DISSERTATION

5-HYDROXYMETHYLCYTOSINE AND ENDONUCLEASE G AS REGULATORS OF HOMOLOGOUS RECOMBINATION

Submitted by

Crystal M. Vander Zanden

Department of Biochemistry and Molecular Biology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2017

Doctoral Committee:

Advisor: P. Shing Ho

Olve Peersen
Santiago Di Pietro
Nick Fisk

ABSTRACT

5-HYDROXYMETHYLCYTOSINE AND ENDONUCLEASE G AS REGULATORS OF HOMOLOGOUS RECOMBINATION

Homologous recombination (HR) is a necessary biological process for all living organisms, and it is especially important for repairing damaged DNA. Improper HR results in DNA damage-related diseases, notably increased likelihood of cancer when HR regulators, such as the human *BRCA1* gene, are impaired. HR is also a tool for biotechnology, giving scientists the power to easily delete or mutate genes and study the effects of those modifications. Recently, the epigenetically modified nucleotide 5-hydroxymethylcytosine (^{5hm}C) was found to regulate vertebrate HR via interaction with the protein endonuclease G (EndoG). In this dissertation, I use biochemical/biophysical methods to elucidate the interaction between ^{5hm}C and EndoG, thus working towards understanding their roles as regulators of recombination. I find that ^{5hm}C forms a unique hydrogen bond to stabilize Holliday junctions, the four-stranded DNA intermediate in HR. ^{5hm}C also induces a global structure change to the junction, increasing protein access to the junction crossover and providing potential for either direct or indirect readout of ^{5hm}C. Further connecting EndoG with recombination, we present the first evidence that EndoG preferentially binds and cleaves Holliday junction DNA, implicating a role for EndoG as a resolvase. I demonstrate that EndoG recognizes ^{5hm}C in the junction context and observe unique cleavage products from EndoG interaction with ^{5hm}C-junctions. These results suggest that EndoG may have a previously unrecognized junction resolvase function and, in this way, play a more direct role in recombination than simply creating double-stranded breaks in duplex DNA to initiate the

HR mechanism. Finally, I present a unique structural feature of vertebrate EndoG that we hypothesize is the basis for ^{5hm}C recognition. I present the structure of mouse EndoG and propose that a two amino acid deletion, conserved in vertebrate EndoG sequences, is associated with unraveling of an α-helix. This structural perturbation positions amino acid side chains to confer ^{5hm}C-sensing ability to all vertebrate EndoG. I expect that these deletion mutations and resulting structural effects co-evolved with the appearance of ^{5hm}C in vertebrate genomes to give EndoG an additional function of recognizing ^{5hm}C in the cell. Overall this work is building onto the understanding of ^{5hm}C and EndoG as markers and regulators of recombination.

ACKNOWLEDGEMENTS

The first person deserving of acknowledgement for the construction of this dissertation is my advisor – Dr. P. Shing Ho. Shing has been an incredible source of knowledge throughout this process, improving my understanding of a variety of topics ranging from quantum mechanics to grammar mechanics of the English language. He has been instrumental to my growth as a scientist. He encouraged me to take on a new project, attempt difficult experiments, write a grant, and be the teaching assistant for the Physical Biochemistry course, even when I was skeptical of my own abilities. I will be forever grateful for the training I received during my PhD, as I believe it has left me well prepared for my career ahead.

I want to thank my Student Advisory Committee: Dr. Olve Peersen, Dr. Santiago Di Pietro, and Dr. Nick Fisk. Olve was especially helpful in answering all of my questions about various biochemistry topics, especially kinetics. I am lucky to have the fortune of earning my PhD in such a supportive and nurturing scientific environment. I have received some form of assistance from probably every single lab in the BMB department, and I appreciate all of their kindnesses.

My students have taught me that I have a strong passion for education, which has been one of the most important lessons I've learned during my PhD. They all deserve acknowledgement for every moment of patience they had for me as I struggled through learning how to effectively communicate. Thank you to the students in Physical Biochemistry (Fall 2012 and 2013), LIFE 203 (Spring 2013), and the various students I tutored during my time here; you have inspired me to pursue a career as an educator. I also appreciate the work of my trainees in

the lab: Amanda, Rhea Kay, Alex, Colleen, and Ethan. You kept me on my toes with your challenging questions and improved my science.

I have learned an incredible amount from my fellow lab mates Megan, Matt, Melissa, Rhea Kay, and Anna-Carin. You have been the best friends and coworkers I could have hoped for. You were always willing to talk out experiments and data with me, and you never hesitated to offer gestures of compassion when I was having a difficult time. I especially appreciate Matt who has been a friend to me since we met at recruitment weekend; he was always able to make me laugh and was there to offer support in times of struggle. I appreciate the strong friendships I've formed with others in and outside of the department. You inspire me with your strength, and I'm grateful for the bravery you showed to share your stories and triumphs with me.

Finally I want to thank my family, for they have been by my side the longest. I'm eternally grateful to all of my parents (Yvette & Allan, and Paul & Patty) for being supportive and providing the environment that got me here today. You taught me how to work hard, and I hope someday I can repay your sacrifices. Nicole has always been my best friend and ally, and she offers unconditional love in spite of my selfish time sacrifices towards my own education. You are an incredible person, and I'm proud to call you my sister. To AJ, Justin, Bryanna, and Joy – thank you so much for letting me be a part of your life. It's amazing to watch you grow up, and I appreciate the weekly phone calls more than you can know. Thank you Nik for honestly thinking my science is cool; I'm not sure you realize how special that makes you. You have been incredibly supportive of my career and aspirations. You're amazing.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Significance and Background

Homologous recombination (HR) is the process of exchanging genetic information between two separate pieces of DNA, and it is a natural phenomenon in all living organisms and viruses. HR is used in the cell to repair damaged DNA[1], thus preventing DNA lesions from incorrectly replicating and potentially becoming cancerous. In meiosis, DNA undergoes HR to recombine genetic information to generate offspring diversity during reproduction[2]. HR is also implicated in the restart of stalled DNA replication forks[3]. When DNA replication is erroneous, the cell invokes HR to remove the incorrect sequence and replace it with the proper nucleotides. Finally, HR is naturally involved in horizontal gene transfer as one organism transfers portions of its genetic information to another[4].

Deficiencies in HR cause diseases in humans. One classic example is related to the breast cancer related genes *BRCA1* and *BRCA2*. The protein products of these genes are responsible for promoting proper HR, and when they are mutated or deleted the result is a higher likelihood of cancer[5]. Another regulator of recombination is the protein RecQ and its absence results in excessive cellular HR, leading to cancer-related diseases such as Bloom's syndrome, Werner's syndrome, and Rothmund-Thomson syndrome[6,7]. Insufficient recombination during meiosis results in incorrect chromosome separation in sex cells, ultimately leading to Down syndrome[8].

Aside from promoting healthy cellular behavior, HR is also a useful tool to recombine genetic information in laboratory research. Jack Szostak first developed HR technology using plasmids to induce recombination in yeast[9]. In 2007, Mario Capecchi, Martin Evans and Oliver Smithies won the Nobel Prize in Physiology or Medicine for the development of mouse embryonic gene targeting technology via HR[10]. HR is commonly used in the lab to introduce a genetic change, and gives scientist the power to study the effects of particular gene mutations, deletions, and insertions. HR is an essential biological process and a powerful tool for scientific research, so it is crucial that we gain a complete understanding of HR regulation in the cell. In this dissertation, I present original structural and biophysical studies that clarify the roles of the epigenetic DNA marker 5-hydroxymethylcytosine (^{5hm}C) and the enzyme Endonuclease G (EndoG) in HR.

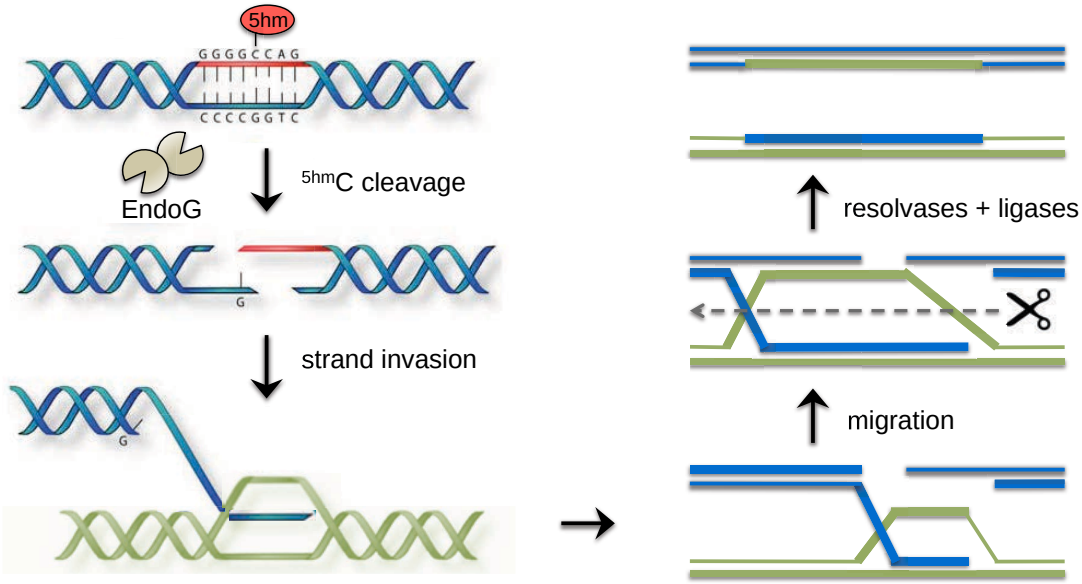
Mechanistically, HR is initiated by a double-strand break on one of the participating DNA duplexes. The 5' ends of the broken strands are degraded, and one of the remaining 3' ends invades an unbroken homologous double stranded DNA partner. The strands of the homologous DNA partner separate and instead cross over to anneal with the respective single strand ends of the broken partner. The DNA forms a 4-stranded intermediate structure, the Holliday junction[11], which enables the crossover and exchange of genetic information. In a process called branch migration, DNA nucleotides break away from their original partners and instead cross over to pair with their homologous neighbor. Branch migration will continue to move the Holliday junction through the needed amount of base pairs until it has reached its destination to complete the desired task (gene transfer, lesion repair, etc.). Branch migration halts, and DNA-cutting proteins (resolvases) will bind and resolve the junction crossover back into two separate pieces of DNA.[12]

Holliday junctions exist in two main conformations: open-X and stacked-X structures[13,14]. The junction takes the open-X structure as it migrates through the DNA sequence during HR. The open-X junction is not thermodynamically stable, and the only published crystal structures of the open-X junction are with assisted stabilization from a bound protein. This structure is, however, more optimized for junction migration because it is able to isoenergetically break and form new base pairs. In contrast, the stacked-X junction is topologically trapped, but quite thermodynamically stable with key structural stabilization elements determined via crystallography[15]. The stacked-X junction is stabilized by essential hydrogen bonds at the junction crossover between cytosine amines and oxygens in the phosphate backbone. The necessity for these hydrogen bonds incurs sequence specificity at the core of the junction – sequences with a RYC trinucleotide (where R is a purine, Y is a pyrimidine) junction core are the most stable[16]. The biological interplay between open-X and stacked-X junctions is not fully understood, but one thought is that stacked-X junctions provide a stable substrate for protein binding[17]. The hypothesis is that the open-X junction will migrate through the desired crossover region until an RYC motif passes into the junction. The stabilizing RYC core will convert the junction to a stacked-X conformation, thus halting migration and providing a stable substrate for resolvases to bind. Cytosines are commonly subjected to epigenetic modification, and so one question is whether epigenetically modified cytosines are also able to stabilize a stacked-X junction core? This is particularly interesting in light of ^{5hm}C, the recently rediscovered epigenetic modification that has been linked to recombination.

^{5hm}C is an epigenetic modifier present on up to one percent of cytosines in the mammalian genome[18]. Although ^{5hm}C DNA is not present in all eukaryotes, it seems to be a conserved feature of vertebrate organisms[19–21]. Natural occurrence of ^{5hm}C is achieved through

oxidative conversion from 5-methylcytosine (^{5m}C) via the ten-eleven translocation (Tet) family of proteins[22]. The Tet enzymes also catalyze further oxidation to formyl- and carboxyl-cytosines, which can be returned to canonical cytosine via removal of fully oxidized species[23]. In 2009, ^{5hm}C was rediscovered as an important mammalian epigenetic modification when Heintz et al. found significant levels of ^{5hm}C present in Purkinje neurons[24]. Breakthroughs in sequencing technology have revealed ^{5hm}C is present at varying concentrations in a plethora of tissue types[25], even dynamically appearing in DNA to play a role in embryonic development[26,27]. Aside from regulating Purkinje neuron development, ^{5hm}C has since been implicated in a variety of functions including 5-methylcytosine metabolism[28], transcription regulation[26,29,30], and DNA recombination. The recombination role was first suggested after a high number of hydroxymethylated cytosines were found in G/C rich recombination hotspots, suggesting the ^{5hm}C was perhaps a marker for recombination[26]. Robertson *et al.* established a more direct link when they discovered that ^{5hm}C promotes recombination via interaction with the protein EndoG[31]. The current model is that EndoG specifically binds the sequence 5'-GGGG^{5hm}CCAG-3'/5'-CTGGCCCC-3' to create double-strand breaks that will initiate homologous recombination (Fig. 1.1). This finding was quite unexpected as EndoG was previously only thought to bind and cleave G/C rich regions with no particular sequence specificity.

EndoG was first discovered in mammals as a non-specific endonuclease with a preference to cleave G/C rich DNA sequences[32–34]. Later, EndoG was found as a conserved protein in all eukaryotes[35] and determined to function via a NHN motif found in many nucleases[36,37]. Mitochondrial EndoG is thought to promote mtDNA replication by cleaving RNA/DNA hybrids to generate primers for replication[39], although this function is still debated. In healthy cells the bulk of EndoG remains in the mitochondria[38], but mass quantities of EndoG



Adapted from Robertson *et al.*, Nucleic Acids Research, 2014; **42**(21): p. 13280-13293.

Figure 1.1: Model for ^{5hm}C and EndoG in homologous recombination. As originally proposed by Robertson *et al.*, EndoG generates double stranded breaks via recognition of ^{5hm}C in a sequence specific context. Double strand breaks promote strand invasion into a homologous DNA neighbor, which initiates the junction crossover. The Holliday junction migrates through the desired length of DNA, and resolvases cleave the junction back into separate DNA strands. Broken and damages strands are finally repaired with ligases and polymerases. In this dissertation we present a structure of mouse EndoG that accounts for ^{5hm}C recognition. We also propose an additional role for EndoG and ^{5hm}C in recombination that invokes Holliday junction recognition.

are trafficked to the nucleus to degrade DNA during apoptosis[40], recognizing chromatin as a substrate[41]. Under non-apoptotic conditions, however, a basal level of EndoG is found regularly present in the nucleus[42]. EndoG has been implicated in several recombination scenarios including Herpes Simplex Virus[43], myeloid/lymphoid leukemia break point clusters[44], immunoglobulins[45], transfected plasmid recombination[46], and genome maintenance through HR[47]. The assumption is that EndoG binds to G/C rich double stranded DNA and creates double stranded breaks to initiate these recombination events.

The work of Robertson *et al.* was the first instance of EndoG recognizing a specific sequence feature, and the first discovery of interactions between ^{5hm}C and EndoG. ^{5hm}C is only present in vertebrates while EndoG is conserved in all eukaryotes, which begs the question of whether vertebrate EndoG has a unique structural feature to specifically recognize ^{5hm}C? Furthermore, despite the thematic appearance of EndoG in DNA recombination, it is unknown whether EndoG recognizes Holliday junctions as a DNA substrate. Finally, we wanted to explore how ^{5hm}C might impact Holliday junctions to serve as a structural or energetic marker for recombination, and whether EndoG recognizes ^{5hm}C in the junction context.

1.2 Aims of this Project

In this work, we aim to understand how ^{5hm}C promotes recombination via EndoG by asking these specific questions:

- Does ^{5hm}C impact Holliday junction structure and stability and can it serve as a marker in HR?
- Does EndoG cleave Holliday junctions as a method to promote recombination and does EndoG recognize ^{5hm}C in the junction context?

- Is there a structural feature conserved among vertebrates that allows EndoG to recognize ^{5hm}C DNA?

The first objective was to establish the Holliday junction as a system to study energies of molecular interactions at the junction core. This analysis would allow direct comparison of the structure and stability of C- vs. ^{5hm}C-stabilized Holliday junctions. Furthermore, I provide evidence of continuity between crystallographic structure and solution state energies for the Holliday junction system. In chapter 2, I discuss Holliday junction cores stabilized by halogen bonds (X-bonds), weak intermolecular interactions analogous to hydrogen bonds. The X-bond is a short electrostatic interaction between a negative acceptor and the positive crown of a polarized covalently bound halogen. In this case the positive halogen crown on a halogenated uracil acts as the donor, replacing the positive amine of a cytosine that typically stabilizes junctions. In this work I use crystallography and differential scanning calorimetry (DSC) to measure the energy of the X-bond stabilized junction compared to the H-bond stabilized junction. Furthermore we isolate specific energies that describe the relative H-bond and X-bond strengths. We show that the junction crystal structures are representative of the solution state energies determined by DSC. The methodology established for halogen bonds in this chapter is shown to be applicable to probe epigenetic effects on junction stability.

