Nucleic Acids Research, 2020 **1**
doi: 10.1093/nar/gkaa117

Structural adaptation of vertebrate endonuclease G for 5-hydroxymethylcytosine recognition and function

Crystal M. Vander Zanden[1,†], Ryan S. Czarny[1,†], Ethan N. Ho[1,†], Adam B. Robertson[2] and P. Shing Ho[1,*]

[1]Department of Biochemistry & Molecular Biology, Colorado State University, Fort Collins, CO 80523-1870, USA and [2]Department of Molecular Microbiology, Oslo University Hospital, Sognsvannsveien 20, NO-0027 Oslo, Norway

Received November 17, 2019; Revised February 09, 2020; Editorial Decision February 12, 2020; Accepted February 19, 2020

ABSTRACT

Modified DNA bases functionally distinguish the taxonomic forms of life—5-methylcytosine separates prokaryotes from eukaryotes and 5-hydroxymethylcytosine (^{5hm}C) invertebrates from vertebrates. We demonstrate here that mouse endonuclease G (*m*EndoG) shows specificity for both ^{5hm}C and Holliday junctions. The enzyme has higher affinity (>50-fold) for junctions over duplex DNAs. A ^{5hm}C-modification shifts the position of the cut site and increases the rate of DNA cleavage in modified *versus* unmodified junctions. The crystal structure of *m*EndoG shows that a cysteine (Cys69) is positioned to recognize ^{5hm}C through a thiol-hydroxyl hydrogen bond. Although this Cys is conserved from worms to mammals, a two amino acid deletion in the vertebrate relative to the invertebrate sequence unwinds an α-helix, placing the thiol of Cys69 into the *m*EndoG active site. Mutations of Cys69 with alanine or serine show ^{5hm}C-specificity that mirrors the hydrogen bonding potential of the side chain (C–H < S–H < O–H). A second orthogonal DNA binding site identified in the *m*EndoG structure accommodates a second arm of a junction. Thus, the specificity of *m*EndoG for ^{5hm}C and junctions derives from structural adaptations that distinguish the vertebrate from the invertebrate enzyme, thereby thereby supporting a role for ^{5hm}C in recombination processes.

INTRODUCTION

Branch points along the tree of life are defined genetically by the sequence and base modifications of the nucleotides in DNA (1). Patterns of GC-rich transcriptional elements, for example, have been shown to delineate the three domains of life—bacteria from archaea from eukarya (2). Conversely, 5-methylcytosine (^{5m}C, Figure 1) serves as an innate immune system in bacteria to protect genomes against viral invasion, while in eukaryotes this base modification is an inherited epigenetic marker that regulates gene transcription (3). Its function in archaea DNA, however, remains largely uncharacterized (4). The oxidized variant of ^{5m}C, 5-hydroxymethylcytosine (^{5hm}C Figure 1), has recently been recognized as an epigenetic marker in vertebrates, but serves no known genetic function in invertebrate organisms (5). We show here that the endonuclease G enzyme from mouse (*m*EndoG) specifically recognizes ^{5hm}C in the context of the four-stranded Holliday junction, which supports its role in promoting recombination at ^{5hm}C-sites (6). Finally, comparisons of the crystal structures of *m*EndoG to those of its invertebrate orthologs show how an enzyme has structurally adapted to confer function to ^{5hm}C as an epigenetic marker in vertebrate organisms.

^{5hm}C was initially identified as an intermediate along the oxidative demethylation pathway of ^{5m}C (7–12). Recent studies, however, point to a broader range of cellular functions, leading ^{5hm}C to now be considered the 'sixth' nucleotide and as an additional epigenetic signal in the genomes of vertebrates (13). This oxidized base has been associated with the development of the liver, with ^{5hm}C levels being 10-fold higher in the adult compared to the fetal organ (14), while other studies suggest that ^{5hm}C is necessary for stem cell renewal (15,16). Genome-wide mapping of the 'hydroxymethylome' shows that ^{5hm}C accumulates in gene bodies and, to a lesser extent, at gene promoters (17–20). Furthermore, ^{5hm}C has been shown to accumulate at sites of DNA damage and at recombination hotspots (18,21), suggesting a role in facilitating recombination repair of damaged DNA. The function of ^{5hm}C in directing site-specific recombination (6), however, is a recent discovery and, therefore, its mechanistic and structural role in recombination has yet to be resolved in detail.

Endonuclease G (EndoG, Gene ID: 6679647) was initially identified as a mitochondrial endonuclease in mammals (22–24) that is encoded in the nucleus (25), but or-

*To whom correspondence should be addressed. Tel: +1 970 491 0569; Email: Shing.Ho@Colostate.edu
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

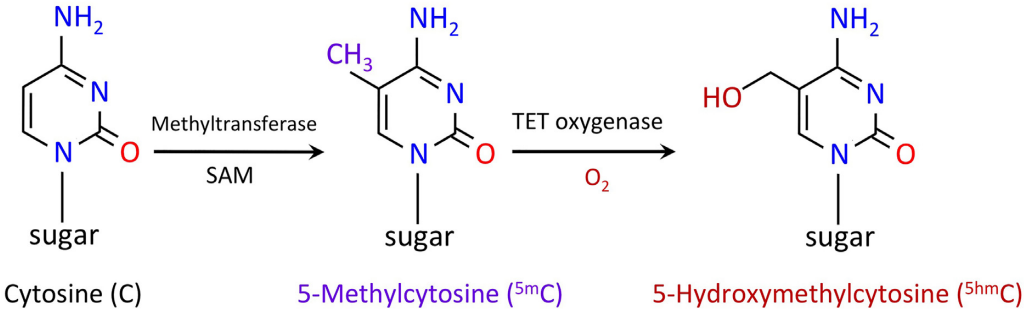

Figure 1. Epigenetic modifications to cytosine (C) bases. Methylation of cytosine at the 5-position by S-adenosylmethionine (SAM) dependent DNA methyltransferase results in 5-methylcytosine (^{5m}C), representing 80% of C's in CpG islands and 3% of all cytosines in eukaryotic genomes (3). Oxidation of ^{5m}C by ten-eleven translocation (TET) methylcytosine dioxygenase results in 5-hydroxymethylcytosine (^{5hm}C) (7–12), found in 0.1–1.3% of C's in vertebrate genomes (5).

thologs have subsequently been found to be present in all eukaryotes (26). The enzyme is characterized by a NHN tripeptide motif, which is common in the active sites of many nucleases (27,28), and has a preference for cleaving G/C-rich DNA. EndoG is primarily found in the mitochondria (25), but a basal level of EndoG is present in the nucleus (29). The enzyme is thought to promote replication of mitochondrial DNA by cleaving RNA/DNA hybrids to generate primers (30), although this function remains debated. During apoptosis, however, large quantities of EndoG are trafficked to the nucleus to degrade chromatin DNA (31,32).

Vertebrate EndoG has been implicated in promoting sequence-mediated recombination (33) and recombination related processes (34–36), while the invertebrate orthologs have no known function in recombination. Robertson, *et al.* (6) demonstrated that the mouse enzyme (*m*EndoG) induces double-strand breaks in ^{5hm}C sequences and promotes conservative recombination events in cells. These studies first showed that mouse liver nuclear extracts (LiNEs) contained an activity that specifically cleaves ^{5hm}C-modified DNA sequences, resulting in the generation of distinct cleavage products. The ^{5hm}C-specific cutting activity in LiNEs was subsequently identified as coming from *m*EndoG, with the purified enzyme having the same specificity for cutting ^{5hm}C-modified over unmodified DNA as the LiNE.

Analysis of the DNA cleavage products indicated that 5′-GGGG***^{5hm}C***CAG-3′ serves as the cognate sequence for LiNE and *m*EndoG (6), leading to the conclusion that *m*EndoG shows both a sequence and a positional preference for cutting ^{5hm}C-modified DNAs. Furthermore, ^{5hm}C-modified DNAs were seen to generate four-times more recombination products than the unmodified DNA in an *in vitro* assay. The preference for modified DNAs was negated when the assay was repeated in the presence of an EndoG specific inhibitor (*f*EndoGI from *Drosophila*) (6). Thus, *m*EndoG was seen to promote ^{5hm}C-dependent recombination.

Robertson, *et al.* (6), proposed a mechanism from their studies in which *m*EndoG promotes ^{5hm}C-specific recombination by recognizing and cleaving ^{5hm}C-modified duplex DNA. The resulting double-strand breaks would progress through a standard recombination mechanism with strand exchange across adjacent duplexes to form Holliday junction cross-overs, culminating in resolution of the junction and ligation to close nicks. The inactive mutant of the enzyme (*i-m*EndoG), however, does not discriminate between ^{5hm}C-modified and unmodified DNA substrates for binding (6), leading to the study here aimed at determining how *m*EndoG shows specificity for ^{5hm}C in the context of this recombination mechanism.

Most 'reader' proteins for ^{5hm}C also recognize ^{5m}C, and bind the unoxidized base as well as or better than the oxidized form (37). An exception is the UHRF2-SRA domain, which shows specificity for ^{5hm}C over ^{5m}C. The UHRF2-SRA domain recognizes ^{5hm}C by flipping the modified base out from the DNA duplex into a specific binding pocket of the protein (38), a mechanism that mimics that of ^{5m}C reader proteins (37). The structures of the *m*EndoG orthologs from *C. elegans* and from *Drosophila* have previously been determined (39,40), but neither show any pockets that could accommodate a flipped-out base, which could explain why these are non-specific nucleases. The question, therefore, is what structural adaptations allow *m*EndoG to recognize ^{5hm}C-modified DNAs while the invertebrate orthologs are nonspecific.

