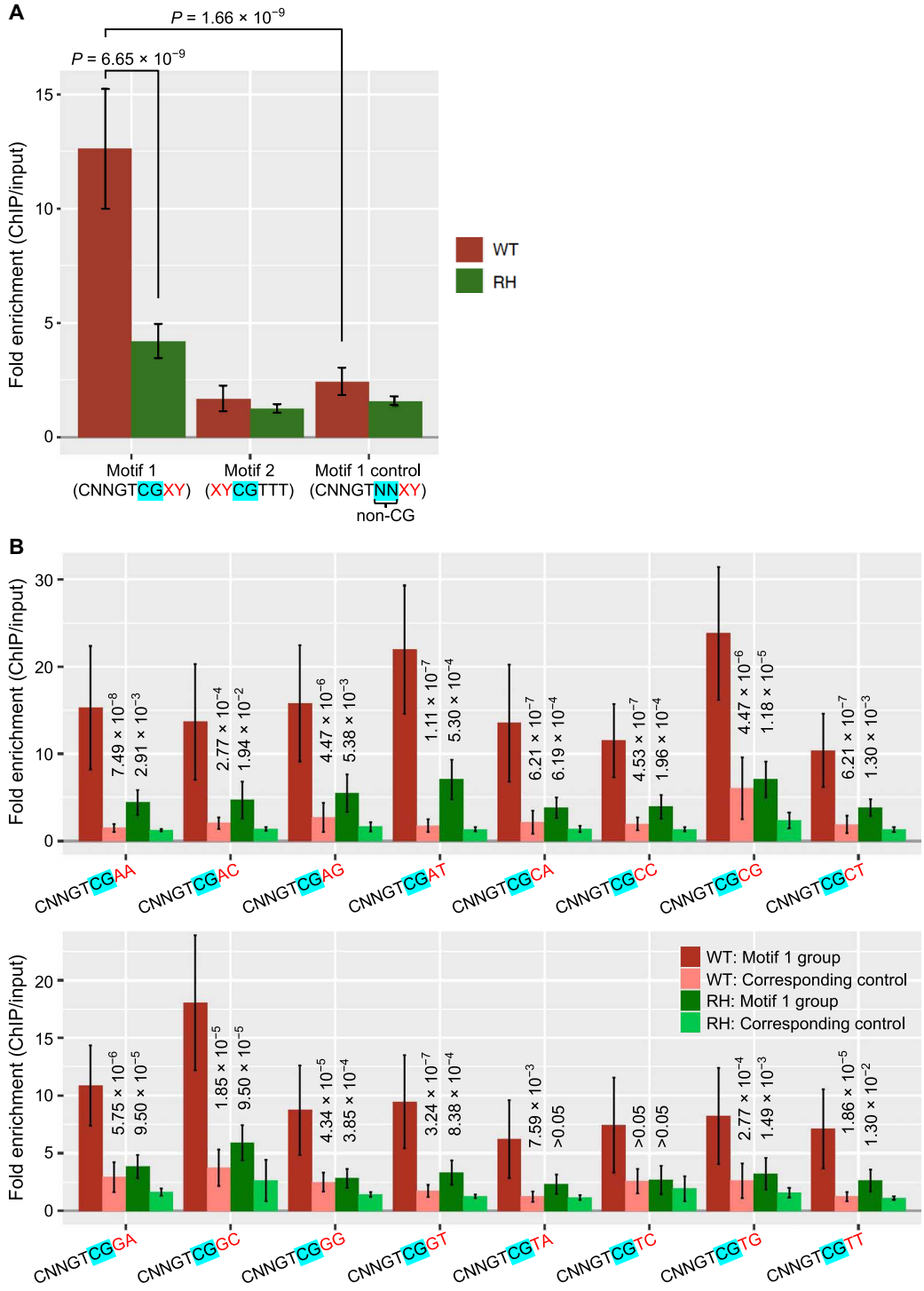


Fig. 8. ChIP-seq analysis reveals binding of motif 1 by mCDCA7. (**A**) Fold enrichment [means ± 1.96 × SE; 95% CI (confidence interval)] of reads containing all the 256 (4N × 4N × 4X × 4Y) motif 1 sequences, all the 16 (4X × 4Y) motif 2 sequences, and all the 2304 (256 × 9) motif 1 control sequences in unmapped reads. Motif 1 control sequences are identical to motif 1 sequences except that the CG dinucleotide critical for CDCA7 binding is replaced by non-CG (AA, AC, AT, GA, GC, GT, TA, TC and TT) dinucleotides. The Wilcoxon rank sum test (one-sided) was used to calculate *P* values. (**B**) Fold enrichment (means ± 1.96 × SE; 95% CI) of reads containing the 16 (4N × 4N) groups of motif 1 sequences with different XY combinations on stem S2, as well as their corresponding non-CG–containing control sequences, in unmapped reads. The two WT (WT-4 and WT-9) or RH (RH-5 and RH-10) replicates were combined. The *P* values above the pink and dark green bars indicate comparisons with the corresponding red bars and were calculated using the Wilcoxon rank sum test (one-sided).

enrichment. In addition, the transient and cell cycle–dependent binding of CDCA7 to the non-B DNA structure may mean that standard ChIP protocols are not optimal for capturing all binding events.

DISCUSSION

Repetitive sequences, which are prone to giving rise to non-B DNA (*56*), are major components of eukaryotic genomes, making up, for instance, over 50% of the human genome (*48*–*50*). Thus, the potential of genomic DNA to form non-B structures is enormous. Despite decades of investigations, the biological effects of non-B-form DNA, as well as the molecular mechanisms involved, are not well characterized. In this study, we demonstrate that the CDCA7 CRD forms a special DNA binding domain that recognizes a CpG-containing non-B DNA, providing a plausible explanation for the functional specificity of CDCA7 in the regulation of DNA methylation and shedding light on the pathogenesis of ICF syndrome. Unlike two other known methyl-CpG binding domains—the SET and RING–associated (SRA) domain of UHRF1, which binds specifically hemi-methylated CpG (*57*), and the methyl-CpG–binding domain (MBD) of MeCP2, which binds fully methylated CpG (*58*)—the CRD of CDCA7 preferentially binds strand-specific hemi-methylated CpG in the context of a non-B DNA. Furthermore, the SRA domain uses base flipping for binding an extra-helical 5mC in an aromatic cage, whereas the MBD and the CDCA7 CRD bind the intra-helical 5mC using an arginine residue in a methyl-Arg-Gua triad (Fig. 6E).