In chapter 3, we ask how ^{5hm}C impacts junction structure and stability. We present the crystal structure of a junction with ^{5hm}C in the core, showing that ^{5hm}C forms specific H-bonds to replace the typical bonds formed by canonical cytosine at the core. ^{5hm}C also confers global structure shape and broadens the angle relating the two arms of the junction. We apply the differential scanning calorimetry method established in chapter 2 to learn about how ^{5hm}C

thermodynamically impacts junction stability. We find that ^{5hm}C is more enthalpically stable, although it suffers an entropic penalty and, as a result, the net change in free energy is very small compared to C-stabilized junctions. We conclude that ^{5hm}C incorporation into junctions is stable, and potentially provides a mechanism for protein recognition via either direct or indirect readout.

The next question was whether EndoG recognizes junctions to promote recombination, and whether ^{5hm}C plays a role in that recognition. In chapter 4, we find that EndoG does indeed recognize junctions, and they are a preferred substrate over duplex DNA. ^{5hm}C is not needed for junction recognition, but does induce a unique cleavage profile. These results together suggest a role for ^{5hm}C as a marker in EndoG-mediated recombination. Finally, we present the crystal structure of mouse EndoG and discover a helix to loop transition relative to invertebrate EndoG structures; we believe this structural change is the source of vertebrate EndoG preference for ^{5hm}C DNA. The loop structure is caused by a two amino acid deletion that is conserved for all vertebrate species, suggesting all vertebrates are capable of recognizing ^{5hm}C via EndoG. ^{5hm}C is only found in vertebrate species, thus we propose that EndoG co-evolved to accommodate the presence of ^{5hm}C.

Overall this work supplies necessary information to understand how 5^{hm}C and EndoG work together to promote DNA recombination and expands on the original model presented by Robertson *et al*. This exciting avenue of research elucidates a new role for both EndoG and ^{5hm}C and is pioneering an unexplored aspect of HR. A mechanistic understanding of this process is the foundation to correct erroneous HR in disease and develop better biotechnology. We work towards a complete understanding of ^{5hm}C as a marker for HR with hopes to eventually manipulate HR in a more targeted and specific way.

REFERENCES

(1) Orr-Weaver, T. L., and Szostak, J. W. (1983) Yeast recombination: the association between double-strand gap repair and crossing-over. *Proc. Natl. Acad. Sci. U. S. A. 80*, 4417–4421.

(2) Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J., and Stahl, F. W. (1983) The double-strand-break repair model for recombination. *Cell 33*, 25–35.

(3) Petermann, E., and Helleday, T. (2010) Pathways of mammalian replication fork restart. *Nat. Publ. Gr. 11*, 683–687.

(4) Cromie, G. A. (2009) Phylogenetic ubiquity and shuffling of the bacterial RecBCD and AddAB recombination complexes. *J. Bacteriol. 191*, 5076–5084.

(5) McCabe, N., Turner, N. C., Lord, C. J., Kluzek, K., Białkowska, A., Swift, S., Giavara, S., O'Connor, M. J., Tutt, A. N., Zdzienicka, M. Z., Smith, G. C. M., and Ashworth, A. (2006) Deficiency in the repair of DNA damage by homologous recombination and sensitivity to poly(ADP-ribose) polymerase inhibition. *Cancer Res. 66*, 8109–8115.

(6) Bugreev, D. V, Yu, X., Egelman, E. H., and Mazin, A. V. (2007) Novel pro- and anti-recombination activities of the Bloom's syndrome helicase. *Genes {&} Dev. 21*, 3085–3094.

(7) Hu, Y., Raynard, S., Sehorn, M. G., Lu, X., Bussen, W., Zheng, L., Stark, J. M., Barnes, E. L., Chi, P., Janscak, P., Jasin, M., Vogel, H., Sung, P., and Luo, G. (2007) RECQL5/Recql5 helicase regulates homologous recombination and suppresses tumor formation via disruption of Rad51 presynaptic filaments. *Genes Dev. 21*, 3073–3084.

(8) Lamb, N. A., Yu, K., Shaffer, J., Feingold, E., and Sherman, S. (2005) Association between Maternal Age and Meiotic Recombination for Trisomy 21. *Am J Hum Genet 76*, 91–99.

(9) Orr-Weaver, T. L., Szostak, J. W., and Rothstein, R. J. (1981) Yeast transformation: a model system for the study of recombination. *Proc. Natl. Acad. Sci. U. S. A. 78*, 6354–8.

(10) (2007) The Nobel Prize in Physiology or Medicine 2007. *Nobel Media AB*.

(11) Holliday, R. (1964) A mechanism for gene conversion in fungi. *Genet. Res. 5*, 282–304.

(12) Haber, J. E. (2014) Genome Stability - DNA repair and recombination (Scholl, S., Ed.) 1st ed. Garland Science, New York.

(13) Lilley, D. M. J. (2000) Structures of helical junctions in nucleic acids. *Q. Rev. Biophys. 33*, 109–159.

(14) Hays, F. A., Schirf, V., Ho, P. S., and Demeler, B. (2006) Solution formation of Holliday junctions in inverted-repeat DNA sequences. *Biochemistry 45*, 2467–2471.

(15) Eichman, B. F., Vargason, J. M., Mooers, B. H. M., and Ho, P. S. (2000) The Holliday junction in an inverted repeat DNA sequence: Sequence effects on the structure of four-way junctions. *Proc. Natl. Acad. Sci. 97*, 3971–3976.

(16) Voth, A. R., Hays, F. a, and Ho, P. S. (2007) Directing macromolecular conformation through halogen bonds. *Proc. Natl. Acad. Sci. U. S. A. 104*, 6188–93.

(17) Khuu, P. A., Voth, A. R., Hays, F. A., and Ho, P. S. (2006) The stacked-X DNA Holliday junction and protein recognition. *J. Mol. Recognit. 19*, 234–42.

(18) Szwagierczak, A., Bultmann, S., Schmidt, C. S., Spada, F., and Leonhardt, H. (2010) Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic DNA. *Nucleic Acids Res. 38*, e181.

(19) Delatte, B., Wang, F., Ngoc, L. V., Collignon, E., Bonvin, E., Deplus, R., Calonne, E., Hassabi, B., Putmans, P., Awe, S., Wetzel, C., Kreher, J., Soin, R., Creppe, C., Limbach, P. A., Gueydan, C., Kruys, V., Brehm, A., Minakhina, S., Defrance, M., Steward, R., and Fuks, F.

(2016) Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. *Science (80-.). 351*, 282–285.

(20) Diotel, N., Yohann, M., Coumailleau, P., Gueguen, M.-M., Serandour, A. A., Salbert, G., and Olivier, K. (2017) 5-Hydroxymethylcytosine Marks Postmitotic Neural Cells in the Adult and Developing Vertebrate Central Nervous System. *J. Comp. Neurol. 525*, 478–497.

(21) Raddatz, G., Guzzardo, P. M., Olova, N., Rosado, M., and Rampp, M. (2013) Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proc. Natl. Acad. Sci. 110*, 8627–8631.

(22) Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L. M., Liu, D. R., Aravind, L., and Rao, A. (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science 324*, 930–5.

(23) 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*, 1300–3.

(24) Kriaucionis, S., and Heintz, N. (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science 324*, 929–30.

(25) Li, W., and Liu, M. (2011) Distribution of 5-hydroxymethylcytosine in different human tissues. *J. Nucleic Acids 2011*, 1–5.

(26) Stroud, H., Feng, S., Morey Kinney, S., 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.

(27) Ruzov, A., Tsenkina, Y., Serio, A., Dudnakova, T., Fletcher, J., Bai, Y., Chebotareva, T., Pells, S., Hannoun, Z., Sullivan, G., Chandran, S., Hay, D. C., Bradley, M., Wilmut, I., and De

Sousa, P. (2011) Lineage-specific distribution of high levels of genomic 5-hydroxymethylcytosine in mammalian development. *Cell Res. 21*, 1332–1342.

(28) Guo, J. U., Su, Y., Zhong, C., Ming, G., and Song, H. (2011) Hydroxylation of 5-Methylcytosine by TET1 Promotes Active DNA Demethylation in the Adult Brain. *Cell 145*, 423–434.

(29) Robertson, J., Robertson, A. B., and Klungland, A. (2011) The presence of 5-hydroxymethylcytosine at the gene promoter and not in the gene body negatively regulates gene expression. *Biochem. Biophys. Res. Commun. 411*, 40–43.

(30) Wen, L., Li, X., Yan, L., Tan, Y., Li, R., Zhao, Y., Wang, Y., Xie, J., Zhang, Y., Song, C., Yu, M., Liu, X., Zhu, P., Li, X., Hou, Y., Guo, H., Wu, X., He, C., Li, R., Tang, F., and Qiao, J. (2014) Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain *15*, 1–17.

(31) 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.

(32) Ruiz-carrilio, A., and Renaud, J. (1987) Endonuclease G: a (dG)n* (dC)n-specific DNase from higher eukaryotes *6*, 401–407.

(33) Cummings, O. W., King, C., Holden, J. A., and Low, L. (1987) Purification and Characterization of the Potent Endonuclease in Extracts of Bovine Heart Mitochondria. *J. Biol. Chem. 262*, 2005–2015.

(34) Cote, J., Renaud, J., and Ruiz-Carrillo, A. (1989) Recognition of (dG)n. (dC)n Sequences by Endonuclease G *264*, 3301–3310.

(35) Schäfer, 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*, 217–28.

(36) 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*, 1–14.

(37) Kieper, J., Lauber, C., Gimadutdinow, O., Urbańska, 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.

(38) Low, R. L. (2003) Mitochondrial Endonuclease G function in apoptosis and mtDNA metabolism: a historical perspective. *Mitochondrion 2*, 225–236.

(39) Côté, J., and Ruiz-Carrillo, A. (1993) Primers for Mitochondrial DNA Replication Generated by Endonuclease G. *Science (80-.). 261*, 765–769.

(40) 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.

(41) 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*, 48404–9.

(42) 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.

(43) 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. 103*, 8995–9000.

(44) Gole, B., Baumann, C., Mian, E., Ireno, C. I., and Wiesmüller, L. (2014) Endonuclease G initiates DNA rearrangements at the MLL breakpoint cluster upon replication stress. *Oncogene*.

(45) 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.

(46) 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. 50*, 252–261.

(47) Büttner, S., Carmona-Gutierrez, D., Vitale, I., Castedo, M., Ruli, D., Eisenberg, T., Kroemer, G., and Madeo, F. (2007) Depletion of endonuclease G selectively kills polyploid cells. *Cell Cycle 6*, 1072–1076.

CHAPTER 2

DETERMINING THERMODYNAMIC PROPERTIES OF MOLECULAR INTERACTIONS FROM SINGLE CRYSTAL STUDIES[1]

The concept of single crystals of macromolecules as thermodynamic systems is not a common one. However, it should be possible to derive thermodynamic properties from single crystal structures, if the process of crystallization follows thermodynamic rules. We review here an example of how the stabilizing potentials of molecular interactions can be measured from studying the properties of DNA crystals. In this example, we describe an assay based on the four-stranded DNA junction to determine the stabilizing potentials of halogen bonds, a class of electrostatic interactions, analogous to hydrogen bonds, that are becoming increasing recognized as important for conferring specificity in protein–ligand complexes. The system demonstrates how crystallographic studies, when coupled with calorimetric methods, allow the geometries at the atomic level to be directly correlated with the stabilizing energies of molecular interactions. The approach can be generally applied to study the effects of DNA sequence and modifications of the thermodynamic stability of the Holliday junction and, by inference, on recombination and recombination dependent processes.

2.1 Introduction

‘‘A picture is worth a thousand words’’. Although the origin of this phrase remains unresolved (being variously attributed to a Japanese philosopher or as an old Chinese proverb, commonly ascribed to Confucius), it has become the mantra in macromolecular crystallography—the ‘‘picture’’ or structure of a protein or nucleic acid, or complex among or

[1] Previously published as: "Determining thermodynamic properties of molecular interactions from single crystal studies" Vander Zanden, C.M., Carter, M., Ho, P.S. *Methods* 64 (2013) 12-18.

between them can tell us so much about function. Of course, the absolute requirement of a crystal in crystallography immediately limits the picture to be a static one, at least for those parts that we can see (parts that are truly dynamic are invisible to our X-ray vision). This has lead to the general perception that one cannot learn anything about thermodynamics from crystallographic studies on single crystals. A crystal is a solid that is assembled by regularly repeating the same molecule over- and-over again into a well-defined crystal lattice. This does not mean, however, that every molecule or even the atoms of the molecule are identical in every way, or even held entirely static. We review here an example of an approach to quantitatively determine the energies of halogen bonds, exploiting the isomerization of a DNA junction in the assay. As a result, we show that indeed crystallography can tell us much about the energies of molecular inter- actions that are important for the assembly and stability of nucleic acids.

2.2 Structure-energy relationships of biological halogen bonds

In this case study, we will characterize the stabilizing potentials of halogen bonds and their correlations with specific geometries in a biological system. Halogen bonds (X-bonds), formerly known as charge-transfer bonds[1], are analogous to H-bonds[2] in that they are directional, primarily electrostatically driven molecular interactions that help to define the specificity of ligands against their protein targets[3–8] and to drive macromolecular conformation[9]. The X-bond is formed when the electropositive crown of a polarizable halogen (as a consequence of depopulating the p_z-atomic orbital when forming a covalent σ-bond to, for example, a carbon[10]) interacts with an electron-rich acceptor, such as an oxygen, nitrogen, or sulfur. The stabilizing potential of the X-bond depends on the degree of polarization of the halogen (I > Br > Cl > F), the electron withdrawing ability of the molecule that is halogenated, the electronegative potential of the acceptor atom, and the distance and angle of approach of the acceptor to the halogen (Fig.

2.1). In order to develop empirical models that can be accurately applied to predict the electrostatic behavior of halogens, including the existence of X-bonds, we must determine how the energies of interaction are related to the geometry of the interacting atoms. For this assay, we will describe how a Holliday junction construct (Fig. 2.2) can be designed to determine the geometries of X-bonds from single crystal studies, and their associated energies of stabilization relative to a competing H-bond in crystals[9,11] and in solution[11,12].

For this set of studies, we developed a DNA Holliday junction as the model system to compete an X-bond against an H-bond in defining the stability and the isomer conformation (conformer) of the four-stranded complex. The form of the Holliday junction under physiological salt concentrations is the stacked-X form[13], in which the four helical arms are paired-up and stacked to form essentially two near continuous columns of standard B-form DNA, interrupted by the cross-over of one strand from each duplex to the adjacent duplex (Fig. 2.2). The cross-over strands form a tight U-turn, which topologically locks the junction in place. The cross-over is stabilized by a set of intrastrand interactions localized at the $N_6Py_7C_8$ trinucleotide core of the general sequence motif 5'- $CCG_3Pu_4N_5N_6Py_7C_8GG$-3' (where Py is a pyrimidine (C or T), Pu is a purine (G or A), and $G_3Pu_4N_5$ are nucleotides that maintain the inverted repeat symmetry of the overall sequence)[14,15]. In particular, an H-bond from the cytosine of C_8 to the phosphate of Py_7 is essential, while an electrostatic interaction (including an H-bond) from Py_7 to N_6 is important for bending the phosphodeoxyribose backbone into the U-turn of the crossing strands and specifying the sequence dependent stability of the junction in the crystal (Fig. 2.2)[14,16] and in solution[17].

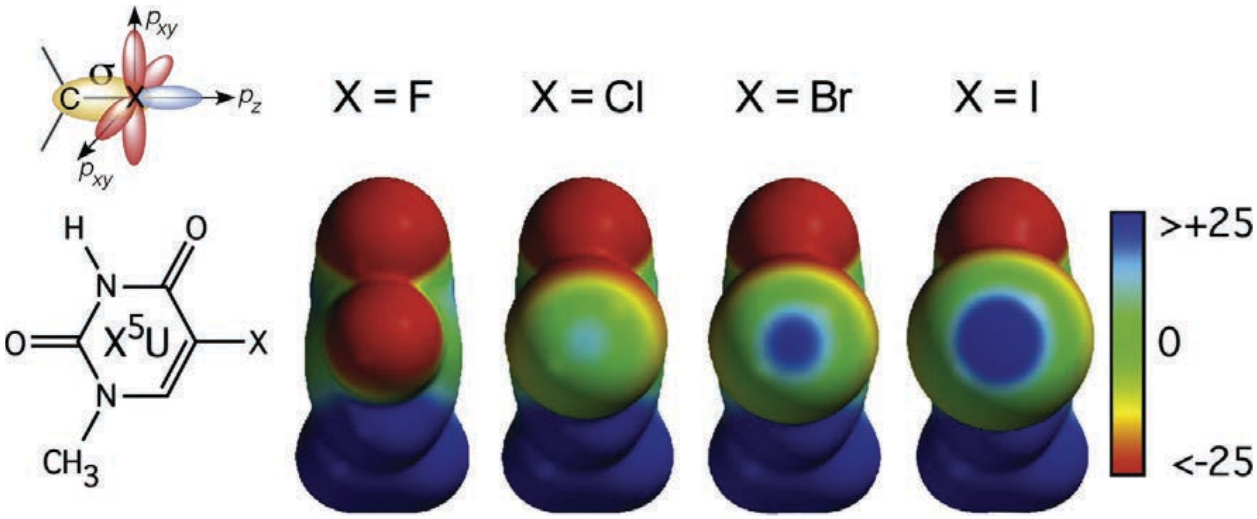


Figure 2.1: Halogen polarization and halogen bonding. Halogen bonds (X-bonds) can be ascribed as the generation of a positive crown (a σ-hole) resulting from the formation of a covalent (σ) bond and the subsequent depopulation of the halogen's p_Z-orbital[10]. The σ-hole is pronounced with larger, more polarizable halogens (F < Cl < Br < I), and is enhanced when the halogen is attached to more electron withdrawing groups (in this case, a uracil base). Looking from the halogen down the axis of the C-X bond, electrostatic potential is shown mapped onto the surface of each halogenated uracil using a color scale ranging from + 25 kcal/mol (blue) to -25 kcal/mol (red)[3].