In order to address this question of how *m*EndoG recognizes ^{5hm}C and thereby promotes sequence dependent recombination, we started by determining the binding specificity of the inactive *i-m*EndoG for various DNA conformations and their ^{5hm}C-modified variants. The active *m*EndoG was then used to determine how ^{5hm}C affects enzymatic activity. Results from these studies led to the conclusion that in the context of a DNA Holliday junction, *m*EndoG shows specificity for ^{5hm}C not in binding, but in cleaving efficiency and in defining cleavage patterns along the DNA. The single crystal structures of *i-m*EndoG with and without DNA, in comparison to the *C. elegans* and *Drosophila* orthologs (39,40), reveal the molecular details that allow the vertebrate enzyme to recognize the modified base and the four-armed junction. A short α-helix at the DNA-binding site in the structures of invertebrate enzymes was seen to be unwound in *m*EndoG, which is proposed to result from a two amino acid deletion in the vertebrate enzyme's sequence relative to those of the invertebrates. Consequently, a conserved Cys residue is repositioned into the active site of *m*EndoG to recognize ^{5hm}C through hydrogen bonding be-

tween the Cys thiol and the hydroxyl of the oxidized base. This structural model for ^{5hm}C recognition does not require base flipping and is supported by mutational studies on the Cys residue. Finally, an additional site was found in the crystal structure *i-m*EndoG that provides a platform to accommodate an orthogonal arm from a Holliday junction. Thus, our studies on *m*EndoG reveal structural adaptations that allow this vertebrate enzyme to recognize ^{5hm}C in the context of Holliday junctions, thereby providing insight into how this endonuclease promotes ^{5hm}C-specific recombination.

MATERIALS AND METHODS

Expression and purification of inactive *i-m*EndoG H97A

A catalytically inactive H97A mutant of mouse endonuclease G (*i-m*EndoG) lacking the N-terminal mitochondrial localization sequence was cloned into a pMal-c2 vector as a fusion protein with an N-terminal maltose binding protein (MBP) and tobacco etch virus (TeV) protease substrate linker. *Escherichia coli* BL21-CodonPlus(DE3)-RIPL strain (B F– ompT hsdS(rB– mB–) dcm+ Tetr gal λ(DE3) endA Hte (CamR)) were transformed with the expression construct in 2XYT media containing 100 μg/ml ampicillin and 25 μg/ml chloramphenicol and induced with 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). With the mitochondrial localization sequence deleted from the N-terminus, the numbering of the amino acids in this study is shifted by 41 residues relative to the full-length wild type enzyme. For example, *i-m*EndoG is referred to as H97A, instead of H138A mutant as in the literature (6).

The MBP-tagged fusion *i-m*EndoG was purified from the cleared lysate on an amylose resin column (New England Biolabs), digested with TeV protease, run through a HiTrap Heparin HP column (GE Healthcare), with final separation of the digested *i*-mEndoG from the MBP tag on an amylose resin column. Trace fusion protein was removed by loading the protein solution onto a gravity-fed Sephadex G-100 column equilibrated in buffer containing 50 mM Tris at pH 8.0, 50 mM NaCl, 1 mM $MgCl_2$ and 0.07% volume of β-mercaptoethanol. The fully digested *i-m*EndoG was concentrated with a 10 kDa MWCO Amicon centrifugal concentrator (Millipore-Sigma) and stored at -80°C.

Expression and purification of active *m*EndoG

Active *m*EndoG is a general nuclease and thus lethal in cellular expression systems. Consequently, we designed a system in which *m*EndoG is coexpressed with the EndoG inhibitor (gene *cg4930*) from *Drosophila melanogaster* (*d*EndoGI), following a protocol similar to that developed for the expression of the active fly enzyme (41). A truncated version of the *m*EndoG gene (with the sequence from maltose binding protein appended to the N-terminus in place of the mitochondrial localization sequence) was inserted into a pET-28A plasmid downstream of a T7 promoter site and upstream of the sequence coding for *d*EndoGI. Transformed BL21 codon+ cells were grown in 1 L of 2XYT media with kanamycin (52 mg), chloramphenicol (33 mg), and 1% (w/v) dextrose (for toxic expression), and induced with IPTG. Cell lysates were loaded on an MBPTrap column (GE Health Care Life Sciences), and the *m*EndoG/*d*EndoGI complex eluted with 20 mM maltose. The inhibitor was released from the enzyme by dialysis overnight against 10 mM EDTA to remove Mg^{2+} at the complex interface. A second MBPTrap column followed by 16-hour incubation in 1:100 gram ratio of TeV protease to protein released *m*EndoG from the attached MBP. Finally, a Heparin column was used to bind *m*EndoG and eluted with 0.5 M NaCl to obtain purified enzyme.

Nuclease assays

Synthetic DNA oligos with 5′ dye labels were purchased from Midland Certified Reagent Company, Bio-Synthesis Inc., and LGC Biosearch Technologies, and were purified as previously described (42). DNA duplexes and Holliday junctions were annealed by combining the appropriate single-stranded oligonucleotides, each at 7 μM, in 0.2× TBE, 10 mM $MgCl_2$, 20 mM NaCl. Annealing reactions were incubated at 90°C for 20 min and slowly cooled to room temperature over 2 h. Fully annealed junctions were separated from incompletely assembled duplexes by native polyacrylamide gel electrophoresis (PAGE).

Nuclease assays were performed as follows. Recombinant wild-type *m*EndoG (50 nM in 50 mM Tris pH 8.0, 50 mM NaCl, 1 mM $MgCl_2$, 0.07% β-mercaptoethanol) was incubated with 100 nM DNA substrate in 20 mM Tris pH 7.5, 4% glycerol, 10 mM β-mercaptoethanol, 0.1% Triton X-100. The reaction was run for 10 min at 37°C, then quenched with 0.02% sodium dodecyl sulfate (SDS) and 1.5 μM proteinase K. To further ensure nuclease inactivation after quenching, the reaction solution was incubated at 50°C for 30 min. Negative controls without protein added were prepared similarly. Dye-labeled products were separated by native and denaturing PAGE, and imaged using a Typhoon FLA 9500 gel imager (GE Healthcare). Gel bands were quantified using ImageJ (43).

Electrophoretic mobility shift assays

Samples containing 70 nM of various DNA constructs with Cy5-labeled DNA substrate in their annealing buffer (see Nuclease Assay above) and varying concentrations of catalytically inactive *i-m*EndoG were mixed on ice in a buffer containing 20 mM HEPES pH 7.9, 5 mM EDTA, 10 mM β-mercaptoethanol, 4% glycerol and 1% Triton X-100. Reactions were incubated at 298 K for 15 min. Protein–DNA complexes were resolved with native PAGE and imaged as described above.

FRET kinetic cleavage assays

Assays for *m*EndoG cleavage of fluorescently labeled DNAs using a Förster Resonance Energy Transfer (FRET) assay were carried out using a plate reader (Perkin Elmer Victor 3) held at 37°C with the DNA constructs 7 and 8 (Supplemental Figure S1) labeled with FITC and Cy5. Junctions with and without the ^{5hm}C modification in 50 mM Tris pH 8.0, 50 mM NaCl, 1 mM $MgCl_2$, 0.07% β-mercaptoethanol were diluted to 50 nM with nanopure water. Reactions were initiated by the addition 15 μl of 500 nM protein to 20 μl of

DNA, resulting in final concentrations of 0.22 μM protein and ∼25 nM DNA. The fluorescence signals from both the FRET donor and acceptor (excited at 485 nm, fluorescence detected from 650 to 750 nm) were collected over ∼90 min total time at 2–3 min intervals (delay time from first data point collection for each dataset was ∼10–12 s). The actual DNA concentration was determined from the Cy5 fluorescence after completion of each reaction, and this value was used to determine the concentration corrected FRET signal for each run. The data were imported into KaleidaGraph to determine the pseudo-first order rate constants, and the rate of cleavage was determined from the concentration corrected FRET signal. The assay was repeated in triplicate for each DNA and enzyme construct and the errors for rates were determined as the standard deviations of the means values.

Crystallization of *i-m*EndoG H97A

Crystals of *i-m*EndoG without DNA were grown at 16°C by hanging drop vapor diffusion, with a reservoir solution containing 25% isopropanol, 0.2 M $MgCl_2$, and 0.1 M HEPES pH 7.6. The hanging drop containing pure *i-m*EndoG (93 μM) and G^{5hm}C dinucleotide (750 nM) was diluted in an equal volume of reservoir solution (no dinucleotide was observed in the final structure). The structure of *i-m*EndoG bound to DNA was obtained from co-crystallization with decanucleotide DNA at 16°C in a sitting drop vapor diffusion setup with a reservoir solution containing 0.1 M sodium citrate tribasic dihydrate pH 5.5 and 22% PEG 1000. The sitting drop containing *i-m*EndoG (94 μM) and *d*(CCGGCGCCGG) (178 μM) was diluted in an equal volume of reservoir solution.

X-ray diffraction data collection and structure determination

X-ray diffraction data was collected at the Advanced Light Source (ALS) Beamline 4.2.2 at Berkeley National Laboratory. Diffraction data was reduced, integrated and indexed using XDS (44) (Supplemental Table S2). Structure phasing was performed in PHENIX Phaser (45) by molecular replacement using the X-ray crystal structure of *C. elegans* EndoG homologue, CPS-6 (PDB 35SB), as a starting structure. Refinement was performed iteratively using PHENIX Autobuild and Refine, and manual adjustments to the initial structure were made using Coot (46,47).