Hypomethylation of satellite repeats, most notably HSat2 on chr 1 and 16, HSat3 on chr 9, and αSat in centromeric regions (*23*), is considered the primary defect in ICF syndrome, which presumably underlies other cellular and clinical features, such as centromeric instability and antibody deficiency. Thus, elucidating the roles of ICF-related genes—*DNMT3B*, *ZBTB24*, *CDCA7*, and *HELLS*—in the regulation of DNA methylation is fundamentally important for understanding the pathophysiology of the disease. Previous studies suggest that CDCA7, by recruiting HELLS to heterochromatin, is a key player that controls DNA methylation at satellite repeats (*34*, *37*, *38*). However, little is known about how CDCA7 specifically recruits HELLS to heterochromatin. On the basis of our findings, we propose that CRD-mediated binding of the CpG-containing non-B DNA confers CDCA7 the specificity to regulate DNA methylation. First, the CDCA7 CRD binds the non-B DNA in a highly specific manner and the identified ICF3 missense mutations—all in the CRD—disrupt the binding. Second, the peri/centromeric regions in human cells have great potential to form the specific non-B DNA. Bioinformatic analysis of the two sequence motifs in the complete human genome assembly (*48*–*50*) revealed that, while motif 1 is distributed (more or less evenly) throughout the genome, motif 2 is highly enriched in centromeric αSat of all chromosomes and is also abundant in HSat2 in the pericentromeric regions of chr 1 and 16, two of the chromosomes most frequently affected in ICF syndrome (Fig. 5 and fig. S14). Intriguingly, some motif 2s are present in the 17-bp CENP-B box (*54*). In vitro evidence suggests that CpG methylation in the CENP-B box inhibits CENP-B binding (*59*). It would be interesting to determine the possible interplay between CDCA7 and CENP-B in αSat methylation, as well as centromere formation and functions. Third, ChIP-seq analysis suggests that mCDCA7 binds motif 1 in mESCs. Fourth, WT CDCA7, but not ICF3 mutants, is concentrated in constitutive heterochromatin foci during DNA replication, when negative supercoiling and ssDNA are created, both favoring the formation of non-B DNA structures (*56*). Given that the sequences of satellite DNA repeats are not conserved in the mouse and human genomes and, yet, CDCA7 foci were detected in both mouse and human cells (Fig. 1 and fig. S2), we speculate that the sequence motifs that form the non-B DNA are also abundant in heterochromatin regions in the mouse genome. As an example, we found that a 23-nt sequence on mouse chr 3 could be bound by mCDCA7 CRD in vitro (fig. S8F). However, the distribution of the two motifs in peri/centromeric regions in the mouse genome remains to be determined, as many repetitive sequences have yet to be assembled. Last, when introduced in cells, a hemi-methylated non-B DNA (Hemi-F) that is preferentially bound by the CRD can inhibit the formation of CDCA7 foci and induce hypomethylation of centromeric satellite repeats, whereas a similar non-B DNA (Hemi-R) that cannot be bound by CDCA7 shows no effects (Fig. 7). The results support the notion that CRD-mediated binding of the specific hemi-methylated non-B DNA contributes to the concentration of CDCA7 in constitutive heterochromatin during DNA replication.

Our observation that CDCA7 CRD–mediated binding of the non-B DNA is differentially affected by strand-specific asymmetric CpG methylation (Fig. 6) provides insights into the mechanism involved in methylation of peri/centromeric satellite repeats. On the basis of our results, CDCA7 would preferentially bind the non-B DNA structure formed with methylated motif 1 and unmethylated motif 2. One scenario is that, during DNA replication, when motif 1, from neighboring and/or distant genomic regions, and motif 2 form the non-B DNA structure, the motif-1 CpG methylation marks (on parental strands) would serve as templates for the methylation of motif 2 (on newly synthesized daughter strands) in peri/centromeric regions. We envisage two related roles for CDCA7: i) temporarily stabilizing the non-B structure; and ii) recruiting HELLS to peri/centromeric regions, where it performs chromatin remodeling and/or recruits components of the DNA methylation machinery to facilitate DNA methylation (*29*–*32*). As fully methylated non-B DNA would disrupt the binding by CDCA7, the 5mC templates in motif 1 and the methylation machinery could be propagated to neighboring motif 2s. Conceivably, the methylation marks deposited in motif 2s may also spread to neighboring CpG sites, further contributing to the efficiency in methylating satellite repeats. This proliferated process would be severely compromised in ICF syndrome because of DNMT3B inactivation (in ICF1), inhibition of *CDCA7* expression (in ICF2 due to *ZBTB24* mutations), alterations in the CDCA7 CRD (in ICF3), or disruption of *HELLS* (in ICF4). While our results can explain the loss of methylation in most peri/centromeric satellite repeats in ICF syndrome, it is worth noting that motifs 1 and 2 are not enriched in HSat3 on chr 9 (Fig. 5B), which is also hypomethylated in ICF syndrome, albeit to a lesser extent compared to other satellite repeats (*23*). Given that different types of ICF syndrome show both common and distinct changes in DNA methylation patterns (*23*, *60*, *61*), it is possible that different mechanisms are involved in methylating different regions.

MATERIALS AND METHODS

Plasmid constructs

The HA-mCDCA7 construct was described previously (*38*). The HA-hCDCA7 and HA-mHELLS constructs were generated by cloning synthesized human *CDCA7* cDNA (accession: NP_665809.1) and polymerase chain reaction (PCR)–amplified mouse *Hells* cDNA

(Accession: NM_008234.3), respectively, into the *pCAG-HA-IRESBlast* vector (*62*). The green fluorescent protein (GFP)–mCDCA7 construct was generated by cloning mouse *Cdca7* cDNA into the *pEGFP-C1* vector (Clontech). The GST-mCDCA7 CRD (pXC2025) and GST-hCDCA7 CRD (pXC2205) constructs were generated by cloning the corresponding CRD fragments (mouse: residues 241 to 382; human: residues 235 to 371) into the *pGEX-6P-1* vector (Amersham). Mutations and deletions in mCDCA7 or hCDCA7 were introduced by PCR-based mutagenesis. The primers used and the synthesized human *CDCA7* cDNA are listed in table S5. All constructs were verified by DNA sequencing.

Cell culture, transfection, and generation of stable cell lines

NIH3T3, HeLa and human embryonic kidney (HEK) 293 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% FBS, 2 mM L-glutamine (Gibco), penicillin (50 U/ml), and streptomycin (50 μg/ml). WT and *Cdca7*$^{-/-}$ J1 mESCs (*38*) were maintained on gelatin-coated petri dishes in serum-containing medium [DMEM supplemented with 15% FBS, 0.1 mM nonessential amino acids, 0.1 mM β-mercaptoethanol, penicillin (50 U/ml), streptomycin (50 μg/ml), and leukemia inhibitory factor (10^3 U/ml)]. Transfection was performed using Lipofectamine 2000 (Invitrogen). Stable NIH3T3 cell lines expressing HA-tagged mCDCA7 or the RH mutant were generated by transfecting the corresponding construct, followed by 7 days of selection with Blasticidin S HCl (Gibco).

Co-IP and IF

Co-IP and IF were performed as described previously (*63*). To map the CDCA7 domain that mediates the interaction with HELLS, HA-mHELLS, and GFP-mCDCA7 proteins (full-length, the R315H mutation, and the ΔLZ and ΔCRD deletions) were coexpressed in HEK293 cells, the cell lysates were immunoprecipitated with GFP antibody (Abcam, ab1218), and the precipitated proteins were immunoblotted with HA antibody [Cell Signaling Technology (CST), #3724]. The localization patterns of WT and mutant mCDCA7 and hCDCA7 were determined by IF analysis of NIH3T3 and HeLa cells, respectively, that expressed HA-tagged CDCA7 proteins. Constitutive heterochromatin foci were marked by DAPI-bright spots and/or stained with H3K9me3 antibody (CST, #13969). Statistical analysis of CDCA7 localization patterns was done using one-way analysis of variance (ANOVA).

Cell synchronization and cell cycle analysis

To assess CDCA7 localization patterns during the cell cycle, NIH3T3 cells stably expressing HA-mCDCA7 were synchronized as described previously (*64*). Briefly, cells were first arrested at the G_0-G_1 phase by culturing them in DMEM medium with 0.5% FBS for 48 hours, then 10% FBS was added to induce the cells to reenter the cell cycle, and, at different time points (6 to 20 hours) following serum induction, some cells were stained with propidium iodide for cell cycle analysis by flow cytometry and other cells were analyzed for CDCA7 localization by IF with HA antibody.

DNA methylation analysis

Methylation of minor satellite DNA in mESCs was analyzed by Southern blotting after digestion of genomic DNA with the methylation-sensitive enzyme *Hpa*II, as described previously (*65*, *66*).