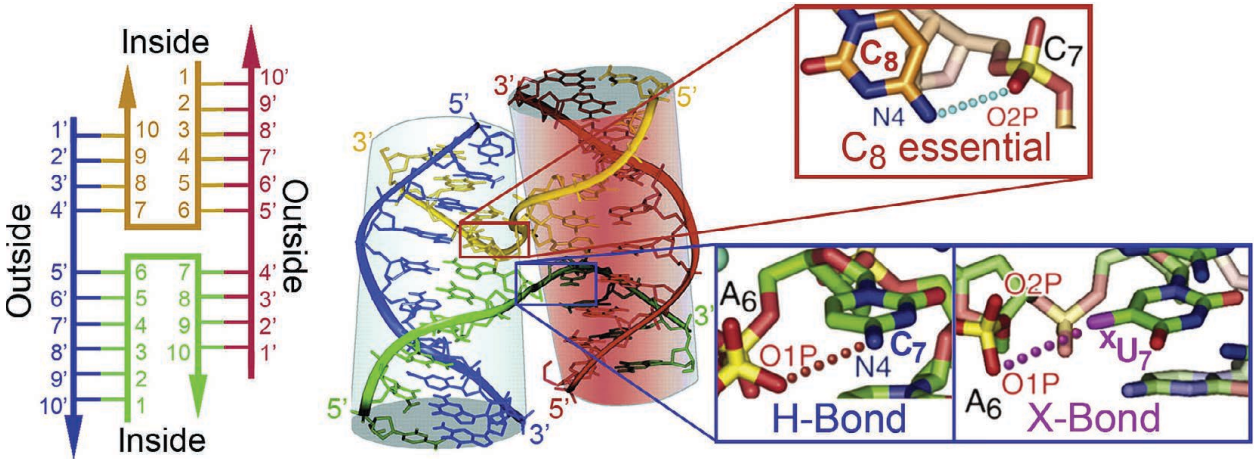


Figure 2.2: Molecular interactions that stabilize DNA junctions. DNA Holliday junctions in inverted repeat sequences of the type 5'-$CCG_3Pu_4N_5N_6Py_7C_8GG$-3' are stabilized by electrostatic interactions. In particular, an H-bond from C_8 to the preceding phosphate is essential, while an H-bond or X-bond from Py_7 to its preceding phosphate group is important for forcing the DNA backbone to form the tight U-turn at the inside strand of the stacked-X form of the junction[14,16].

An important component of these crystals in relationship to the X-bond assay is that the lattice is stabilized primarily by end-to-end stacking of the stacked B-DNA duplex arms of the junctions, forming essentially a continuous sheet of DNAs in the *b*–*c* plane of the *C*2 unit cell[14]. The interactions between sheets are mediated by cations that bridge across the backbones of the DNAs. Consequently, the important intrastrand interactions that stabilize the junction are well away from the lattice interactions, while the lattice interactions are largely independent of DNA sequence, as long as the terminal base pairs remain the same. Thus, this study assumes that the lattice interactions will be identical for all constructs, and that the population of conformations observed in the crystals will reflect their distributions in solution (*i.e.*, there will be no crystal dependent discrimination among different conformers).

For the current assay, a junction was constructed from two different, non-inverted repeat DNA strands, with both strands including the H-bond that is essential at the C_8 position. The strands differ primarily at the Py_7 nucleotide (and the associated Pu_3), with the standard C_7 at this position to provide an H-bond, while the complementary strand places a halogenated uracil base (^{X}U) at this position to potentially form an X-bond to the DNA backbone[9,11], where X is any of the common halogens (F, Cl, Br, or I, Table 2.1). In this mixed sequence construct, the junction can adopt one of two conformers, one that is stabilized by the H-bond (H-isomer) and the alternative that is stabilized by the X-bond (X-isomer) (Fig. 2.3). Aside from these stabilizing interactions, the two isomers are completely isoenergetic, so in effect, the isomer form is determined explicitly by the relative strengths of the two competing interactions. For example, if the two competing interactions had the same energy, the population would occupy both isomers equally; however, if the two energies differ significantly, then the junction will be seen in one form or the other, or as a weighted average of the two forms.

Table 2.1: List of DNA constructs to study X-bonding interactions. The DNA constructs are named according to the number of halogenated uracil (^{X}U) containing strands (X-strand) relative to cytosine H-bonding strands (H-strands) that form a junction. The potential X-bonding and H-bonding nucleotides (italics) results in defined molar ratios of X-bonding halogenated uracil and H-bonding cytosine bases (X:H).

DNA construct	Sequences	X:H
	H-Strand: (CCGGTA***C***CGG)$_2$	
F2J	X-Strand: (CCGGTA^{F}***U***CGG)$_2$	2:2
Cl1J	H-Strand: (CCGGTA***C***CGG)$_2$	1:2
	X-Strand: (CCGGTA^{Cl}***U***CGG + CCGGTA***U***CGG)	
Cl2J	H-Strand: (CCGGTA***C***CGG)$_2$	2:2
	X-Strand: (CCGGTA^{Cl}***U***CGG)$_2$	
Br1J	H-Strand: (CCGGTA***C***CGG)$_2$	1:2
	X-Strand: (CCGGTA^{Br}***U***CGG + CCGGTA***U***CGG)	
Br2J	H-Strand: (CCGGTA***C***CGG)$_2$	2:2
	X-Strand: (CCGGTA^{Br}***U***CGG)$_2$	
I1J	H-Strand: (CCGGTA***C***CGG)$_2$	1:2
	X-Strand: (CCGGTI^{I}***U***CGG + CCGGTA***U***CGG)	
I2J	H-Strand: (CCGGTA***C***CGG)$_2$	2:2
	X-Strand: (CCGGTA^{I}***U***CGG)$_2$	
H2J	H-Strand: (CCGGTA***C***CGG)$_2$	0:2
	X-Strand: (CCGGTA***U***CGG)$_2$	

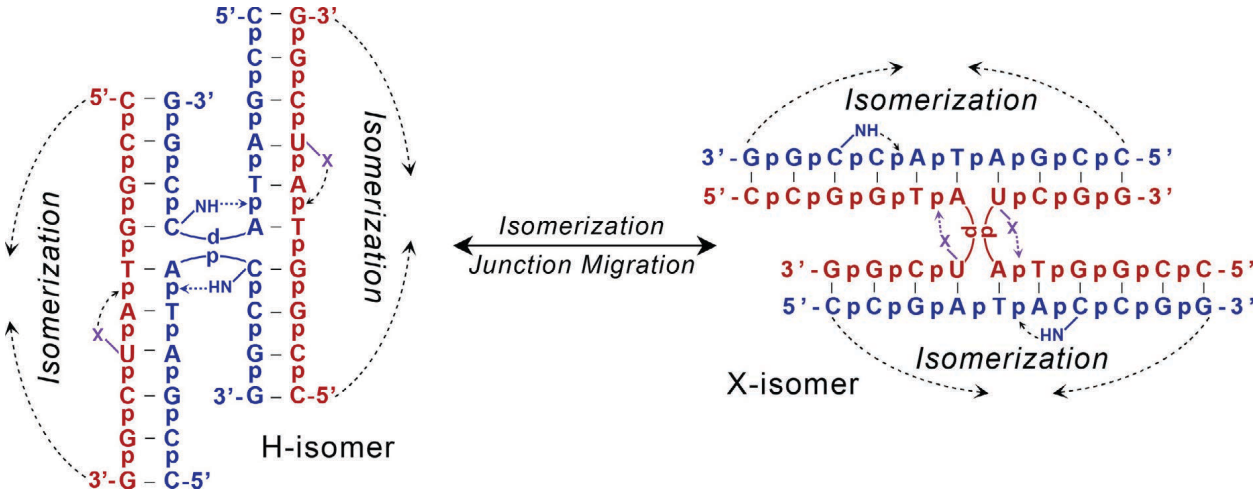


Figure 2.3: H-bond and X-bond driven isomerization of DNA junctions. A junction to assay the energy of an X-bond relative to a competing H-bond was constructed from two DNA sequences, one that contains a cytosine at the Py_7 position, while the complementary strand contains a halogenated uracil ($^{X}U_7$) at this position. The competing interactions direct the junction to adopt either the isomer form that is stabilized by standard H-bonds (H-isomer) or by X-bonds (X-isomer). Aside from these specific interactions, the H- and X-isomers are isoenergetic, allowing free isomerization between the two forms, thus creating an energetic competition between the H-bond and X-bond for control of the overall junction conformation.

An example of an X-bond that is nearly identical in its stability to the competing H-bond is that of the chlorine X-bond (Cl-bond)[11]. Chlorine is intermediate in its polarizability and, therefore, is expect to be intermediate in its X-bonding potential. The position of the chlorine, either on the inside crossing strand (labeled $^{Cl}U_7$) or outside strand (labeled $^{Cl}U'_7$, where the prime is used to indicate that this is at the outside strand position) specifies whether the junction is in the X- or H-isomer form, respectively. Thus, the isomeric form, and the stabilizing interaction, can be determined crystallographically by $^{Cl}U_7$ in the X- or $^{Cl}U'_7$ in the H-isomer. In doing so, we can estimate the difference in energy between the Cl-bond and the competing H-bond by analyzing the distribution of isomeric forms, since only the Py_7 of the trinucleotide core is varied.

For this analysis, we can define the difference in energy of the X- *vs* H-isomer form ($\Delta E_{IsoX-IsoH}$) according to Eq. (1)—this energy is not explicitly that of the molecular interactions.

Eq. (1) $$\Delta E_{IsoX-IsoX} = -RT \ln\left(\frac{\%IsoX}{\%IsoH}\right)$$

Again, the assumption here is that the population distribution of isomers in the crystal mirrors the distribution in solution, since we do not expect any differences in the crystal lattice energies.

The isomer form (H- or X-isomer) can be distinguished in the crystal structures of the two constructs by analyzing the electron density maps from the single crystal structures at the Py_7 nucleotide positions of the outside and the inside strands, along with the Pu_4 position of the complementary strands of the junction. In this analysis, the structures are initially refined with ambiguous base pairs at the Py_7•Pu_4 positions (essentially as U•A base pairs). The 2Fo-Fc density (Fig. 2.4) and Fo-Fc difference density maps are then analyzed to determine whether there is residual positive density ~1.5 Å from the C5 carbon of the Py_7 pyrimidine base (evidence for a halogenated uracil base) and from the C2 carbon of the Pu_4 purine base (evidence for an N2

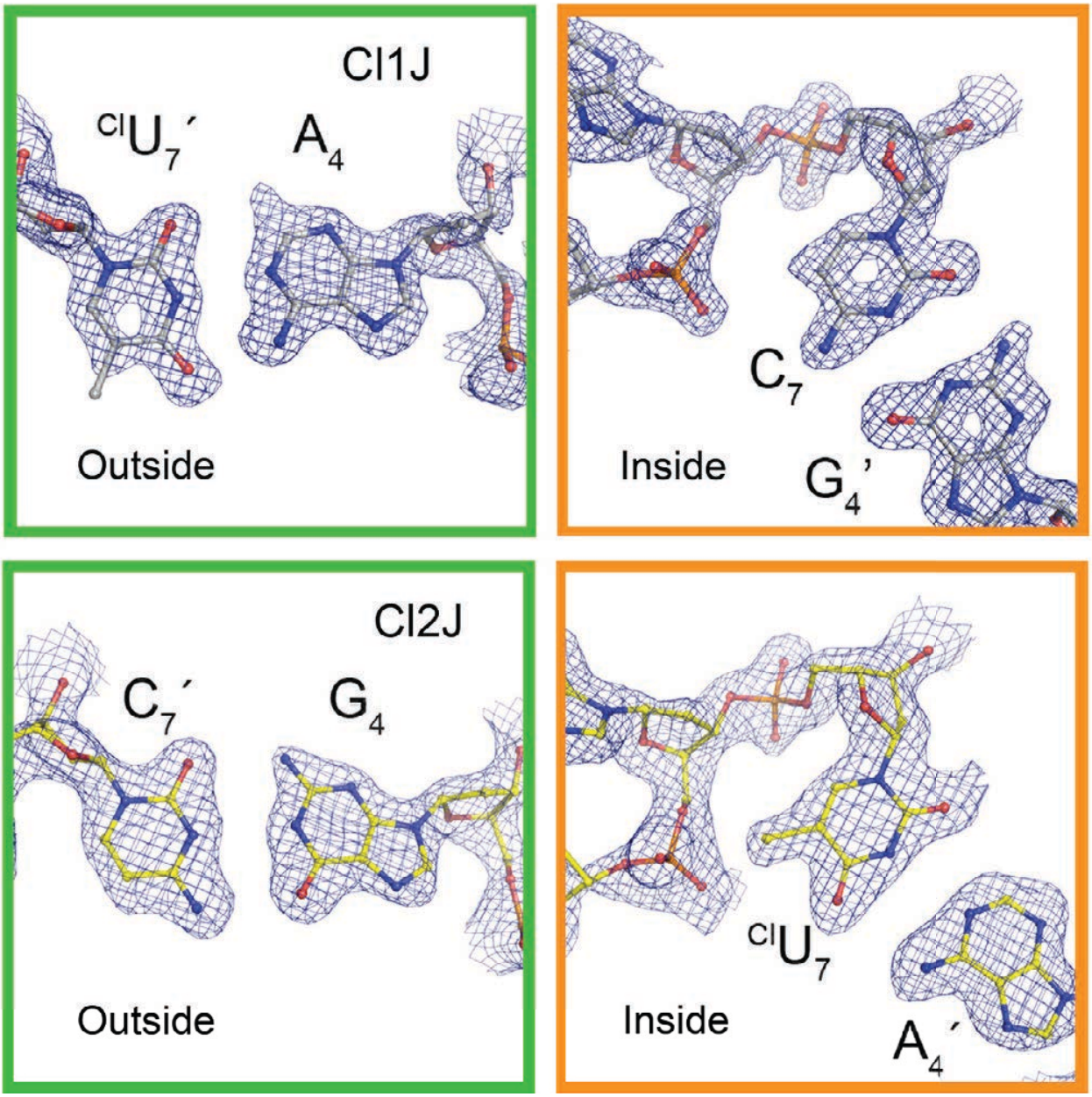


Figure 2.4: Electron density maps of chlorinated base pairs in the Cl1J and Cl2J DNA junctions. 2Fo-Fc electron density maps are shown for the potentially chlorinated nucleotides and their base pairs, where 100% H-isomer would have the halogen completely occupying the outside position (green box) and 100% X-isomer would have it in the inside position (orange box) forming an X-bond and stabilizing the overall structure. The maps show more chlorine density on the outside for the Cl1J structure (one X-bond competing against two H-bonds) than for the Cl2J structure (two X-bonds and two H-bonds).

amino group of a guanine base). The presence of the excess density at Py_7 and the absence of excess density at Pu_4 would indicate that this is a $^{Cl}U_7 \cdot A_4$ base pair, while the absence of excess density at Py_7 and presence of density at Pu_4 would indicate that this is a $C_7 \cdot G_4$ base pair. The X-isomer would place the ^{Cl}U of the $^{Cl}U_7 \cdot A_4$ base pair at the inside crossing strand, while the H-isomer places the halogenated uracil on the outside strand. The structures can then be refined as fully either the X- or H-isomer form, and the electron density maps reanalyzed to determine whether there is residual positive or negative difference density, indicating a mixture of isomers in the crystal structure.

We designed two DNA constructs in which one potential X-bond competes against two H-bond (Cl1J construct) or two X-bonds competes against two H-bonds (Cl2J construct). Analysis of the electron density maps indicated that the Cl1J construct was primarily H-isomer, while the Cl2J junction was primarily X-isomer. This suggested that the Cl-bond is only slightly more stabilizing than the H-bond, out competing the H-bond in a one-to-one competition, but not capable of winning in a one-to-two competition.

2.2.1 Quantitative conformer analysis by crystallographic occupancy titration

In order to estimate the actual stabilization energy of the Cl- bond relative to the H-bond, we need to accurately quantify the percent composition of each isomer through an occupancy titration procedure[11]. The initial step for this quantitation is to refine the structure of the junction using a model containing all the nucleotides that are common to both conformers, but with the four $Py_7 \cdot Pu_4$ base pairs that are unique to each isomer left as ambiguous (*i.e.*, as unhalogenated U•A base pairs). Once the structure of this ambiguous model has been fully refined, with solvent added around the common nucleotides and maximum-likelihood convergence, the next step is to apply isomer-specific modeling and attempt to determine the percent of each isomer in the

overall composition of the structure, which is the percent occupancy of the X- and H-isomers (%*IsoX* and %*IsoH*) in Eq. (1).