Minimization and energy calculations

Energy minimization and equilibration calculations were performed in Amber18 and AmberTools18 (48). A starting structure was constructed in PyMOL (49) by superimposing the crystal structure of *i-m*EndoG H97A onto the structure of the *C. elegans* EndoG homologue CPS-6 bound to single-stranded DNA (ssDNA) sequence 5′-TTTTT-3′ (PDB 5GKP, RMSD = 5.4 Å (40)). The coordinates of the CPS-6 protein were removed, leaving only *m*EndoG and the ssDNA substrate. Starting structures for other ssDNA sequences were generated by mutating the appropriate nucleotide bases in Coot (47). Starting structures were solvated in an octahedral box with TIP3P water, and the system was set to charge neutral with the addition of Mg^{2+} or Cl^- ions. Minimizations were performed in three steps, first allowing only solvent to minimize, then minimizing hydrogen atoms only, and finally minimizing all atoms in the structure. Equilibration calculations were also performed in three steps, first heating the system from 0 to 300 K over 20 ps with weak restraints on atoms in the protein-DNA complex, then maintaining a constant 300 K for 20 ps with the same restraints, and finally maintaining a constant temperature at 300 K for 100 ps with no restraints. Equilibration calculations were performed in triplicate.

RESULTS

The goal of this study was to determine the structural adaptations that support ^{5hm}C recognition in the vertebrate *m*EndoG enzyme, which consequently distinguish it functionally from its invertebrate homologues. In these studies, we first characterized the substrate preference of the enzyme in terms of the binding and recognition preference of the inactive *i-m*EndoG. We then compared the cleavage efficiency of active *m*EndoG on various unmodified and ^{5hm}C-modified DNA substrates (Supplemental Figure S1) as standard duplex or four-stranded Holliday junction structures (Supplemental Figure S2). Finally, crystal structures were determined for *i-m*EndoG in the apo- and DNA-complexed forms, which, when compared to those of the fly and worm homologues, provide the structural rationale for the vertebrate enzyme's recognition and functional specificity.

*i-m*EndoG preferentially binds Holliday junction over duplex DNA

To determine the substrate specificity of *m*EndoG, we first asked what its binding preference is for duplex compared with other forms of DNA, and how ^{5hm}C modification affects this binding specificity. A 20 bp double-stranded DNA substrate (containing the previously reported recognition sequence 5′-GGGGCCAG-3′ (6)) was titrated with the catalytically inactive H97A mutant *i-m*EndoG and the binding affinity for the DNA analyzed by electrophoretic mobility shift assays (EMSA, Supplemental Figure S3a). This EMSA analysis showed that <50% of the DNA substrate was shifted from the duplex band at enzyme concentrations up to 50 μM, leading to an estimated dissociation constant (K_D) of *i-m*EndoG for duplex DNA that is ≥75 μM. The ^{5hm}C-modified duplex substrate showed no observable difference in *i-m*EndoG's affinity for the DNA duplex (Supplemental Figure S3b). These results are not consistent with previous reports that the K_m values are 4-fold higher for ^{5hm}C-modified compared to unmodified DNAs (6). Our binding studies, therefore, indicate that binding affinity does not account for the enzyme's functional specificity for ^{5hm}C-modifications in the context of double-helical DNA, as consistent with that reported previously (6).

Given its activity in promoting recombination (6), we asked whether *m*EndoG has a higher affinity for binding Holliday junctions over duplex DNA, and whether its specificity for ^{5hm}C is manifest in this context. The DNA binding assay repeated with a four-armed junction as the substrate (Figure 2) showed that *i-m*EndoG has a significantly

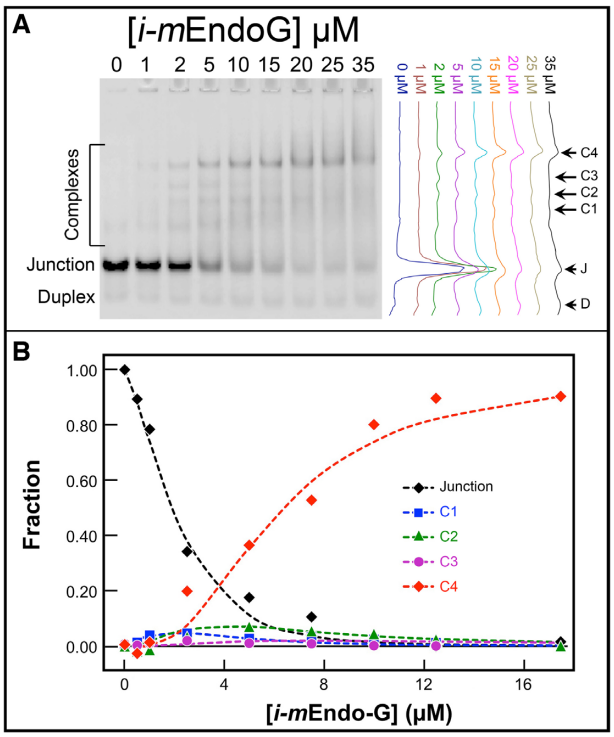

Figure 2. Binding assays of inactive H97A mutant of mouse EndoG (*i-m*EndoG) with unmodified junction DNA. (**A**) Representative gel image from titration of junction DNA with increasing concentrations of *i-m*EndoG, showing electrophoretic mobility shifts that are consistent with the formation of sequentially larger protein-DNA complexes. Scans of each lane of the gel showed four distinct complexes (C1, C2, C3 and C4) emerging with the loss of the junction (J) band. The normalized area under each scan profile was interpreted as indicative of the fraction of DNA distributed between free junction and each complex that entered the gel. (**B**) The normalized fractions of all junction species as a function of *i-m*EndoG concentration was globally fit to a binding model in which the DNA provides four independent sites for protein binding. The calculated curves for the global fit of all data are shown for each species, with K_D values of 10 μM for formation of the C1 complex, 1 μM for C2, 10 μM for C3 and 0.3 μM for C4 (overall R^2 of 0.99 for the global fit of the data). Errors from three independent replicates indicate that the K_D values are accurate to one digit.

higher affinity for this recombination intermediate than for duplex DNAs. The appearance of lower mobility bands in the gels indicated the formation of four successively larger complexes with increasing concentrations of *i-m*EndoG, which we interpreted as one, two, three, and finally four protein molecules binding to the four-armed junction. From a global analysis of the appearance and disappearance of these complexes, we derived a binding model in which the protein shows two distinct sets of affinities for the junction arms, with the first and third binding steps being lower affinity (~10 μM K_D) and the second and fourth steps being higher affinity (~0.3 and 1 μM K_D, respectively). The overall affinity as estimated from analysis of the loss of the junction band resulted in a $K_D = 1.8 \pm 0.6$ μM. The effective overall binding affinity for the equivalent ^{5hm}C-modified junction, however, was not significantly different from the unmodified junction, although it is still ~50-fold higher than that for duplex DNAs.

The affinity of *i-m*EndoG for single-stranded DNAs was low, comparable to that of duplex, while that of three-armed junctions was similar to the four-armed constructs (data not shown). Thus, the enzyme shows preference for binding DNA substrates based on conformation, but this structural preference does not distinguish ^{5hm}C in either the duplex or junction contexts.

^{5hm}C defines cutting efficiency and site specificity by active *m*EndoG

The disparity between the equilibrium K_D values measured here and the K_M values reported previously (6) suggests that ^{5hm}C plays a differential role in the kinetics of *m*EndoG activity. This may be due in part to the prior published studies being performed using enzymes purified from liver nuclear extracts, while the current studies use recombinantly expressed and purified enzyme. Thus, it was important to determine whether the active *m*EndoG enzyme functionally discriminates between unmodified and ^{5hm}C-modified DNAs. To address this question, unmodified or ^{5hm}C-modified duplexes and unmodified or modified four-armed junction substrates were incubated with active *m*EndoG, with the cleavage products resolved by native PAGE analysis. The cleavage efficency for each DNA construct was measured as the loss of fluorescently-labeled substrate bands. The complementary strands of the duplex (DA and DB, Supplemental Figure S1) and the junction (JA and JB strands, Supplemental Figure S1) were also monitored fluorescently in order to determine the sequence as well as conformational dependencies of the cleavage reactions (Figure 3, Supplemental Table S1). By visual inspection, it was clear that the preferred substrate for *m*EndoG cleavage is the junction DNA, which showed significant loss of the 40 bp junction substrate along with concomitant increase in the 20 bp band and some increase in the ~10 bp band as products for all the substrates. It is particularly noteworthy that the primary cleavage products from the junction substrates were the 20 bp duplex DNAs, indicating that the enzyme cuts across the cross-over structure of the junction; this observation is consistent with *m*EndoG potentially acting as a DNA junction resolvase.

The cutting efficiencies were further analyzed by quantifying the fluorescence intensities of the substrate and product PAGE bands and comparing these intensities for each construct in the presence *versus* the absence of the enzyme (Figure 4A, Supplemental Table S1). From this more detailed analysis, we see that the unmodified and modified duplexes are very poor substrates for *m*EndoG, with only 5–10% DNA cut and no discrimination among the two strands. Thus, the enzyme does not appear to show any sequence preference in duplex DNA.

The quantitative analysis confirms that junctions are the preferred substrate for *m*EndoG, with the ^{5hm}C-modified construct being the most efficiently cleaved substrate, at ~10-fold greater efficiency compared to any of the duplex constructs. For the unmodified junction, the enzyme preferentially cuts the strand with the GGGCCAG sequence (JA) over the strand in the absence of this sequence (JB), although it is not entirely clear why that is the case, since we are monitoring the loss of substrate. For the ^{5hm}C-modified

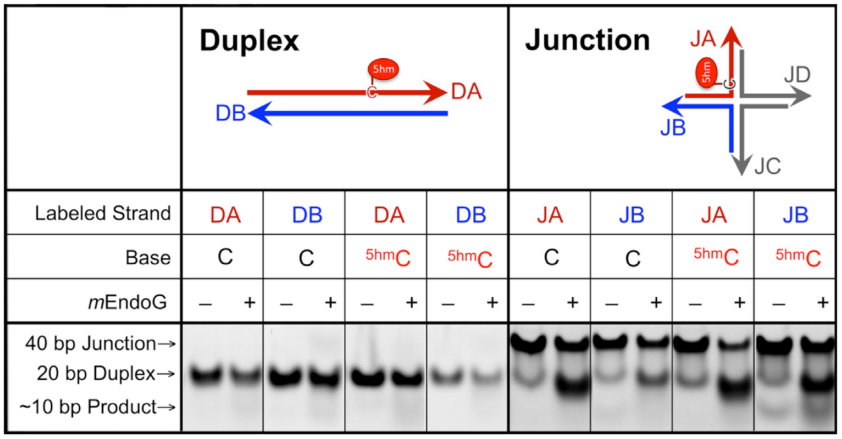

Figure 3. Native polyacrylamide gel electrophoresis (PAGE) analysis of *m*EndoG cleavage of duplex and junction DNA constructs. Representative gel image of a single experiment comparing the cleavage efficiency of *m*EndoG for unmodified and ^{5hm}C-modified duplex or junction DNA substrates. In each construct, only the one strand (DA of the duplex and JA for the junction) is modified, although either the A or B stands (Supplemental Figure S1) are labeled in order to monitor strand cleavage. The efficiency of cutting for any particular construct is quantified by the intensity of substrate band (as the loss of the 20 bp duplex or of the 40 bp junction) and of the smaller cleavage product band(s) in the presence of enzyme (*m*EndoG+) versus its absence (*m*EndoG-). Experiments were repeated four times.

