Molecular basis of DNA recognition by the HMG-box-C1 module of Capicua

Jonathan Webb[1], Jeremy J.M. Liew[1], Andrew D. Gnann[1], Khandan Ilkhani[2], MacKenzie Patterson[1&], Sayantanee Paul[2%], Marta Forés[3], Gerardo Jiménez[3,4], Alexey Veraksa[2], and Daniel P. Dowling[1]*

[1] Chemistry Department, University of Massachusetts Boston, Boston, MA, 02125, USA

[2] Biology Department, University of Massachusetts Boston, Boston, MA, 02125, USA

[3] Instituto de Biología Molecular de Barcelona-Consejo Superior de Investigaciones Científicas (CSIC), Parc Cientific de Barcelona, Barcelona, 08028, Spain

[4] Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, 08010, Spain

[&] Present address: Department of Biochemistry, Brandeis University, Waltham, MA, 02453, USA

[%] Present address: Department of Discovery Oncology, Genentech Inc., South San Francisco, CA, 94080, USA

* Corresponding author: Daniel Dowling

Lead contact: Daniel Dowling, Daniel.dowling@umb.edu

E-mail: daniel.dowling@umb.edu

https://doi.org/10.1016/j.str.2025.08.018.

Summary
The HMG-box protein Capicua (CIC) is a conserved transcriptional repressor with key functions in development and disease. CIC binding of DNA requires both its HMG-box and a separate domain called C1. How these domains cooperate to recognize specific DNA sequences is not known. Here, we report the crystal structure of the human CIC HMG-box and C1 domains complexed with a DNA oligomer containing a consensus octameric binding site. We find that both domains adopt tri-helical structures that pack against opposite sides of the DNA helix. The C1 domain folds into a helix-turn-helix (HTH) structure, inserting into the DNA major groove to enhance affinity. We investigate the system using molecular dynamics simulations and binding assays that interrogate the observed HMG-box and C1 domain interface and prominent cancer variants. Our results reveal a unique bipartite DNA-binding module and provide insights into the effects of cancer and domain interface mutations.

Introduction

The HMG-box protein Capicua (CIC) is a tumor suppressor frequently inactivated in oligodendroglioma, gastric adenocarcinoma and other cancers.[1-7] Originally identified in *Drosophila*, CIC is highly conserved in evolution and exists in two isoforms, Short (CIC-S) and Long (CIC-L), which acting redundantly or not, control numerous cellular and developmental processes, acting as a sequence-specific transcriptional repressor.[6,8-16] CIC often functions as a transcriptional repressor downstream of Receptor Tyrosine Kinase (RTK) signaling pathways, which once activated lead to phosphorylation and inactivation of CIC and, consequently, to derepression of its target genes.[5,17-23] This connection to RTK signaling means that oncogenic RTK activation can similarly lead to CIC inactivation and derepression of CIC targets such as the *ETV1/4/5* family of proto-oncogenes.[5,24,25] Aberrant transcriptional activity of CIC is implicated in other clinical disorders, particularly in Spinocerebellar Ataxia Type 1 and other neurobehavioral syndromes.[15,26,27]

CIC contains an HMG-box domain of the Sox type[28] but appears to employ its own mode of DNA binding (Figure 1A).[29] Unlike Sox proteins, which typically bind DNA as monomers, homodimers, or together with partner proteins that recognize adjacent DNA sites,[30] CIC primarily binds to isolated TGAATGAA-like octameric sites independently of other factors. However, *Drosophila* Cic can also cooperatively bind to suboptimal binding sites together with Dorsal, an NF-κB family transcription factor.[16] Remarkably, CIC associates with both types of DNA sites by using two separate domains: the N-terminal HMG-box and a C-terminal C1 domain conserved in all CIC proteins;[16,29] how this binding occurs, however, is not known. The importance of the HMG-box and C1 domains for CIC DNA binding is highlighted by the fact that they are both hotspots for inactivating mutations in oligodendroglioma and other tumors (Figure 1B).[1,31] Also, both domains appear to be required for the activity of oncogenic CIC-DUX4 fusions that bind to and activate (rather than repress) CIC targets and cause Ewing-like sarcomas.[25,29,32] Rapid release of CIC from DNA has been suggested as a likely first step involved in CIC downregulation by RTK signaling, although the molecular details of this process are lacking.[33] Therefore, structural studies identifying the mechanism of HMG-box-C1 DNA binding are needed; these studies will advance our understanding of CIC function and regulation and may offer future avenues for blocking CIC-DUX4 activity.

Here, we explore the structural features of how CIC binds DNA through its HMG-box and C1 domains. We report the crystal structure of the human CIC HMG-box fused to the C1 domain by a short protein linker (an HMG-box-C1 module), in complex with a DNA site from the *ETV5* promoter. We find that the C1 domain adopts a helix-turn-helix

(HTH) structure that is necessary for increasing both the affinity and sequence specificity of the HMG-box towards the target site. Molecular dynamics (MD) simulations were employed to probe fluctuations of the HMG-box and HTH domains when bound to DNA, and biochemical studies were performed to explore the effects of mutants from the COSMIC database and the interdomain interface on DNA binding. We observe an effect on binding by certain cancer and interface mutations of the HMG-box and C1 domains for DNA binding, and we test the dependence of these effects on linker length. This study provides a molecular depiction of how the HMG-box can be coupled with an HTH domain to bind a highly invariant octameric sequence in DNA.

Results and Discussion

Overall structure of the CIC DNA binding domains in complex with DNA

To structurally characterize the DNA-binding domains of CIC, an expression construct (CIC^{min18}) including the HMG-box and C1 domains of the human CIC protein joined by an 18-residue linker was employed (Figure 1A and Table S1). A similar construct with a 31-residue linker (CIC^{min31}) was previously shown to be functional (Table S1).[29] The linker was shortened to 18 residues to further decrease the size of the expression construct (186 residues) in comparison to full length CIC (1,608 residues in the short isoform), which was predicted to facilitate crystallization of the protein-DNA complex. We first assessed the ability of CIC^{min18} to bind an *ETV5* promoter containing the optimal TGAATGAA sequence by performing electrophoretic mobility shift assay (EMSA) experiments with a 30mer DNA oligonucleotide probe labeled with IRDye 700 (Figure 2A). The DNA probe shows the expected mobility shift for a protein-DNA complex when incubated with increasing amounts of CIC^{min18}. We next incubated CIC^{min18} with a 30mer DNA labeled probe replacing the CIC binding site with the random sequence GTCGCTGC and saw no mobility shift (Figure 2A), confirming that sequence specificity is maintained for the CIC^{min18} construct. Dynamic light scattering (DLS) experiments of CIC^{min18} showed a homogenous species with a radius of hydration (R_H) of ~ 3.5 nm (Figure S1A), thus opening the way to crystallization trials.

Crystallization of CIC^{min18} in complex with an 18-mer DNA oligonucleotide containing the TGAATGAA sequence was achieved using a 3:1 ratio of protein:DNA (Figure S1B), detailed in the Methods Details section. The 2.95-Å resolution crystal structure of the CIC^{min18} DNA complex was solved by molecular replacement using a model of the HMG-box with DNA (PDB 6JRP),[34] and X-ray data collection and refinement statistics are presented in Table 1. The crystal contains one CIC^{min18}-DNA complex per asymmetric unit in space group $P2_12_12_1$ with a Matthew's coefficient of 2.84 $Å^3$/Da. Clear electron density is observed for the HMG-box and C1 domains, as well as the 18-mer DNA (Figure S1C,D), whereas the N- and C-termini and the interdomain linker are disordered within the crystal structure. The final model includes the HMG-box (His33–Lys106 in PDB 7M5W | His199–Lys272 in Capicua (GenBank ID AAK73515.1)) and C1 (Pro118–Ala180 in PDB 7M5W | Pro1459–Ala1521 in Capicua) domains and the entire 18-mer duplex DNA. Unless otherwise indicated, the numbering used below corresponds to the human CIC-S isoform (see Table S1 for sequence numbering). Because the linker region lacks electron density, presumably due to flexibility, we explored if the orientations of the HMG-box and C1 domains bound to DNA are consistent with the linker length. Using BioLuminate,[35] the 18-residue linker could be built between the positioned HMG-box and C1 domains without clashing of the linker with the two domains or the DNA (Figure 2B), indicating that CIC^{min18} is bound to one TGAATGAA sequence within the crystal lattice.

The CIC HMG-box binds the minor groove of DNA

The CIC HMG-box domain adopts a canonical HMG-box fold consisting of three α-helices arranged in an L-shaped configuration, interacting with the minor groove of the DNA along the octameric sequence (Figure 2B). Helix-H1 (HMG^{H1}) is inserted within the minor groove of DNA, helix-H2 (HMG^{H2}) packs along the DNA phosphodiester backbone, and helix-H3 (HMG^{H3}) makes interactions with both the phosphodiester backbone and the minor groove. Binding of the HMG-box induces a ~66° bend of the DNA calculated using Curves+,[36] expanding the average minor groove width at the bend to 12 Å. Similar binding modes are observed in the nearest structural homologs, SOX9 (PDB 4S2Q)[37] and SOX18 (PDB 4Y60),[38] with root mean square deviations (r.m.s.d.) of ~1.0 Å for 70 aligned Cα atoms and shared sequence identities of 36% and 39%, respectively (Table S2 and Figure S2A,B).[39] Another structural homolog is the lymphoid enhancing factor, LEF1 (PDB 2LEF),[40] with an r.m.s.d. of ~1.3 Å and sequence identity of 32% (Figure S2A,B). In comparing to the nonspecific HMG-box domain, an alignment with HMGB1 (PDB 2GZK)[41] yielded an r.m.s.d. of ~1.6 Å and sequence identity of 30% (Figure S2A,C). Therefore, similar overall DNA binding modes are observed in both specific and non-specific HMG-box domains.

The CIC C1 domain adopts a HTH fold and interfaces with the HMG-box

Unlike the HMG-box, the C1 domain shares no sequence identity with known protein structures. To our surprise, electron density for the C1 domain revealed a tri-helical HTH fold with helices arranged in a right-handed helical bundle (Figure 2B). In comparison to traditional HTH proteins,[42] an extended loop is observed between the first and second helices, and a short 3_{10} helix is observed within the loop segment between the second and third helices. At the center of the helical bundle is a hydrophobic core of nonpolar and aromatic residues. The first half of helix-H1 ($C1^{H1}$) is positioned to interact with the DNA phosphodiester backbone. Helix-H2 ($C1^{H2}$) is in an antiparallel orientation to $C1^{H1}$, leading to helix-H3 ($C1^{H3}$), which is inserted into the major groove of the DNA (Figure 2C). The major groove, where C1 is bound, is distorted as a result of HMG-box binding and is narrower than standard B-form DNA, with an average groove width of 7.7 Å in comparison to 12.2 Å outside of the protein binding site (calculated using CURVES+)[36]. The groove depth increases from ~5 Å to 8.5 Å in the protein-binding region. The classic HTH domain is known to bind to B-form DNA,[42] and our structure of the CIC-C1 domain reveals how a HTH domain can bind to bent DNA.

The orientation of the C1 domain positions the loop preceding $C1^{H2}$ and the C-terminal portion of $C1^{H3}$ closest to HMG^{H2} (Figure 2C). This orientation of domains places HMG-Y239 against a hydrophobic pocket within C1 (residues M1519, F1483, F1484, and F1478) and the DNA. HMG-K232 is positioned within van der Waals distance to the C1 domain as well as hydrogen bonding distance to the backbone C=O group of C1-P1485. These interactions wrap the CIC^{min18} construct around the DNA and may assist in binding. The number of interactions between the HMG-box and C1 domains is nevertheless small, which is consistent with previous EMSA experiments showing that, at low concentrations, separately expressed HMG-box and C1 domains cannot efficiently bind to DNA (even when both are present in the same reaction) unless they are physically joined by a linker.[29] In contrast, there are literature reports of CIC HMG box binding to DNA containing the *ETV5* promoter sequence with affinities ranging from low nM[43] to μM.[34] A comparison of EMSA conditions reveals that those studies omitted poly dIdC for reduction of nonspecific DNA binding, which could explain the observed discrepancy. Poly dIdC was included in all EMSA reactions presented here and in the Forés et al. study.[29] Therefore, it is likely that high concentrations of the HMG-box can lead to a DNA-bound state as observed in the reported structure of the CIC HMG-box alone with DNA. The HMG-box domains align with an r.m.s.d. of 0.5 Å for all Cα atoms, therefore

the presence of the C1 domain does not greatly impact the conformation of the HMG-box domain. However, it remains unclear if DNA is bound nonspecifically under such experimental conditions, which we explore below.