To determine %*IsoX* and %*IsoH* of a construct, the re! ned ambiguous structure is used to construct two models, one with the halogenated uracils of the $^{X}U_7 \cdot A_4$ base pairs at the inside crossing strand (X-isomer) and the other with the halogenated uracils on the outside strand (H-isomer). One approach to calculating the *%IsoX* and %*IsoH* at this point might be to re! ne the partial occupancies of the all atoms in the $Py_7 \cdot Pu_4$ base pairs, or of just the halogens and the N2 amino of the purines in each model; however, we found that this approach resulted in very large variations in occupancies, and with results that were often times incomprehensible (for example, where the sum of the occupancies for the two models were signi! cant higher or lower than 100%). We, therefore, needed to develop a more robust approach to determining the %*IsoX* and %*IsoH* for the DNA junction system.

Our alternative approach makes the assumption that the sum of %*IsoX* and %*IsoH* will be 100% (this is a standard coupled occupancy analysis), and that the true ratio of conformers will be re" ected in the statistics of the crystallographic re! nement, in particular the R_{free} value[18]. In this analysis, a combined model that includes the two isomer forms in their entirety are re! ned, starting with the occupancies of all atoms, including solvent, of the X-isomer component set at 99% and those of the H-isomer set at 1%. The combined model is fully re! ned and the R_{free}-value recorded (Fig. 2.5). This re! ned model is then used in the next step, with the occupancy of the X-isomer reduced by 2% to 97% and the H-isomer increased accordingly, and the new composite model re! ned. This ‘‘titration’’ of the occupancy is repeated until we reach the end point of 1% X- and 99% H-isomers (Fig. 2.5A). To account for any effects of the re! nement process on R_{free}, an equivalent number of re! nements was performed on a 99% X- and 1% H-

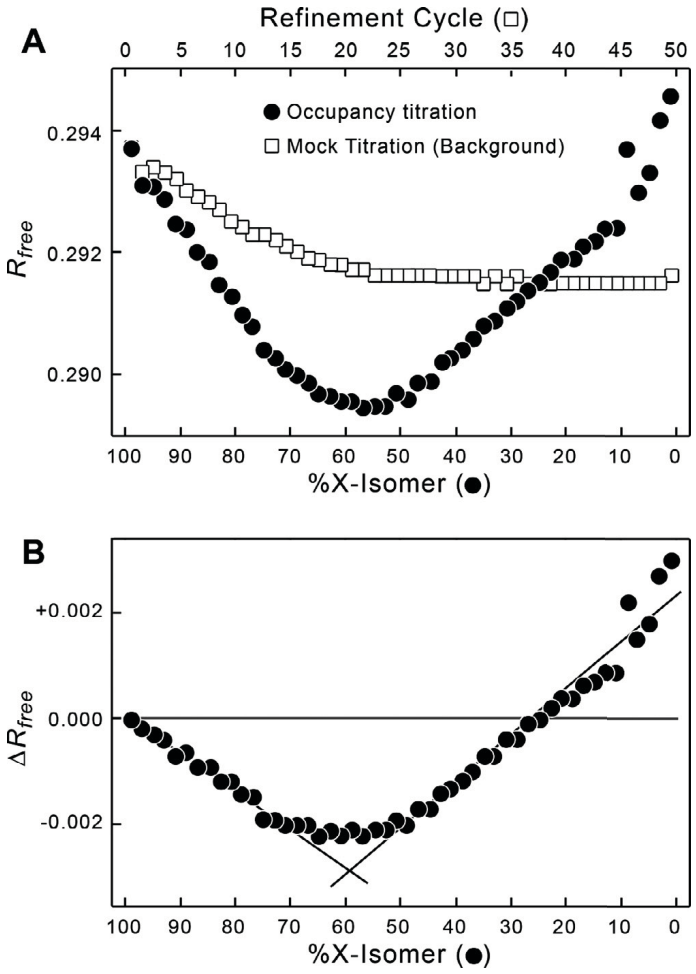


Figure 2.5: Occupancy titration and background correction for Cl2J junction. **A.** The occupancy titration (closed circles) and mock titration (open squares). For the occupancy titration, the DNA construct starts with 99% of the junction in the X-isomer and 1% as the H-isomer (Fig. 2.3), with the % X-isomer reduced by 2% and H-isomer increased by 2% at each titration point. The composite isomer models are refined and the R_{free} values recorded for each titration point. To account for the effect of refinement on the R_{free} values, the initial model is refined without changing the 99% X-isomer/1% H-isomer compositions for the same number of refinement cycles as the actual occupancy titration. This mock titration serves as the background associated with the effect of the refinement process on R_{free}. **B.** Background corrected occupancy titration. The R_{free} values from the mock background titration are subtracted from the occupancy titration data in A. The resulting minimum ΔR_{free} at each %X-isomer indicates the optimum contribution of the X-isomer to the structure in the crystal. The linear regions of the data on both sides of the minimum are fit by linear regression analysis (lines), and the intersection of the two fitted lines is used to quantify the X-isomer to H-isomer ratio.

isomer model, but with no change in the actual occupancies. The R_{free} values from this ‘‘mock’’ titration subtracted from the experimental values (Fig. 2.5B). To account for any potential hysteretic effects, the titration was repeated in the opposite direction starting with 99% H- isomer and decreasing the occupancy to 1% H-isomer (Fig. 2.6). The results for the chlorinated junctions show clear minima for R_{free}-values of both sets of titrations when plotted as a function of % X-isomer, with minimum sitting below 50% X-isomer for the Cl1J and above 50% X-isomer for the Cl2J constructs, consistent with what was observed in the electron density maps.

To determine the actual observed %X-isomer (%$IsoX_{obs}$), the linear portions of the titration data on either side of the minima were fitted with simple linear models using the KaleidaGraph program, and solving for the point of intersection for the equations of the two lines on either side of the R_{free} minimum (Fig. 2.6). The corresponding %H-isomer observed (%$IsoH_{obs}$) was taken as 100%–%$IsoX_{obs}$. The uncertainty in each %$IsoX_{obs}$ was estimated by propagating the errors on the slopes and *y*-intercepts of the linear equations used to fit the titration curves.

2.2.2 Estimating energy of Cl-bonds

With the %$IsoX_{obs}$ and %$IsoH_{obs}$ now determined, the $\Delta E_{IsoX–IsoH}$ can be estimated for each DNA junction[11] according to Eq. (1). From $\Delta E_{IsoX–IsoH}$ and the ratio of Cl-bonds to H-bonds in the Cl1J and Cl2J constructs, we derived an approach to explicitly determine the absolute energies of the Cl-bond and the competing H-bond. This analysis starts with the assumption that the Cl-bonds in both constructs are nearly identical in structure and energy.

The structures are indeed nearly identical (Cl•••O distances differ by <2.5% and C–Cl•••O angles differ by ~4% from each other). We can, therefore, define the observed $\Delta E_{IsoX–IsoH}$ in

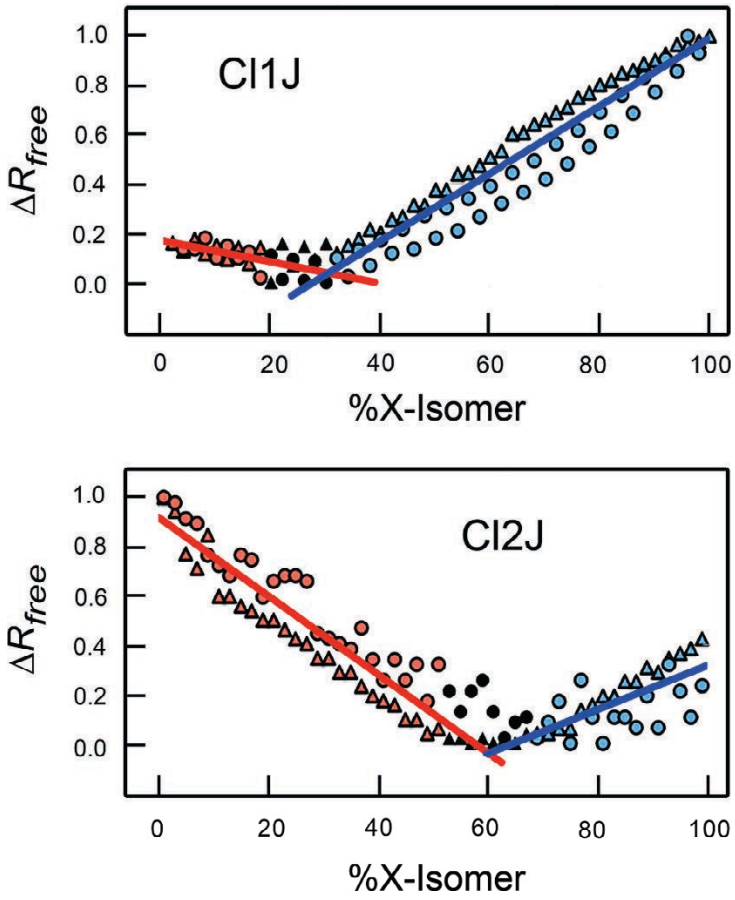


Figure 2.6: Occupancy titrations of Cl1J and Cl2J DNA constructs. The titrations show minima for ΔR_{free} values as a function of % X-isomer (note: the % X-isomer scale of this figure is reversed and the ΔR_{free} values normalized between 0 and 1 relative to that in Fig. 2.5). The % X-isomer at the ΔR_{free} minima of the two isomers of were determined by least squares fitting of the linear portions of each titration curve (red symbols represent % X-isomer values < ΔR_{free} minimum shown as, blue symbols for those > minimum, and black symbols for values not used in the calculations). The actual % X-isomer value of each construct was calculated from the slopes and y-intercepts of the two lines. The titrations repeated in the ascending (from 0% to 100% X-isomer, circles) and descending (from 100% to 0%, triangles) directions.

terms of the contributions of Cl-bonds and H-bonds for the Cl1J ($\Delta E_{IsoX–IsoH(Cl1J)}$, Eq. (2)) and Cl2J ($\Delta E_{IsoX–IsoH(Cl2J)}$, Eq. (3)) constructs.

Eq. (2) $$\Delta E_{IsoX-IsoH(Cl1J)} = E_{Cl-bond} - 2E_{H-bond}$$

Eq. (3) $$\Delta E_{IsoX-IsoH(Cl2J)} = 2E_{Cl-bond} - 2E_{H-bond}$$

It is now a simple matter to solve for the two unknowns ($E_{Cl\text{-}bond}$ and $E_{H\text{-}bond}$) using the two simultaneous Eqs. (2) and (3). At this point, you recognize the flaw in this simple setup, which is that $\%IsoX_{obs}$ likely does not reflect the actual % X-isomer for the singly chlorinated junction in the Cl1J construct. That is because the junction, when annealed from two H-bonding strands, one potential Cl-bond forming $^{Cl}U_7$ containing strand, and one nonhalogenated U_7 containing strand would statistically form three different species: the completely unchlorinated junction that can only form two H- bonds (H2J), the fully chlorinated junction that is identical to Cl2J, and the singly chlorinated Cl1J junction, in ratios of 1:1:2. If we normalize the statistical ratios, we see that the H2J:Cl1J:Cl2J components contribute in 0.25:0.5:0.25 fractional ratios to the overall percent of X-isomer observed ($\%IsoX_{obs}$) for the Cl1J construct Eq. (4). The resulting contributions to the observed energy difference between the X- and H-isomers will also be defined by these fractional ratios Eq. (5).

Eq. (4) $$\%IsoX_{obs} = 0.25\left(\%X_{Cl2J}\right) + 0.5\left(\%X_{Cl1J}\right) + 0.25\left(\%X_{H2J}\right)$$

Eq. (5) $$\Delta E_{IsoX-IsoH(Cl1J)} = 0.25\left(\Delta E_{IsoX-IsoH(Cl2J)}\right) + 0.5\left(\Delta E_{IsoX-IsoH(Cl1J)}\right) + 0.25\left(\Delta E_{IsoX-IsoH(H2J)}\right)$$

$$= 0.25(2E_{Cl-bond} - 2E_{H-bond}) + 0.5(E_{Cl-bond} - 2E_{H-bond}) + 0.25(0 - 2E_{H-bond})$$

Eq. (5) reduces to exactly Eq. (2), which means that we were in fact correct in assuming that values for $E_{Cl\text{-}bond}$ and $E_{H\text{-}bond}$ could be solved using the two simultaneous equations of 2 and

3. The resulting values for $E_{Cl\text{-}bond}$ and $E_{H\text{-}bond}$ of -0.79 ± 0.12 kcal/mol and -0.64 ± 0.07 kcal/mol, respectively, indicate that the Cl-bond is very similar in energy to that of the competing H-bond, yet slightly more negative as expected from analysis of the electron density maps.

2.2.3 Estimating energies of F-bonds, Br-bonds, and I-bonds

With the energy of the H-bond determined, we can use the $\%IsoX_{obs}/\%IsoH_{obs}$ and resulting $\Delta E_{IsoX\text{-}IsoH}$ for each of the other halogens to determine their X-bonding potential. Using the equivalent of Eq. (1), the stabilizing energies were estimated to be -0.52 ± 0.06 kcal/mol for the F-bond, -2.28 ± 0.11 kcal/mol for the Br-bond in the Br1J construct, and at least -2.1 kcal/mol for the I-bond in the I1J construct[11]. We could not estimate the energy of the shorter Br-bond of Br2J nor the I-bond in I2J, because their $\%IsoX_{obs}$ was ≥95% or, within the error of the method, essentially entirely X-isomer[9,11]. To determine these values, and as a validation of the assumptions and energies from the crystallographic assay, we developed a differential scanning calorimetry (DSC) method to determine the Br-bond and I-bond energies in solution from the Br2J and I2J constructs.

The DSC assay compares the thermodynamics of melting the DNA constructs in their four-stranded junction form relative to their duplex forms. The assumption is that this comparison eliminates the contributions of base stacking along the DNA helical arms, and localizes the observed differences in energy between junction and duplex ($\Delta E_{J\text{-}D}$) to the stabilizing interactions at the $N_6Py_7C_8$-trinucleotide core of the junction. The $\Delta E_{J\text{-}D}$ for the H-bonded H2J junction construct is subtracted from the $\Delta E_{J\text{-}D}$ for the Br2J and I2J junctions, which further eliminates the contributions from the N_6 and C_8 nucleotides of the core, localizing the differences in the observed energies to specifically the interactions at Py_7. Since the H2J does not have the potential for any stabilizing X-bond interactions at Py_7, the resulting energies can be

attributed specifically to the Br-bond and I-bond of the Br2J and I2J constructs. We recognize that the halogens may also affect the base stacking interactions at Py_7; however, the halogens in these structures extend beyond the neighboring stacks and, therefore, should not be a dominant factor. In addition, these two large halogens are also known to be hydrophobic. To estimate the contribution of solvent effects on the solution-state energies, we calculated the solvent free energies (SFE) from the solvent accessible surfaces and atomic solvation parameters[19–21] for each halogen in models built for the X- and H-isomeric forms[9,11]. The resulting SFE indicates that burying the bromine or iodine into their respective junctions would contribute on average -0.4 kcal/mol to the stabilization of the X-isomers[11]. Thus, the DSC assay should provide a reasonable estimate for the thermodynamic terms responsible for the stabilizing potentials of the Br-bond and I-bond in solution.

The DSC melting profiles were measured using a TA Instruments Nano DSC with the pressure held constant at 3.0 atm; thus, the energy measured is the enthalpies of melting (ΔH_m). Each DNA sample was run against buffer in a heating cycle from 0°C to 90°C at a scanning rate of 1°C/min with an equilibrium time of 900 s. The analyses were repeated at least three times for each sample. Data were analyzed using the NanoAnalyze software from TA Instruments (TA Instruments, New Castle, DE) in the same manner as previously published[12], with the best fit determined according to the standard deviation of the fit.

Each DNA construct was analyzed at multiple DNA concentrations. At low DNA concentrations, DSC profiles were fit using a standard single-component, two-state model for a duplex melting to single-stranded DNA (Fig. 2.7A). As the DNA concentration increased, the profiles became broader and shifted to higher melting temperature (T_m), indicating the formation of the junction—this is consistent with the concentration dependent formation of junctions for this type

of sequence in solution[12,17]. The high concentration profiles were best fit applying a two-component, two-state model (Fig. 2.7B). The lower T_m component was similar to that of the DNA duplex, indicating that, for these constructs the duplex and junction forms coexist, and the equilibrium between the forms is slow[17] compared to the rate of melting. The ΔH_m for melting the duplex form of each construct was taken as the average ΔH_m for DNA concentrations from 15 to 20 μM along the low temperature component from the two component analysis of data at [DNA] >100 μM (Table 2.2). The ΔH_m of the junction form of each construct was taken as the average of the higher temperature component of the analyses.

The ΔH_m at the associated T_m allowed calculation of the ΔS_m at the T_m according to the Eq. (6), assuming that $\Delta G_m = 0$ at the melting temperature. All ΔH_m and ΔS_m were extrapolated to a reference temperature of 25°C (ΔH^{25C} and ΔS^{25C}) using the heat capacities for each melting profile. This then allowed the calculation of ΔG^{25C}.