GST-CDCA7 CRD protein expression and purification

Recombinant GST fusion proteins were expressed and purified as described previously (*36*). Plasmids expressing mCDCA7 or hCDCA7 CRD fragments were transformed into *Escherichia coli* strain BL21-codon-plus (DE3)-RIL. Bacteria were grown in LB broth at 37°C until the log phase [optical density at 600 nm (OD_{600}) = 0.4 to 0.5], when the temperature was lowered to 16°C and 25 μM $ZnCl_2$ was added. When OD_{600} reached 0.8, 0.2 mM isopropylthio-β-galactoside was added, followed by continuing growth for 20 hours at 16°C. Cells were lysed by sonication in lysis buffer [20 mM tris-HCl (pH 7.5), 250 mM NaCl, 5% glycerol, 0.5 mM tris (2-carboxyethl) phosphine (TCEP), and 25 μM $ZnCl_2$]. The lysate was further treated with polyethylenimine solution (Sigma-Aldrich), added drop-by-drop into the lysate to a final concentration of 0.3% (v/v) while stirring on an ice bath. After removal of the debris by centrifugation, the supernatant was loaded onto a 5-ml GSTrap column (GE Healthcare). The resin was washed by lysis buffer, and the bound protein was eluted in 100 mM tris-HCl (pH 8.0), 250 mM NaCl, 5% glycerol, 0.5 mM TCEP, and 20 mM glutathione (reduced form). For SELEX and EMSA, the GST-CRD fusion proteins were used.

Protein purification was further carried out at 4°C through a multi-column chromatography protocol for ITC and crystallography. The GST tag was removed by digestion with PreScission protease (produced in-house). The cleaved proteins were dialyzed in a buffer consisting of 20 mM Hepes (pH 7.0), 0.1 M NaCl, 0.5% glycerol, and 0.5 mM TCEP and then loaded onto columns of HiTrap Q HP (5 ml) and HiTrap Heparin HP (5 ml) (GE Healthcare) connected in tandem. After washing with the same buffer, the Q column was disconnected, and the target protein was eluted from the Heparin column by 0.1 to 1 M NaCl gradient. For each protein, the peak fractions eluted from the Heparin column were pooled and loaded onto a second GSTrap column, from which the flow-through was collected, concentrated, and loaded onto a HiLoad 16/60 Superdex S200 column (GE Healthcare) equilibrated with 20 mM tris-HCl (pH 7.5), 150 mM NaCl, 5% glycerol, and 0.5 mM TCEP. The protein fractions were pooled, concentrated, and stored at −80°C before use. We observed that purifying the protein without protease inhibitors resulted in a degraded, shorter form of hCDCA7, referred to as hCDCA7(S).

To prevent degradation, cells were lysed via sonication in a lysis buffer that was fortified with 0.1 M phenylmethylsulfonyl fluoride and Pierce protease inhibitor tablets. The supernatant obtained from centrifuging the cell lysates was then loaded onto a 5-ml GSTrap column (GE Healthcare). Following this, the eluted GST-fusion protein was cleaved using PreScission protease. The protein was further purified using a tandem column setup, which included both HiTrap Q HP (5 ml) and HiTrap Heparin HP (5 ml) columns (GE Healthcare). Elution from the HiTrap Q column was achieved using a gradient ranging from 0.1 to 1 M NaCl. Subsequently, a subtract GSTrap column was used to remove residual GST and any uncleaved fusion protein. The flow-through, containing hCDCA7(L), underwent further purification via a sizing exclusion column, as previously described.

Systematic evolution of ligands by exponential enrichment

To identify potential DNA sequences recognized by the CDCA7 CRD, SELEX was performed according to the originally reported procedures (*67*, *68*), with modifications. GST-mCRD conjugated on

glutathione Sepharose 4B beads and a synthetic ssDNA library (Thermo Fisher Scientific, NC1108024) of random 30-mer sequences flanked by 23 nt at each end (5′-TAG GGA AGA GAA GGA CAT ATG AT-3′ and 5′-TTG ACT AGT ACA TGA CCA CTT GA-3′) were incubated in binding buffer [50 mM tris-HCl (pH 7.4), 150 mM NaCl, bovine serum albumin (BSA, 0.1 mg/ml), 3 mM dithiothreitol (DTT), 20 μM ZnSO4, and salmon sperm DNA (5 μg/ml)] for 1 hour at room temperature (RT). The beads were washed with binding buffer for three times, and bound ssDNA was eluted in water by boiling for 5 min, followed by snap cooling on ice for 3 min. The DNA was extracted by phenol/chloroform (25:24) and used as the template for PCR amplification—with primers complementary to the 23-nt flanking sequences (see table S5)—to obtain the ssDNA pool for the next round of selection. The PCR products were extracted by phenol/chloroform (25:24), heated at 95°C for 5 min, and then snap-cooled on ice. To minimize nonspecific binding of DNA species, we applied the counter-selection step using GST-mCRD:RH mutant before GST-mCRD was used in each cycle. After six rounds of selection, the PCR products were subcloned into the *pCR2.1-TOPO* TA cloning vector (Thermo Fisher Scientific, #450641), and 30 clones were sequenced.

Electrophoretic mobility shift assay

GST-CRD fusion proteins (20 nM) were incubated with DNA probes (10 nM) in binding buffer [2.5% glycerol, 1 mM $MgCl_2$, 0.5 mM EDTA, 0.5 mM DTT, BSA (0.1 mg/ml), 20 μM $ZnSO_4$, 50 mM NaCl, and 10 mM tris-HCl (pH7.5)] for 30 min at RT. Then, the samples were subjected to electrophoresis through 5% native polyacrylamide gel in 0.5× tris-borate-EDTA (TBE) buffer and imaged with 9410 Typhoon variable mode imager (GE Healthcare).

In a separate experimental setup, we investigated the interaction between DNA and the two forms of hCDCA7, namely hCDCA7(L) and hCDCA7(S). For this purpose, 20 nM of DNA was incubated with a series of twofold dilutions of the hCDCA7 protein, starting at a concentration of 320 nM. This incubation took place in a buffer containing 150 mM NaCl, 20 mM tris-HCl at pH 7.5, 1 mM DTT, 10% glycerol, and 0.05% NP-40, and was conducted on ice for 30 min. Following the incubation, the samples were subjected to electrophoresis using an 8% native polyacrylamide gel. The electrophoresis was run at a constant voltage of 150 V in an ice-cold 0.5× TBE buffer, and this process lasted for 40 min. After electrophoresis, the gel was stained using Sytox Green nucleic acid stain (catalog no.: S7020) at a 1:20,000 dilution in water. This staining procedure was carried out at RT for 10 min. The gel was imaged using a 9410 Typhoon variable mode imager (GE Healthcare) on the Cy2 channel to visualize the results.

Isothermal titration calorimetry

The ITC experiments were performed on a Microcal PEAQ-ITC instrument (Malvern) at 25°C. The protein was diluted to 40 to 50 μM and dialyzed in a buffer consisting of 20 mM tris-HCl (pH 7.5) and 150 mM NaCl. In some cases, 5% glycerol and 0.5 mM TCEP were included in the buffer. Then, 150 μl of protein sample was loaded into syringe. DNA was diluted to 5 μM using the same buffer, and 300 μl of DNA sample was loaded into the sample cell. The titration protocol was the same for all the measurements, which was composed of a single initial injection of 0.2 μl of protein, followed by 19 injections of 2 μl protein into DNA samples, the intervals between injections was set to 300 s and a reference power is 8 μcal s^{-1}. Curve fitting to a single-site binding model was performed by MicroCal PEAQ-ITC.