The C1 domain is structurally similar to other HTH domains, including the FF domain (a type of HTH domain characterized by two conserved phenylalanine residues located in the first and third helices) of the glucocorticoid receptor DNA-binding factor 1 (PDB 2K85),[44] with an r.m.s.d. of 2.3 Å for 60 aligned Cα atoms despite only 12% sequence identity (Tables S3 and S4).[45] A similar alignment score is observed for other HTH family members including the homeodomain (Table S3). The FF domain, however, contains an inserted 3_{10} helix between the 2nd and 3rd helices, similar to the C1 domain. Interestingly, the FF domain binds phosphopeptides,[42] and the FF domain from CA150 has been implicated in DNA binding.[46] No structures of the FF domain bound to DNA or to a phosphopeptide are currently available; however, NMR titration analyses of the FF domain from human HYPA/FBP11 support a conserved role of this domain in phosphopeptide binding.[47] Since CIC repressor activity is inhibited through RTK-induced phosphorylation, it is tempting to consider if the structural similarity of the C1 and FF domains might implicate C1 in recognizing a phosphopeptide. Residues within CIC^{min18} at the DNA-binding interface create a complementary electrostatic surface for DNA binding (Figure 2D); however, an electrostatic potential mapping of the C1 domain shows the surface away from the DNA is largely nonpolar, except for the presence of R1465 from H1, R1496 of H2, and D1499 within the 3_{10}-helix. The lack of positive electrostatic features outside of the DNA-binding surface implies that the C1 domain cannot simultaneously bind to DNA and a phosphorylated peptide, and the surface electrostatic features of the HMG-box domain similarly lack extensive positive electrostatic features outside of a minor patch near R213 of HMG^{H1} and K244 and K248 of HMG^{H3} (Figure 2D). Therefore, if the C1 domain is somehow involved in phosphopeptide binding as in the FF domain, binding to DNA and a possible phosphopeptide would be mutually exclusive, suggesting the primary role of C1 is in DNA binding (but raising the possibility that such binding could be inhibited if C1 were able to recognize other CIC motifs phosphorylated by the RTK pathway; see also below).

The HMG-box and C1 domains of CIC provide a stable DNA-binding interface

The observed binding mode of the HMG-box and C1 domains to bent DNA with a narrower major groove suggests that binding of each domain is necessary to increase the binding affinity of the other domain. In fact, neither domain can bind without the other at lower concentrations that mimic physiological conditions and contain poly dIdC to reduce nonspecific DNA binding,[29] which suggests that the intact CIC protein must maintain functional orientations of both the HMG-box and C1 domains to achieve robust DNA binding. When bound to DNA, the HMG-box and C1 domains bury approximately 1275 $Å^2$ and 517 $Å^2$, respectively, with 220 $Å^2$ buried between the two domains. To explore the stability of the CIC^{min18}-DNA complex, motions of the complex on the ns time scale were calculated from three separate 250 ns runs of MD simulation within Schrodinger's BioLuminate using three different starting models prepared as detailed in the Materials and Methods. All models had a computationally generated interdomain linker. As expected, the root mean square fluctuation (r.m.s.f.) values of loops in both domains are greater than the ordered α-helices, and the protein termini and interdomain linker have the highest r.m.s.f. values (Figures 3A and S3A). The lower r.m.s.f. for the α-helices of the HMG-box and C1 domain suggests a stable binding conformation. Helix $C1^{H3}$ demonstrates the lowest r.m.s.f. values, consistent with its placement towards the

narrow major groove, at the interface with the phosphodiester backbone of the upper strand and at the interface with the HMG-box. Interactions between CIC^{min18} and DNA are stable over the course of the 250 ns trajectories.

We further used MD simulations to examine the influence of the interdomain linker on the CIC^{min18}-DNA structure. The r.m.s.d. values of helices, from both the HMG-box and C1-domain, over the 250 ns trajectories are shown in Figure 3B. The structures appear to equilibrate within the first half of the trajectories (Figure S3A), and an alignment of initial and final states shows similar positioning of the HMG-box and C1-domains (Figure S3B). Interestingly, the linker behaves differently in the three performed simulations, suggesting it does not pose large restraints on DNA binding (Figure S3C). The helices surrounding the linker, HMG^{H3} and C1^{H1}, are more mobile over the course of the trajectories and have higher r.m.s.f. values compared to helices that have greater contacts with the DNA (Figures S3A). Therefore, the CIC^{min18}-DNA complex structure is likely unaffected by the linker, suggesting that both domains are freely moving and aligning during binding.

Molecular determinants of CIC^{min18} binding to DNA

A majority of the interactions between the HMG-box and DNA are mediated by helices HMG^{H1} and HMG^{H2}, in addition to residues just N-terminal of HMG^{H1} (Figure 4A,B). HMG^{H1} bends the oligomer DNA by packing F207 and M208 as a wedge between subsequent DNA bases in the minor groove of the octameric sequence (Figure 4C), extending the rise and twist of the DNA to 5.3 Å and 50.8°. Two residues N-terminal to the HMG^{H1} helix, R201 and R202, wrap around the DNA to enable possible hydrogen-bond and electrostatic contacts (Figure 4D). R201 is directed towards phosphates of the backbone, but R202 is positioned more closely to the first two nucleotides of the octameric sequence (T6^{U} and dG7^{U}). Within HMG^{H1}, K212 interacts with the phosphodiester backbone whereas R215 is positioned to interact with T10^{U} (Figure 4E). Residues N227, R228, and S231 form the interface between HMG^{H2} and the oligomer (Figure 4F). N227 can form hydrogen bond contacts with T10^{U} and dG11^{U} and S231 forms a hydrogen bonding interaction with the N3 nitrogen of dA9^{L}. By contrast, HMG^{H3} seems to play a minor role in binding with only a few contacts, including H250, K257, and N205 (from the N-terminus) to the phosphodiester backbone of the lower strand (Figure 4G).

Although the HMG-box domains from CIC and the related factors SOX18 and LEF1 similarly bend DNA and use many of the same residues for their nucleobase and phosphodiester interactions (Figure S2A), comparison reveals a few notable differences in DNA-binding interactions (Figure S4). LEF1, a member of the TCF/LEF family mediating both activation and repression of transcription, employs an additional methionine (M14),[40] compared to the methionine and phenylalanine pair in CIC and SOX18,[38] to intercalate within its DNA binding site in the T-cell receptor-alpha enhancer region. CIC lacks contributions to the electrostatic interface that are observed in SOX18 and LEF1,[38,40] such as H29 and R73 (Figure S4B and S4C). However, CIC does form an interaction with T7^{L} via R228, which is an alanine in both SOX18 and LEF1 (Figure S4D).[38,40] This residue interacts with the carbonyl of T7^{L} and may support binding of the octameric sequence by recognizing a pyrimidine at this position of the lower strand.

Importantly, our structure suggests why the HMG-box alone does not allow for binding of CIC to DNA under physiological conditions. In the CIC^{min18} structure, the C1 domain makes several non-sequence specific contacts that may increase binding affinity of CIC to DNA. Most of these contacts originate from the C1^{H3} helix and are exclusively with the phosphodiester backbone within the major groove, opposite of the HMG-box

(Figure 5A,B). These interactions are within the lower strand of the octameric sequence as well as the flanking sequence. Residues Q1508, R1512, R1515, and Q1516 are within hydrogen-bonding distance of $dA9^{L}$-$T11^{L}$ (Figure 5C). The $C1^{H3}$ helix also forms contacts with the flanking sequence phosphodiester backbone, along with R1464, R1471, and K1510 of $C1^{H1}$ (Figure 5D). K1510 is also within distance to form a hydrogen-bonding network with E1513. Therefore, these interactions likely support an important role of $C1^{H3}$ in contacting the lower strand of the CIC site via the major groove, and the HMG box provides specific interactions with the nucleobases of the TGAATGAA sequence.

DNA specificity of CIC^{min18} is driven by an unpredicted DNA binding mode for a HTH domain

Clear electron density enabled unambiguously building the entire 18-mer duplex DNA. To test the proper placement of DNA, refinement was attempted using DNA in the flipped orientation. Refinement of the flipped DNA led to an increase in R-free of ~2.6% and difference electron density peaks (Figure S5A-C), supporting the DNA directionality in our final model. The orientation of CIC in our structure positions the intercalating residues of the HMG-box domain, M208 and F207, between $dA8^{U}$-dA^{9U} and $T10^{L}$-$T11^{L}$, respectively. Surprisingly, in the previously reported structure of the CIC HMG-box alone with DNA (PDB ID 6JRP),[34] the octameric sequence is flipped and shifted by 1 bp with respect to the DNA positioning in the CIC^{min18} complex. In this flipped and shifted orientation, the HMG-box wedge residues F207 and M208 disrupt base stacking between the dGdA/dCT of the 3′ half of the octameric CIC binding site (Figure S5D), which would be less favorable in comparison to intercalation within the dAdA/TT sequence. Since our previous DNA binding studies revealed that the HMG-box alone does not effectively interact with DNA in the presence of poly dIdC,[29] the above differences in binding mode led us to wonder if the HMG-box-C1 module used here might enhance DNA binding specificity in comparison to the HMG-box-only structure.

To address C1 contributions to DNA specificity, we analyzed movement of the C1 domain over the 250 ns simulations discussed above. From the solved structure, the C1 domain is positioned with $C1^{H3}$ along the major groove making most of the contacts with the DNA (Figure 5), and this binding mode is maintained during the trajectory (Figure S3B). Contacts between the C1 domain and the DNA are strictly with the DNA phosphodiester backbone throughout MD simulations, either by hydrogen bonding with polar residues such as R1464, K1510, R1512, and R1515, or potentially by water-mediated contacts with Q1508, E1513, and Q1516 (Figure 5C and 5D). E1513 makes a water-mediated contact between its carboxylate and the phosphate group of $T10^{U}$ in the consensus sequence. R1512, R1515, Q1508, and Q1516 make contacts with the phosphate groups of the $dA9^{L}$-$dC12^{L}$. These C1 interactions lack hydrogen-bonding specificity with the octameric sequence; however, Q1516 is positioned near $T11^{L}$, which would clash with the observed $dA8^{U}$ in the 6JRP structure (Figure S5E). Previous work showed that a TGAA**C**GAA oligomer, which contains a T to dC mutation, had greatly reduced binding.[29] Although we observe no sequence-specific interactions to this base position, the mutation is located within the distorted bend of the DNA oligo. Therefore, the paired guanosine of the opposite strand would introduce a clash of its C6 carbonyl with the C4 carbonyl of the displaced thymidine from the site of intercalation (Figure S5F). Lack of nucleobase-specific interactions suggests that the C1 domain provides stability in the bound structure through increased interactions with the DNA and interactions with HMG^{H2} along the DNA backbone, and the C1 domain further adds specificity by destabilizing certain DNA sequences within the bent DNA binding mode.

CIC^{min18} and CIC^{min31} bind the *ETV5* promoter with nM binding affinities

To complement the structure of CIC^{min18} for understanding the effects of prevalent cancer mutations within the HMG-box and C1 domains, the DNA binding affinity of CIC^{min18} was first assessed in comparison to the previously reported CIC^{min31} construct used in Forés et al. (2017).[29] Because the interdomain linker in CIC^{min31} is 13 residues longer than for CIC^{min18}, it was important to determine if the linker length has any effect on the DNA binding behavior. Binding was studied by microscale thermophoresis (MST) using bacterially expressed proteins and a 30-mer deoxyoligonucleotide sequence, containing the CIC binding site and flanking bases from the ETV5 promoter, labeled with Cy5. Recombinant proteins were assessed by SDS PAGE (Figure S6A) and circular dichroism (CD) (Figure S6B and S6C), demonstrating the expected α-helical fold with minima near 208 nm and 222 nm. Attempts to use the EMSA buffer containing 50 mM NaCl (as in Figure 2A) in MST experiments led to protein aggregation observable in DLS (Figure S6D). As protein aggregation can complicate MST data analysis, all MST experiments were performed in PBS buffer that minimizes aggregation. Interestingly, the CIC^{min18} construct shows cooperativity in binding (Table 2 and Figure S7), whereas the CIC^{min31} construct shows standard 1:1 binding (Figure S12A). Therefore, the Hill model was used to compare EC_{50} and Hill coefficient (n) values. CIC^{min18} bound 10-fold tighter than CIC^{min31} (Table 2) with a Hill coefficient of 5. Nonetheless, nanomolar binding affinities for both the CIC^{min18} and CIC^{min31} constructs allowed us to assay protein variants for changes in binding affinity and cooperativity.