Eq. (6) $$\Delta S_m = \frac{\Delta H_m}{T_m}$$

The energy differences between the junction and duplex forms at the reference temperature ($\Delta H_{J\text{-}D}{}^{25°C}$, $\Delta S_{J\text{-}D}{}^{25°C}$, $\Delta G_{J\text{-}D}{}^{25°C}$) were simply calculated by subtracting the associated energy values of the duplex from the junction forms, normalized to molar concentrations of duplex. For each construct analyzed (H2J, Br2J, and I2J), the trinucleotide core of the junction was seen to be highly stabilizing in terms of the enthalpy, but these interactions resulted in significant loss of entropy, as one would expect from the concept of enthalpy–entropy compensation.

When the $\Delta H_{J\text{-}D}{}^{25°C}$, $\Delta S_{J\text{-}D}{}^{25°C}$, $\Delta G_{J\text{-}D}{}^{25°C}$ of the non-X-bonding H2J construct are subtracted from those of the Br-bond stabilized Br2J and I-bond stabilized I2J constructs, we can

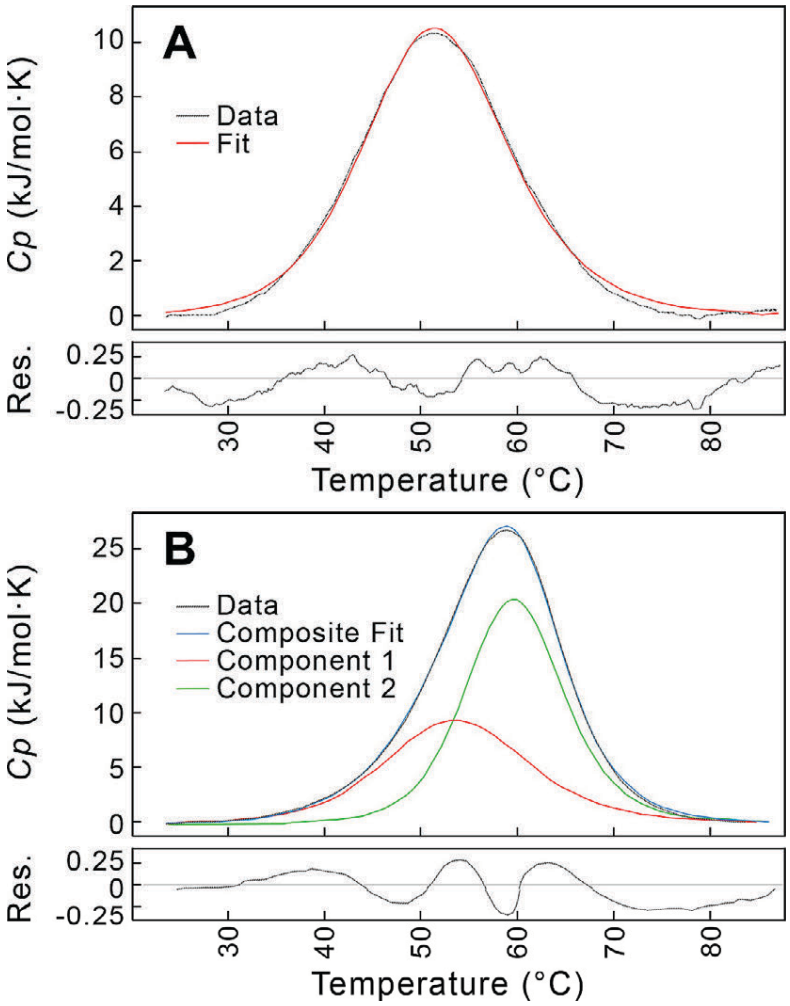


Figure 2.7: Differential scanning calorimetry (DSC) traces for melting of Br2J construct as a function of DNA concentration. **A.** Duplex melting at low DNA concentration. The Br2J construct, annealed at a concentration of 40 μM, is melted and the heat capacity under constant 3 atm pressure (C_p) monitored as a function of temperature (dotted black line). The DSC trace can be fit to a single component, two-state transition model (red curve), with a melting temperature (T_m) at 51.7°C and ΔH_m = 42.2 kJ/mol (ΔH_m values are normalized per mole of DNA duplex). This was interpreted as the melting of the DNA as a duplex to single-strands. The residual (Res.) between the data and the fitted curve is shown in the panel below. **B.** Duplex and junction melting at high DNA concentration. The DSC trace of the Br2J construct, annealed at a concentration of 300 μM, is best fit by a two component, two-state transition model. The first component of this model (red curve) has a T_m = 53.8°C and ΔH_m = 43.5 kJ/mol, and therefore was interpreted as the melting of the duplex DNA. The second component (green curve) has a higher T_m = 59.8°C and ΔH_m = 65.3 kcal/mol, and was interpreted as the melting of the junction form of the DNA. The residual (Res.) between the data and the composite of the two component fitted curves (blue curve) is shown in the panel below.

Table 2.2: Thermodynamics of melting X- and H-bonded DNA junctions. The melting of each DNA construct was monitored by DSC for the melting temperature (T_m) and the enthalpy of melting (ΔH_m), with the entropy of melting (ΔS_m) calculated from ΔH_m and the T_m. The low T_m parameters were assigned to the duplex and high T_m component to the junction forms of the DNA, and each parameter normalized to the molar concentration of duplex DNA. The melting enthalpy and entropy are extrapolated to the reference temperature of 25°C applying the heat capacity of each construct to determine the difference in enthalpy ($\Delta H_{(J-D)}^{25°C}$) and entropy ($\Delta S_{(J-D)}^{25°C}$) between the junction and duplex forms, and the associated Gibb's free energy ($\Delta G_{(J-D)}^{25°C}$). The parameters for the stability ($\Delta\Delta H_{(X-H)}^{25°C}$, $\Delta\Delta S_{(X-H)}^{25°C}$, and $\Delta\Delta G_{(X-H)}^{25°C}$) of the X-bonds in the Br2J and I2J constructs were determined by subtracting the melting parameters for each from those of H2J.

Construct	H2J	Br2J	I2J
Thermodynamic parameters for melting of duplex and junction forms			
T_m (Duplex)	49.9 ± 0.3°C	50.5 ± 0.2°C	52.7 ± 0.3°C
T_m (Junction)	56.5 ± 0.2°C	58.3 ± 0.3°C	58.6 ± 0.1°C
$\Delta H_{(J-D)}^{25\,°C}$ (kcal/mol)	13.1 ± 0.9	16.7 ± 0.9	19.0 ± 0.6
$\Delta S_{(J-D)}^{25\,°C}$ (cal/mol K)	43 ± 3	39 ± 3	55 ± 2
$\Delta G_{(J-D)}^{25\,°C}$ (kcal/mol)	0.3 ± 1.3	5.1 ± 1.3	-2.6 ± 0.8
Thermodynamic stability of X-bonds versus H-bonds (normalized)			
$\Delta\Delta H_{(X-H)}^{25\,°C}$ (kcal/mol)	NA	-3.6 ± 1.3	-5.9 ± 1.1
$\Delta\Delta S_{(X-H)}^{25\,°C}$ (cal/mol K)	NA	4.2 ± 4	-12.0 ± 4
$\Delta\Delta G_{(X-H)}^{25\,°C}$ (kcal/mol)	NA	-4.9 ± 1.8	-2.3 ± 1.6

estimate the thermodynamic terms of each X-bond type. For the I-bond, the resulting $\Delta\Delta G_{X\text{-}H}^{25°C}$ of -2.3 kcal/mol is very similar to the energy of stabilization $E_{X\text{-}bond} \leq$ -2.1 kcal/mol estimated from the crystallographic assay. The $\Delta\Delta G_{X\text{-}H}^{25°C}$ of -4.8 kcal/mol for the Br-bond in solution is about twice that of the $E_{X\text{-}bond} \leq$ -2.28 kcal/mol in the crystal system; however, we must recall that the $E_{X\text{-}bond}$ from the crystal assay was for the longer X-bond of the Br1J construct, while the solution studies assayed the shorter X-bond of the Br2J construct. This shorter Br-bond was expected to be about twice as stabilizing, which is what is observed when comparing the two assays. Thus, we can conclude that the energies determined from the crystallographic analyses reflect the free energies of these X-bonds in solution.

2.3 Concluding remarks

We presented here descriptions of a system where energies of interactions can be derived from X-ray diffraction studies on single crystals. In this setup, we exploit a DNA system to determine the geometries and the associated energies of halogen bonds. The power of this assay is that they uniquely provide a direct link between the atomic level structures and their associated energies of the molecular interactions—no other approach provides such a concerted characterization of structure-energy relationships in biomolecules.

We started with the statement that one ‘‘cannot derive thermodynamic properties from crystal structures.’’ From the X-bond study, we see that the thermodynamic properties of DNA junctions determined in crystals reflect the free energies of interaction in solution. Thus, we have validated our assumption that the distribution of conformations in a crystal mirrors that of the crystallization solution. Furthermore, the initial statement certainly cannot be an absolute truism, specifically for systems where conformations are discrete and when the molecular interactions that confer these conformations are distant from and thus unaffected by crystal lattice contacts.

This differs from an earlier study, where we had taken advantage of the crystal lattice contacts to determine the structures and relative stabilities of reverse base pairs in DNA[22]. The two studies together (on the energetics of reverse base pairs and of halogen bonds), however, demonstrate that one can, in fact, derive thermodynamic properties of a macromolecule from single crystal X-ray diffraction studies.

We do not, however, claim here that every crystal system can be used to characterize energetic properties. The thermodynamic questions must be very well defined and focused. In our studies, we designed experiments that took advantage of the physicochemical features that were unique to their respective crystal systems. In short, we took what the systems gave us, and no more. The method presented here emphasizes the complementary nature of crystallography with various other solid-state or solution-state techniques. In short, when the experiments are properly designed, such crystallographic/thermodynamic studies can be very revealing, and have been instrumental in helping us to develop and test empirical force fields for various classes of molecular interactions[23].

Finally, the specific DNA junction system described here could further be adapted to study effects that would be directly related to the Holliday junction and recombination. The additional applications could include, for example, the effects of sequence[16], epigenetic markers such as methylation[24], and miss-pairs on the stability of the junction[25].

2.3.1 Funding

This work was funded in part by a grant from the National Science Foundation (CHE-1152494) and from Colorado State University.

REFERENCES

(1) Hassel, O. (1970) Structural Aspects of Inter-Atomic Charge Transfer Bonding. *Science (80-.). 170*, 497–502.

(2) Metrangolo, P., Neukirch, H., Pilati, T., and Resnati, G. (2005) Halogen bonding based recognition processes: A world parallel to hydrogen bonding. *Acc. Chem. Res. 38*, 386–395.

(3) Auffinger, P., Hays, F. A., Westhof, E., and Ho, P. S. (2004) Halogen bonds in biological molecules. *Proc. Natl. Acad. Sci. U. S. A. 101*, 16789–94.

(4) Voth, A. R., and Ho, P. S. (2007) The role of halogen bonding in inhibitor recognition and binding by protein kinases. *Curr. Top. Med. Chem. 7*, 1336–48.

(5) Xu, Z., Liu, Z., Chen, T., Chen, T., Wang, Z., Tian, G., Shi, J., Wang, X., Lu, Y., Yan, X., Wang, G., Jiang, H., Chen, K., Wang, S., Xu, Y., Shen, J., and Zhu, W. (2011) Utilization of halogen bond in lead optimization: a case study of rational design of potent phosphodiesterase type 5 (PDE5) inhibitors. *J. Med. Chem. 54*, 5607–11.

(6) Lu, Y., Shi, T., Wang, Y., Yang, H., Yan, X., Luo, X., Jiang, H., and Zhu, W. (2009) Halogen bonding--a novel interaction for rational drug design? *J. Med. Chem. 52*, 2854–2862.

(7) Ibrahim, M. A. A. (2011) Molecular Mechanical Study of Halogen Bonding in Drug Discovery.

(8) Scholfield, M. R., Zanden, C. M. Vander, Carter, M., and Ho, P. S. (2013) Halogen bonding (X-bonding): a biological perspective. *Protein Sci. 22*, 139–52.

(9) Voth, A. R., Hays, F. A., and Ho, P. S. (2007) Directing macromolecular conformation through halogen bonds. *Proc. Natl. Acad. Sci. U. S. A. 104*, 6188–93.

(10) Clark, T., Hennemann, M., Murray, J. S., and Politzer, P. (2007) Halogen bonding: the

sigma-hole. Proceedings of "Modeling interactions in biomolecules II", Prague, September 5th-9th, 2005. *J. Mol. Model. 13*, 291–6.

(11) Carter, M., Voth, A. R., Scholfield, M. R., Rummel, B., Sowers, L. C., and Ho, P. S. (2013) Enthalpy–Entropy Compensation in Biomolecular Halogen Bonds Measured in DNA Junctions. *Biochemistry 52*, 4891–4903.

(12) Carter, M., and Ho, P. S. (2011) Assaying the Energies of Biological Halogen Bonds. *Cryst. Growth Des. 11*, 5087–5095.

(13) Lilley, D. M. J. (2000) Structures of helical junctions in nucleic acids. *Q. Rev. Biophys. 33*, 109–159.

(14) Eichman, B. F., Vargason, J. M., Mooers, B. H. M., and Ho, P. S. (2000) The Holliday junction in an inverted repeat DNA sequence: Sequence effects on the structure of four-way junctions. *Proc. Natl. Acad. Sci. 97*, 3971–3976.

(15) Eichman, B. F., Ortiz-Lombardía, M., Aymamí, J., Coll, M., and Ho, P. S. (2002) The inherent properties of DNA four-way junctions: Comparing the crystal structures of holliday junctions. *J. Mol. Biol. 320*, 1037–1051.

(16) Hays, F. A., Teegarden, A., Jones, Z. J. R., 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*, 7157–62.

(17) Hays, F. A., Schirf, V., Ho, P. S., and Demeler, B. (2006) Solution formation of Holliday junctions in inverted-repeat DNA sequences. *Biochemistry 45*, 2467–2471.

(18) Brünger, A. T. (1992) Free R-Value - A Novel Statistical Quantity for Assessing the Accuracy of Crystal-Structures. *Nature 355*, 472–475.

(19) Eisenberg, D., and McLachlan, A. D. (1986) Solvation energy in protein folding and

binding. *Nature 319*, 199–203.

(20) Kagawa, K., Howell, M. L., Tseng, K., and Ho, P. S. (1993) Effects of base substituents on the hydration of B- and ZDNA: Correlations to the B- to Z-DNA transition. *Nucleic Acids Res. 21*, 5978–5986.

(21) Kagawa, T. F., Stoddard, D., Zhou, G. W., and Ho, P. S. (1989) Quantitative analysis of DNA secondary structure from solvent-accessible surfaces: the B- to Z-DNA transition as a model. *Biochemistry 28*, 6642–51.

(22) Mooers, B. H. M., Eichman, B. F., and Ho, P. S. (1997) The Structures and Relative Stabilities of d(GÁ G) Reverse Hoogsteen, d(G ÁT) Reverse Wobble, and d(G Á C) Reverse Watson-Crick Base-pairs in DNA Crystals. *J. Mol. Biol. 269*, 796–810.

(23) Carter, M., Rappé, A. K., and Ho, P. S. (2012) Scalable Anisotropic Shape and Electrostatic Models for Biological Bromine Halogen Bonds. *J. Chem. Theory Comput. 8*, 2461–2473.

(24) Vargason, J. M., and Shing Ho, P. (2002) The effect of cytosine methylation on the structure and geometry of the holliday junction: The structure of d(CCGGTACm5CGG) at 1.5 ?? resolution. *J. Biol. Chem. 277*, 21041–21049.

(25) Ortiz-Lombardia, M., Gonzalez, A., Eritja, R., Aymami, J., Azorin, F., and Coll, M. (1999) Crystal Structure of a {DNA} {Holliday} Junction. *Natsb 6*, 913–917.

CHAPTER 3

EFFECT OF HYDROXYMETHYLCYTOSINE ON THE STRUCTURE AND STABILITY OF HOLLIDAY JUNCTIONS[2]

5-Hydroxymethylcytosine (^{5hm}C) is an epigenetic marker that has recently been shown to promote homologous recombination (HR). In this study, we determine the effects of ^{5hm}C on the structure, thermodynamics, and conformational dynamics of the Holliday junction (the four-stranded DNA intermediate associated with HR) in its native stacked-X form. The hydroxymethyl and the control methyl substituents are placed in the context of an amphimorphic G^{x}CC trinucleotide core sequence (where ^{x}C is C, ^{5hm}C, or the methylated ^{5m}C), which is part of a sequence also recognized by endonuclease G to promote HR. The hydroxymethyl group of the ^{5hm}C junction adopts two distinct rotational conformations, with an in-base-plane form being dominant over the competing out-of-plane rotamer that has typically been seen in duplex structures. The in-plane rotamer is stabilized by a more stable intramolecular hydrogen bond to the junction backbone. Stabilizing hydrogen bonds (H-bonds) formed by the hydroxyl substituent in ^{5hm}C or from a bridging water in the ^{5m}C structure provide approximately 1.5−2 kcal/mol per interaction of stability to the junction, which is mostly off set by entropy compensation, thereby leaving the overall stability of the G^{5hm}CC and G^{5m}CC constructs similar to that of the GCC core. Thus, both methyl and hydroxymethyl modifications are accommodated without disrupting the structure or stability of the Holliday junction. Both ^{5hm}C and ^{5m}C are shown to open the structure to make the

[2] Previously published as: "Effect of Hydroxymethylcytosine on the Structure and Stability of Holliday Junctions" Vander Zanden, C.M., Rowe, R.K., Broad, A.J., Robertson, A.B., Ho, P.S. *Biochemistry* 55 (2016) 5781-9.

junction core more accessible. The overall consequences of incorporating ^{5hm}C into a DNA junction are thus discussed in the context of the specificity in protein recognition of the hydroxymethyl substituent through direct and indirect readout mechanisms.