DNA binding by pull-down and ELISA

To demonstrate binding of the non-B DNA by full-length CDCA7, we performed pull-down and ELISA experiments using biotinylated DNA and HA-tagged full-length mCDCA7 or the RH mutant stably expressed in NIH3T3 cells. The cells were lysed (4×10^6 cells/ml) in lysis buffer [50 mM tris HCl (pH 7.5), 150 mM NaCl, 0.1% NP-40, 5 mM EDTA, 5 mM EGTA, and 15 mM $MgCl_2$] containing protease inhibitor cocktail, sonicated, and, after centrifugation, the supernatants were collected.

Pull-down experiments were performed by incubating cell lysates (2 hour at 4°C) either with biotinylated DNA probes that had been preconjugated to streptavidin agarose beads (Millipore, #16-126) (method 1) or with unconjugated biotinylated DNA probes, followed by incubation (1 hour at 4°C) with streptavidin agarose beads (method 2). The beads were washed three times with lysis buffer, and the pulled-down proteins were immunoblotted with HA antibody (CST, #3724).

ELISA was performed in the following steps, with three times of washing with tris-buffered saline containing 0.05% v/v Tween-20 (TBST) after each step: i) coating 96-well ELISA plate wells with neutravidin [10 mg/liter in phosphate-buffered saline (PBS), 100 μl per well, 1 hour at RT]; ii) coating the same wells with biotinylated DNA (10 mg/liter in PBS, 100 μl per well, 1 hour at RT); iii) incubating with cell lysates (100 μl per well, 2 hour at RT); iv) blocking with blocking buffer (1% BSA in TBST, 100 μl per well, 1 hour at RT); v) incubating with mouse monoclonal HA antibody (Abclonal, AE008, 1:3000 in blocking buffer, 100 μl per well, overnight at 4°C); vi) incubating with horseradish peroxidase–conjugated goat anti-mouse IgG (SouthernBiotech, #1030-05, 1:5,000 in blocking buffer, 100 μl per well, 1 hour at RT); and vii) adding 1-Step Ultra TMB-ELISA substrate solutions (Thermo Fisher Scientific, #34028, 100 μl per well) and, after 30 min in the dark at RT, stopping the reactions with ELISA stop solution (Thermo Fisher Scientific, SS03, 100 μl per well) and measuring absorbance at 450 nm using a microplate reader (BioTek Synergy H1).

Thermal stability assay

The stability of the non-B structure was tested by a thermal stability assay. In brief, various non-B DNA probes, as well as ssDNA and dsDNA (B DNA) controls, were mixed with iTaq Universal SYBR Green Supermix (Bio-Rad, #1725120, 10 μl 2× Supermix +10 μl DNA at 1 μM) at RT. The reactions were incubated in a thermal cycler at temperatures ranging from 25° to 95°C, with 0.5°C increments (5 s at each temperature), and fluorescence was monitored. The melting temperature (T_m) of each probe was determined on the basis of the melt curve.

Crystallography

The protein-DNA complex was prepared by mixing the purified mCDCA7 CRD (residues 244 to 382) with the 36-mer, 34-mer, 32-mer, 26-mer, or 32-mer 5mC oligonucleotides [annealed in 10 mM tris-HCl (pH 7.5), 50 mM NaCl] in a 1:1.2 ratio following by 1 hour incubation on ice. An Art Robbins Gryphon Crystallization Robot was used to set up screens of the sitting drop of 0.4 μl at ~19°C via vapor diffusion. The complex crystal with 36-mer oligonucleotides [Protein Data Bank (PDB) 8TLE and 8TLF] were obtained under

the condition of 0.2 M $MgCl_2$, 0.1 M tris-HCl (pH 8.5), and 25% polyethylene glycol (PEG) 3350. The complex crystal with 34-mer (PDB 8TLG) or 32-mer (PDB 8TLH) oligonucleotides were grown under the condition of 0.1 M $MgCl_2$, 0.1 M tris-HCl (pH 8.5), and 25% PEG3350. The complex crystal with 26-mer oligonucleotides (PDB 8TLL) were obtained under the condition of 0.1 M $MgCl_2$, 0.1 M tris-HCl (pH 8.5), and 23% PEG3350. The complex crystal with 32-mer 5mC oligonucleotides (PDB 8TLJ) were obtained under the condition of 0.1 M $MgCl_2$, 0.1 M bis-tris (pH 6.2), and 25% PGE3350. The complex crystal of hCDCA7 CRD with 32-mer 5mC oligonucleotides (PDB 8TLK) were grown under the condition of 0.2 M $MgCl_2$, 0.1 M bis-tris (pH 6.0), and 25% PEG3350.

Crystals were flash frozen using 20% (v/v) ethylene glycol as the cryoprotectant. Resulting crystallographic datasets were processed with HKL2000 (*69*). Two scaled files were output with one file combining Bijvoet pairs and the other keeping them separate. The dataset for the first structure (PDB 8TLE) was examined for single-wavelength anomalous dispersion phasing using the PHENIX Xtriage module, which reported severe anisotropy but also a good anomalous signal to ~4 Å. The PHENIX AutoSol module (*70*) readily found three zinc atom positions to give an interpretable map with initial figure of merit of 0.31 and gave a density-modified map with an R-factor of 0.42 (table S1). The initial electron density showed recognizable molecular features of the β strands and α helices and DNA bases and backbone. Reinserting the zinc positions into AutoSol and using the full resolution of the dataset gave a better map allowing for our initial model build. This resulting structure was used for molecular replacement in the PHENIX PHASER module (*71*) for the other structures. All structure refinements were performed by PHENIX Refine (*72*), with 5% randomly chosen reflections for validation by R-free values. Manual (re)building with COOT (*73*) was conducted carefully between refinement cycles. Structure quality was analyzed during PHENIX refinements and validated by the PDB validation server (*74*). Molecular graphics were generated using open-source PyMOL (http://pymol.org/pymol).

Chromatin immunoprecipitation sequencing

To identify genome-wide CDCA7-binding sites, ChIP-seq analysis was performed with HA antibody using *Cdca7*$^{-/-}$ mESCs reconstituted with HA-tagged WT mCDCA7 or the RH mutant (two biological replicates per genotype). For each sample, ~5 million mESCs were used for ChIP. ChIP-seq libraries were constructed using a KAPA HyperPrep kit (Roche) and sequenced in a 50-bp single-read run on Illumina HiSeq2500 instrument (Illumina).

ChIP and input fastq files were processed using Trim Galore! (version 0.4.1) (https://github.com/FelixKrueger/TrimGalore/issues/25) and cutadapt (version 1.6) (*75*) to remove low-quality reads and trim Illumina adapter sequences. Reads mapping to Phix, Mycoplasma, and human (hg38) genomes were then excluded from analysis to eliminate potential contamination. The Phix sequence was downloaded from Illumina iGenomes (https://support.illumina.com/sequencing/sequencing_software/igenome.html), and sequences of diverse Mycoplasma species were obtained from the NCBI genome database (https://ncbi.nlm.nih.gov/datasets/genome/). The remaining reads were mapped to the mouse genome mm10 and divided into mm10-mapped, mm10-multimapped, and mm10-unmapped read sets. All the mappings were done using Bowtie (version 1.1.2) (*76*) with the following parameters: "-v 2 -m 1 --best --strata". Uniquely mapped reads, after deduplication, were used for peak calling with MACS2 (version V2.1.1.20160309) (*55*).