Interrogating cancer mutations within the CIC DNA-binding domains

To better understand the impact on DNA binding of cancer-associated CIC mutations[48] (Figures 1B and Table S5), the prominent variants R201W and R215W were selected. Both R201W and R215W disrupt CIC activity[29,49]. Therefore, these variants were first tested by EMSA using proteins produced by in vitro transcription/translation (IVT) (Figure 6). The CIC^{min31} construct was chosen to provide direct comparisons to previously reported EMSA mutagenesis results,[29] and IVT protein samples were validated by western blotting. Intriguingly, R201 is positioned along the backbone of the DNA, whereas R215 is positioned to interact with nucleobases within the *ETV5* site. R215W was found to dramatically decrease DNA binding of the CIC^{min31} construct (Figure 6), in line with previous mutational observations for this construct[29] and for the HMG-box alone.[34] The CIC^{min31}:R201W variant, however, displayed greater binding to DNA by EMSA (Figures 6 and S8A), which contrasts with the previous reports.[29,34]

We next explored DNA binding by MST to further quantify changes in binding affinity. MST data for R201W in both the CIC^{min18} and CIC^{min31} constructs also show slightly stronger binding in comparison to WT, with a decreased Hill coefficient of 0.85 for CIC^{min18}:R201W (Table 2 and Figure S7). The observed DNA binding for CIC^{min18}:R201W is consistent with the solvent exposed location of R201 (Figure 4D) in the CIC^{min18} structure, and the observed negative cooperativity suggests differences in DNA binding as a result of the R201W mutation. The R215W protein produced aggregates when recombinantly expressed within the CIC^{min18} construct, therefore only CIC^{min31}:R215W was explored by MST. Surprisingly, the recombinant CIC^{min31}:R215W protein bound slightly tighter to DNA than CIC^{min31}:WT (Table 2 and Figure S7). These results point out unexpected differences between EMSA and MST methods for R215W,

which suggests changes in binding rates that are not captured by MST. Despite these observations, cellular assays with full-length CIC:R201W[24,49] and CIC-R215W[50] both showed decreased DNA binding and reduced repressor activity. Therefore, future experiments with longer CIC constructs are likely needed to better understand the full effects of these mutations on the structure and function of CIC, although these observations are intriguing to consider for future bioengineering of minimal DNA-binding domains in the future.

There are also several mutations found in the COSMIC database within the C1 domain. Missense mutations primarily affect R1465, R1471, V1474, R1512, and R1515 (Table S5). The most frequent mutations are arginine to methionine, cysteine, or leucine. Residue V1474 is located on $C1^{H3}$ and is wedged between the first and third helices of the C1 domain; therefore, mutations of V1474 to similarly non-polar residues leucine, glycine, and phenylalanine may alter hydrophobic packing between the helices of the C1 domain and decrease DNA binding and overall stability. R1512 and R1515 are the most frequently mutated residues of the C1 domain within the COSMIC database, and the structure of CIC^{min18} with DNA provides a clear explanation: loss of the positive electrostatic interactions and hydrogen bonding capability of either R1512 or R1515 would severely disrupt the protein-DNA interface, thus decreasing binding and regulatory activity. To test this hypothesis, we produced the R1512H and R1515H variants within CIC^{min18}, and the R1515H/L variant within CIC^{min31}. Replacement of arginine with histidine is a more conservative mutation, and similar binding is observed in comparison to WT for the CIC^{min18} construct (Table 2), albeit with a decreased Hill coefficient. The CIC^{min31} R1515H/L variants, however, show the weakest DNA binding observed by EMSA (Figure 6C) and MST (Table 2 and Figure S7 for the R1515H variant). The observed migration shift for the CIC^{min31}R1515H/L variants bound to DNA in comparison to WT (Figure 6C) is likely due to the weakened binding of the C1 domain, yielding a less compact structure. These findings highlight the C1 domain's role in increasing affinity for the TGAATGAA consensus sequence. Mutation results for R1515 are consistent with the previous report that the R1515L variant in CIC^{min31} is unable to effectively bind DNA in an EMSA assay.[29] Cellular studies with full length CIC:R1515H report decreased repression and increased cytoplasmic distribution of this variant.[49] Taken together, these results suggest that although histidine may be able to make polar interactions within the shorter linker CIC^{min18}, these interactions are not enough to maintain DNA binding for a repressor function of CIC^{min31} and full length CIC.

Mutations at the interdomain interface affect DNA binding

Having confirmed that known cancer mutations can affect DNA binding ability, we next explored the importance of the interdomain interface on binding DNA within CIC^{min18} and results are presented in Table 2 and Figure S7. Residues K232 and Y239 within the HMG-box were individually mutated to alanine residues to determine the effect of loss of the polar and packing interactions between the two domains. M1519 within the C1 domain was mutated to an aspartate to introduce a negatively charged residue into an aromatic pocket. Residues K232A, Y239A and M1519D showed weaker DNA binding with smaller Hill coefficients compared to WT. These data reveal that the polar and packing interactions between the HMG-box and C1 domains are important for supporting the binding mode observed in the crystal structure. We lastly wanted to test if the linker length affects the importance of these interface residues. Therefore, we generated CIC^{min31}:Y239A and M1519D. The Y239A variant bound slightly weaker than WT

CIC^{min31} in the EMSA assay (Figure 6), consistent with the CIC^{min18}:Y239A result, but CIC^{min31}:Y239A yielded insoluble protein, preventing MST analysis for the longer linker construct. Surprisingly, CIC^{min31}:M1519D bound tighter to DNA than WT in MST and with a similar affinity in EMSA (Table 2 and Figure 6B). That M1519D had opposite effects on binding depending on the linker length suggests the HMG-box and C1 domains of CIC^{min31}:M1519D may arrange differently on the DNA. Nonetheless, the observed effects on DNA binding with the longer interdomain linker support the identified interdomain interface. These studies identify important interdomain interactions, and future experiments will be needed that explore possible conformational flexibility within the full-length CIC protein.

Based on the observations for different cancer and interdomain mutations, we could not rule out that DNA binding specificity or protein stability might be altered by a mutation. Therefore, we tested if any of the generated mutants might demonstrate more promiscuity and bind tighter to a randomized octameric sequence by EMSA. No binding was observed to the randomized sequence (Figure S8B). To determine if stability was affected, we chose to explore the CIC^{min31} construct because larger binding changes were observed for its variants in comparison to CIC^{min18}. Thermal melts of CIC^{min31} and its reported mutants did not show large deviations, although the mutations R1515H and M1519D decreased the T_m from 62 °C to 58 °C (Figure S6E). Therefore, the explored mutations do not greatly alter DNA-binding specificity or intrinsic protein stability, at least in the context of our minimal constructs. These mutational studies reveal that both the HMG-box and C1 domains must be functional to have appropriate binding to the TGAATGAA sequence, and the interdomain linker length alters binding cooperativity and limits the magnitude of mutational effects on DNA binding, which may guide future bioengineering of related DNA-binding constructs. Intriguingly, derepression of CIC targets is linked to CIC phosphorylation by RTK signaling, and known sites of phosphorylation are outside of the observed DNA binding domains.[9] A recent study suggested that multisite phosphorylation of the central regions of full-length CIC may result in a disruption of an intramolecular interaction between the two parts of the protein harboring the HMG-box and the C1 domain, leading to a decrease in DNA binding.[51] Therefore, phosphorylation in full length CIC may affect the conformational landscape of the HMG-box and C1 domains, in turn altering CIC DNA-binding.

In summary, we have presented the structure of the CIC HMG-box and C1 bi-partite module bound to DNA, revealing a new binding mode of the HMG-box and C1-domain and providing new molecular detail into how known cancer-related mutations and interdomain interaction mutations would affect DNA binding. The C1 domain adopts a HTH fold that resembles the overall structure of the FF domain, uncovering a previously unknown mechanism of DNA binding in which an HMG-box and a HTH domain act together to recognize a specific octameric target site. Interestingly, whereas the HMG-box is an ancient domain present in a diverse set of proteins, the C1 domain is restricted to the CIC family of proteins, suggesting that the latter originated by genetic drift within a pre-existing HMG-box protein. This would have allowed a highly selective mechanism of recognition that is not found among other HMG-box factors, but which is nevertheless susceptible to inactivation by mutations affecting either of those two domains, as seen in CIC-related cancers. Finally, since oncogenic CIC-DUX chimeras rely on the same DNA binding mechanism for activating instead of repressing CIC target genes, the structure

and affinity studies we report here may facilitate the development of strategies to interfere with CIC-DUX activity in CIC-rearranged sarcomas.

Resource Availability:

Lead contact: Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Daniel Dowling (Daniel.dowling@umb.edu).

Materials availability: Plasmids generated in this study are available from the lead contact.

Data and code availability:

- The coordinates of the reported crystal structure have been deposited in the Protein Data Bank with accession code 7M5W for public availability.
- This paper does not report original code.
- All data reported in this paper will be shared by the corresponding author upon reasonable request.

Acknowledgements

The authors kindly acknowledge the MIT Structural Biology Core facility and Staff Scientist Robert Grant, and the Biophysical Instrumentation Core at the University of Massachusetts Boston, funded by the Massachusetts Life Sciences Center. A.V. is funded by the National Institutes of Health grants GM141843 and GM158116. G.J. is funded by the Spanish Government (grants BFU2017-87244-P and PID2020-119248GB-I00) and ICREA. This work was funded in part by the University of Massachusetts Boston Office of Global Programs (to D.P.D.).

Author contributions

Investigation and Data Curation, J.W., J.J.M.L., A.D.G., M.P., K.I., S.P., and M.F.; Formal Analysis, J.W., J.J.M.L., A.D.G. and D.P.D.; Validation, J.W., J.J.M.L., A.D.G., K.I., S.P., M.F., G.J., A.V. and D.P.D.; Writing – Original Draft, J.W., G.J., A.V., and D.P.D.; Writing – Review & Editing, J.W., G.J., A.V., and D.P.D.; Funding Acquisition, D.P.D., G.J. and A.V.; Resources, G.J., A.V., and D.P.D.; Supervision, G.J., A.V., and D.P.D.

Declaration of Interests

S.P. is currently employed by Genentech, Inc., South San Francisco, CA, USA, and does not hold any Roche shares. Her contributions to this work were made while she was a graduate student at the University of Massachusetts. The other authors declare no competing interests.

Main Figure Titles and Legends

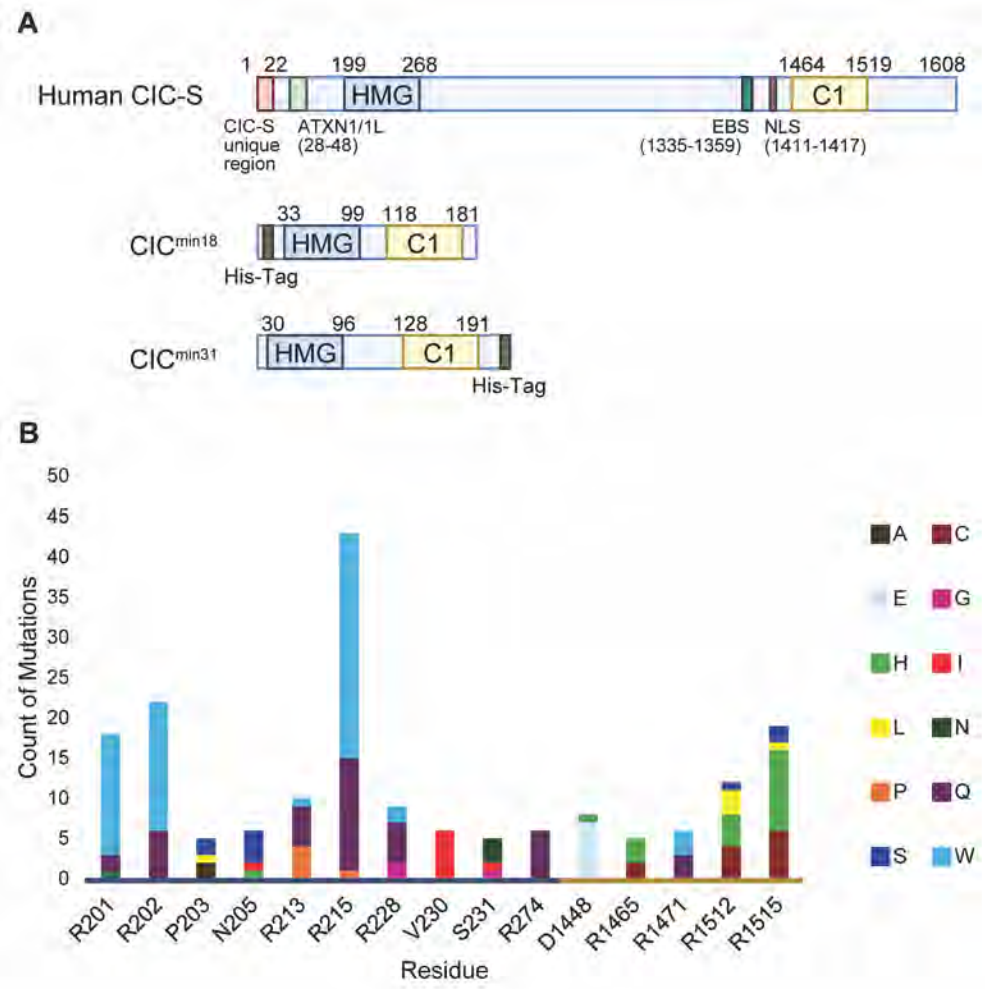

Figure 1. Capicua is a transcriptional repressor found to contain multiple cancer-associated mutations within its DNA-binding domains. **(A)** The CIC-S isoform is shown, displaying its two DNA-binding domains, the HMG-box domain (residues 201 – 269) and C1 domain (residues 1459 – 1521). Also shown are the ataxin-l-like binding region (ATXN1/1L), the ERK binding site (EBS), and nuclear localization signal (NLS). Boundaries of the Human CIC-S protein are defined based on reference.[52] This work employed a minimal construct (CIC^{min18}) containing the HMG-box and C1 domains joined by an 18-mer linker, in addition to a 31-mer linker construct (CIC^{min31}). The DNA binding domain boundaries are defined based on this structural work. **(B)** Mutations associated with cancer patients from the COSMIC database were identified within CIC, and those mutations that lie within the CIC DNA-binding domains are displayed. Displayed data have been culled to show only residues that contain at least 5 reported occurrences of a missense mutation.