3.1 Introduction

Epigenetic modifications to DNA are now recognized as a complementary mechanism for expanding and regulating genomic information. For example, 5-methylcytosine (^{5m}C) serves as a mark to target gene silencing in eukaryotes: misregulation of specific gene silencing events can be hugely detrimental, even fatal, to an organism's development[1]. A complex system of proteins help to regulate ^{5m}C levels, including modifying the DNA *de novo* in response to an external stimulus[2] and maintenance inheritance of the methylation fingerprint from a previous generation of cells[3]. Although other DNA modifications, adenine methylation[4] and N4-methylcytosine[5], are found in genomes[6], most of the epigenetic research on the mammalian genome has been focused on determining the effects and regulatory mechanisms of ^{5m}C. We show here that 5-hydroxymethylcytosine (^{5hm}C), an epigenetic marker recently shown to promote recombination[7], affects the structure and stability of the DNA Holliday junction.

5-Hydroxymethylcytosine (^{5hm}C) is a modified base that was first reported in animal cells in 1972[8], but it has recently seen renewed interest when Heintz observed its presence in Purkinje neurons[9]. The ten-eleven translocation (Tet) family of dioxygenases generates ^{5hm}C in the cell by oxidizing 5-methylcytosine (^{5m}C)[10], which can further convert ^{5hm}C to increasingly oxidized formyl- and carboxyl-cytosines[11]. Standard bisulfite sequencing analysis cannot distinguish between ^{5hm}C and ^{5m}C[12]. New methods, including Tet-assisted bisulfite sequencing and others, have allowed ^{5hm}C to be mapped in genomic and physiological contexts[13–17], which has resulted in a new surge of interest in the effects of ^{5hm}C

on biological processes (~94% of all ^{5hm}C-related papers have been published since 2009, according to the Web of Science[18]). The initial mapping of ^{5hm}C onto specific genomic regions, tissue types, and development stages in both normal and cancerous cells[19–24] has implicated ^{5hm}C's involvement in gene regulation[25–27], in brain development[19,28,29], in regulation of ^{5m}C levels[30], in embryonic development[10,21,26], and potentially in regulating homologous recombination (HR) events[7,26].

The evidence of its role in HR came initially from the observation that ^{5hm}C's were enriched with GC-rich regions[26], which are associated with recombination hot spots[15,31]. This theory was further strengthened recently by the studies of Robertson et al.[7], which demonstrated that ^{5hm}C promotes homologous recombination in a sequence-dependent manner. This effect was seen to be mediated by endonuclease G (EndoG), specifically through recognition and binding of the sequence 5′-GGGG^{5hm}CCAG-3′ / 5′ -CTGGCCCC-3′ to induce double-strand breaks that then trigger the actions of the cell's recombination machinery. The question we raise here is whether and how ^{5hm}C affects the structure and stability of the Holliday junction, the four-stranded DNA structure that is the intermediate formed during homologous recombination events[32].

The formation of Holliday junctions has been shown to be sequence-dependent in crystals[33] and in solution[34]. Junctions exist in two functional forms: the open-X and stacked-X structures[35]. The open-X form takes a classical "cruciform DNA" shape and allows the junction to isoenergetically migrate along stretches of DNA sequence during HR. This form of the junction is seen under low-salt conditions in DNA only constructs, or in complex with proteins that require migration of the junction to locate a specific recombination site (as in the RuvABC DNA repair system[36]).

The stacked-X junction is essentially two continuous duplexes interrupted by the crossovers that connect the adjacent duplexes. The stacked-X form is observed in DNAs under high-salt conditions and, because it is topologically locked and cannot migrate, is seen in complexes with sequence independent resolvases (such as the T7 bacteriophage endonuclease I[37] or the T4 bacteriophage endonuclease VII[38]). The crystal structures of DNA only constructs have revealed that the stacked-X junction is stabilized by a trinucleotide core, a three-nucleotide sequence that defines the crossover point between adjacent duplexes of the junction. The sequence preference within this trinucleotide core is A > G > C at the first position, C > T at the second, and C required at the third[33]. The specificity at the second and third positions is attributed to a unique set of hydrogen bonds (H-bonds) that form between the amino group of the cytosine bases and oxygens from the adjacent residue's backbone phosphate group, helping to mitigate the inter-phosphate electrostatic repulsion along the DNA backbone as it makes the tight U-turn that connects the two adjoining duplexes of the junction.

It was interesting to us that the Endo G recognition sequence identified by Robertson et al.[7] contained the sequence motif G^{5hm}CC, a ^{5hm}C-modified version of the GCC trinucleotide core shown previously to stabilize junctions[33]. This hydroxymethyl group is potentially positioned to displace the H-bond that stabilizes the stacked-X junction structure[39]. The question we posed is whether the hydroxymethyl group introduced at this position would sterically interfere with this important interaction or, because it is an H-bond donor itself, supplant this interaction, and how these perturbations would affect the conformation and stability of the junction as a whole. As a control, we compare the effects of ^{5hm}C with the methylated variant (^{5m}C), which would have similar steric effects but cannot form an H-bond.

3.2 Methods

Oligonucleotides: DNAs were designed as self-complementary decanucleotide sequences in the 5′ -CCGGCGXCGG- 3′ motif (X is C, ^{5m}C, or ^{5hm}C), previously shown to form junctions in the presence of monovalent and divalent cations[33].

Oligonucleotides were purchased from Midland Certified Reagent Co. with the 5′-dimethoxytrityl (DMT) protecting group intact and remaining attached to the CPG solid support bead to facilitate purification. The CPG was removed by suspension in ammonium hydroxide, and the full-length products were isolated by reverse phase high-performance liquid chromatography on a C18 column, taking advantage of the additional hydrophobicity of the 5′-DMT. The DMT group was cleaved by resuspending the oligos in 3% acetic acid and the final product desalted by size exclusion chromatography off a Sephadex G-25 column.

Crystallography and Structure Analysis: Crystals were grown in sitting drop trays, with 8−10 μL sample volumes containing 0.78 mM DNA (not annealed), 25 mM sodium cacodylate (pH 7.0), calcium chloride (ranging from 1 to 15 mM), and spermine (ranging from 0.1 to 2.0 mM), and equilibrated against a reservoir solution of 25% 2-methyl-2,4-pentanediol (MPD). These crystallization drop conditions were chosen for screening because of their propensity to yield both duplex and junction DNA crystals[33].

Data were collected using a Rigaku Compact Home Lab equipped with a PILATUS detector; HLK3000[40] was used to index, integrate, and scale the data. The structures were determined by molecular replacement [using the GCC core junction as the starting search model, Protein Data Bank (PDB) 1P4Y[33]] and subsequently refined using Phenix[41]. Standard Phenix occupancy refinement routines were used to determine the occupancy of each rotamer for the ^{5hm}C hydroxyl group (deposited under PDB entries 5DSA and 5DSB for the ^{5m}C and

^{5hm}C structures, respectively). DNA structure measurements (rise, twist, slide, etc.) were performed with the CURVES+ DNA structure analysis program[42], and junction structure parameters (J_{roll} and J_{twist}) were calculated according to the methods described by Watson *et al*[43].

Melting Profiles Determined by Differential Scanning Calorimetry (DSC): DSC samples were prepared by annealing 25 μM DNA in 15 mM calcium chloride and 50 mM sodium cacodylate (pH 7.0) at 90 °C for 20 min and allowed to slowly cool over 2h. The DNA melting data were collected using a TA Instruments Nano DSC apparatus with equilibration for 900s and scanning from 5 to 105 °C at a rate of 1 °C/min at a constant pressure of 3.0 atm. Melting temperatures (T_m) and enthalpies of melting (ΔH_m) were determined by fitting the data with TA Nano Analyze software using a two-component (junction and duplex), two-state scaled model. Each construct was measured through at least 18 replicates. Melting energies were extrapolated to a standard temperature of 25 °C, and the duplex melting energies were subtracted from the junction to determine the stabilization energy of the junction core[44,45].

Quantum Mechanical (QM) Calculations: QM calculations were performed using Gaussian09[46] at the Møller−Plesset 2 (MP2) level, using the 6-31++G** basis set. Cyclohexane (ε = 2) was chosen as the solvent to mimic the semi-sequestered and hydrophobic environment of the junction core, and a counterpoise (BSSE) correction was applied from a gas phase calculation. Geometry scanning calculations (5° increments) were first performed on the in-context dinucleotide (G_6-$^{5hm}C_7$) to determine the minimum energy orientation of the hydrogen from the ^{5hm}C' s hydroxyl group for each isomer resolved in the crystal structure. To determine the relative rotamer stability of the crystallographic structures, energy calculations were performed on the isolated ^{5hm}C bases, including the optimized

hydrogen positions for the in-context crystal structure (Fig. 3). Dimethyl phosphate was chosen to mimic the DNA backbone in the calculation to determine the $Phos_6$—$^{5hm}C_7$ hydrogen bond (H-bond) energy for each ^{5hm}C rotamer.

3.3 Results and Discussion

The initial premise of this study is that the hydroxyl group of ^{5hm}C could form an H-bond to supplant or supplement an interaction that had previously been shown to stabilize and, thus, infer sequence specificity to the four-stranded Holliday junction. We have determined the structural effects of this epigenetic modification by determining the crystallographic structure of ^{5hm}C in the self-complementary sequence d(CCGGC***G^{5hm}CC***GG). We designed this sequence motif around a GCC trinucleotide core (in bold italics), which had previously been shown to be amphimorphic (capable of forming either B-DNA duplexes or four-stranded HJs, depending on cations)[33]; thus, this sequence, as opposed to a strictly junction-forming ACC core, would be very sensitive to any destabilizing effects of the substituents on the junction. In addition, we apply differential scanning calorimetry to correlate the structural effects on the overall stability of the junction and interpret these energies in terms of contributions of the molecular interactions on the enthalpic and entropic effects locally and globally. A parallel study of the methylated sequence d(CCGGC***G^{5m}CC***GG) allowed us to distinguish between contributions from steric and hydrogen bonding interactions.

3.3.1 ^{5hm}C and ^{5m}C modifications are structurally accommodated in the Holliday junction core.

The first observation from the overall crystal structures of d-(CCGGC***G^{5hm}CC***GG) and d(CCGGC***G^{5m}CC***GG) is that both the bulky methyl and hydroxymethyl substituent

groups are accommodated at the key stabilization trinucleotide of the stacked-X form of the HJ (Table 1 and Fig. 3.1). Both sequences conform to the overall conformation of the stacked-X junction as seen in previous crystal structures[33], with the DNA forming two sets of nearly continuous double helices, interrupted by the crossing of the phosphodiester bond at nucleotide 6, which connects the helices to form the four-stranded junction (Fig. 3.1). The assembly of these self-complementary sequences results in junctions in which the methyl or hydroxymethyl modifications sit at two unique nucleotide positions and, thus, experience two unique structural environments. In each structure, either a ^{5m}C or a ^{5hm}C base sits on a continuous strand of an uninterrupted B-type duplex region, while the second similarly modified base sits at the crossing strand that joins the two duplexes of a junction. The two positions allow us to compare and contrast the effects of each modification in the conformation of the HJ relative to that of a standard B-DNA duplex within the same structure.

The global structures of HJs are described by the geometric relationships between the two sets of interconnected double helices (Table 2, where J_{twist} and J_{roll} define the angular relationships of the helical axes and in the plane perpendicular to their axes, respectively)[43], which reflect the accessibility of the junction crossover to the environment. A comparison of G^{5hm}CC and G^{5m}CC to the parent GCC[33] sequences (referring to the core trinucleotide of each sequence) shows progressively larger J_{twist} and J_{roll} values as the size of the substituent group (H to CH_3 to CH_2OH) increases, resulting in a more open and potentially more accessible overall structure of the junction.

A more detailed analysis of the crystal structures (Table 2) shows that the methylated and hydroxymethylated bases at the N_7 position adopt geometries associated with more "ideal" B-DNA double helices than that of the unmodified junction. In particular, the ^{5hm}C·G

Table 3.1: Crystallographic parameters and refinement statistics for $G^{5hm}CC$ and $G^{5m}CC$ core Holliday junction structures.

	$G^{5hm}CC$ Core Junction	**$G^{5m}CC$ Core Junction**
Crystallographic Parameters		
PDB Entry	5DSB	5DSA
Wavelength (Å)	1.5418	1.5418
Resolution range (Å) *	30.67 - 1.50 (1.55 - 1.50)	30.59 - 1.69 (1.75 - 1.69)
Space group	C2	C2
Unit cell dimensions		
a (Å)	66.07	65.97
b (Å)	24.77	24.54
c (Å)	38.06	37.99
β (deg)	111.82	111.98
Total reflections	366673	260245
Unique reflections *	9344 (863)	6158 (511)
Multiplicity	6.7 (2.2)	6.0 (2.1)
Completeness (%) *	98 (91)	93.82 (88)
Mean I/σ(I) *	23.47 (2.04)	51.00 (3.781)
Wilson B-factor	19.65	20.53
R_{merge} *	0.029 (0.418)	0.068 (0.245)
R_{pim} *	0.036 (0.478)	0.023 (0.173)
R_{meas} *	0.041 (0.592)	0.072 (0.302)
$CC_{1/2}$	0.999	0.952
CC*	1	0.988
Refinement Statistics		
Reflections used for R-free	4.99%	4.85%
R_{cryst} *Ŧ	0.2506 (0.4023)	0.2480 (0.3408)
R_{free} *Ŧ	0.2883 (0.4302)	0.2775 (0.3648)
Number non-hydrogen atoms	507	495
RMSD for bond lengths (Å)	0.011	0.009
RMSD for bond angles (deg)	1.28	1.19
Average B-factor	25.33	27.31
Macromolecules	24.37	26.42
Solvent	29.70	31.56

**Values for the highest-resolution shell are given in parentheses*
Ŧ Values for R_{cryst} and R_{free} for the current structures are within 1 standard deviation of published B-type duplex and junction DNA structures (Hays, et al., 2005, Proc. Natl. Acad. Sci., 102: 7157-7162).

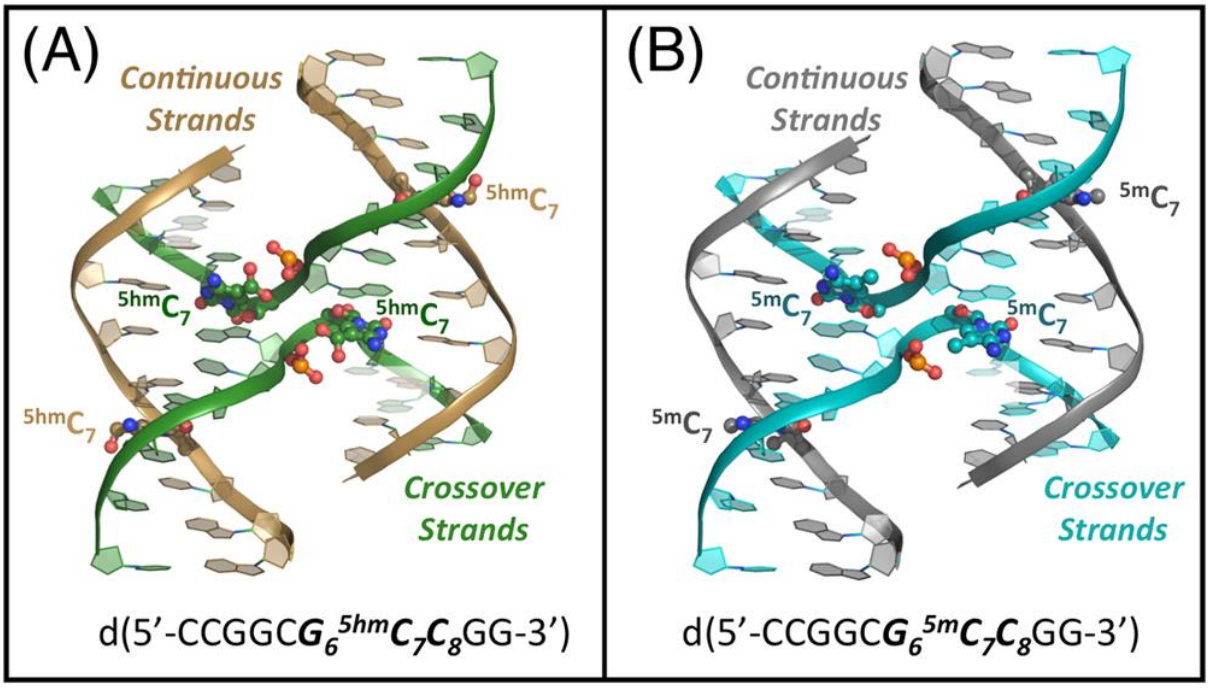
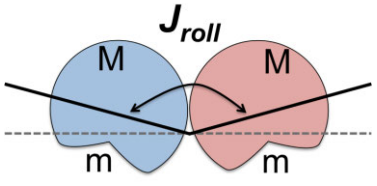
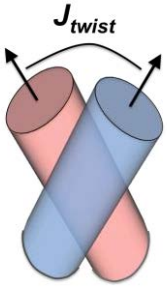


Figure 3.1: Comparison of the structures of 5-hydroxymethylcytosine (^{5hm}C) and 5-methylcytosine (^{5m}C) in DNA Holliday junctions. **(A)** Crystal structure of the hydroxymethylated sequence d- (CCGGCG$_6$^{5hm}C$_7$C$_8$GG) with the DNA backbones traced as ribbons (colored gold for the outside continuous strands and green for the junction-crossing strands). The ^{5hm}C bases, along with the phosphate groups to which they are H-bonded, are rendered as ball-and-stick models, with the carbon atoms of the nucleotides along the continuous strand colored gold and those at the junction colored green. The ^{5hm}C bases on the crossover strands have hydroxyl groups that occupy two rotamer conformations, both shown in the image. **(B)** DNA backbone of the methylated sequence d(CCGGCG$_6$^{5m}C$_7$C$_8$GG) traced as ribbons (gray along the outside continuous strands and blue along the junction-crossing strands). The ball-and-stick models of the ^{5m}C bases are colored gray on the continuous strands and blue on the junction-crossing strands.