We compared the enrichment of motif 1 (CNNGTCGXY, with 256 possible instances due to the 4 possible nucleotides at positions N, X, and Y) and motif 2 (XYCGTTT, with 16 possible instances due to the 4 possible nucleotides at positions X and Y) in WT versus RH. In addition, we compared motif 1 against control sequences where the CG dinucleotide critical for CDCA7 binding was replaced by non-CG dinucleotides, resulting in 2304 possible instances (256 multiplied by 9 possible dinucleotide replacements). Motif 1 and motif 2, as well as motif 1 control sequences, were mapped to the read groups (mapped, multimapped, and unmapped) for each sample. The percentage of reads containing a motif/control sequence in the corresponding read group was calculated. ChIP/input ratios were then determined as the percentage ratios, and mean values were found for the various sets of sequences indicated in Fig. 8 and fig. S16. The Wilcoxon rank sum test (one-sided) was used to calculate *P* values. The ChIP-seq data have been deposited in the GEO database (accession number: GSE255395; token: sfobuieyznivzkp).

Bioinformatics analysis of motifs 1 and 2 in the human genome

The CENP-A ChIP-seq dataset (SRR766736) was downloaded from the GenBank Short Read Archive. T2T centromere/satellite (CenSat) data (chm13v2.0_censat_v2.0.bed) were accessed from the Telomere-to-Telomere (T2T) Consortium CHM13 project site (https://github.com/marbl/CHM13). Motifs 1 and 2 sequences and the CENP-A ChIP sequences were aligned to the T2T genome (T2T-CHM13v2.0), and the aligned counts on both strands were binned at 100 kb to generate the coverage plots showing chromosome-wide motif alignments and CENP-A ChIP fragment alignments. The CenSat annotation map was produced using the CenSat bed file and assigning the color of the annotation to each 25-kb bin region across the peri/centromere regions, except for chr Y where the bins spanned the entire chromosome.

Supplementary Materials

The PDF file includes:

Figs. S1 to S17
Tables S1 and S5
Legends for tables S2 to S4

Other Supplementary Material for this manuscript includes the following:

Tables S2 to S4

REFERENCES AND NOTES

1. A. Bansal, S. Kaushik, S. Kukreti, Non-canonical DNA structures: Diversity and disease association. *Front. Genet.* **13**, 959258 (2022).
2. B. Wittig, S. Wölfl, T. Dorbic, W. Vahrson, A. Rich, Transcription of human c-myc in permeabilized nuclei is associated with formation of Z-DNA in three discrete regions of the gene. *EMBO J.* **11**, 4653–4663 (1992).
3. H. Li, J. Xiao, J. Li, L. Lu, S. Feng, P. Dröge, Human genomic Z-DNA segments probed by the Z alpha domain of ADAR1. *Nucleic Acids Res.* **37**, 2737–2746 (2009).
4. G. Biffi, D. Tannahill, J. McCafferty, S. Balasubramanian, Quantitative visualization of DNA G-quadruplex structures in human cells. *Nat. Chem.* **5**, 182–186 (2013).
5. S. Kendrick, H. J. Kang, M. P. Alam, M. M. Madathil, P. Agrawal, V. Gokhale, D. Yang, S. M. Hecht, L. H. Hurley, The dynamic character of the BCL2 promoter i-motif provides a mechanism for modulation of gene expression by compounds that bind selectively to the alternative DNA hairpin structure. *J. Am. Chem. Soc.* **136**, 4161–4171 (2014).
6. V. S. Chambers, G. Marsico, J. M. Boutell, M. di Antonio, G. P. Smith, S. Balasubramanian, High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.* **33**, 877–881 (2015).