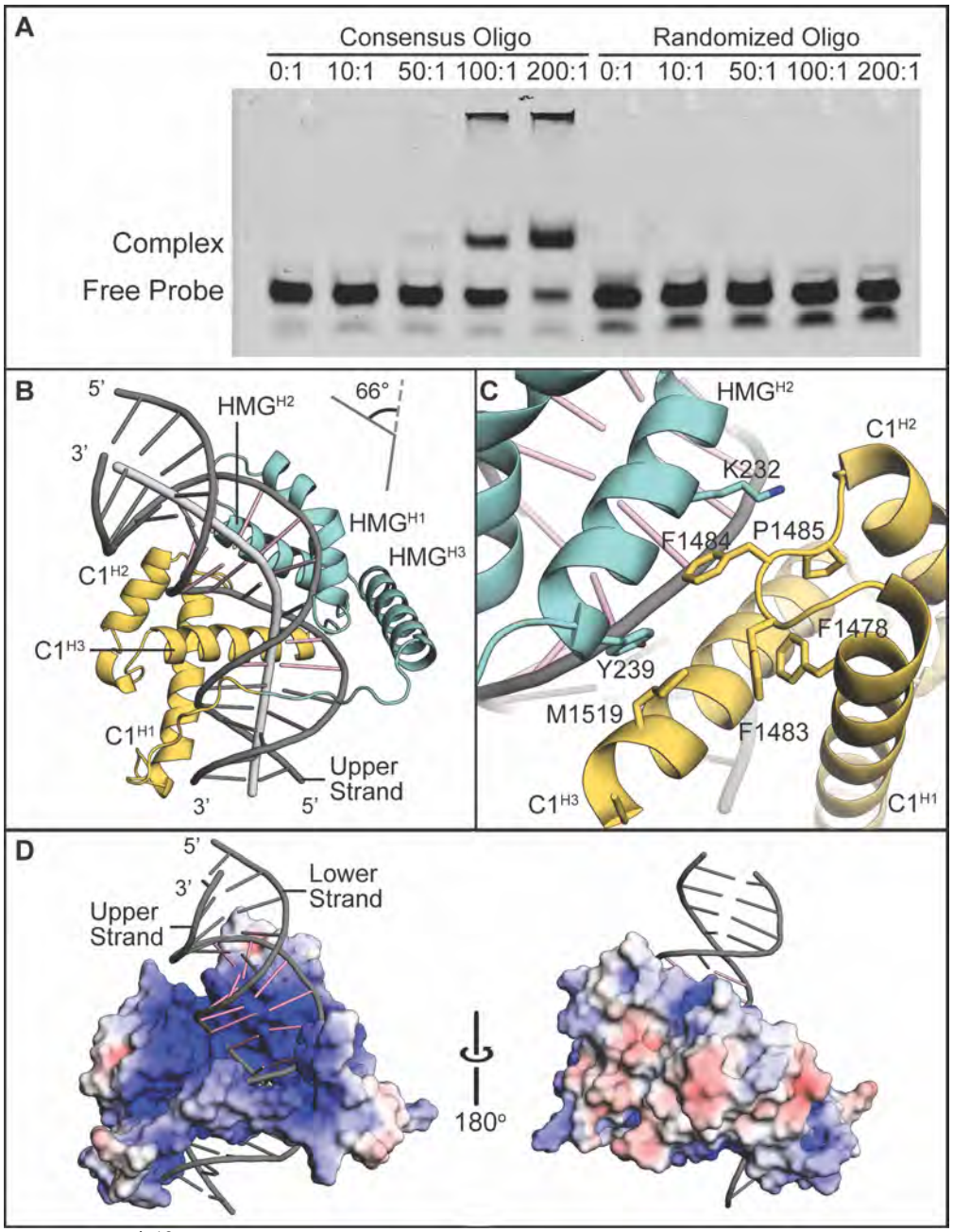

Figure 2. CIC^{min18} utilizes coordinated binding of the HMG-box and C1 domains to the core octameric DNA sequence. **(A)** EMSA of CIC^{min18} with oligomer containing either the *ETV5* promoter or a random sequence shows specific binding at increasing ratios of protein:oligomer. **(B)** The overall structure of the CIC^{min18} complex shows bending of bound DNA by 66°, represented by a line along the center axis. The C1-domain is colored in yellow, the HMG-box is colored in teal, and the DNA is labeled in pink for nucleotides of the TGAATGAA sequence and grey for the flanking sequence. **(C)** Shown are the interfacing residues between the HMG-box and C1 domains. **(D)** The electrostatic surface of the CIC^{min18} protein reveals an electropositive region for the binding of DNA. Electrostatics were calculated for CIC^{min18} separately from DNA. The electrostatic surface of the protein is colored as a red to blue heat map for -5 to +5 k_BT/e_c. See also Figure S1.

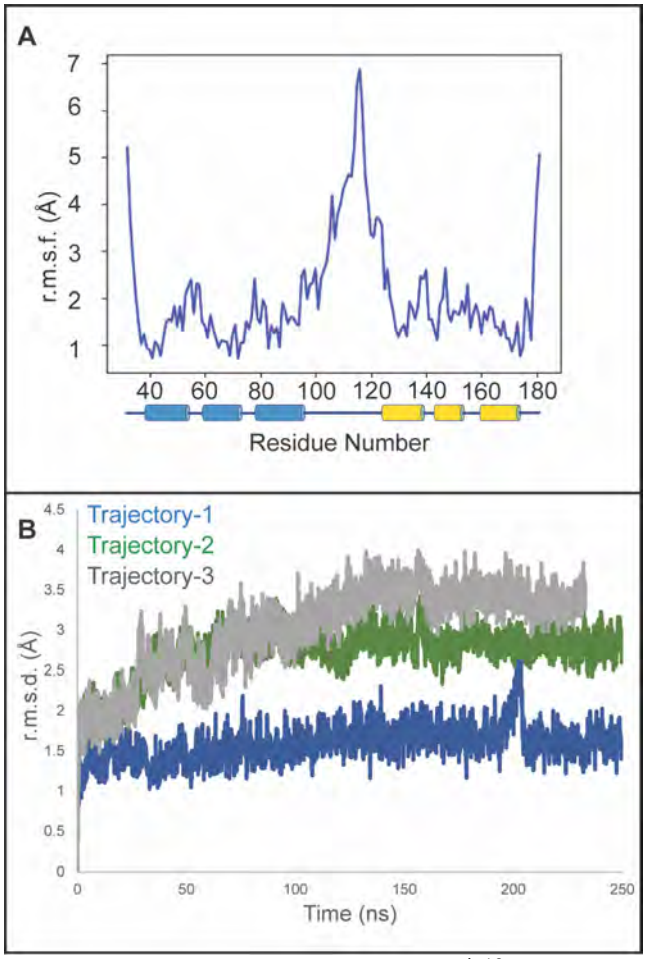

Figure 3. Molecular dynamics simulations of the CIC^{min18} structure show stable binding of the DNA oligomer. **(A)** The r.m.s.f. values calculated over a representative 250 ns trajectory shows the linker region as being highly flexible relative to the HMG-box and C1 domains (represented by cartoons of blue and yellow cylinders for helices, respectively). See Figure S3 for information for all MD simulations. **(B)** The r.m.s.d. values calculated from the alignment of Cα protein atoms of helices only to the initial frame for the three trajectories, colored green, grey, and blue, indicate equilibration of the CIC^{min18} complex.

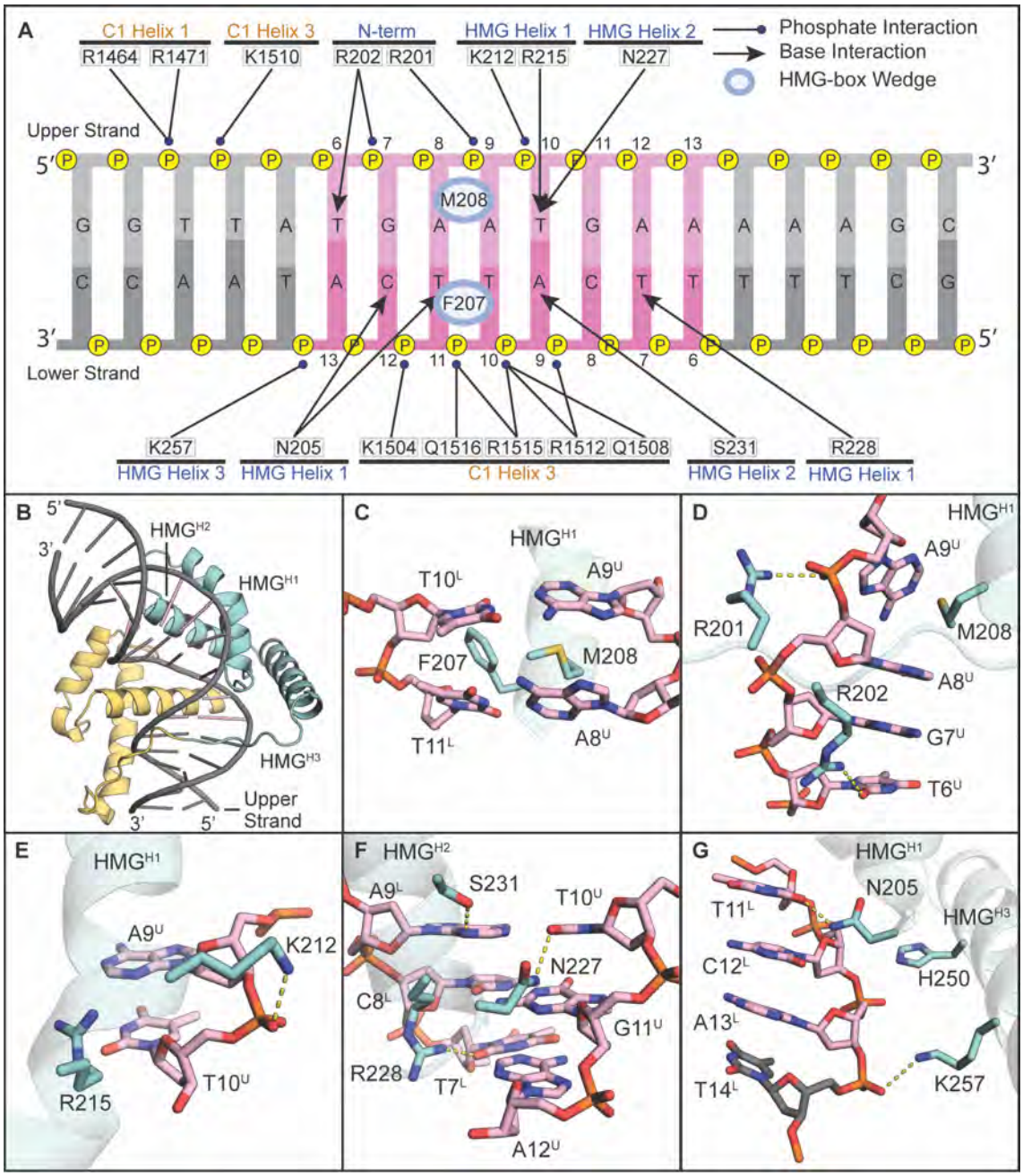

Figure 4. The crystal structure of CIC^{min18} details specific interactions by the HMG-box within the octameric sequence of the DNA. Superscripts indicate upper (U) or lower (L) DNA strands. Interactions within 3.3 Å are shown as yellow dashed lines. **(A)** A summary of interactions between CIC^{min18} and DNA are shown. **(B)** Crystal structure of CIC^{min18} with the HMG-box (blue) helices and C1 domain (yellow) shown to help orient panels c - g. DNA is colored as in Figure 2. **(C)** F207 and M208 form the intercalating wedge between two AT pairs of the DNA in the octameric sequence. Image is the 180° rotation of panel b along the vertical axis. **(D)** Residues R201 and R202, at the N-terminal extension of the HMG-box, wrap around the phosphodiester backbone of the first TGAA DNA sequence. **(E)** and **(F)** Residues from the first two helices of the HMG-box domain form several polar interactions with the upper strand just after the intercalation by F207 and M208. **(G)** Helix HMG^{H3} forms several interactions with the phosphodiester backbone.