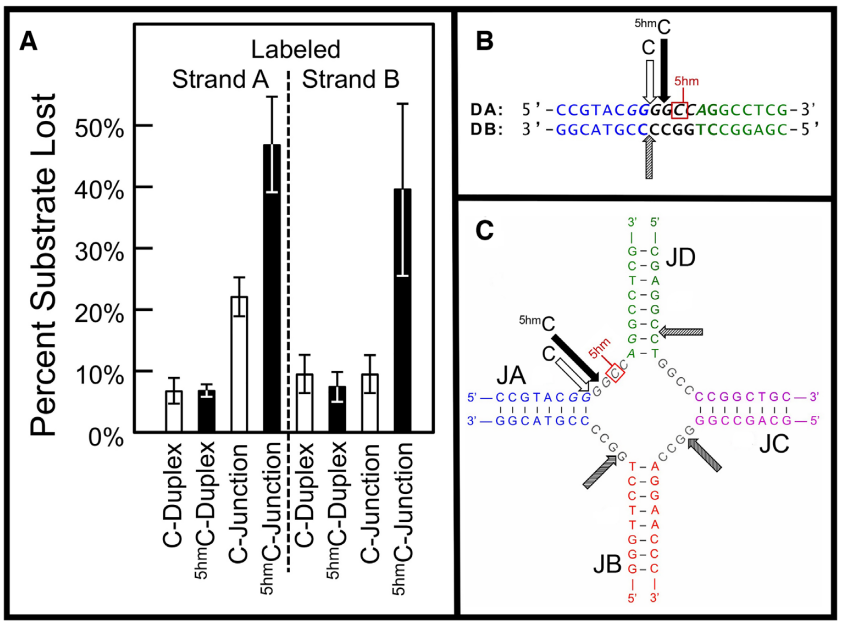

Figure 4. Representative images of gel analysis with efficiencies and cut sites of *m*EndoG cleavage for duplex and junction DNAs. (**A**) Percent of unmodified (open bars) or ^{5hm}C-modified (solid bars) duplex or junction DNAs lost as substrates from cleavage by *m*EndoG, as monitored by fluorescent Cy5 label on strand A (with the *m*EndoG cognate sequence) or strand B (Figure 3, Supplemental Table S1). Percentages in loss of substrates were determined relative to amount of starting material in the absence of protein. Error bars represent standard deviations of the mean from four replicate experiments. (**B**) Major products from cleavage of duplex DNA by *m*EndoG from denaturing gel analysis (Supplemental Figure S5). The major cut sites for the labeled DA strand are indicated for the unmodified duplex construct (open arrow, C) and the ^{5hm}C-modified construct (solid arrow, ^{5hm}C). The DB labeled strand showed the same cut site (grey arrow) for both the unmodified and modified constructs. Complete cutting profiles are shown in Supplemental Figure S6. (**C**) Major cut sites on the junction DNA, labeled as in B. Complete cutting profiles are shown in Supplemental Figure S7.

junction, however, both the JA and JB strands are efficiently cut, suggesting that the modified base also enhances cleavage on the opposing strand of the junction cross-over. Thus, both the junction conformation and the 5hmC modification of the DNA substrate enhance the endonuclease activity of *m*EndoG.

To show that this specificity in cleavage was not an artifact from monitoring single time points, we developed a Förster Resonance Energy Transfer (FRET) assay to determine the rate of cutting over time. In this assay, the unmodified and ^{5hm}C-modified junctions with Cy5 and fluorescein (FITC) chromophore labels (constructs 7 and 8, respectively, Supplemental Figure S1) were incubated with *m*EndoG and the loss of fluorescence signal from the FITC emission resulting from the excitation of the Cy5 FRET donor was monitored over time. The loss of the FRET sig-

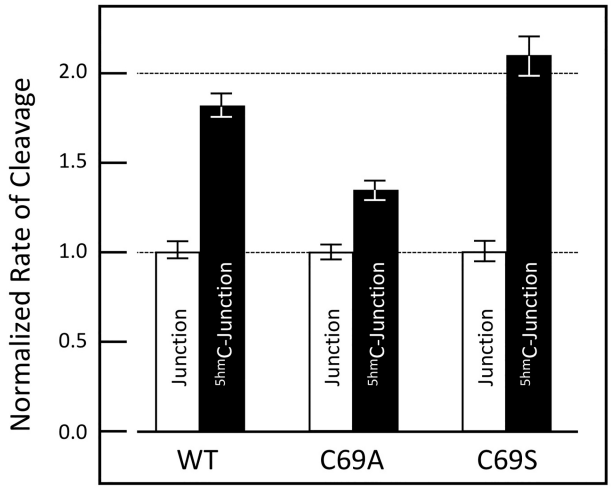

Figure 5. Specificity of *m*EndoG enzymes for cleavage of unmodified and ^{5hm}C-modified junction DNA substrates. Rate of cleavage from FRET time-course assays (Supplemental Figure S4) of unmodified junction (open bars, construct 7, Supplemental Figure S1) or ^{5hm}C-modified junction (solid bars, construct 8, Supplemental Figure S1) are compared for the wild type (WT), Cys69 to Ala mutant (C69A), and the Cys69 to Ser mutant (C69S). Rates were normalized to the rate of cleavage of the unmodified junction for each form of the enzyme. Error bars are shown for the standard deviation of the mean for three experimental replicates for each enzyme and substrate construct.

nal was indicative of cleavage of the junction, which separates the FRET donor on one arm of the junction from the acceptor on an opposing arm. The resulting kinetic traces were fit as pseudo first-order processes, resulting in pseudo first-order rate constants (Supplemental Figure S4). The resulting analyses showed that the rate constant for *m*EndoG cleavage was ∼1.8-fold higher for the ^{5hm}C-modified over the unmodified junction (Figure 5).

The cut sites of the various DNA substrates were resolved at the nucleotide level through denaturing PAGE analysis of the *m*EndoG cleavage products (Figure 4b-c, and Supplemental Figure S5). The various products were quantified, with the fluorescence intensities normalized for total counts across all substrate and product bands in each gel lane (Supplemental Figures S6 and S7). In order to estimate the relative amounts of each product, the normalized intensities of the lane lacking enzyme were subtracted from the companion lane containing enzyme for each DNA construct (we note, however, that these are estimates). For the unmodified duplex substrate, the loss in the 20 nucleotides (nt) DA-strand substrate resulted in appearance of primary cleavage products of 8, 12 and 13 nt in length relative to the 5′-label. Incorporating ^{5hm}C onto the DA strand also resulted in *m*EndoG cleavage fragments of 12 and 13 nt, but the 8 nt product was replaced by primary product that is 9 nt in length (one nt closer to the ^{5hm}C nucleotide). The primary cleavage product for the complementary DB-strand of the duplex is a 12 nt fragment for both the unmodified and ^{5hm}C-modified constructs, with the latter showing a slight increase in cutting efficiency.

Denaturing PAGE analysis of the *m*EndoG cleavage products on the junction substrate shows that the JA strand behaves similarly to the DA strand of the duplex (Supplemental Figure S7). The major cut site is shifted from being 2 nt to 1 nt from the ^{5hm}C-nucleotide and the cutting efficiency increases at that site over the alternative sites. We should note, however, that the denaturing gel analyses provides information only on relative efficiency of cleavage among the various potential cut sites within one strand (as seen with the JA strand), and not absolute cleavage efficiencies along the entire duplex or junction (thus the cleavage efficiency cannot be compared between strands of the junction). Similar to the complementary DB-strand of the duplex, the other three strands of the junction show the same pattern of cleavage products for both the modified and unmodified junction substrates. Thus, the ^{5hm}C-specificity for *m*EndoG is associated with the catalytic function of the enzyme in increasing the efficiency of cutting of the DNA substrate and specifying the point of cleavage 1 nt closer to the ^{5hm}C nucleotide position along the cognate sequence relative to that of the unmodified substrates. When coupled with its significantly higher affinity for junctions over duplex DNAs, the results support a conclusion that *m*EndoG is a ^{5hm}C and junction specific endonuclease.

Structure of *i-m*EndoG and comparisons to homologs

In order to understand the ^{5hm}C-specificity at the molecular level, we determined the single-crystal structure of *i-m*EndoG to 2.1 Å resolution (Figure 6A, Supplemental Table S2). The *i-m*EndoG crystallizes as a dimer and is overall very similar to the previously determined structures of the invertebrate orthologs from *C. elegans* and *Drosophila* (39,40) (Figure 6). The most immediate difference in the mouse structure was that the N-termini are swapped across the two subunits of the dimer, which is not seen in either of the two invertebrate enzymes. Excluding the swapped N-termini, the root-mean-square-deviation (RMSD) of backbone atoms was 0.68 Å (for 341 aligned out of 472 total residues) between *i-m*EndoG and the *C. elegans* ortholog and 0.60 Å (for 357 aligned residues) to *Drosophila*. The RMSD comparing the structures of the two invertebrate enzymes was 0.77 Å.