7. F. Kouzine, D. Wojtowicz, L. Baranello, A. Yamane, S. Nelson, W. Resch, K. R. Kieffer-Kwon, C. J. Benham, R. Casellas, T. M. Przytycka, D. Levens, Permanganate/S1 nuclease footprinting reveals non-B DNA structures with regulatory potential across a mammalian genome. *Cell Syst.* **4**, 344–356.e7 (2017).
8. L. F. M. Passalacqua, M. T. Banco, J. D. Moon, X. Li, S. R. Jaffrey, A. R. Ferré-D'Amaré, Intricate 3D architecture of a DNA mimic of GFP. *Nature* **618**, 1078–1084 (2023).
9. A. Bacolla, R. D. Wells, Non-B DNA conformations, genomic rearrangements, and human disease. *J. Biol. Chem.* **279**, 47411–47414 (2004).
10. R. D. Wells, Non-B DNA conformations, mutagenesis and disease. *Trends Biochem. Sci.* **32**, 271–278 (2007).
11. S. Kasinathan, S. Henikoff, Non-B-form DNA is enriched at centromeres. *Mol. Biol. Evol.* **35**, 949–962 (2018).
12. V. S. P. Patchigolla, B. G. Mellone, Enrichment of non-B-form DNA at *D. melanogaster* centromeres. *Genome Biol. Evol.* **14**, evac054 (2022).
13. Q. Liu, C. Yi, Z. Zhang, H. Su, C. Liu, Y. Huang, W. Li, X. Hu, C. Liu, J. A. Birchler, Y. Liu, F. Han, Non-B-form DNA tends to form in centromeric regions and has undergone changes in polyploid oat subgenomes. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2211683120 (2023).
14. G. Wang, K. M. Vasquez, Z-DNA, an active element in the genome. *Front. Biosci.* **12**, 4424–4438 (2007).
15. V. Brazda, R. C. Laister, E. B. Jagelská, C. Arrowsmith, Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol. Biol.* **12**, 33 (2011).
16. T. Oyoshi, T. Masuzawa, Modulation of histone modifications and G-quadruplex structures by G-quadruplex-binding proteins. *Biochem. Biophys. Res. Commun.* **531**, 39–44 (2020).
17. M. Hulten, Selective somatic pairing and fragility at lq12 in a boy with common variable Immuno deficiency. *Clin. Genet.* **14**, 294 (2008).
18. L. Tiepolo, P. Maraschio, G. Gimelli, C. Cuoco, G. F. Gargani, C. Romano, Multibranched chromosomes 1, 9, and 16 in a patient with combined IgA and IgE deficiency. *Hum. Genet.* **51**, 127–137 (1979).
19. M. M. Hagleitner, A. Lankester, P. Maraschio, M. Hultén, J. P. Fryns, C. Schuetz, G. Gimelli, E. G. Davies, A. Gennery, B. H. Belohradsky, R. de Groot, E. J. A. Gerritsen, T. Mattina, P. J. Howard, A. Fasth, I. Reisli, D. Furthner, M. A. Slatter, A. J. Cant, G. Cazzola, P. J. van Dijken, M. van Deuren, J. C. de Greef, S. M. van der Maarel, C. M. R. Weemaes, Clinical spectrum of immunodeficiency, centromeric instability and facial dysmorphism (ICF syndrome). *J. Med. Genet.* **45**, 93–99 (2008).
20. C. M. Weemaes, M. J. D. van Tol, J. Wang, M. M. van Ostaijen-ten Dam, M. C. J. A. van Eggermond, P. E. Thijssen, C. Aytekin, N. Brunetti-Pierri, M. van der Burg, E. Graham Davies, A. Ferster, D. Furthner, G. Gimelli, A. Gennery, B. Kloeckener-Gruissem, S. Meyn, C. Powell, I. Reisli, C. Schuetz, A. Schulz, A. Shugar, P. J. van den Elsen, S. M. van der Maarel, Heterogeneous clinical presentation in ICF syndrome: Correlation with underlying gene defects. *Eur. J. Hum. Genet.* **21**, 1219–1225 (2013).
21. F. Kiaee, M. Zaki-Dizaji, N. Hafezi, A. Almasi-Hashiani, H. Hamedifar, A. Sabzevari, A. Shirkani, Z. Zian, F. Jadidi-Niaragh, F. Aghamahdi, M. Goudarzvand, R. Yazdani, H. Abolhassani, A. Aghamohammadi, G. Azizi, Clinical, immunologic and molecular spectrum of patients with immunodeficiency, centromeric instability, and facial anomalies (ICF) syndrome: A systematic review. *Endocr. Metab. Immune Disord. Drug Targets* **21**, 664–672 (2021).
22. Z. Ying, S. Hardikar, J. B. Plummer, T. Hamidi, B. Liu, Y. Chen, J. Shen, Y. Mu, K. M. McBride, T. Chen, Enhanced CD19 activity in B cells contributes to immunodeficiency in mice deficient in the ICF syndrome gene Zbtb24. *Cell. Mol. Immunol.* **20**, 1487–1498 (2023).
23. G. Velasco, G. Grillo, N. Touleimat, L. Ferry, I. Ivkovic, F. Ribierre, J. F. Deleuze, S. Chantalat, C. Picard, C. Francastel, Comparative methylome analysis of ICF patients identifies heterochromatin loci that require ZBTB24, CDCA7 and HELLS for their methylated state. *Hum. Mol. Genet.* **27**, 2409–2424 (2018).
24. R. S. Hansen, C. Wijmenga, P. Luo, A. M. Stanek, T. K. Canfield, C. M. R. Weemaes, S. M. Gartler, The DNMT3B DNA methyltransferase gene is mutated in the ICF immunodeficiency syndrome. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 14412–14417 (1999).
25. M. Okano, D. W. Bell, D. A. Haber, E. Li, DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell* **99**, 247–257 (1999).
26. G. L. Xu, T. H. Bestor, D. Bourc'his, C. L. Hsieh, N. Tommerup, M. Bugge, M. Hulten, X. Qu, J. J. Russo, E. Viegas-Péquignot, Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* **402**, 187–191 (1999).
27. J. C. de Greef, J. Wang, J. Balog, J. T. den Dunnen, R. R. Frants, K. R. Straasheijm, C. Aytekin, M. van der Burg, L. Duprez, A. Ferster, A. R. Gennery, G. Gimelli, I. Reisli, C. Schuetz, A. Schulz, D. F. C. M. Smeets, Y. Sznajer, C. Wijmenga, M. C. van Eggermond, M. M. van Ostaijen-Ten Dam, A. C. Lankester, M. J. D. van Tol, P. J. van den Elsen, C. M. Weemaes, S. M. van der Maarel, Mutations in ZBTB24 are associated with immunodeficiency, centromeric instability, and facial anomalies syndrome type 2. *Am. J. Hum. Genet.* **88**, 796–804 (2011).
28. P. E. Thijssen, Y. Ito, G. Grillo, J. Wang, G. Velasco, H. Nitta, M. Unoki, M. Yoshihara, M. Suyama, Y. Sun, R. J. L. F. Lemmers, J. C. de Greef, A. Gennery, P. Picco, B. Kloeckener-Gruissem, T. Güngör, I. Reisli, C. Picard, K. Kebaili, B. Roquelaure, T. Iwai, I. Kondo, T. Kubota, M. M. van Ostaijen-ten Dam, M. J. D. van Tol, C. Weemaes, C. Francastel, S. M. van der Maarel, H. Sasaki, Mutations in CDCA7 and HELLS cause immunodeficiency-centromeric instability-facial anomalies syndrome. *Nat. Commun.* **6**, 7870 (2015).
29. K. Dennis, T. Fan, T. Geiman, Q. Yan, K. Muegge, Lsh, a member of the SNF2 family, is required for genome-wide methylation. *Genes Dev.* **15**, 2940–2944 (2001).
30. H. Zhu, T. M. Geiman, S. Xi, Q. Jiang, A. Schmidtmann, T. Chen, E. Li, K. Muegge, Lsh is involved in de novo methylation of DNA. *EMBO J.* **25**, 335–345 (2006).
31. J. Ren, V. Briones, S. Barbour, W. Yu, Y. Han, M. Terashima, K. Muegge, The ATP binding site of the chromatin remodeling homolog Lsh is required for nucleosome density and de novo DNA methylation at repeat sequences. *Nucleic Acids Res.* **43**, 1444–1455 (2015).
32. M. Han, J. Li, Y. Cao, Y. Huang, W. Li, H. Zhu, Q. Zhao, J. D. J. Han, Q. Wu, J. Li, J. Feng, J. Wong, A role for LSH in facilitating DNA methylation by DNMT1 through enhancing UHRF1 chromatin association. *Nucleic Acids Res.* **48**, 12116–12134 (2020).
33. H. Wu, P. E. Thijssen, E. de Klerk, K. K. D. Vonk, J. Wang, B. den Hamer, C. Aytekin, S. M. van der Maarel, L. Daxinger, Converging disease genes in ICF syndrome: ZBTB24 controls expression of CDCA7 in mammals. *Hum. Mol. Genet.* **25**, 4041–4051 (2016).
34. C. Jenness, S. Giunta, M. M. Müller, H. Kimura, T. W. Muir, H. Funabiki, HELLS and CDCA7 comprise a bipartite nucleosome remodeling complex defective in ICF syndrome. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E876–E885 (2018).
35. J. J. Thompson, R. Kaur, C. P. Sosa, J. H. Lee, K. Kashiwagi, D. Zhou, K. D. Robertson, ZBTB24 is a transcriptional regulator that coordinates with DNMT3B to control DNA methylation. *Nucleic Acids Res.* **46**, 10034–10051 (2018).
36. R. Ren, S. Hardikar, J. R. Horton, Y. Lu, Y. Zeng, A. K. Singh, K. Lin, L. D. Coletta, J. Shen, C. S. Lin Kong, H. Hashimoto, X. Zhang, T. Chen, X. Cheng, Structural basis of specific DNA binding by the transcription factor ZBTB24. *Nucleic Acids Res.* **47**, 8388–8398 (2019).
37. M. Unoki, H. Funabiki, G. Velasco, C. Francastel, H. Sasaki, CDCA7 and HELLS mutations undermine nonhomologous end joining in centromeric instability syndrome. *J. Clin. Invest.* **129**, 78–92 (2019).
38. S. Hardikar, Z. Ying, Y. Zeng, H. Zhao, B. Liu, N. Veland, K. McBride, X. Cheng, T. Chen, The ZBTB24-CDCA7 axis regulates HELLS enrichment at centromeric satellite repeats to facilitate DNA methylation. *Protein Cell* **11**, 214–218 (2020).
39. J. E. Prescott, R. C. Osthus, L. A. Lee, B. C. Lewis, H. Shim, J. F. Barrett, Q. Guo, A. L. Hawkins, C. A. Griffin, C. V. Dang, A novel c-Myc-responsive gene, JPO1, participates in neoplastic transformation. *J. Biol. Chem.* **276**, 48276–48284 (2001).
40. R. C. Osthus, B. Karim, J. E. Prescott, B. D. Smith, M. McDevitt, D. L. Huso, C. V. Dang, The Myc target gene JPO1/CDCA7 is frequently overexpressed in human tumors and has limited transforming activity in vivo. *Cancer Res.* **65**, 5620–5627 (2005).
41. Y. Goto, R. Hayashi, T. Muramatsu, H. Ogawa, I. Eguchi, Y. Oshida, K. Ohtani, K. Yoshida, JPO1/CDCA7, a novel transcription factor E2F1-induced protein, possesses intrinsic transcriptional regulator activity. *Biochim. Biophys. Acta* **1759**, 60–68 (2006).
42. R. M. Gill, T. V. Gabor, A. L. Couzens, M. P. Scheid, The MYC-associated protein CDCA7 is phosphorylated by AKT to regulate MYC-dependent apoptosis and transformation. *Mol. Cell. Biol.* **33**, 498–513 (2013).
43. J. Guiu, D. J. M. Bergen, E. de Pater, A. B. M. M. K. Islam, V. Ayllón, L. Gama-Norton, C. Ruiz-Herguido, J. González, N. López-Bigas, P. Menendez, E. Dzierzak, L. Espinosa, A. Bigas, Identification of Cdca7 as a novel Notch transcriptional target involved in hematopoietic stem cell emergence. *J. Exp. Med.* **211**, 2411–2423 (2014).
44. P. R. Jimenez, C. Martin-Cortazar, O. Kourani, Y. Chiodo, R. Cordoba, M. P. Dominguez-Franjo, J. M. Redondo, T. Iglesias, M. R. Campanero, CDCA7 is a critical mediator of lymphomagenesis that selectively regulates anchorage-independent growth. *Haematologica* **103**, 1669–1678 (2018).
45. C. Martin-Cortazar, Y. Chiodo, R. P. Jimenez, M. Bernabe, M. L. Cayuela, T. Iglesias, M. R. Campanero, CDCA7 finely tunes cytoskeleton dynamics to promote lymphoma migration and invasion. *Haematologica* **105**, 730–740 (2020).
46. L. Gold, SELEX: How it happened and where it will go. *J. Mol. Evol.* **81**, 140–143 (2015).
47. D. L. Grady, R. L. Ratliff, D. L. Robinson, E. C. McCanlies, J. Meyne, R. K. Moyzis, Highly conserved repetitive DNA sequences are present at human centromeres. *Proc. Natl. Acad. Sci. U.S.A.* **89**, 1695–1699 (1992).
48. S. Nurk, S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, S. J. Hoyt, M. Diekhans, G. A. Logsdon, M. Alonge, S. E. Antonarakis, M. Borchers, G. G. Bouffard, S. Y. Brooks, G. V. Caldas, N. C. Chen, H. Cheng, C. S. Chin, W. Chow, L. G. de Lima, P. C. Dishuck, R. Durbin, T. Dvorkina, I. T. Fiddes, G. Formenti, R. S. Fulton, A. Fungtammasan, E. Garrison, P. G. S. Grady, T. A. Graves-Lindsay, I. M. Hall, N. F. Hansen, G. A. Hartley, M. Haukness, K. Howe, M. W. Hunkapiller, C. Jain, M. Jain, E. D. Jarvis, P. Kerpedjiev, M. Kirsche, M. Kolmogorov, J. Korlach, M. Kremitzki, H. Li, V. V. Maduro, T. Marschall, A. M. McCartney, J. McDaniel, D. E. Miller, J. C. Mullikin, E. W. Myers, N. D. Olson, B. Paten, P. Peluso, P. A. Pevzner, D. Porubsky, T. Potapova, E. I. Rogaev, J. A. Rosenfeld, S. L. Salzberg, V. A. Schneider, F. J. Sedlazeck, K. Shafin, C. J. Shew, A. Shumate, Y. Sims, A. F. A. Smit, D. C. Soto, I. Sović, J. M. Storer, A. Streets, B. A. Sullivan, F. Thibaud-Nissen, J. Torrance,