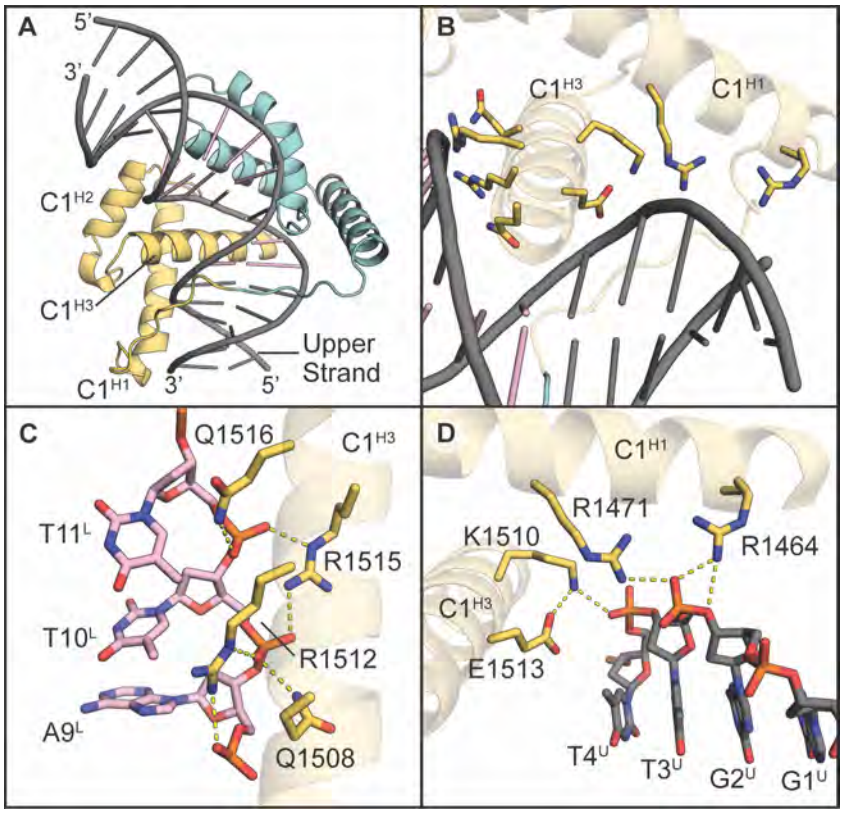

Figure 5. The crystal structure of CIC^{min18} details specific interactions by the C1 domain within and around the octameric sequence of the DNA. Superscripts indicate upper (U) or lower (L) DNA strands, as shown in Figure 4A. Interactions within 3.3 Å are shown as yellow dashed lines. **(A)** Crystal structure of CIC^{min18} showing the C1-domain in yellow, the HMG-box in blue, and the interdomain linker, built with BioLuminate,[35] between the two domains. DNA is colored as in Figure 2. **(B)** The C1 domain positions the third helix within the major groove of the DNA, where it forms interactions with the DNA backbone **(C)** within and **(D)** outside of the octameric recognition sequence.

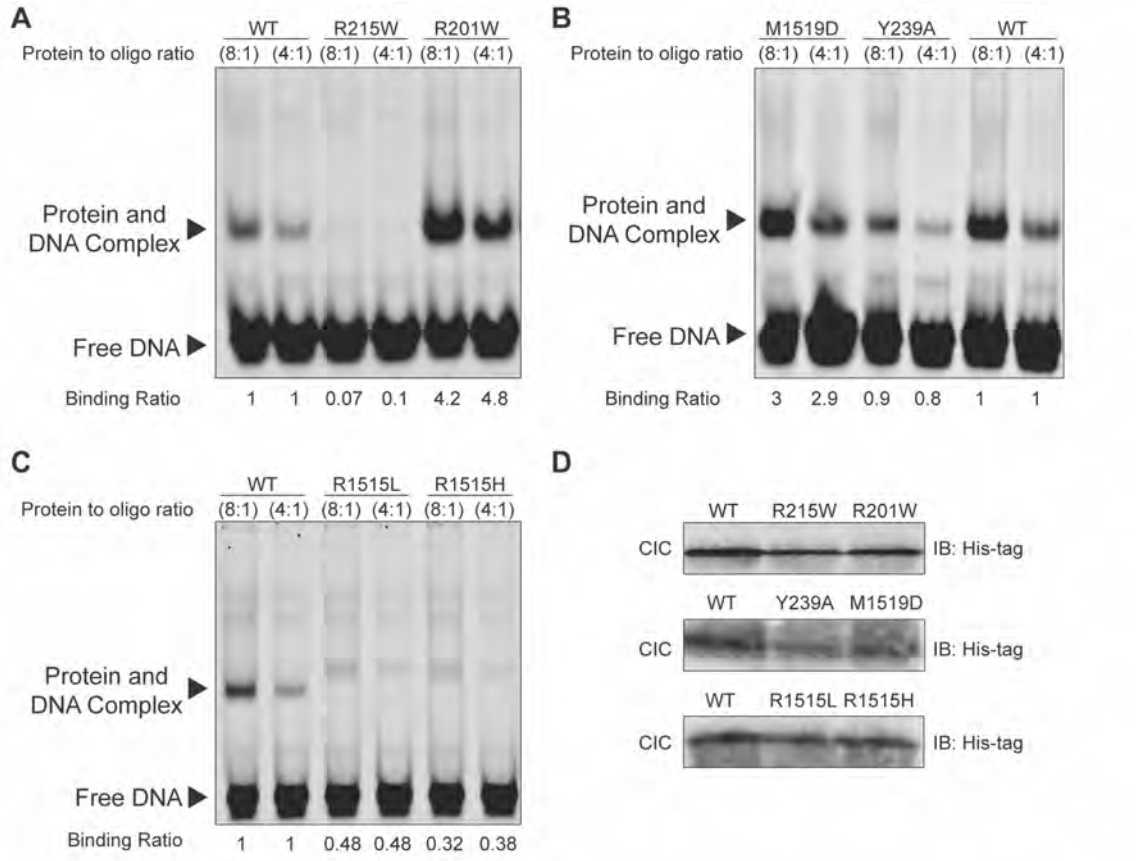

Figure 6. EMSA of IVT-produced CIC^{min31} mutants shows varied effects on binding between protein and DNA oligo. **(A)** EMSA of CIC^{min31} WT and HMG-box mutants against a DNA probe, containing the CIC consensus sequence TGAATGAA, shows decreased binding by the R215W variant and increased binding by the R201W variant compared to WT. **(B)** EMSA of CIC^{min31} WT and interface mutants against the same DNA probe shows reduced binding by the Y239A variant but increased binding by the M1519D variant. **(C)** EMSA of CIC^{min31} WT and C1-domain mutants against the DNA probe shows decreased binding and a migration shift by both C1-domain mutants. **(D)** Corresponding western blots of CIC^{min31} expressed by IVT show similar levels of expression between each mutant. Binding ratios were determined by ImageJ and are relative to the WT of the same protein-to-oligo ratio in their respective gels. Raw intensity values of protein/DNA complexes in EMSA were normalized by the intensity of the protein bands observed on western blots. See also Figure S8.

Table 1. Crystallographic table.

	CIC^{min18} complexed with DNA
Resolution range (Å)	23.45 - 2.95 (3.13 - 2.95) [*]
Space group	*P* 2_1 2_1 2_1
Unit cell	46.892 79.915 106.963 90 90 90
Total reflections	79037(12044)
Unique reflections	8838 (1349)
Multiplicity	8.9 (6.6)
Completeness (%)	98.8 (97.3)
Mean I/sigma(I)	22.44 (3.66)
Wilson B-factor (Å^2)	49.76
R-sym [#]	0.083 (0.506)
*CC*1/2 [&]	0.999 (0.970)
Refinement	
Reflections used in refinement	8769 (2705)
Resolution (Å)	23.84 – 2.95 (3.38 – 2.95)
R-work [$]	0.2093 (0.2471)
R-free [$]	0.2434 (0.3135)
Number of non-hydrogen atoms	1912
macromolecules	1896
metal ion	1
solvent	15
RMS(bonds) (Å)	0.003
RMS(angles) (°)	0.46
Ramachandran favored (%) [†]	98.51
Ramachandran allowed (%)	1.49
Ramachandran outliers (%)	0
Rotamer outliers (%)	0.83
Clashscore	1.15
Average B-factor (Å^2)	67.44
macromolecules	67.42
ligands	92.74
solvent	49.27
Number of TLS groups	*4*

[*] Highest resolution shell is shown in parentheses

[#] $R_{sym} = \Sigma_{hkl}\Sigma_i|I_i(hkl) - \langle I(hkl)\rangle|/\Sigma_{hkl}\Sigma_i I_i(hkl)$.

[&] $CC^* = [2CC_{1/2}/(1+CC_{1/2})]^{1/2}$ where $CC_{1/2}$ is the correlation between two random half datasets containing half of the measured intensities for each unique reflection and CC* approximates the correlation coefficient for a noise-free dataset.

[$] $R_{work} = \Sigma|F_{obs}(hkl) - F_{calc}(hkl)|/\Sigma|F_{obs}(hkl)|$, where $F_{obs}(hkl)$ and $F_{calc}(hkl)$ are the observed and calculated structure factor amplitudes of ~95% of the reflections used for refinement. R_{free} was calculated from the ~5% of total reflections that were omitted from the refinement.

[†] Ramachandran percentages generated using MolProbity.[53]

Table 2. Binding of CIC constructs to 30-mer oligo determined by MST. Purified recombinant proteins were assayed at 20 °C for DNA binding against a 1 nM concentration of annealed 30-mer Cy5-labeled DNA oligonucleotides containing the *ETV5* promoter sequence. See also Figures S7 and S8.

Sample	EC50 (nM) [a]	Hill Coefficient[b]	Relative Binding Strength[c]
CIC^{min18}:WT	**15.5 [14.2 – 16.9]**	**5.00**	**1**
CIC^{min18}:R201W	**10.3 [7.6 – 14.0]**	**0.85**	**1.5**
CIC^{min18}:K232A	**26.3 [22.8 – 30.4]**	**1.80**	**0.6**
CIC^{min18}:Y239A[d]	**55.0 [46.4 – 65.3]**	**1.56**	**0.3**
CIC^{min18}:R1512H	**14.6 [11.9 – 18.0]**	**1.54**	**1.1**
CIC^{min18}:R1515H	**16.7 [12.1 – 23.2]**	**1.10**	**0.9**
CIC^{min18}:M1519D	**50.3 [44.8 – 56.5]**	**2.48**	**0.3**
CIC^{min31}:WT	**226 [160 – 319]**	**1**	**1**
CIC^{min31}:R201W	**171 [145 – 203]**	**1**	**1.3**
CIC^{min31}:R215W[e,f]	**151 [118 – 194]**	**1**	**1.5**
CIC^{min31}:R1515H	**1431 [1184 –1729]**	**1**	**0.1**
CIC^{min31}:M1519D	**32.2 [24.6 – 42.2]**	**1**	**7.0**

[a.] EC50 values calculated using the Hill model are reported with 95% confidence interval in brackets from merged datasets of three independent replicates. [b.] Hill coefficients of the CIC^{min31} construct were fixed to 1 as no cooperativity was observed. [c.] Change in binding strength calculated relative to WT of the respective construct. [d.] The corresponding CIC^{min31}:Y239A variant was insoluble. [e.] The corresponding CIC^{min18}:R215W variant was insoluble. [f.] The increased binding affinity for R215W by MST counters the observed EMSA results in Figure 6, suggesting changes in this variant not captured by equilibrium binding tests with MST. The signal to noise ratios for all data were between 21 and 45.

STAR Methods

Lead contact: Daniel Dowling, Daniel.dowling@umb.edu

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

T7 Express *E. coli* cells were purchased from New England Biolabs (NEB) and were used for plasmid amplification and recombinant protein expression. Bacterial growth conditions can be found in method details.