In comparing the crystal structures in greater detail, we observed both similarities and also significant conformational differences between *i-m*EndoG and the invertebrate proteins at their active sites. An important common feature is that an Mg^{2+} ion is seen coordinated at essentially the same position, coordinated by N128 and E136 in *i-m*EndoG (both amino acids are conserved in all three structures). The position of this catalytically important cation defines the cut site of the DNA backbone within the active site. There were, however, significant differences found in the protein secondary structures that define the active site and, therefore potentially provide specificity for ^{5hm}C binding.

The DNA binding pockets within the active sites of the invertebrate proteins are defined by four distinct α-helices, labeled as α1, α2, α3 and α4 (from the N- to C-termini, Figure 6C). The *C. elegans* protein was crystallized with a T_5 pentanucleotide, which was resolved in the complex (40). With the high similarity among the various protein structures, we applied a least squares alignment of the Cα carbons of all residues to compare the active sites of the mouse and fly structure to that of the DNA-bound worm complex

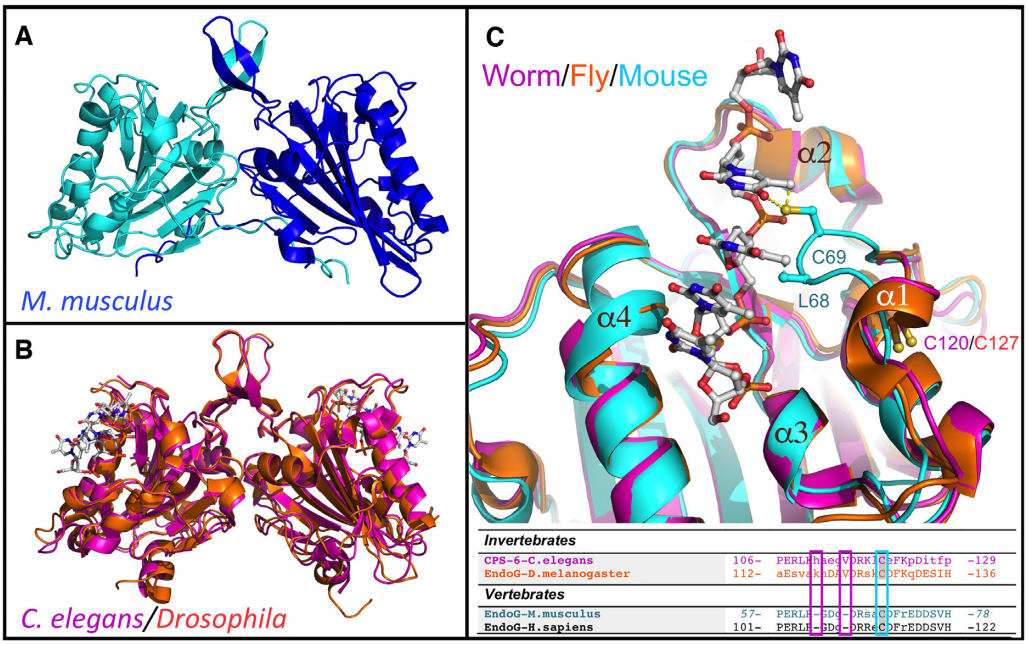

Figure 6. Crystal structures of *i-m*EndoG in comparison to the invertebrate orthologs. (**A**) Structure of mouse *i-m*EndoG in the absence of DNA is a dimer (cyan and blue). (**B**) Superposition of worm (magenta) and fly (orange) orthologs of EndoG (39,40). Superpositions of protein structures in this figure were performed using the ALIGN function of the PyMOL program (59) to overlay the Cα-carbons of the protein backbones for the aligned amino acid sequences. (**C**) Superposition of mouse (cyan), worm (magenta), and fly (orange) structures, focusing on the active site (the A-site for DNA binding). As with **b**, the protein structures from the three organisms were superimposed based on the Cα-carbons of the protein backbones for the aligned amino acid sequences. The α-helices that define the active site are labeled as α1, α2, α3 and α4, from the N- to C-termini. The T_5 pentanucleotide strand from the *C. elegans* structure (ball-and-stick models, with carbons colored white) remained fixed to its position and conformation in the co-crystal structure to show its relationship to the active site of all three proteins. The side chains of the C69 Cys residue and the equivalent C120 and C127 are shown as ball-and-stick models. The amino acid sequences of invertebrate and vertebrate EndoG orthologs are shown for the active site region, with the conserved Cys highlighted by a blue box and the two deleted residues in the vertebrate sequences (relative to the invertebrates) highlighted by magenta boxes.

in order to identify those amino acids that are important for DNA binding and recognition (Figure 6C). The primary protein-DNA contacts within all of the overlaid complexes are between side chains of basic amino acids (primarily Arg) and the phosphate backbone of the DNA, consistent with the lack of specificity of the invertebrate enzymes.

A significant conformational deviation seen in the DNA binding pocket of the *i-m*EndoG structure is that the five-residue α1 helix is unwound into a three-residue loop. The aligned sequences for the vertebrate and invertebrate enzymes suggest that the unwinding of α1 can be attributed to two amino acids being deleted at positions immediately preceding the helix (highlighted by the magenta boxes in Figure 6C). We propose that these deletions reduce the number of residues available to span across this stretch of the vertebrate protein. In order to maintain the conserved conformational features on either side of this region, the *i-m*EndoG structure must unwind the helix of the invertebrate structures. The other secondary structure elements (α2, α3 and α4) remain intact in *i-m*EndoG, with nearly all of the Arg/Lys conserved structurally and available to make contact with the DNA backbone.

C69 confers ^{5hm}C specificity and defines the DNA cut site in *m*EndoG

The consequence of unwinding α1 into a loop in *i-m*EndoG is that C69, a Cys residue that is conserved across animal EndoGs, is repositioned in the mouse enzyme to recognize ^{5hm}C at the major groove surface of the DNA. In both invertebrate structures (39,40), the Cys is positioned with its side chain orientation away from the DNA pocket, but in the *i-m*EndoG structure, it points into the pocket (Figure 6C). In the current structure, this Cys actually forms a disulfide bond to an adjacent protein within the crystal lattice. However, as with other free sulfhydryl groups in protein crystals (50,51), we interpret this interaction as an opportunistic disulfide bond, formed simply because the thiols are available to make the bond when exposed in this conformation and in this particular crystal lattice. We assert, therefore, that the unwinding of the α1 helix is a consequence of the two residue deletion in the sequence, and not from the intermolecular disulfide bond.

With the *i-m*EndoG structure superimposed onto the protein-T_5 complex of *C. elegans*, we see that the –SH of C69 makes contact (within 1.8 Å) with the methyl group and O4 oxygen at the major groove of one thymine base. In addition, L68 of the loop makes contact (within 1.3 Å) with the methyl of an adjacent T base. The putative scissile phosphate (closest to the catalytic Mg^{2+} cation) immediately precedes the L68 contacted nucleotide. Thus, C69 is positioned to recognize ^{5hm}C through an H-bond with the hydroxyl of the modified base and thereby specifies the cleavage site for the modified DNA to be 1 nt upstream of the ^{5hm}C- nucleotide, as defined by the positioning of C69 and the catalytic Mg^{2+}. This structural model is consistent with the de-

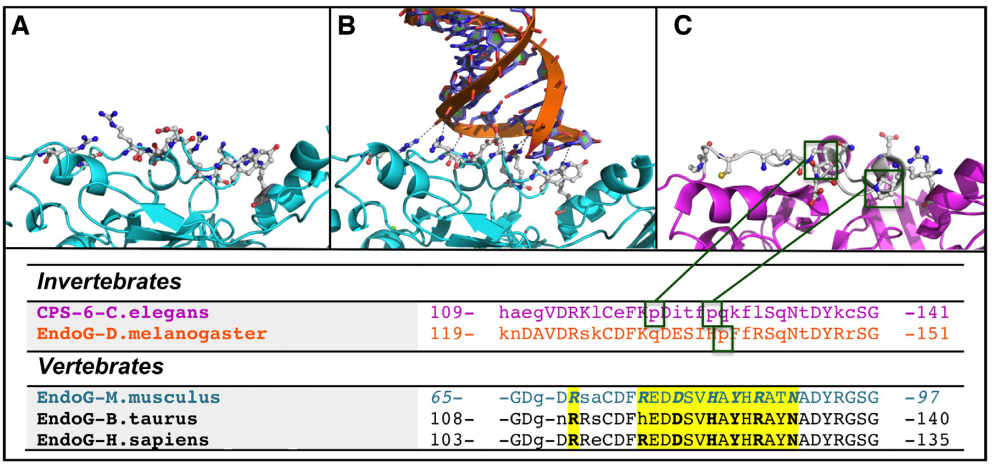

Figure 7. Structures of the orthogonal B-site DNA binding surface. (**A**) The structure of the B-site surface of *i-m*EndoG in the absence of DNA, with residues that interact with the DNA from panel **b** highlighted as ball-and-stick models. (**B**) Structure of *i-m*EndoG from mouse (cyan) in the presence of a decanucleotide DNA sequence. Amino acids that interact (dotted lines) with the DNA are shown as ball-and-stick models. (**C**) Structure of the equivalent surface of the *C. elegans* enzyme (magenta) (40), with the analogous strand highlighted in white and residues as ball-and-stick models. The bottom panel compares the aligned sequences of invertebrate and vertebrate EndoG orthologs, with the sequences of the B-site of the vertebrate enzymes that contact the DNA highlighted in yellow. The Pro residues from the *C. elegans* sequence are highlighted by the green boxes, and mapped to their position in the crystal structure in panel C.

naturing gel analysis, which showed that *m*EndoG cleaved ^{5hm}C-modified DNAs (both duplex and junction) 1 nt upstream of the ^{5hm}C-nucleotide. The absence of the hydroxyl group in unmodified substrates results in the loss of this O–H···S–H H-bond, resulting in the loss of specificity and reduction in efficiency in the cleavage.