Table 3.2: Structural parameters of GCC, $G^{5m}CC$, and $G^{5hm}CC$ core Holliday Junctions. Parameters that describe the helical structure[42] around the modified C_7 cytosine of the crossover GCC trinucleotide core are listed. The standard values for these parameters in B-DNA are shown in parentheses[33]. The overall conformation of the junction is reflected in the parameters J_{roll} and J_{twist} (schematics for these two are shown below, adapted from ref 43).

DNA Core:	***GCC***	***$G^{5m}CC$***	***$G^{5hm}CC$***
Rotational Parameters (deg)			
Helical Twist (34.7)	37.3	34.5	34.4
Propeller Twist (-12.0)	-20.1	-17.6	-14.2
Tilt (-0.62)	-0.7	4.9	5.0
Roll (1.74)	-2.7	0.9	3.6
Buckle (-0.23)	4.8	-0.8	5.3
Opening	1.6	1.6	0.2
Translational Parameters (Å)			
Rise (3.30)	3.52	3.49	3.41
Slide (0.66)	0.84	0.38	0.51
Shear	0.14	-0.10	-0.03
Stretch	0.08	-0.08	-0.07
Stagger	0.24	0.23	0.42
Shift	-0.43	0.75	0.82
Junction Parameters (deg)			
J_{roll}	135.31	143.04	150.46
J_{twist}	39.95	40.81	41.10





pair shows reduced shear, propeller twist, and opening of the base pair (Table 2), and both the methylated and hydroxymethylated base pairs show helical twists that are typical of an ~10.4 bp/turn repeat as compared to those of the overwound 9.7 bp/turn repeat for the unmodified structure. These analyses suggest that the direct H-bonding from C_7 to the phosphate of G_6 that stabilizes the unmodified GCC structure[33] (Fig. 3.2A) induces distortions to the natural geometric tendencies of stacked B-DNA base pairs, and that methylation or hydroxymethylation at this base helps to relieve some of the local conformational stress by breaking the direct H-bonding interaction of the amine.

In the ^{5m}C structure, the direct N4-amino to phosphate oxygen H-bond is now displaced by the methyl group and is replaced by a water-mediated interaction (Fig. 3.2B). In the G^{5hm}CC structure, the hydroxymethyl substituent was seen to occupy two distinct rotamer conformations (Fig. 3.2C). The major rotamer form (R1, representing approximately two-thirds of the structure) sits in the plane of the cytosine base and is H-bonded to O5′ of nucleotide G6. The minor rotamer (R2, which accounts for one-third of the structure) is rotated 112° out of plane, in a position similar to the conformation of ^{5hm}C on the outside continuous strand (Fig. 3.2D) and to previous rotamers seen in B-DNA duplexes[47,48]. In the R2 rotamer, the OH forms an H-bond to the non-linkage oxygen of the G6 phosphate and is bridged to the N4 amino group of the cytosine base by a water. Similar waters are seen coordinated to the N4 amine and hydroxyl on the continuous strand ^{5hm}C residue. Thus, both rotamer forms of ^{5hm}C place the OH in position to form an H-bond that replaces the standard interaction of the N4 amine. The question is what factors determine which conformation is dominant.

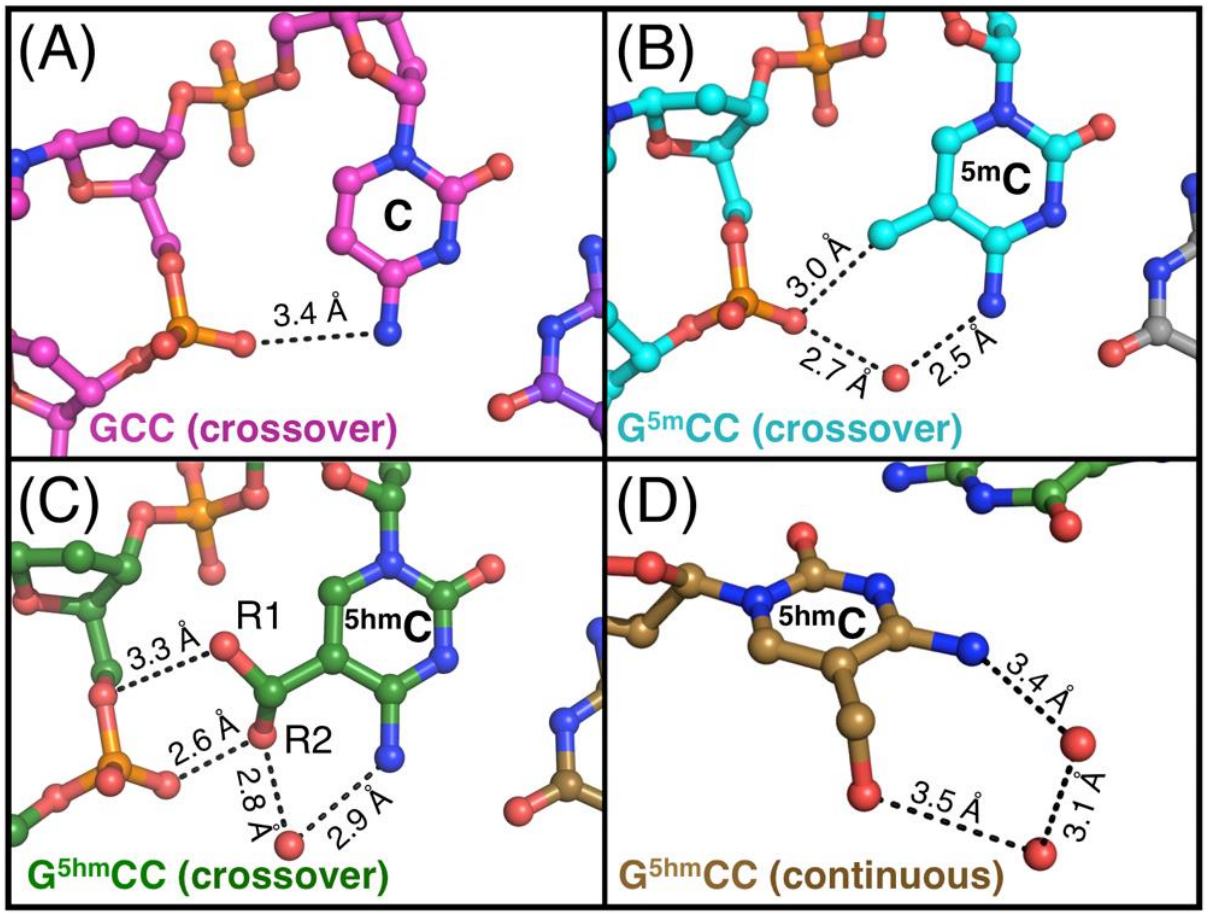


Figure 3.2: Structures of (A) GCC, (B) $G^{5m}CC$, and (C and D) $G^{5hm}CC$ trinucleotide cores of DNA junctions. **(A)** GCC core structure (PDB entry 1P4Y33) stabilized by an H-bond from the N4-amine of the C_7 base to the neighboring G_6 phosphate. No waters are observed within H-bonding distance of the bases (waters in subsequent panels shown as red spheres). **(B)** The methyl of $^{5m}C_7$ sterically interferes with the amine's direct H-bond, which is replaced by an H-bonded water that bridges the amine and the phosphate. The methyl group is within H-bonding distance of the phosphate, likely indicating a weak attractive force. **(C)** $^{5hm}C_7$ stabilizes the junction core by displacing the amine to allow the hydroxyl group to H-bond with the G_6 phosphate. The hydroxyl group is observed in two orientations, with the dominant rotamer in the plane of the base and the minor rotamer 112° out of plane. The rotamers interact with two different oxygens on the phosphate. A water is held in place by H-bonds to the hydroxyl and amine in the minor form. **(D)** ^{5hm}C on the continuous (not junction-stabilizing) strand adopts an out-of-plane hydroxyl position, similar to those seen in previous B-DNA structures. Two waters are within H-bonding distance of the base.

3.3.2 Hydroxymethyl rotamers in the ^{5hm}C junction

It is clear from previous structures of ^{5hm}Cs in duplex DNAs[47,48] that there is a rotational bias to position the hydroxyl substituent in an out-of-plane geometry. From quantum mechanical studies[47], the perpendicular out-of-plane rotamer is the global energy minimum and is ~2 kcal/mol more stable than an in-plane form, which sits at a local minimum (Fig. 3.3). This torsional preference explains why the ^{5hm}C along the continuous strand of the junction is in the out-of-plane geometry, as seen in the structures of the DNA duplexes (Table 3). Spingler *et al*.[47] suggested that the bridging waters add very little to the preference of this torsionally preferred rotamer.

The relatively small difference in occupancy between R1 and R2 at the junction's crossover in the current structure suggests that additional interactions, in this case H-bonds to the phosphate group, could readily shift the rotamer preference. We therefore applied an MP2 calculation on the two rotamer forms of ^{5hm}C at the H-bonding geometries seen in the crystal structures, using a dimethyl phosphate as a model for the H-bond acceptor of the junction backbone (Fig. 3.4), to compare their H-bonding energies ($\Delta E_{H\text{-}Bond}$). From this calculation, we estimate an ~3.5 kcal/mol difference in $\Delta E_{H\text{-}Bond}$ that favors the R1 rotamer over the R2 rotamer. The two contributing energies (intrinsic torsional energy vs H-bonding energy) oppose each other, resulting in an overall preference for R1 of ~1.5 kcal/ mol, which would explain the approximate 2-fold preference for this rotamer in the crystal structure.

3.3.3 Energetic effects of hydroxymethyl and methyl substituents in solution

With the atomic details elucidated, we then asked whether and how the various interactions observed in the crystal structures (the direct hydroxyl H-bonds in G^{5hm}CC and the water-mediated H-bond seen in the G^{5m}CC structures) confer stability to the HJ in

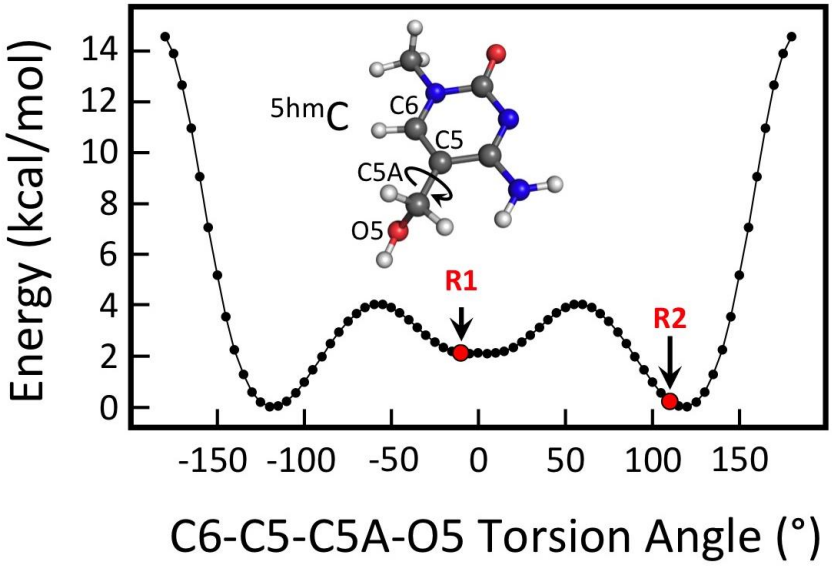


Figure 3.3: Energies for rotational isomers (rotamers) of an isolated 5-hydroxymethylcytosine (^{5hm}C, inset). Quantum mechanical energies at the MP2 level were calculated as the hydroxymethyl substituent is rotated around the C5-C5A bond (the C6-C5-C5A-O5 torsion angle labeled along the horizontal axis). Torsion angles for rotamers observed in the crystal DNA junction structure are shown as red circles. The dominant R1 rotamer (-12.0°) falls into a calculated local energy minimum, while the minor R2 rotamer (111.4°) falls in a global energy minimum well that is close to the lowest energy rotation (at 119°).

Table 3.3: Torsion angles (degrees) relating atoms C6, C5, C5A, O5 of the ^{5hm}C bases in crossover and continuous strands in the current junction structure, and in B-DNA duplexes from the literature. A torsion angle of 0° indicates the hydroxyl is in plane with the base and pointed towards the glycosidic bond, while a 180° angle points the hydroxyl towards the N4 amine group. Positive angles place the hydroxyl above the plane of the base in the 5' direction and negative angles are in the 3' direction.

Junction DNA (PDB entry 5DSB)	
Crossover Strand (R1)	-12.0
Crossover Strand (R2)	111.4
Continuous Strand	92.3
Published out of plane rotamers	
4HLI[a]	132.8
4GLC[a]	126.9
4GLC[a]	96.1
4GLH[a]	115.7
4GLH[a]	109.8
4I9V[b]	111.0
4I9V[b]	112.1
Average (Standard Deviation)	**114.9 (12.0)**
Published rotamer outliers	
4HLI[a]	72.6
4I9V[b]	24.7

[a]Renciuk, D. *et al. Nucleic Acids Res.* 2013 (ref 47)
[b]Szulik, M.W. *et al.*, *Biochemistry* 2015 (ref 48)

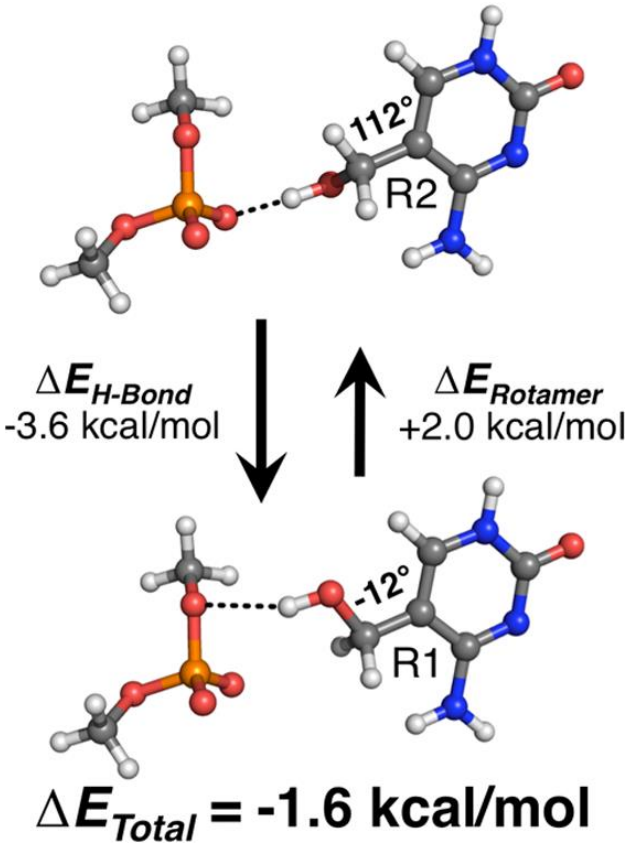


Figure 3.4: Comparison of H-bond energies ($\Delta E_{H\text{-}Bond}$) and rotamer energies ($\Delta E_{Rotamer}$) between
the major (R1, bottom) and minor (R2, top) conformations of the hydroxymethyl substituent in the ^{5hm}C structure. Quantum mechanical (QM) energies were calculated on small molecule models of the junction core (^{5hm}C and dimethyl phosphate), constructed from atomic coordinates taken from the crystal structure. The isolated ^{5hm}C base has a 2.0 kcal/mol energy preference for the R2 rotamer (112°) in bond rotation energy. However, the H-bonding interaction energy was calculated to favor the R1 rotamer (−12°) by −3.6 kcal/mol (signs of the energy terms are defined as the difference $E_{R1} - E_{R2}$). In summation, the dominant R1 rotamer is favored by an overall energy ($\Delta E_{Total} = \Delta E_{Rotamer} + \Delta E_{H\text{-}Bond}$) of −1.6 kcal/mol.

solution. We had previously shown that the sequence-dependent formation of HJs identified in crystals translates well to the stability of junctions in solution[34]. In the study presented here, we can directly apply differential scanning calorimetry (DSC) to determine the effects of these molecular interactions on the melting energies and, thus, tease out their effects on the stabilization of the four-stranded junction[45,49].