METHOD DETAILS

Materials

Molecular biology materials were purchased from NEB unless otherwise indicated. Crystallization reagents were purchased from Hampton Research. Other chemicals were purchased from commercial suppliers and used without further purification.

Protein overexpression and purification of wild type CIC^{min18}

A minimal DNA-binding construct (HMG-box-C1, termed CIC^{min18}) containing the HMG-box (D190 to N273) and C1 (K1459 to Q1529) domains of human Capicua (GenBank ID AAK73515.1) was generated with an 18-amino acid linker containing 15 residues flanking helix H3 of the HMG-box, a phenylalanine insertion for increased UV absorption to aid protein purification, and 2 residues that precede H1 of the C1 domain. CIC^{min18} was generated by GenScript as a codon-optimized gene sequence for recombinant protein expression in *E. coli* (Figure S1A). The commercially synthesized DNA product was digested with NdeI and BamHI in CutSmart Buffer and purified using a Monarch PCR and DNA Cleanup Kit. The pET28a vector was similarly digested, treated with alkaline phosphatase, and purified on a 1.0% (w/v) agarose gel. The digested vector was extracted using the Monarch DNA Gel Extraction Kit, and the digested CIC^{min18} DNA and pET28a were ligated together in a 5:1 ratio using T4 DNA ligase, yielding $pET28^{cicmin18}$. Ligated product was used to transform NEB5α cells. A single colony was grown overnight in 5 mL Miller lysogeny (LB-Miller) medium supplemented with 50 µg/mL kanamycin sulfate. DNA was isolated using a Monarch Plasmid Miniprep Kit, and Sanger DNA sequencing was performed by Genewiz.

Chemically competent T7 Express cells were transformed with $pET28^{cicmin18}$ and plated on LB-Miller agar supplemented with 50 µg/mL kanamycin. A starter culture originating from a single colony was used to inoculate 8 × 1 L of autoclaved LB-medium with 50 µg/ml kanamycin and grown at 37 °C while shaking at 250 rpm until an optical density (OD_{600}) of 0.50 was reached. WT culture was induced by adding IPTG to a concentration of 0.1 mM and the cells were grown for 4 h at 37 °C shaking at 250 rpm. Cells were harvested via centrifugation at 4,816 × g for 30 min at 4 °C, and the resulting cell pellet was resuspended in Buffer A (50 mM Tris (pH 8.0), 500 mM sodium chloride, and 10% (v/v) glycerol) and supplemented with an EDTA-free Pierce Protease Inhibitor tablet. The resuspended cells were kept on ice and lysed by sonication using a Branson Digital 250 Sonifier with five cycles of 3 min active sonics (40% duty, output level 7) and 3 min rest. Cell debris was removed by centrifugation at 20,000 × g and 4 °C for 1 h. The supernatant was consecutively filtered through 0.45 µm and 0.22 µm PES syringe filters prior to loading onto a 5 ml HisTrap nickel affinity column (GE Healthcare) preequilibrated with Buffer A. Unbound protein was removed using an isocratic step with 8% Buffer B (50 mM Tris (pH 8.0), 500 mM sodium chloride, 10% (v/v) glycerol, and 1 M imidazole), and bound CIC^{min18} was then eluted with a linear gradient from 8 to 50% Buffer B. Fractions containing CIC^{min18} were collected and dialyzed using a slide-a-lyzer mini dialysis cassette (3500 MWCO) (Pierce) into Buffer A supplemented with 1 mM

DTT. The nickel affinity chromatography step was repeated once more with dialyzed CIC^{min18} prior to purification on a HiLoad Superdex 75 16/600 column (GE Healthcare) equilibrated with 50 mM Tris (pH 8.0), 300 mM sodium chloride, 10% (v/v) glycerol, and 1 mM DTT. The size and purity of the eluted protein was determined by SDS–PAGE, and protein was concentrated using a 3 kDa MWCO Vivaspin centrifugal concentrator (MilliporeSigma). Final protein concentration was assessed using 5,5-dithio-bis-(2-nitrobenzoic acid) (ThermoFisher) and a standard curve of glutathione.[54,55] Protein was aliquoted, flash frozen in liquid nitrogen, and stored at –80 °C.

CIC^{min18} and CIC^{min31} Mutagenesis, Overexpression, and Purification

Point mutations were introduced using designed mutagenic inverse primers (Table S6), from NEBaseChanger, and WT CIC^{min18} DNA isolated from the Monarch Plasmid Miniprep Kit. Mutagenic primers were ordered from ThermoFisher and suspended in distilled water. Q5 PCR was performed using 11.25 ng of WT template in a reaction mixture using Q5 reaction buffer, Q5 GC-rich enhancer buffer, and 1.25 µL of 1.0 µM forward and reverse primers. Mutagenic PCR was run for 35 cycles consisting of a 10 s 98 °C denaturation step, a 30 s annealing step, a 3.5-min 72 °C elongation step, and a final 2 min elongation at 72 °C. PCR products were confirmed via agarose gel electrophoresis and methylated template DNA was digested via DpnI. PCR products were purified via the Monarch PCR & DNA Cleanup Kit. PCR products containing the point mutation were phosphorylated and ligated via reaction with T4 PNK and T4 DNA Ligase in T4 DNA Ligase Buffer. Ligation was performed overnight for at least 16 h and plasmids were used for transformations directly. Incorporations of mutants and intact reading frames were confirmed via Sanger Sequencing (Quintara Biosciences). Methods for the expression and purification of recombinant protein were the same as for WT CIC^{min18}, except overexpression was achieved using 0.5 mM IPTG.

A non-codon optimized CIC^{min31} R201W sequence was obtained from Forés et al.[29] and cloned into a pET32a expression vector containing an in-frame C-terminal 6-His tag. NdeI and BamHI HF (NEB) restriction enzymes were used to create complementary N and C-terminal sticky overhangs in the pET32a vector and insert. Insert containing CIC^{min31} R201W was annealed into pET32a multicloning site and ligated using T4 PNK and T4 DNA Ligase in T4 DNA Ligase Buffer. CIC^{min31} WT was generated via site-directed mutagenesis using mutagenic primers (Table S6). Remaining CIC^{min31} point mutations were then introduced into the CIC^{min31} WT in pET32a using designed mutagenic inverse primers (Table S6) from NEBaseChanger. Overexpression of CIC^{min31} mutants was achieved using 0.5 mM IPTG. Cell lysis and lysate clarification was achieved similar to CIC^{min18} WT protein. Protein was loaded onto a 5 mL HisTrap nickel affinity column (GE Healthcare) preequilibrated with Buffer A, and unbound protein was removed using an isocratic step with 8% Buffer B. Bound CIC^{min31} was then eluted with an isocratic gradient of 40% Buffer B. Fractions containing CIC^{min31} were collected, concentrated using a 3 kDa MWCO Vivaspin centrifugal concentrator (MilliporeSigma),and injected onto a HiPrep 16/60 Sephacryl S-200 HR column (Cytiva) equilibrated with 50 mM Tris (pH 8.0), 300 mM sodium chloride, 10% (v/v) glycerol, and 1 mM DTT. The size and purity of the eluted protein was determined by SDS–PAGE, and protein was concentrated using a 3 kDa MWCO Vivaspin centrifugal concentrator (MilliporeSigma). Protein was aliquoted, flash frozen in liquid nitrogen, and stored at –80 °C.

Circular Dichroism (CD)

CD spectra were measured with a J-1500 circular dichroism spectrophotometer (Jasco Incorporated) with 1 mm pathlength quartz microcuvette, using an HTCD autosampler with samples held at 4 °C. Protein samples were prepared by buffer exchange

using a Biospin P-6 gel column (Bio-Rad) preequilibrated in 50 mM phosphate buffer (pH 8.0), 300 mM $(NH_4)_2SO_4$, and 2% (v/v) glycerol. Samples were diluted to approximately 0.75 mg/mL and loaded via the HTCD autosampler. Spectra were taken at 25 °C using a CD scale of 200 mdeg/0.1 dOD with a D.I.T. of 2 s and scan rate of 50 nm/s. Scans were performed between 260 nm and 180 nm. Buffer-subtracted spectra were plotted using Excel.

CD thermal ramp spectra of CIC^{min31} mutants were collected with 1 mm pathlength quartz cuvettes. Protein samples were diluted to 0.75 mg/mL in PBS solution containing 0.05% Tween 20. The temperature was increased from 4 °C to 90 °C using a 1 °C/min ramp rate with spectra collected every 2 °C steps, monitoring for stability in temperature for a minimum of 1 minute prior to collection. Spectra were collected between 260 nm and 180 nm. Buffer-subtracted spectra were plotted using Excel. Melting temperatures were determined by plotting buffer-subtracted molar ellipticity at 208 nm over 20 °C to 80 °C, fitting a Gompertz sigmoidal function against the data, and finding the maxima by calculating the second derivative of the fitted curve in MATLAB.

Microscale thermophoresis (MST)

MST was performed using the NanoTemper Monolith NT.115 instrument (NanoTemper Technologies) with PicoRed detection channel. Cy5-labeled DNA derived from the ETV5 promoter (5′-Cy5-GGCGTTTTTTATGAATGAAAAACGTCCTCC-3′) and the unlabeled reverse complementary sequence were ordered from IDT Inc. with HPLC purification. Oligomer was annealed by incubating at 95° C for 2 min and then decreasing temperature by 17 °C in 11.5 min steps until a final temperature of 25 °C was reached for 11.5 min. WT and mutant CIC^{min18} were diluted to a maximum concentration of 4 μM using PBS supplemented with 0.05% (v/v) TWEEN 20. A 1:1 dilution series was created for each sample using the dilution buffer, and equivalent volumes of the Cy5-oligo titrant were added to each solution. Solutions were allowed to incubate at RT for 5 min before each solution was loaded into a Monolith NT.115 Premium Capillary (NanoTemper Technologies). Optimized percent MST power was determined automatically for each variant and the mode was used for each triplicate trace. LED power was maintained at a low setting and samples were kept at 20 °C within the instrument. Analysis was performed using a default 5 s before IR-on, IR-on for 30 s, and 5 s after IR-off. Data were analyzed and final figures were generated using MO.Affinity Analysis 3 software (NanoTemper Technologies). Data from 5 s intervals were chosen based on quality of data and fit of binding model within MO.Affinity Analysis 3 software. All 5 s intervals used to generate binding curves were from within the first 15 s of IR exposure.

Electrophoretic mobility shift assay (EMSA), in vitro transcription/translation (IVT), and western blotting

Fluorescent electrophoretic mobility shift assay was performed using the wild type and mutant CIC^{min31} protein fragments synthesized with the TNT T7 Quick Coupled Transcription/Translation system (Promega) according to the manufacturer's instructions (see above for details of CIC^{min31} DNA construct preparation). Oligonucleotides were labeled with IRDye 700 (LI-COR). The sequence with a CIC binding site was derived from the *ETV5* promoter (5′-GGCGTTTTTTATGAATGAAAAACGTCCTCC-3′), and the 5′-GGCGTTTTTTAGTCGCTGCAAACGTCCTCC-3′ sequence was used for testing specificity. Oligonucleotides were diluted in annealing buffer (10 mM Tris-HCl (pH 7.5), 50 mM NaCl, 1 mM EDTA) to the final concentration of 20 pmol/μL. For

annealing, 20 µL of forward and 20 µL of reverse oligonucleotide were mixed in a microcentrifuge tube, placed into a beaker with boiling water, and allowed to cool to room temperature. Annealed oligonucleotides were further diluted to a working concentration of 0.05 pmol/µL. Binding reactions were carried out in a total volume of 20 µL containing 10 mM Tris-HCl (pH 7.5), 50 mM KCl, 3.5 mM DTT, 0.25% Tween 20, 1 µg poly(dI·dC), 0.05 pmol of annealed oligo, and 1 µL or 2 µL of IVT-synthesized CIC^{min31} protein. Reactions were incubated for 30 min at room temperature in the dark. Samples were mixed with 1 µL of 10× Orange loading dye (LI-COR) and loaded on a 4% 0.5× TBE native polyacrylamide gel. The gel was run in the dark for 30 min and scanned with the LI-COR Odyssey imaging system.

For western blot analysis, 5 µL of synthesized CIC protein were added to 20 µL of 4x SDS buffer and heated for 5 min at 95 °C. Samples were loaded on the 12% polyacrylamide SDS gel and transferred onto a PVDF membrane at 100 volts for 90 min at 4 °C. A 6x-His Tag Monoclonal Antibody was used as primary antibody (1:250, ThermoFisher), and goat anti-mouse (1:500, LI-COR) was used as secondary antibody for the western blot.