To test this thesis, we superimposed the coordinates of the DNA from the *C. elegans* complex into *i-m*EndoG and replaced the bases of the T_5 sequence to construct a GGG^{5hm}CC strand within the DNA pocket (Supplemental Figure S8). The resulting complex, after equilibrium molecular dynamics followed by geometry/energy optimization simulations, shows an O–H···S H-bond from the ^{5hm}C hydroxyl group to C69 of *i-m*EndoG, and an S–H···O H-bond from C69 to the phosphate backbone of the DNA. When the hydroxymethyl substituent is removed (as a GGGCC sequence), the cytosine forms a weaker C–H···S H-bond, although the S–H···O H-bond to the phosphate backbone remains intact. In addition, the C-base becomes unstacked from the remainder of the bases of the DNA strand—a highly unstable conformation. Thus, the thiol of C69 primarily serves as an H-bond acceptor to the hydroxymethyl group, thereby conferring specificity for the ^{5hm}C-modified base, and participates in a weaker interaction as an H-bond donor to the DNA backbone. Finally, an *i-m*EndoG structural model constructed with the unmodified cognate DNA sequence shifted by 1 nt (GGGGC sequence) results an S–H···N H-bond from C69 to the N7 of the guanine base (in position replacing the ^{5hm}C). Overall, these structural models support the observation that *m*EndoG cleaves DNA 1 nt downstream of the ^{5hm}C-modified base.

To provide experimental support that an H-bond from Cys69 confers ^{5hm}C specificity, we replaced this residue with an alanine (a C69A mutant) and measured the rate of cleavage of ^{5hm}C-modified *versus* unmodified junction using the FRET time-course assay, as described above (Figure 5). The C69A mutant showed only a 1.3-fold higher rate for the cleavage of the ^{5hm}C-modified over unmodified junction, as compared to the 1.8-fold difference seen in the wild type enzyme. The same DNA cleavage assay applied to the C69S mutant, where the thiol substituent of the side chain is replaced by a hydroxyl group, showed a 2.1-fold higher rate for cleavage of the modified over unmodified junction. Overall, the ^{5hm}C-specificity follows the expected trend for the H-bonding potential of the amino acid side chain (Ala-C–H << Cys-S–H < Ser-O–H), suggesting that the interaction is a (C/S/O)···O type H-bond, although we do not know which is the donor and which the acceptor in the interaction. Thus, an H-bond from amino acid residue Cys69 was shown to account for the ^{5hm}C specificity for DNA cleavage by *m*EndoG.

Structure of *i-m*EndoG-DNA complex: a second site for DNA binding

In an attempt to determine the structure of *i-m*EndoG in complex with DNA, we crystallized the protein with the self-complementary sequence CCGGCGCCGG (Supplemental Table S3). The X-ray data phased with only the coordinates of the protein showed residual density in the $F_o - F_c$ difference electron density map that is consistent with a DNA duplex bridging across two *i-m*EndoG dimers in the crystal lattice (Supplemental Figure S9). The DNA modeled into the density was a double-helix in the A-DNA conformation (Figure 7), consistent with the conformation that is favored by such GC-rich sequences (52).

The DNA, however, does not sit in the active-site pocket (which we will at this point call the A-site), as seen in the *C. elegans* complex structure (40) and conserved in both the *Drosophila* (39) and mouse structures. Instead, it was bound at a surface (we will call this the B-site) that is adjacent, but geometrically orthogonal to the A-site loop. While

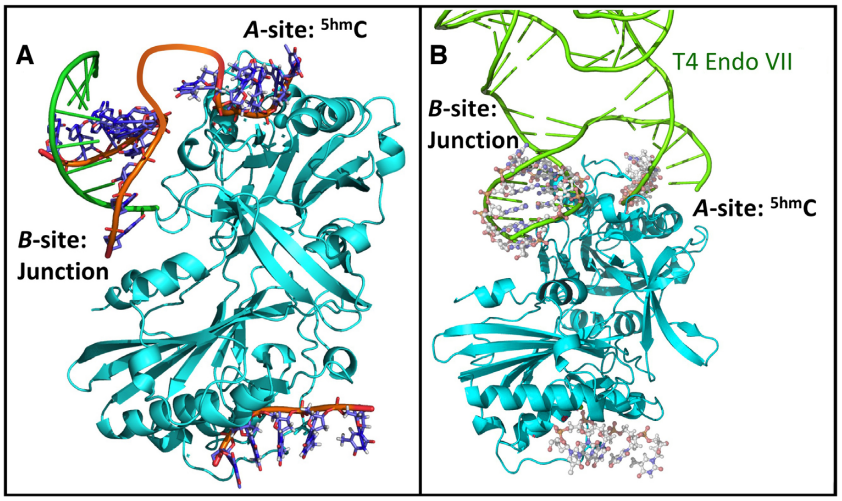

Figure 8. Relationship between the A- and B-sites for DNA binding in *i-m*EndoG. The T_5 DNA from the *C. elegans* crystal structure was superimposed into the A-site of the *i-m*EndoG-DNA complex. (**A**) A simple model connecting the 5′ end of the T_5 strand in the A-site (responsible for ^{5hm}C recognition) to the 3′-end in the B-site (orange backbone, atoms of base pairs in ball-and-stick), providing binding to an additional arm of a DNA junction. The complementary strand of the duplex that does not make contact with the B-site is shown in green, with the bases represented by sticks for clarity. (**B**) Model of DNA junction (green) from the structure of T4 endonuclease VII mapped on the *i-m*EndoG-DNA complex. The phosphodeoxyribose backbone of the T4 Endo VII junction DNA (green ribbon, with bases shown as sticks) (53) was superimposed onto the T_5 DNA (ball-and-stick model) in the A-site and 5 base pairs of the DNA duplex (ball-and-stick model) in the B-site of the complex from panel A, with no additional optimization or modeling.

the amino acid sequence of the B-site surface is conserved in vertebrate EndoGs, it is highly variable in the invertebrates (Figure 7). Similar to the A-site, the interactions of the B-site surface to the DNA are primarily charge-charge and H-bond interactions with the phosphate backbone, with the exceptions of an H-bond from D75 to N4 at the major groove of the C_7 base and a stacking interaction from H78 to the terminal G_{10} base of the DNA duplex (Figure 7B and Supplemental Table S3). This set of interactions suggests that the B-site shows no sequence specificity for binding DNA.

The conformations of the amino acids in the B-site are nearly identical in the structures of the DNA-complexed and apo-enzymes, suggesting that these residues are poised to bind DNA in a non-specific manner. The analogous solvent-facing amino acids of the *C. elegans* and *Drosophila* protein structures (39,40) are not positioned to interact with a DNA fragment, even though there are several basic amino acids that are conserved between the vertebrate and invertebrate sequences. The structural differences that preclude formation of a B-site may be associated with the Pro residues within the invertebrate sequences in this region. The *m*EndoG B-site, therefore, is a surface that is adjacent and geometrically orthogonal to the conserved active A-site and is well-positioned to serve as a second non-specific DNA binding site. The A-site, therefore, recognizes and cuts ^{5hm}C-modified DNA strands, while the B-site is an orthogonal surface that accommodates an adjacent duplex arm of a four-stranded DNA junction. Together, the two sites provide a structural rationale for *m*EndoG's preference for binding and cutting ^{5hm}C-modified junctions.

DISCUSSION

We have shown here that *m*EndoG has functional and structural properties consistent with an endonuclease that is specific for ^{5hm}C in the context of a DNA junction. The enzyme shows significantly higher affinity for four-armed junctions over duplex and single-stranded DNAs. The ^{5hm}C-modified nucleotide, however, does not explicitly increase the binding affinity, but instead increases the efficiency of cutting and specifies the site of cleavage for the modified DNAs (of both duplexes and junctions). An H-bond between the ^{5hm}C-base and the thiol side chain of C69 is proposed to be the structural rationale for ^{5hm}C-recognition by *m*EndoG. In the vertebrate enzyme, deletion of two amino acids from the invertebrate sequence unwinds the α1-helix, resulting in a loop that places the thiol of Cys69 into the DNA binding pocket. The presence of the B-site provides a platform for binding an orthogonal DNA duplex, potentially accounting for *m*EndoG's unique preference for efficiently binding and cutting junctions.

The relationship between the A- and B-sites in conferring junction binding can be appreciated by considering a structural model of the *i-m*EndoG with DNA placed in both DNA-binding sites. The 3′-end of the DNA in the A-site can be readily connected to the 5′-end of one strand from the duplex in the B-site (Figure 8A). The most direct path connecting the DNA strands between the sites requires a bridge that spans a 13.5 Å gap, the equivalent of two nucleotides (each spanning ∼7 Å). The two DNA binding sites can also be bridged using the atomic coordinates of the DNA junction from the T4 endonuclease VII crystal structure (53). In this alternative model complex, the backbone phosphates of the T4 endonuclease VII DNA junction are aligned with phosphates of the single-stranded T_5 DNA from the *C. elegans* structure placed into the A-site and the DNA duplex observed in the B-site of the *i-m*EndoG structures (Figure 8B). This *m*EndoG-junction model complex shows that the two binding sites of the vertebrate enzyme can accommodate a structurally characterized DNA junction. It should be noted, however, that this model has not been evaluated in

terms of energetic or structural feasibility and should, therefore, be considered in that context.