J. Wagner, B. P. Walenz, A. Wenger, J. M. D. Wood, C. Xiao, S. M. Yan, A. C. Young, S. Zarate, U. Surti, R. C. McCoy, M. Y. Dennis, I. A. Alexandrov, J. L. Gerton, R. J. O'Neill, W. Timp, J. M. Zook, M. C. Schatz, E. E. Eichler, K. H. Miga, A. M. Phillippy, The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
49. N. Altemose, G. A. Logsdon, A. V. Bzikadze, P. Sidhwani, S. A. Langley, G. V. Caldas, S. J. Hoyt, L. Uralsky, F. D. Ryabov, C. J. Shew, M. E. G. Sauria, M. Borchers, A. Gershman, A. Mikheenko, V. A. Shepelev, T. Dvorkina, O. Kunyavskaya, M. R. Vollger, A. Rhie, A. M. McCartney, M. Asri, R. Lorig-Roach, K. Shafin, J. K. Lucas, S. Aganezov, D. Olson, L. G. de Lima, T. Potapova, G. A. Hartley, M. Haukness, P. Kerpedjiev, F. Gusev, K. Tigyi, S. Brooks, A. Young, S. Nurk, S. Koren, S. R. Salama, B. Paten, E. I. Rogaev, A. Streets, G. H. Karpen, A. F. Dernburg, B. A. Sullivan, A. F. Straight, T. J. Wheeler, J. L. Gerton, E. E. Eichler, A. M. Phillippy, W. Timp, M. Y. Dennis, R. J. O'Neill, J. M. Zook, M. C. Schatz, P. A. Pevzner, M. Diekhans, C. H. Langley, I. A. Alexandrov, K. H. Miga, Complete genomic and epigenetic maps of human centromeres. *Science* **376**, eabl4178 (2022).
50. A. Rhie, S. Nurk, M. Cechova, S. J. Hoyt, D. J. Taylor, N. Altemose, P. W. Hook, S. Koren, M. Rautiainen, I. A. Alexandrov, J. Allen, M. Asri, A. V. Bzikadze, N. C. Chen, C. S. Chin, M. Diekhans, P. Flicek, G. Formenti, A. Fungtammasan, C. Garcia Giron, E. Garrison, A. Gershman, J. L. Gerton, P. G. S. Grady, A. Guarracino, L. Haggerty, R. Halabian, N. F. Hansen, R. Harris, G. A. Hartley, W. T. Harvey, M. Haukness, J. Heinz, T. Hourlier, R. M. Hubley, S. E. Hunt, S. Hwang, M. Jain, R. K. Kesharwani, A. P. Lewis, H. Li, G. A. Logsdon, J. K. Lucas, W. Makalowski, C. Markovic, F. J. Martin, A. M. Mc Cartney, R. C. McCoy, J. McDaniel, B. M. McNulty, P. Medvedev, A. Mikheenko, K. M. Munson, T. D. Murphy, H. E. Olsen, N. D. Olson, L. F. Paulin, D. Porubsky, T. Potapova, F. Ryabov, S. L. Salzberg, M. E. G. Sauria, F. J. Sedlazeck, K. Shafin, V. A. Shepelev, A. Shumate, J. M. Storer, L. Surapaneni, A. M. Taravella Oill, F. Thibaud-Nissen, W. Timp, M. Tomaszkiewicz, M. R. Vollger, B. P. Walenz, A. C. Watwood, M. H. Weissensteiner, A. M. Wenger, M. A. Wilson, S. Zarate, Y. Zhu, J. M. Zook, E. E. Eichler, R. J. O'Neill, M. C. Schatz, K. H. Miga, K. D. Makova, A. M. Phillippy, The complete sequence of a human Y chromosome. *Nature* **621**, 344–354 (2023).
51. H. F. Willard, J. S. Waye, Hierarchical order in chromosome-specific human alpha satellite DNA. *Trends Genet.* **3**, 192–198 (1987).
52. S. M. McNulty, B. A. Sullivan, Alpha satellite DNA biology: Finding function in the recesses of the genome. *Chromosome Res.* **26**, 115–138 (2018).
53. A. Gershman, M. E. G. Sauria, X. Guitart, M. R. Vollger, P. W. Hook, S. J. Hoyt, M. Jain, A. Shumate, R. Razaghi, S. Koren, N. Altemose, G. V. Caldas, G. A. Logsdon, A. Rhie, E. E. Eichler, M. C. Schatz, R. J. O'Neill, A. M. Phillippy, K. H. Miga, W. Timp, Epigenetic patterns in a complete human genome. *Science* **376**, eabj5089 (2022).
54. H. Masumoto, H. Masukata, Y. Muro, N. Nozaki, T. Okazaki, A human centromere antigen (CENP-B) interacts with a short specific sequence in alphoid DNA, a human centromeric satellite. *J. Cell Biol.* **109**, 1963–1973 (1989).
55. Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, X. S. Liu, Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
56. C. Mellor, C. Perez, J. E. Sale, Creation and resolution of non-B-DNA structural impediments during replication. *Crit. Rev. Biochem. Mol. Biol.* **57**, 412–442 (2022).
57. H. Hashimoto, J. R. Horton, X. Zhang, M. Bostick, S. E. Jacobsen, X. Cheng, The SRA domain of UHRF1 flips 5-methylcytosine out of the DNA helix. *Nature* **455**, 826–829 (2008).
58. M. J. Sperlazza, S. M. Bilinovich, L. M. Sinanan, F. R. Javier, D. C. Williams Jr., Structural basis of MeCP2 distribution on non-CpG methylated and hydroxymethylated DNA. *J. Mol. Biol.* **429**, 1581–1594 (2017).
59. Y. Tanaka, H. Kurumizaka, S. Yokoyama, CpG methylation of the CENP-B box reduces human CENP-B binding. *FEBS J.* **272**, 282–289 (2005).
60. C. Wijmenga, R. S. Hansen, G. Gimelli, E. J. Björck, E. G. Davies, D. Valentine, B. H. Belohradsky, J. J. van Dongen, D. F. C. M. Smeets, L. P. W. J. van den Heuvel, J. A. F. M. Luyten, E. Strengman, C. Weemaes, P. L. Pearson, Genetic variation in ICF syndrome: Evidence for genetic heterogeneity. *Hum. Mutat.* **16**, 509–517 (2000).
61. Y. L. Jiang, M. Rigolet, D. Bourc'his, F. Nigon, I. Bokesoy, J. P. Fryns, M. Hultén, P. Jonveaux, P. Maraschio, A. Mégarbané, A. Moncla, E. Viegas-Péquignot, DNMT3B mutations and DNA methylation defect define two types of ICF syndrome. *Hum. Mutat.* **25**, 56–63 (2005).
62. S. J. Kim, H. Zhao, S. Hardikar, A. K. Singh, M. A. Goodell, T. Chen, A DNMT3A mutation common in AML exhibits dominant-negative effects in murine ES cells. *Blood* **122**, 4086–4089 (2013).
63. J. Dan, P. Rousseau, S. Hardikar, N. Veland, J. Wong, C. Autexier, T. Chen, Zscan4 inhibits maintenance DNA methylation to facilitate telomere elongation in mouse embryonic stem cells. *Cell Rep.* **20**, 1936–1949 (2017).
64. E. W. Lam, R. J. Watson, An E2F-binding site mediates cell-cycle regulated repression of mouse B-myb transcription. *EMBO J.* **12**, 2705–2713 (1993).
65. N. Veland, Y. Lu, S. Hardikar, S. Gaddis, Y. Zeng, B. Liu, M. R. Estecio, Y. Takata, K. Lin, M. W. Tomida, J. Shen, D. Saha, H. Gowher, H. Zhao, T. Chen, DNMT3L facilitates DNA methylation partly by maintaining DNMT3A stability in mouse embryonic stem cells. *Nucleic Acids Res.* **47**, 152–167 (2019).
66. Y. Zeng, R. Ren, G. Kaur, S. Hardikar, Z. Ying, L. Babcock, E. Gupta, X. Zhang, T. Chen, X. Cheng, The inactive Dnmt3b3 isoform preferentially enhances Dnmt3b-mediated DNA methylation. *Genes Dev.* **34**, 1546–1558 (2020).
67. C. Tuerk, L. Gold, Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505–510 (1990).
68. A. D. Ellington, J. W. Szostak, In vitro selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).
69. Z. Otwinowski, D. Borek, W. Majewski, W. Minor, Multiparametric scaling of diffraction intensities. *Acta Crystallogr. A* **59**, 228–234 (2003).
70. T. C. Terwilliger, P. D. Adams, R. J. Read, A. J. McCoy, N. W. Moriarty, R. W. Grosse-Kunstleve, P. V. Afonine, P. H. Zwart, L. W. Hung, Decision-making in structure solution using Bayesian estimates of map quality: The PHENIX AutoSol wizard. *Acta Crystallogr. D Biol. Crystallogr.* **65**, 582–601 (2009).
71. A. J. McCoy, R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni, R. J. Read, Phaser crystallographic software. *J. Appl. Cryst.* **40**, 658–674 (2007).
72. J. J. Headd, N. Echols, P. V. Afonine, R. W. Grosse-Kunstleve, V. B. Chen, N. W. Moriarty, D. C. Richardson, J. S. Richardson, P. D. Adams, Use of knowledge-based restraints in phenix.refine to improve macromolecular refinement at low resolution. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 381–390 (2012).
73. P. Emsley, K. Cowtan, Coot: Model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).
74. R. J. Read, P. D. Adams, W. B. Arendall III, A. T. Brunger, P. Emsley, R. P. Joosten, G. J. Kleywegt, E. B. Krissinel, T. Lütteke, Z. Otwinowski, A. Perrakis, J. S. Richardson, W. H. Sheffler, J. L. Smith, I. J. Tickle, G. Vriend, P. H. Zwart, A new generation of crystallographic validation tools for the protein data bank. *Structure* **19**, 1395–1412 (2011).
75. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17**, 10–12 (2011).
76. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