Crystallization and data collection

To obtain a protein-DNA complex, an 18-mer DNA oligonucleotide containing the ETV5 promoter (5′-GGTTATGAATGAAAAACC-3′) and its reverse complement were purchased from ThermoFisher with HPLC purification. DNA was resuspended in ultrapure water and quantified by absorbance at 260 nm using a NanoDrop 2000c spectrophotometer. CIC^{min18} in complex with the 18-mer DNA oligonucleotide was screened for crystallization conditions using a Crystal Phoenix robot (Art Robbins Instruments) within the MIT crystallization facility. Initial sparse matrix screening that was performed at a near equimolar ratio of protein:DNA yielded DNA only crystals; therefore, an optimized 3:1 ratio of protein:DNA was used to select for crystals of the protein-DNA complex as the protein alone did not crystallize in sparse matrix screening attempts. Using a final concentration of 7.1 mg/mL CIC^{min18}, crystallization conditions were identified using the sitting drop vapor diffusion method with the Kerafast Protein-Nucleic Acid Complex Crystal Screen. Crystals appeared under multiple conditions, and diffraction quality crystals leading to structure determination were obtained in 16% (w/v) PEG 8000, 0.1 M 2-morpholinoethanesulfonic acid (MES) (pH 6.0), 0.1 M $CaCl_2$, and 0.1 M NaCl. Crystals were cryoprotected with the reservoir solution supplemented with final concentrations of 20% (v/v) glycerol and 20% (w/v) PEG 8000 and cryocooled in liquid nitrogen.

Structure determination and refinement

X-ray diffraction data were collected at a wavelength of 1.54178 Å at the MIT crystallization facility on a rotating copper anode X-ray generator (Micromax 007-HF) equipped with a Saturn 944+ detector and cryostream 800 (Oxford Cryosystems). Data were collected with 0.5° oscillations at 100 K. Diffraction intensities were indexed to space group $P2_12_12_1$, integrated, and scaled with the XDS program suite.[56] Data processing details are presented in Table 1.

The structure of the CIC^{min18}–DNA complex was solved by molecular replacement (MR) using chain A of PDB ID 6JRP as a search probe.[34] MR was performed in PHASER[57] within Phenix,[58] yielding a solution with overall LLG and TFZ scores of 445 and 20.8, respectively. Clear electron density for the full HMG-box domain was observed with remaining difference electron density for the 18-mer DNA and the C1 domain. Model building was performed in COOT,[59] and refinement was carried out in

Phenix.[58] Initial refinement steps included simulated annealing with minimization and individual ADP refinement. Subsequent refinement steps included minimization, individual ADP refinement, and translation/libration/screw (TLS) motion refinement with separate TLS groups defining the HMG-box, the C1-domain, and each DNA strand. Waters and a calcium ion were added towards the end of refinement. The model was validated using composite omit maps, and Ramachandran angles were assessed using MolProbity.[53] Final model refinement statistics are presented in Table 1. The final model includes the HMG-box (His33–Pro99) and C1 (Pro118–Ala181) domains of the protein and the entire 18-mer duplex DNA.

The CIC^{min18} linker was generated using the Crosslink Proteins tool in BioLuminate (Schrödinger Inc.). The linker was generated between Lys272 and Pro1459 of the HMG-box and C1 domain, respectively. The sequence between these two residues, from the CIC^{min18} construct, was provided as the crosslinker and a simple *de novo* loop creation was executed using Prime[35] with implicit solvent.

Molecular dynamics

MD simulations were performed through the Desmond suite of BioLuminate 2021.[60-63] Three MD simulations of CIC^{min18} complexed with DNA (PDB ID 7M5W) were performed, using three separately prepared starting models containing minor conformation differences in hydrophobic packing residues in the HMG-box near the HMG-box and C1-domain interface. The structure was preprocessed using the Protein Preparation Wizard tool available in the Biologics suite of BioLuminate (2021)[64] to add hydrogens and optimize hydrogen-bonding. The resulting model was refined to decrease potential energy until the heavy-atom r.m.s.d. reached 0.3 Å, at which point the refinement was stopped. The minimized structure was solvated in SPC water and neutralized by the addition of 19 Na^+ ions. The structure was heated for 2 ps at 300 K and 1.01325 bar using the Nose-Hoover chain thermostat method, Martyna-Tobias-Klein barostat method, and isotropic coupling. After heating, the simulations were conducted for 250 ns using the NPT ensemble and randomized initial velocities. The OPLS4 force field was used for all simulations.[65] For replicate MD simulations, starting models were varied by change of start seed in the BioLuminate software, selecting a different simulation path each time.[66] The first MD simulation ran for a length of 250 ns with 100 ps frames and the second and third for a length of 250 ns and 233 ns with 25 ps frames.

Analysis of trajectories was done using the Simulation Event Analysis (SEA) module within BioLuminate 2021. For r.m.s.d. calculations, protein selections were limited to helices of the HMG-box and C1-domains. The trajectories were fit to the initial frame of the respective trajectory after the initial equilibration of the MD simulation. SEA was used to generate r.m.s.f. and r.m.s.d. analyses for the three trajectories.

Figure generation

Protein figures were generated using PyMOL[35] and electrostatic surface calculations were performed with the APBS plugin.[67] Figures containing DNA nucleobases are labeled based on the purine base (GATC), and buried surface area calculations are performed using the PISA server.[68] MD graphs were plotted using BioLuminate,[35] figures were assembled using Adobe Illustrator, and multiple sequence alignment figures were generated using Geneious Prime 2022.0.1 (https://www.geneious.com).

QUANTIFICATION AND STATISTICAL ANALYSIS

X-ray crystallography data collection and refinement statistics are summarized in Table 1. Software used, statistical tests, and number of replicates are described in the figure legends.

Supplemental Information

Document S1. Figures S1-S8 and supplemental references.

Table S1. Sequences for the minimal CIC constructs, related to Figures 4, S1, S7, and S8

Table S2. Nearest structural homologs to the CIC HMG box domain, related to Figures 2 and S2

Table S3. Alignment of the CIC C1 domain with FF domains, related to Figure 2

Table S4. Nearest structural homologs to the CIC C1 domain, related to Figure 2

Table S5. Cancer-associated CIC mutations, related to Figures 6 and S7 and Table 2

Table S6. Primers used for mutagenesis in this work, related to Star Methods

References:

1. Bettegowda, C., Agrawal, N., Jiao, Y., Sausen, M., Wood, L.D., Hruban, R.H., Rodriguez, F.J., Cahill, D.P., McLendon, R., Riggins, G., et al. (2011). Mutations in CIC and FUBP1 contribute to human oligodendroglioma. Science *333*, 1453-1455. 10.1126/science.1210557.

2. Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of gastric adenocarcinoma. Nature *513*, 202-209. 10.1038/nature13480.

3. Icgc Tcga Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82-93. 10.1038/s41586-020-1969-6.

4. Kim, J.W., Ponce, R.K., and Okimoto, R.A. (2021). Capicua in human cancer. Trends Cancer *7*, 77-86. 10.1016/j.trecan.2020.08.010.

5. Okimoto, R.A., Breitenbuecher, F., Olivas, V.R., Wu, W., Gini, B., Hofree, M., Asthana, S., Hrustanovic, G., Flanagan, J., Tulpule, A., et al. (2017). Inactivation of Capicua drives cancer metastasis. Nat. Genet. *49*, 87-96. 10.1038/ng.3728.

6. Simon-Carrasco, L., Grana, O., Salmon, M., Jacob, H.K.C., Gutierrez, A., Jiménez, G., Drosten, M., and Barbacid, M. (2017). Inactivation of Capicua in adult mice causes T-cell lymphoblastic lymphoma. Genes Dev. *31*, 1456-1468. 10.1101/gad.300244.117.

7. Tan, Q., Brunetti, L., Rousseaux, M.W.C., Lu, H.C., Wan, Y.W., Revelli, J.P., Liu, Z., Goodell, M.A., and Zoghbi, H.Y. (2018). Loss of Capicua alters early T cell development and predisposes mice to T cell lymphoblastic leukemia/lymphoma. Proc. Natl. Acad. Sci. U. S. A. *115*, E1511-E1519. 10.1073/pnas.1716452115.

8. Ajuria, L., Nieva, C., Winkler, C., Kuo, D., Samper, N., José Andreu, M., Helman, A., González-Crespo, S., Paroush, Z., Courey, A.J., and Jiménez, G. (2011). Capicua DNA binding sites are general response elements for RTK signaling in *Drosophila*. Development *138*, 915-924. 10.1242/dev.057729.

9. Dissanayake, K., Toth, R., Blakey, J., Olsson, O., Campbell, D.G., Prescott, A., and Mackintosh, C. (2011). Erk/p90RSK/14-3-3 signalling impacts on expression of PEA3 Ets transcription factors via the transcriptional repressor capicúa. Biochem J. *433*, 515-525. 10.1042/Bj20101562.

10. Jiménez, G., Guichet, A., Ephrussi, A., and Casanova, J. (2000). Relief of gene repression by Torso RTK signaling: role of capicua in *Drosophila* terminal and dorsoventral patterning. Genes Dev. *14*, 224-231.

11. Jiménez, G., Shvartsman, S.Y., and Paroush, Z. (2012). The Capicua repressor--a general sensor of RTK signaling in development and disease. J. Cell Sci. *125*, 1383-1391. 10.1242/jcs.092965.

12. Lee, Y., Fryer, J.D., Kang, H., Crespo-Barreto, J., Bowman, A.B., Gao, Y., Kahle, J.J., Hong, J.S., Kheradmand, F., Orr, H.T., et al. (2011). ATXN1 protein family and CIC regulate extracellular matrix remodeling and lung alveolarization. Dev. Cell *21*, 746-757. 10.1016/j.devcel.2011.08.017.

13. Park, S., Lee, S., Lee, C.G., Park, G.Y., Hong, H., Lee, J.S., Kim, Y.M., Lee, S.B., Hwang, D., Choi, Y.S., et al. (2017). Capicua deficiency induces autoimmunity and promotes follicular helper T cell differentiation via derepression of ETV5. Nat. Commun. *8*, 16037. 10.1038/ncomms16037.

14. Rodríguez-Muñoz, L., Lagares, C., González-Crespo, S., Castel, P., Veraksa, A., and Jiménez, G. (2022). Noncanonical function of Capicua as a growth termination signal in *Drosophila* oogenesis. Proc. Natl. Acad. Sci. U. S. A. *119*, e2123467119. 10.1073/pnas.2123467119.

15. Lam, Y.C., Bowman, A.B., Jafar-Nejad, P., Lim, J., Richman, R., Fryer, J.D., Hyun, E.D., Duvick, L.A., Orr, H.T., Botas, J., and Zoghbi, H.Y. (2006). ATAXIN-1 interacts with the repressor Capicua in its native complex to cause SCA1 neuropathology. Cell *127*, 1335-1347. 10.1016/j.cell.2006.11.038.

16. Papagianni, A., Forés, M., Shao, W., He, S., Koenecke, N., Andreu, M.J., Samper, N., Paroush, Z., González-Crespo, S., Zeitlinger, J., and Jiménez, G. (2018). Capicua controls Toll/IL-1 signaling targets independently of RTK regulation. Proc. Natl. Acad. Sci. U. S. A. *115*, 1807-1812. 10.1073/pnas.1713930115.

17. Astigarraga, S., Grossman, R., Díaz-Delfín, J., Caelles, C., Paroush, Z., and Jiménez, G. (2007). A MAPK docking site is critical for downregulation of Capicua by Torso and EGFR RTK signaling. EMBO J. *26*, 668-677. 10.1038/sj.emboj.7601532.

18. Futran, A.S., Kyin, S., Shvartsman, S.Y., and Link, A.J. (2015). Mapping the binding interface of ERK and transcriptional repressor Capicua using photocrosslinking. Proc. Natl. Acad. Sci. U. S. A. *112*, 8590-8595. 10.1073/pnas.1501373112.

19. Liao, S., Davoli, T., Leng, Y., Li, M.Z., Xu, Q., and Elledge, S.J. (2017). A genetic interaction analysis identifies cancer drivers that modify EGFR dependency. Genes Dev. *31*, 184-196. 10.1101/gad.291948.116.

20. Paul, S., Yang, L., Mattingly, H., Goyal, Y., Shvartsman, S.Y., and Veraksa, A. (2020). Activation-induced substrate engagement in ERK signaling. Mol. Biol. Cell *31*, 235-243. 10.1091/mbc.E19-07-0355.

21. Simón-Carrasco, L., Jiménez, G., Barbacid, M., and Drosten, M. (2018). The Capicua tumor suppressor: a gatekeeper of Ras signaling in development and cancer. Cell Cycle *17*, 702-711. 10.1080/15384101.2018.1450029.

22. Tseng, A.S., Tapon, N., Kanda, H., Cigizoglu, S., Edelmann, L., Pellock, B., White, K., and Hariharan, I.K. (2007). Capicua regulates cell proliferation downstream of the receptor tyrosine kinase/ras signaling pathway. Curr. Biol. *17*, 728-733. 10.1016/j.cub.2007.03.023.