We are now left with the question of what role the ^{5hm}C-specific junction endonuclease activity of vertebrate EndoG plays in the promotion of homologous recombination. Robertson, *et al.* (6) originally proposed a mechanism for how ^{5hm}C promotes recombination, in which this modified nucleotide determines where and how a double-strand break is created in a DNA duplex to initiate the strand invasion. Such a mechanism is consistent with the observation that ^{5hm}C specifies where the DNA backbone is cleaved relative to the modified nucleotide. This mechanism, however, is not consistent with the observations from this study that the Holliday junction is the preferred substrate for both binding and cleavage efficiency by *m*EndoG.

Alternatively, the preference of *m*EndoG's for cutting Holliday junctions suggests that this enzyme is a junction resolvase. The observation that an approximate 20 bp duplex is the major product from cleavage of 40 bp junctions shows that *m*EndoG is capable of cutting symmetrically across a Holliday junction (Figure 3). Despite the apparent symmetric cuts on the junction observed in native gel analyses, denaturing gel analyses showed that *m*EndoG cutting efficiency is enhanced at the ^{5hm}C modified strand, but not necessarily the opposing strand across the junction. We note that a true resolvase would have equal cutting activity on symmetric strands of the junctions. The current studies do have not determined whether incorporating ^{5hm}C nucleotides symmetrically across opposing arms would result in symmetrically cleaved junctions. In short, further work is required to evoke *m*EndoG's role as a resolvase.

To accommodate *m*EndoG's possible role as a junction resolvase, we propose here an alternative mechanism where a non-specific double-stranded break is created by an endonuclease (not necessarily *m*EndoG). The role of ^{5hm}C would come later in this mechanism, by first pausing the junction at the modified nucleotide, followed by ^{5hm}C-specific resolution of the junction by *m*EndoG. Such a mechanism takes advantage of the stabilization of Holliday junctions by ^{5hm}C (54) and the preference for ^{5hm}C-modified junctions as a substrate for *m*EndoG.

An alternative interpretation is that the junction specificity of *m*EndoG is not associated with mechanisms of recombination, but that the four-armed DNA structure mimics the duplex DNAs exiting the eukaryotic nucleosome structure (55) or duplex crossings of supercoiled DNAs (56). Thus, this enzyme may have the potential to recognize ^{5hm}C-flagged DNA damage sites within the distinctive structural features either of chromosomal DNA in the nucleus or of the supercoiled genome in the mitochondrion.

The ^{5hm}C-specific junction endonuclease function of EndoG is unique to vertebrate organisms. The study here shows that the structural adaptations required to confer recognition for both the modified base and the recombination intermediate requires only minimal perturbations to the sequence of the invertebrate homologues—a two amino acid deletion to confer ^{5hm}C-recognition and replacement of one or more prolines to establish an additional DNA binding platform for a second arm of a four-armed junction. There obviously are other substitutions that are required to evolve the vertebrate functions to this enzyme. Many of the structural components for these various functions are already in place, including the ability to associate with FEN-1 and its homologs to target damaged or perturbed DNAs (57,58). However, the results from studies of the C69A and C69S mutants indicate that the modified base is recognized through an H-bond in the active site.

With the C69 H-bond conferring ^{5hm}C recognition in *m*EndoG, the question is why this protein has retained the Cys from the invertebrate enzyme, when hydroxyl groups are expected to form stronger H-bonds than thiols. One simple answer is that there was no need to replace this Cys, since the thiol provides all of the specificity required for *m*EndoG function. An alternative answer, is that the Cys provides redox control over the recognition of this oxidation dependent DNA modification—a mechanism that is not available to a serine at this position.

The crystal structure of *m*EndoG is a dimer, likely stabilized by a domain swap, that places two symmetric active sites on opposite faces of the dimer. This observation raises the question of whether there is a function for having symmetric but opposing cleavage sites. Perhaps this structural feature is associated with *m*EndoG's apoptotic function where the enzyme is mass trafficked to the nucleus to degrade nuclear DNA, with the opposing binding sites enhancing the enzyme's ability to rapidly cut supercoiled DNA. There remains the possibility that the dimeric structure may be involved in *m*EndoG's function in recombination.

Overall, this study shows how *m*EndoG has structurally deviated from the non-specific endonucleases found in invertebrates into one that shows specificity for modifications in both DNA bases and conformation. Through a set of apparently minor amino acid deletions near the active site pocket and substitutions at an adjacent surface, a vertebrate enzyme becomes distinguished from its invertebrate brethren and confers a functional distinction for ^{5hm}C as a vertebrate specific epigenetic marker.

DATA AVAILABILITY

X-ray data and models have been deposited to the PDB as PDB ID 6NJT and 6JNU.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Science Foundation [MCB-1515521, CBET-1905328 to P.S.H.]; National Institute of General Medical Sciences of the National Institutes of Health [F31GM113580 to C.M.V.Z.]; South-Eastern Norway Regional Health Authority [Helse Sør Øst, Project 2014017 to A.B.R.]; A.B.R. was supported in part as a member of Professor Arne Klungland's research group at the Oslo University Hospital. This material is based on work while P.S.H. was serving at the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Funding for open access charge: NSF.

Conflict of interest statement. None declared.