Acknowledgments: We thank R. Wood, D. Johnson, M. Bedford (MD Anderson Cancer Center), and C. Cardoso (Technische Universität Darmstadt) for discussion; Y. Cao and Y. Zeng (MD Anderson Cancer Center) for technical assistance; and X. Kong (New York University) for assistance of access to 17-ID-1 beamtime. We thank the beamline scientists of Southeast Regional Collaborative Access Team (SER-CAT) at the Advanced Photon Source (APS), Argonne National Laboratory, and 17-ID-1 of the National Synchrotron Light Source II, Brookhaven National Laboratory. **Funding:** This work was financially supported by the National Institutes of Health grant R01AI1214030 (T.C.), National Institutes of Health grant R35GM134744 (X.C.), Cancer Prevention and Research Institute of Texas grant RR160029 (X.C.), Sam and Freda Davis Fund fellowship (Z.Y.), The Cockrell Family Foundation in Houston (J.Z.), National Institutes of Health equipment grants S10_RR25528, S10_RR028976, and S10_OD027000, U.S. Department of Energy contract W-31-109-Eng-38, and National Synchrotron Light Source II resources 17-ID-1 under contract DE-SC0012704. **Author contributions:** Conceptualization: T.C. and X.C. Investigation: S.H., R.R., Z.Y., J.Z., J.R.H., Bigang Liu, L.D.C., and J.D. Bioinformatics analysis: M.D.B., Bin Liu, and Y.L. Supervision: Bin Liu, J.S., X.Z., X.C., and T.C. Writing: T.C. and X.C. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The authors have deposited the x-ray structure (coordinates) and the source data (structure factor file) of CDCA7-DNA to the PDB, and these will be released upon article publication under accession numbers PDB 8TLE (mCDCA7 and oligo 36 nt), PDB 8TLF (mCDCA7 and oligo 36 nt), PDB 8TLG (mCDCA7 and oligo 34 nt), PDB 8TLH (mCDCA7 and oligo 32 nt), PDB 8TLL (mCDCA7 and oligo 26 nt), PDB 8TLJ (mCDCA7 and 5mC oligo in 32 nt), and PDB 8TLK (hCDCA7 and 5mC oligo in 32 nt). The ChIP-seq data have been deposited in the GEO database (accession number: GSE255395).

Submitted 10 June 2024
Accepted 18 July 2024
Published 23 August 2024
10.1126/sciadv.adr0036