23. Wang, B., Krall, E.B., Aguirre, A.J., Kim, M., Widlund, H.R., Doshi, M.B., Sicinska, E., Sulahian, R., Goodale, A., Cowley, G.S., et al. (2017). ATXN1L, CIC, and ETS transcription factors modulate sensitivity to MAPK pathway inhibition. Cell Rep. *18*, 1543-1557. 10.1016/j.celrep.2017.01.031.

24. Bunda, S., Heir, P., Metcalf, J., Li, A.S.C., Agnihotri, S., Pusch, S., Yasin, M., Li, M., Burrell, K., Mansouri, S., et al. (2019). CIC protein instability contributes to tumorigenesis in glioblastoma. Nat. Commun. *10*, 661. 10.1038/s41467-018-08087-9.

25. Kawamura-Saito, M., Yamazaki, Y., Kaneko, K., Kawaguchi, N., Kanda, H., Mukai, H., Gotoh, T., Motoi, T., Fukayama, M., Aburatani, H., et al. (2006). Fusion between CIC and DUX4 up-regulates PEA3 family genes in Ewing-like sarcomas with t(4;19)(q35;q13) translocation. Hum. Mol. Genet. *15*, 2125-2137. 10.1093/hmg/ddl136.

26. Fryer, J.D., Yu, P., Kang, H., Mandel-Brehm, C., Carter, A.N., Crespo-Barreto, J., Gao, Y., Flora, A., Shaw, C., Orr, H.T., and Zoghbi, H.Y. (2011). Exercise and genetic rescue of SCA1 via the transcriptional repressor Capicua. Science *334*, 690-693. 10.1126/science.1212673.

27. Lu, H.C., Tan, Q., Rousseaux, M.W., Wang, W., Kim, J.Y., Richman, R., Wan, Y.W., Yeh, S.Y., Patel, J.M., Liu, X., et al. (2017). Disruption of the ATXN1-CIC complex causes a spectrum of neurobehavioral phenotypes in mice and humans. Nat. Genet. *49*, 527-536. 10.1038/ng.3808.

28. Štros, M., Launholt, D., and Grasser, K.D. (2007). The HMG-box: a versatile protein domain occurring in a wide variety of DNA-binding proteins. Cell. Mol. Life Sci. *64*, 2590-2606. 10.1007/s00018-007-7162-3.

29. Forés, M., Simón-Carrasco, L., Ajuria, L., Samper, N., González-Crespo, S., Drosten, M., Barbacid, M., and Jiménez, G. (2017). A new mode of DNA binding distinguishes Capicua from other HMG-box factors and explains its mutation patterns in cancer. PLoS Genet. *13*, e1006622. 10.1371/journal.pgen.1006622.

30. Kamachi, Y., and Kondoh, H. (2013). Sox proteins: regulators of cell fate specification and differentiation. Development *140*, 4129-4144. 10.1242/dev.091793.

31. Gleize, V., Alentorn, A., Connen de Kerillis, L., Labussiere, M., Nadaradjane, A.A., Mundwiller, E., Ottolenghi, C., Mangesius, S., Rahimian, A., Ducray, F., et al. (2015). CIC inactivating mutations identify aggressive subset of 1p19q codeleted gliomas. Ann. Neurol. *78*, 355-374. 10.1002/ana.24443.

32. Graham, C., Chilton-MacNeill, S., Zielenska, M., and Somers, G.R. (2012). The CIC-DUX4 fusion transcript is present in a subgroup of pediatric primitive round cell sarcomas. Hum. Pathol. *43*, 180-189. 10.1016/j.humpath.2011.04.023.

33. Keenan, S.E., Blythe, S.A., Marmion, R.A., Djabrayan, N.J., Wieschaus, E.F., and Shvartsman, S.Y. (2020). Rapid dynamics of signal-dependent

transcriptional repression by capicua. Dev. Cell *52*, 794-801 e794. 10.1016/j.devcel.2020.02.004.

34. Lee, H., and Song, J.-J. (2019). The crystal structure of Capicua HMG-box domain complexed with the ETV5-DNA and its implications for Capicua-mediated cancers. FEBS J. *286*, 4951-4963. https://doi.org/10.1111/febs.15008.

35. Schrodinger, L. (2015). The PyMOL Molecular Graphics System, Version 1.8.

36. Blanchet, C., Pasi, M., Zakrzewska, K., and Lavery, R. (2011). CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. Nucleic Acids Res. *39*, W68-W73. 10.1093/nar/gkr316.

37. Vivekanandan, S., Moovarkumudalvan, B., Lescar, J., and Kolatkar, P.R. (2015). Crystal structure of HMG domain of the chondrogenesis master regulator, Sox9 in complex with ChIP-Seq identified DNA element. 10.2210/pdb4S2Q/pdb.

38. Klaus, M., Prokoph, N., Girbig, M., Wang, X., Huang, Y.-H., Srivastava, Y., Hou, L., Narasimhan, K., Kolatkar, P.R., Francois, M., and Jauch, R. (2016). Structure and decoy-mediated inhibition of the SOX18/Prox1-DNA interaction. Nucleic Acids Res. *44*, 3922-3935. 10.1093/nar/gkw130.

39. Holm, L. (2020). DALI and the persistence of protein shape. Protein Sci. *29*, 128-140. 10.1002/pro.3749.

40. Love, J.J., Li, X., Case, D.A., Giese, K., Grosschedl, R., and Wright, P.E. (1995). Structural basis for DNA bending by the architectural transcription factor LEF-1. Nature *376*, 791-795. 10.1038/376791a0.

41. Stott, K., Tang, G.S., Lee, K.B., and Thomas, J.O. (2006). Structure of a complex of tandem HMG boxes and DNA. J. Mol. Biol. *360*, 90-104. 10.1016/j.jmb.2006.04.059.

42. Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M., and Iyer, L.M. (2005). The many faces of the helix-turn-helix domain: transcription regulation and beyond. FEMS Microbiol. Rev. *29*, 231-262. 10.1016/j.femsre.2004.12.008.

43. Gao, Y., Tan, D.S., Girbig, M., Hu, H., Zhou, X., Xie, Q., Yeung, S.W., Lee, K.S., Ho, S.Y., Cojocaru, V., et al. (2024). The emergence of Sox and POU transcription factors predates the origins of animal stem cells. Nat. Commun. *15*, 9868. 10.1038/s41467-024-54152-x.

44. Bonet, R., Ruiz, L., Aragón, E., Martín-Malpartida, P., and Macias, M.J. (2009). NMR structural studies on human p190-A RhoGAPFF1 revealed that domain phosphorylation by the PDGF-receptor α requires its previous unfolding. J. Mol. Biol. *389*, 230-237. 10.1016/j.jmb.2009.04.035.

45. Bedford, M.T., and Leder, P. (1999). The FF domain: a novel motif that often accompanies WW domains. Trends Biochem. Sci. *24*, 264-265. Doi 10.1016/S0968-0004(99)01417-6.

46. Lu, M., Yang, J., Ren, Z., Sabui, S., Espejo, A., Bedford, M.T., Jacobson, R.H., Jeruzalmi, D., McMurray, J.S., and Chen, X. (2009). Crystal structure of the three tandem FF domains of the transcription elongation regulator CA150. J. Mol. Biol. *393*, 397-408. 10.1016/j.jmb.2009.07.086.

47. Allen, M., Friedler, A., Schon, O., and Bycroft, M. (2002). The structure of an FF domain from human HYPA/FBP11. J. Mol. Biol. *323*, 411-416. 10.1016/s0022-2836(02)00968-3.

48. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res. *47*, D941-D947. 10.1093/nar/gky1015.

49. Bunda, S., Heir, P., Li, A.S.C., Mamatjan, Y., Zadeh, G., and Aldape, K. (2020). c-Src phosphorylates and inhibits the function of the CIC tumor suppressor protein. Mol. Cancer Res. *18*, 774-786. 10.1158/1541-7786.MCR-18-1370.

50. Weissmann, S., Cloos, P.A., Sidoli, S., Jensen, O.N., Pollard, S., and Helin, K. (2018). The tumor suppressor CIC directly regulates MAPK pathway genes via histone deacetylation. Cancer Res. *78*, 4114-4125. 10.1158/0008-5472.CAN-18-0342.

51. Park, J., Park, G.Y., Lee, J., Park, J., Kim, S., Kim, E., Park, S.Y., Yoon, J.H., and Lee, Y. (2022). ERK phosphorylation disrupts the intramolecular interaction of capicua to promote cytoplasmic translocation of capicua and tumor growth. Front. Mol. Biosci. *9*, 1030725. 10.3389/fmolb.2022.1030725.

52. Lee, Y. (2020). Regulation and function of capicua in mammals. Exp. Mol. Med. *52*, 531-537. 10.1038/s12276-020-0411-3.

53. Williams, C.J., Headd, J.J., Moriarty, N.W., Prisant, M.G., Videau, L.L., Deis, L.N., Verma, V., Keedy, D.A., Hintze, B.J., Chen, V.B., et al. (2018). MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. *27*, 293-315. 10.1002/pro.3330.

54. Winther, J.R., and Thorpe, C. (2014). Quantification of thiols and disulfides. Biochim. Biophys. Acta *1840*, 838-846. 10.1016/j.bbagen.2013.03.031.

55. Aitken, A., and Learmonth, M. (2003). Quantitation and location of disulfide bonds in proteins. Methods Mol. Biol. *211*, 399-410. 10.1385/1-59259-342-9:399.

56. Kabsch, W. (2010). XDS. Acta Crystallogr. D Biol. Crystallogr. *66*, 125-132. 10.1107/s0907444909047337.

57. McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). Phaser crystallographic software. J. Appl. Crystallogr. *40*, 658-674. 10.1107/S0021889807021206.

58. Liebschner, D., Afonine, P.V., Baker, M.L., Bunkoczi, G., Chen, V.B., Croll, T.I., Hintze, B., Hung, L.W., Jain, S., McCoy, A.J., et al. (2019).

Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr. D Struct. Biol. *75*, 861-877. 10.1107/S2059798319011471.

59. Casañal, A., Lohkamp, B., and Emsley, P. (2020). Current developments in *Coot* for macromolecular model building of electron cryo-microscopy and crystallographic data. Protein Sci. *29*, 1055-1064. 10.1002/pro.3791.

60. Beard, H., Cholleti, A., Pearlman, D., Sherman, W., and Loving, K.A. (2013). Applying physics-based scoring to calculate free energies of binding for single amino acid mutations in protein-protein complexes. PLoS One *8*, e82849. 10.1371/journal.pone.0082849.

61. Salam, N.K., Adzhigirey, M., Sherman, W., and Pearlman, D.A. (2014). Structure-based approach to the prediction of disulfide bonds in proteins. Protein Eng. Des. Sel. *27*, 365-374. 10.1093/protein/gzu017.

62. Zhu, K., Day, T., Warshaviak, D., Murrett, C., Friesner, R., and Pearlman, D. (2014). Antibody structure determination using a combination of homology modeling, energy-based refinement, and loop prediction. Proteins *82*, 1646-1655. 10.1002/prot.24551.

63. Bowers, K.J., Chow, D.E., Xu, H., Dror, R.O., Eastwood, M.P., Gregersen, B.A., Klepeis, J.L., Kolossvary, I., Moraes, M.A., Sacerdoti, F.D., et al. (2006). Scalable algorithms for molecular dynamics simulations on commodity clusters. held in Tampa, Florida, 11-17 Nov. 2006. pp. 43-43.

64. Sastry, G.M., Adzhigirey, M., Day, T., Annabhimoju, R., and Sherman, W. (2013). Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. J. Comput. Aided Mol. Des. *27*, 221-234. 10.1007/s10822-013-9644-8.

65. Lu, C., Wu, C., Ghoreishi, D., Chen, W., Wang, L., Damm, W., Ross, G.A., Dahlgren, M.K., Russell, E., Von Bargen, C.D., et al. (2021). OPLS4: Improving force field accuracy on challenging regimes of chemical space. J. Chem. Theory Comput. *17*, 4291-4300. 10.1021/acs.jctc.1c00302.

66. Knapp, B., Ospina, L., and Deane, C.M. (2018). Avoiding false positive conclusions in molecular simulation: The importance of replicas. J. Chem. Theory Comput. *14*, 6127-6138. 10.1021/acs.jctc.8b00391.

67. Dolinsky, T.J., Nielsen, J.E., McCammon, J.A., and Baker, N.A. (2004). PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucleic Acids Res. *32*, W665-667. 10.1093/nar/gkh381.

68. Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. J. Mol. Biol. *372*, 774-797. 10.1016/j.jmb.2007.05.022.