REFERENCES

1. Hug,L.A., Baker,B.J., Anantharaman,K., Brown,C.T., Probst,A.J., Castelle,C.J., Butterfield,C.N., Hernsdorf,A.W., Amano,Y., Ise,K. *et al.* (2016) A new view of the tree of life. *Nat. Microbiol.*, **1**, 16048.
2. Khuu,P., Sandor,M., DeYoung,J. and Ho,P.S. (2007) Phylogenomic analysis of the emergence of GC-rich transcription elements. *Proc. Natl Acad. Sci. U.S.A.*, **104**, 16528–16533.
3. Mulligan,C.J. (2018) Insights from epigenetic studies on human health and evolution. *Curr. Opin. Genet. Dev.*, **53**, 36–42.
4. Couturier,M. and Lindas,A.C. (2018) The DNA methylome of the hyperthermoacidophilic crenarchaeon sulfolobus acidocaldarius. *Front. Microbiol.*, **9**, 137.
5. Huber,S.M., van Delft,P., Mendil,L., Bachman,M., Smollett,K., Werner,F., Miska,E.A. and Balasubramanian,S. (2015) Formation and abundance of 5-hydroxymethylcytosine in RNA. *ChemBioChem*, **16**, 752–755.
6. Robertson,A.B., Robertson,J., Fusser,M. and Klungland,A. (2014) Endonuclease G preferentially cleaves 5-hydroxymethylcytosine-modified DNA creating a substrate for recombination. *Nucleic Acids Res.*, **42**, 13280–13293.
7. Cortellino,S., Xu,J.F., Sannai,M., Moore,R., Caretti,E., Cigliano,A., Le Coz,M., Devarajan,K., Wessels,A., Soprano,D. *et al.* (2011) Thymine DNA glycosylase is essential for active DNA demethylation by linked deamination-base excision repair. *Cell*, **146**, 67–79.
8. Guo,J.U., Su,Y.J., Zhong,C., Ming,G.L. and Song,H.J. (2011) Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. *Cell*, **145**, 423–434.
9. Ito,S., Shen,L., Dai,Q., Wu,S.C., Collins,L.B., Swenberg,J.A., He,C. and Zhang,Y. (2011) Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science*, **333**, 1300–1303.
10. He,Y.F., Li,B.Z., Li,Z., Liu,P., Wang,Y., Tang,Q.Y., Ding,J.P., Jia,Y.Y., Chen,Z.C., Li,L. *et al.* (2011) Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science*, **333**, 1303–1307.
11. Maiti,A. and Drohat,A.C. (2011) Thymine DNA glycosylase can rapidly excise 5-Formylcytosine and 5-carboxylcytosine potential implications for active demethylation of CpG sites. *J. Biol. Chem.*, **286**, 35334–35338.
12. Schiesser,S., Hackner,B., Pfaffeneder,T., Muller,M., Hagemeier,C., Truss,M. and Carell,T. (2012) Mechanism and stem-cell activity of 5-carboxycytosine decarboxylation determined by isotope tracing. *Angew Chem Int Edit*, **51**, 6516–6520.
13. Rusk,N. (2012) The sixth base and counting. *Nat. Methods*, **9**, 646–646.
14. Ivanov,M., Kals,M., Kacevska,M., Barragan,I., Kasuga,K., Rane,A., Metspalu,A., Milani,L. and Ingelman-Sundberg,M. (2013) Ontogeny, distribution and potential roles of 5-hydroxymethylcytosine in human liver function. *Genome Biol.*, **14**, R83.
15. Koh,K.P., Yabuuchi,A., Rao,S., Huang,Y., Cunniff,K., Nardone,J., Laiho,A., Tahiliani,M., Sommer,C.A., Mostoslavsky,G. *et al.* (2011) Tet1 and Tet2 regulate 5-hydroxymethylcytosine production and cell lineage specification in mouse embryonic stem cells. *Cell Stem Cell*, **8**, 200–213.
16. Ficz,G., Branco,M.R., Seisenberger,S., Santos,F., Krueger,F., Hore,T.A., Marques,C.J., Andrews,S. and Reik,W. (2011) Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature*, **473**, 398–U589.
17. Song,C.X., Szulwach,K.E., Fu,Y., Dai,Q., Yi,C.Q., Li,X.K., Li,Y.J., Chen,C.H., Zhang,W., Jian,X. *et al.* (2011) Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat. Biotechnol.*, **29**, 68–72.
18. Stroud,H., Feng,S.H., Kinney,S.M., Pradhan,S. and Jacobsen,S.E. (2011) 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol.*, **12**, R54.
19. Su,Z.X., Han,L. and Zhao,Z.M. (2011) Conservation and divergence of DNA methylation in eukaryotes New insights from single base-resolution DNA methylomes. *Epigenetics-US*, **6**, 134–140.
20. Wu,H., D'Alessio,A.C., Ito,S., Wang,Z.B., Cui,K.R., Zhao,K.J., Sun,Y.E. and Zhang,Y. (2011) Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. *Gene Dev.*, **25**, 679–684.
21. Kafer,G.R., Li,X., Horii,T., Suetake,I., Tajima,S., Hatada,I. and Carlton,P.M. (2016) 5-Hydroxymethylcytosine marks sites of DNA damage and promotes genome stability. *Cell Rep.*, **14**, 1283–1292.
22. Cote,J., Renaud,J. and Ruiz-Carrillo,A. (1989) Recognition of (dG)n.(dC)n sequences by endonuclease G. Characterization of the calf thymus nuclease. *J. Biol. Chem.*, **264**, 3301–3310.
23. Cummings,O.W., King,T.C., Holden,J.A. and Low,R.L. (1987) Purification and characterization of the potent endonuclease in extracts of bovine heart mitochondria. *J. Biol. Chem.*, **262**, 2005–2015.
24. Ruiz-Carrillo,A. and Renaud,J. (1987) Endonuclease G: a (dG)n X (dC)n-specific DNase from higher eukaryotes. *EMBO J.*, **6**, 401–407.
25. Low,R.L. (2003) Mitochondrial Endonuclease G function in apoptosis and mtDNA metabolism: a historical perspective. *Mitochondrion*, **2**, 225–236.
26. Schafer,P., Scholz,S.R., Gimadutdinow,O., Cymerman,I.A., Bujnicki,J.M., Ruiz-Carrillo,A., Pingoud,A. and Meiss,G. (2004) Structural and functional characterization of mitochondrial EndoG, a sugar non-specific nuclease which plays an important role during apoptosis. *J. Mol. Biol.*, **338**, 217–228.
27. Wu,S.L., Li,C.C., Chen,J.C., Chen,Y.J., Lin,C.T., Ho,T.Y. and Hsiang,C.Y. (2009) Mutagenesis identifies the critical amino acid residues of human endonuclease G involved in catalysis, magnesium coordination, and substrate specificity. *J. Biomed. Sci.*, **16**, 6.
28. Kieper,J., Lauber,C., Gimadutdinow,O., Urbanska,A., Cymerman,I., Ghosh,M., Szczesny,B. and Meiss,G. (2010) Production and characterization of recombinant protein preparations of Endonuclease G-homologs from yeast, C. elegans and humans. *Protein Expr. Purif.*, **73**, 99–106.
29. Gerschenson,M., Houmiel,K.L. and Low,R.L. (1995) Endonuclease G from mammalian nuclei is identical to the major endonuclease of mitochondria. *Nucleic Acids Res.*, **23**, 88–97.
30. Cote,J. and Ruiz-Carrillo,A. (1993) Primers for mitochondrial DNA replication generated by endonuclease G. *Science*, **261**, 765–769.
31. Parrish,J., Li,L., Klotz,K., Ledwich,D., Wang,X. and Xue,D. (2001) Mitochondrial endonuclease G is important for apoptosis in C. elegans. *Nature*, **412**, 90–94.
32. Widlak,P., Li,L.Y., Wang,X. and Garrard,W.T. (2001) Action of recombinant human apoptotic endonuclease G on naked DNA and chromatin substrates: cooperation with exonuclease and DNase I. *J. Biol. Chem.*, **276**, 48404–48409.
33. Huang,K.J., Ku,C.C. and Lehman,I.R. (2006) Endonuclease G: A role for the enzyme in recombination and cellular proliferation. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 8995–9000.
34. Gole,B., Baumann,C., Mian,E., Ireno,C.I. and Wiesmuller,L. (2015) Endonuclease G initiates DNA rearrangements at the MLL breakpoint cluster upon replication stress. *Oncogene*, **34**, 3391–3401.
35. Misic,V., El-Mogy,M., Geng,S. and Haj-Ahmad,Y. (2016) [Effect of endonuclease G depletion on plasmid DNA uptake and levels of homologous recombination in hela cells]. *Mol. Biol. (Mosk.)*, **50**, 291–301.
36. Zan,H., Zhang,J., Al-Qahtani,A., Pone,E.J., White,C.A., Lee,D., Yel,L., Mai,T. and Casali,P. (2011) Endonuclease G plays a role in immunoglobulin class switch DNA recombination by introducing double-strand breaks in switch regions. *Mol. Immunol.*, **48**, 610–622.
37. Frauer,C., Hoffmann,T., Bultmann,S., Casa,V., Cardoso,M.C., Antes,I. and Leonhardt,H. (2011) Recognition of 5-hydroxymethylcytosine by the Uhrf1 SRA domain. *PLoS One*, **6**, e21306.
38. Zhou,T., Xiong,J., Wang,M., Yang,N., Wong,J., Zhu,B. and Xu,R.M. (2014) Structural basis for hydroxymethylcytosine recognition by the SRA domain of UHRF2. *Mol. Cell*, **54**, 879–886.
39. Loll,B., Gebhardt,M., Wahle,E. and Meinhart,A. (2009) Crystal structure of the EndoG/EndoGI complex: mechanism of EndoG inhibition. *Nucleic Acids Res.*, **37**, 7312–7320.
40. Lin,J.L., Wu,C.C., Yang,W.Z. and Yuan,H.S. (2016) Crystal structure of endonuclease G in complex with DNA reveals how it nonspecifically degrades DNA as a homodimer. *Nucleic Acids Res.*, **44**, 10480–10490.
41. Temme,C., Weissbach,R., Lilie,H., Wilson,C., Meinhart,A., Meyer,S., Golbik,R., Schierhorn,A. and Wahle,E. (2009) The

drosophila melanogaster gene cg4930 encodes a high affinity inhibitor for endonuclease G. *J. Biol. Chem.*, **284**, 8337–8348.
42. Eichman,B.F., Vargason,J.M., Mooers,B.H.M. and Ho,P.S. (2000) The Holliday junction in an inverted repeat sequence: sequence effects on the structure of four-way junctions. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 3971–3976.
43. Rueden,C.T., Schindelin,J., Hiner,M.C., DeZonia,B.E., Walter,A.E., Arena,E.T. and Eliceiri,K.W. (2017) ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics*, **18**, 529.
44. Kabsch,W. (2010) Xds. *Acta Crystallogr. D*, **66**, 125–132.
45. Adams,P.D., Afonine,P.V., Bunkoczi,G., Chen,V.B., Davis,I.W., Echols,N., Headd,J.J., Hung,L.W., Kapral,G.J., Grosse-Kunstleve,R.W. *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D*, **66**, 213–221.
46. Winn,M.D., Ballard,C.C., Cowtan,K.D., Dodson,E.J., Emsley,P., Evans,P.R., Keegan,R.M., Krissinel,E.B., Leslie,A.G.W., McCoy,A. *et al.* (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr. D*, **67**, 235–242.
47. Nicholls,R.A. (2017) Ligand fitting with CCP4. *Acta Crystallogr. D Struct. Biol.*, **73**, 158–170.
48. Lee,T.S., Cerutti,D.S., Mermelstein,D., Lin,C., LeGrand,S., Giese,T.J., Roitberg,A., Case,D.A., Walker,R.C. and York,D.M. (2018) GPU-Accelerated molecular dynamics and free energy methods in Amber18: Performance enhancements and new features. *J. Chem. Inf. Model.*, **58**, 2043–2050.
49. Janson,G., Zhang,C., Prado,M.G. and Paiardini,A. (2017) PyMod 2.0: improvements in protein sequence-structure analysis and homology modeling within PyMOL. *Bioinformatics*, **33**, 444–446.
50. Evrard,C., Capron,A., Marchand,C., Clippe,A., Wattiez,R., Soumillion,P., Knoops,B. and Declercq,J.P. (2004) Crystal structure of a dimeric oxidized form of human peroxiredoxin 5. *J. Mol. Biol.*, **337**, 1079–1090.
51. Hall,A., Sankaran,B., Poole,L.B. and Karplus,P.A. (2009) Structural changes common to catalysis in the Tpx peroxiredoxin subfamily. *J. Mol. Biol.*, **393**, 867–881.
52. Hays,F.A., Teegarden,A., Jones,Z.J., Harms,M., Raup,D., Watson,J., Cavaliere,E. and Ho,P.S. (2005) How sequence defines structure: a crystallographic map of DNA structure and conformation. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 7157–7162.
53. Biertumpfel,C., Yang,W. and Suck,D. (2007) Crystal structure of T4 endonuclease VII resolving a Holliday junction. *Nature*, **449**, 616–620.
54. Vander Zanden,C.M., Rowe,R.K., Broad,A.J., Robertson,A.B. and Ho,P.S. (2016) Effect of hydroxymethylcytosine on the structure and stability of holliday junctions. *Biochemistry*, **55**, 5781–5789.
55. Hill,D.A. and Reeves,R. (1997) Competition between HMG-I(Y), HMG-1 and histone H1 on four-way junction DNA. *Nucleic Acids Res.*, **25**, 3523–3531.
56. Wendorff,T.J. and Berger,J.M. (2018) Topoisomerase VI senses and exploits both DNA crossings and bends to facilitate strand passage. *eLife*, **7**, e31724.
57. BoseDasgupta,S., Das,B.B., Sengupta,S., Ganguly,A., Roy,A., Dey,S., Tripathi,G., Dinda,B. and Majumder,H.K. (2008) The caspase-independent algorithm of programmed cell death in Leishmania induced by baicalein: the role of LdEndoG, LdFEN-1 and LdTatD as a DNA 'degradesome'. *Cell Death Differ.*, **15**, 1629–1640.
58. Parrish,J.Z., Yang,C., Shen,B. and Xue,D. (2003) CRN-1, a Caenorhabditis elegans FEN-1 homologue, cooperates with CPS-6/EndoG to promote apoptotic DNA degradation. *EMBO J.*, **22**, 3451–3460.
59. DeLano,W. (2002) Pymol: an open-source molecular graphics tool. *CCP4 Newslett. Protein Crystallogr.*, **40**, 82–92.