To determine the effects of hydroxymethyl or methyl modifications on the DSC energies, we take advantage of the concentration dependence for the formation of four-stranded junctions by self-complementary decanucleotides, in which DNAs at lower concentrations show melting parameters of duplexes and at higher concentrations reflect those of junctions[44]. We chose a DNA concentration for our DSC studies that showed both duplex and junctions in solution, thereby allowing us to measure the energies of the DNA species simultaneously (Fig. 3.5). Because the stacked-X junction is essentially composed of two duplexes and the interruption of the crossover region, the difference between junction and duplex DSC energies (scaled per two strands of DNA) isolates the stabilization energy associated with just the interactions at the junction core. In this way, we were able to determine the energetic contributions (ΔH, ΔS, and ΔG) of the core trinucleotides to junction stabilization. Furthermore, by subtracting the thermodynamic values for GCC from those of either $G^{5hm}CC$ or $G^{5m}CC$, we can specifically determine the effect of each substituent at the C_7 nucleobase on the stability of the junction.

DSC melting profiles for each construct were best fit using a two-component analysis, indicating the presence of both duplex and junction DNA in each sample. An analysis of the melting temperatures shows that cytosine methylation has an overall effect of stabilizing the duplex [increased T_m (Table 4)] relative to the unmodified DNA, while hydroxymethylation slightly destabilizes the duplex. We see very similar effects of the substituents on the T_m

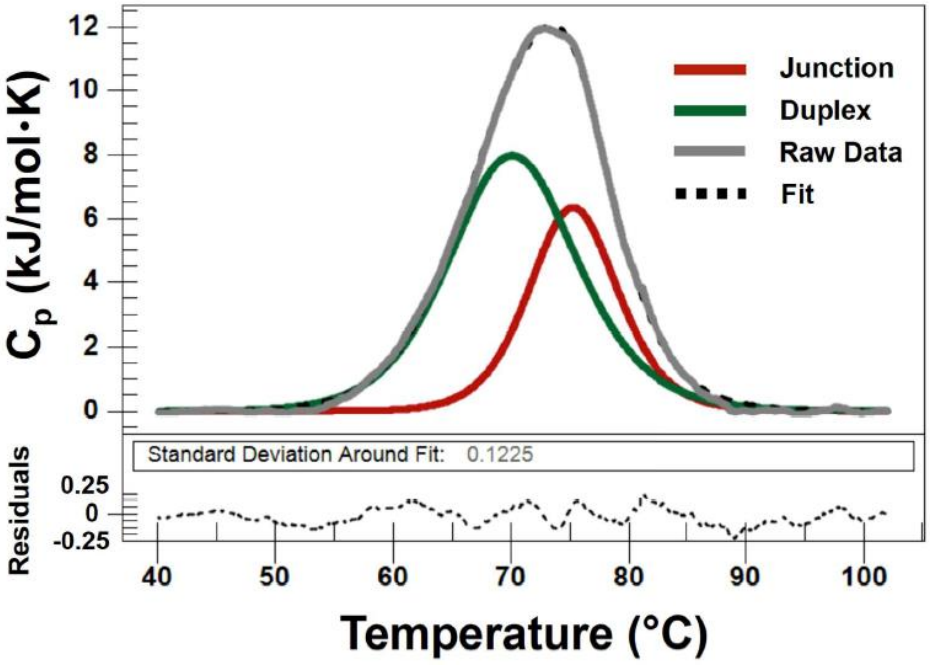


Figure 3.5: Representative DSC melting profile showing the melting data (gray) is fit with a two-component analysis indicating the presence of both junction (red) and duplex (green) DNA. A two-state scaled model was used, which incorporated a weighting term to account for differences in concentration between the duplex and junction populations. The composite fit (black dashes) modeling both junction and duplex melting events shows good agreement with the raw data (standard deviation around fit = 0.1225), and the residuals plot (bottom panel) is randomly distributed around zero. In contrast, fitting the raw data with a single peak resulted in larger and less random distribution of residuals and a 0.3469 standard deviation around the fit, indicating the melting event is indeed best modeled with a two-component fit.

Table 3.4: Melting temperatures (T_m) and melting enthalpies (ΔH_m) measured by DSC of GCC, $G^{5m}CC$, and $G^{5hm}CC$ core DNA constructs in solution. The entropy of melting (ΔS_m) for each construct is calculated from the T_m and ΔH_m by the equation $\Delta S_m = \Delta H_m/T_m$, with the assumption that at the melting temperature, the concentrations of folded and denatured DNA are equal and, thus, $\Delta G_m = 0$[50].

	Duplex			**Junction**			**J–D**
DNA Core:	T_m **(°C)**	ΔH_m **(kcal/mol)**	ΔS_m **(cal/mol)**	T_m **(°C)**	ΔH_m **(kcal/mol)**	ΔS_m **(cal/mol)**	ΔT_m **(°C)**
GCC	68.01 ± 0.11	81.7 ± 0.7	196.9 ± 1.7	73.50 ± 0.08	109.5 ± 1.4	283 ± 2	5.49 ± 0.14
$G^{5m}CC$	70.06 ± 0.10	68.1 ± 1.0	193 ± 3	75.20 ± 0.06	98.7 ± 1.0	285 ± 3	5.14 ± 0.12
$G^{5hm}CC$	65.03 ± 0.09	66.7 ± 0.5	239 ± 2	70.60 ± 0.05	97.3 ± 0.7	316 ± 4	5.57 ± 0.10

values for the junction, where the methyl group is associated with the highest T_m and the hydroxymethyl with the lowest. However, when we subtract the T_ms of each duplex from those of the junctions, we see that the methylcytosine results in a smaller difference in ΔT_m compared to that of the native GCC sequence, suggesting that methylation has a destabilizing effect on the junction. In contrast, this analysis of ΔT_m suggests that hydroxymethylation would have a slightly stabilizing effect on the Holliday junction relative to its duplex. This is consistent with the hydroxyl groups forming additional stabilizing H-bonds to the junction core. The magnitude of the difference in ΔH_m between the junction and duplex forms of the hydroxymethylated G^{5hm}CC construct ($\Delta\Delta H_m$ = 30.6 kcal/mol) is indeed larger than that of the parent GCC (27.8 kcal/mol).

We find that the G^{5m}CC DNA constructs are the most thermally stable (highest T_m) of the species studied. At the T_m, GCC stabilization was the most enthalpically driven, in contrast to the stronger entropic stabilization of G^{5hm}CC. The duplex and junction constructs follow similar trends with respect to their relative melting parameters (Table 4), suggesting the energetic effects of the modified bases are similar in both duplex and junction.

A better measure of the effect of each substituent on the energetics of the junction is to determine the $\Delta\Delta G°$ relative to the duplex at a standard temperature (25°C). To determine the interaction energies of each substituent group in the DNA junction[45,49], we first extrapolate the DSC energies to a common reference temperature (25°C) using the standard relationship (Eqs. 1 and 2). Following those extrapolations, the duplex energies were subtracted from those of the junctions, leaving only the junction core stabilization energy. Finally, the GCC core energy was subtracted from those of the modified cores (G^{5m}CC and G^{5hm}CC), reported as $\Delta\Delta H^{25°C}$, $\Delta\Delta S^{25°C}$, and $\Delta\Delta G^{25°C}$, to narrow the analysis to the specific

interaction energies associated with each modification [methylation or hydroxymethylation (Table 5)].

Eq. (1) $$\Delta H_{ref} = \Delta H_m + \Delta C_p(T_{ref} - T_m)$$

Eq. (2) $$\Delta S_{ref} = \Delta S_m + \Delta C_p \ln\left(\frac{T_{ref}}{T_m}\right)$$

The most immediate observation is that methylation or hydroxymethylation has little effect on the overall free energies ($\Delta\Delta G^{25°C} \approx 0$), indicating that the modified bases cause minimal disruption to the stability of the Holliday junction. However, we observed compensatory enthalpic and entropic effects, which contribute to these very small $\Delta\Delta G^{25°C}$ values. G^{5hm}CC and G^{5m}CC gain 1.5 and 2.0 kcal/mol of enthalpic energy, respectively (calculated per interaction, meaning twice this energy is stabilizing the whole junction), which suggests either stronger core H-bond stabilization or reduction in the level of conformational strain on the residue 7 base pair (Table 2).

The stabilizing enthalpies are compensated by unfavorable energy from entropic terms (−5 or −6 cal mol^{-1} K^{-1}, equivalent to ~1.5 kcal/mol of unfavorable energy at 25 °C) in the modified constructs. We had seen this type of enthalpy−entropy compensation previously in a DNA junction that is stabilized through halogen bonds[49]. In this latter case, we attributed the loss of entropic stabilization to reduced dynamics, as reflected in the smaller *B* factors associated with the nucleotide bases and phosphates that were involved in the stronger molecular interaction. A similar *B* factor analysis of these structures, however, showed that restriction of the conformational dynamics from stronger molecular interactions is not the rationale for the loss of entropy in the G^{5hm}CC and G^{5m}CC structures. A comparison of *B* factors indicates that the modified constructs are more locally dynamic at the junction

Table 3.5: Thermodynamic stabilization of $G^{5m}CC$ and $G^{5hm}CC$ junction cores relative to the GCC junction core. The enthalpic, entropic, and overall free energies (at 25 °C) for each modified construct are listed with the values from the parent construct subtracted ($\Delta\Delta H^{25°C}$, $\Delta\Delta S^{25°C}$, and $\Delta\Delta G^{25°C}$, respectively). Values reflect stabilization per interaction at each crossover strand; therefore, each complete junction structure is stabilized by twice the tabulated energies.

DNA Core	$\Delta\Delta H^{25°C}$ **(kcal/mol)**	$\Delta\Delta S^{25°C}$ **(cal/mol•K)**	$\Delta\Delta G^{25°C}$ **(kcal/mol)**
$G^{5m}CC$ - GCC	-2.1 ± 1.0	-6 ± 3	-0.22 ± 0.13
$G^{5hm}CC$ - GCC	-1.5 ± 0.9	-5 ± 3	-0.01 ± 0.12

core than the unmodified junction (Fig. 3.6). The GCC construct shows the pattern that is typical of H-bond-stabilized junctions, where the *B* factors for nucleotides 6−8 (where the stabilizing H-bonding interactions occur) are lower than that of the overall junction. This same pattern is also seen with the G^{5hm}CC and G^{5m}CC structures; however, the modifications result in atoms at the junction core that are less constrained than those of the GCC structure, particularly at the base of C_7 and the phosphate of position 6, the specific positions involved in the ^{5hm}C or ^{5m}C interactions.

The H-bond from the C_8 amino to the C_7 phosphate that is essential for the stabilization of the junction continues to constrain the dynamics of these interacting groups relative to the overall junction. The methyl and hydroxymethyl modifications, however, do also appear to increase the dynamics of the C_8 nucleotide and the C_7 phosphate, indicating that these substituents do affect the overall conformational dynamics of the entire junction core. It is clear, therefore, that the entropic compensation for the stabilizing enthalpy of folding does not come explicitly from the loss of conformational dynamics of the nucleotides involved in the H-bonding interactions.

3.4 Conclusions

The recent evidence that ^{5hm}C promotes recombination[7,26] prompted us to study the impact of this base modification on the structure and stability of the DNA Holliday junction and consider its potential impact on HR. We modified the C_7 cytosine of the ***G$_6$C$_7$C$_8$*** trinucleotide core (to form ***G^{5hm}CC***) of the sequence d(CCGGC***GCC***GG), a construct that is sensitive to environmental effects to junction stability[33]. As a steric control, we also considered the effects of cytosine methylation at this same cytosine position on the properties of the junction. We show that there is a minimal effect of either the hydroxymethyl or methyl

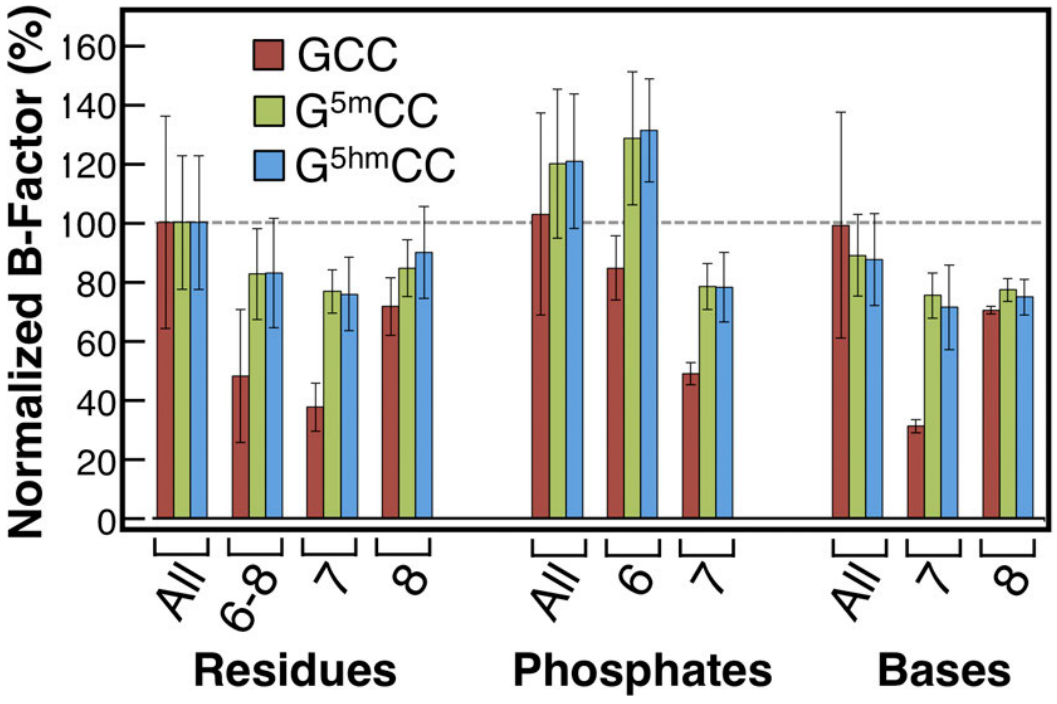


Figure 3.6: Normalized temperature factors of DNA junctions for the unmodified (GCC), methylated ($G^{5m}CC$), and hydroxymethylated ($G^{5hm}CC$) structures. Temperature factors (*B* factors) were normalized to the average value for the nonsolvent atoms in each structure (100% = average), and each structure was normalized on its own scale in a manner independent of the others. Error bars represent the standard deviation of B factors for the atoms in the selected group.

substituent on the overall thermodynamic stability of the junction, although the general structure becomes more open, leaving the trinucleotide core more accessible.

The H-bond from the C_7 base to the G_6 phosphate, which helps define the sequence dependence of junction formation[33,39], is seen to be disrupted in both modified constructs, with the hydroxyl group of the ^{5hm}C providing compensatory H-bonds. In this case, the hydroxymethyl adopts two different rotamer conformations, with the prevalent interaction being associated with a less favorable rotation. We thus see that, although there is a preferred intrinsic rotamer for the ^{5hm}C substituent, as seen here and in previous structures of B-DNA duplexes[47,48], a strong intramolecular interaction can overcome the energy barrier for the hydroxymethyl to adopt a less favored rotation. In the case of the ^{5m}C construct, the lost H-bond of the native GCC core is replaced by a water, which serves to bridge the N4 amino of the cytosine back again to the G_6 phosphate. Such water-mediated H-bonds have been shown to compensate well for direct H-bonds in DNA, for example, in providing stability to GT mismatches relative to standard GC Watson−Crick base pairs[51]. The resulting compensatory H-bonds (directly from the hydroxymethyl of ^{5hm}C or through water mediation in the ^{5m}C construct) resulted in a slight enthalpic stabilization of the GCC trinucleotide core in the junction.

The enthalpic stabilization in both the ^{5hm}C and ^{5m}C construct junctions is counterbalanced by losses of entropic stabilization. We had previously seen this entropy−enthalpy compensation effect when a halogen bond was engineered to stabilize the DNA junction, with the energetically stable halogen bond resulting in a less dynamic junction core[49] (as reflected in the reduced *B* factors of the core). The increased conformational dynamics for the ^{5hm}C- and ^{5m}C-modified junctions, as reflected in crystallographic *B* factor analysis, however, was initially perplexing, as it appears to be in contrast with the decreased

entropy of these constructs as measured by DSC. Clearly, the entropic penalty for folding is not associated with reduced conformational dynamics resulting from stabilization of the junction core, specifically by the methyl or hydroxymethyl groups. For the ^{5hm}C base, some entropy loss may be attributed to constraining the hydroxyl substituent to the two specific rotamer conformations required to form the H-bonds to the junction backbone, which would impose an entropic penalty relative to the range of energetically favorable rotamers observed for the unconstrained ^{5hm}C base (Table 3). As the C and ^{5m}C do not have multiple rotational states available, the H-bond conformation is not a constraint. One likely explanation is a change in the solvent entropy due to constrained water molecules around the junction core. In the G^{5m}CC crystal structure, we observe a highly structured water molecule bound near the junction core, and this water is absent in the native GCC core structure. Similarly, a highly structured water is observed in the ^{5hm}C junction, but in this case, the water bridges the N4 amino and the OH of the hydroxymethyl substituent and, thus, does not help to stabilize the overall junction.

The G^{5hm}CC and G^{5m}CC junction cores are quite different from the GCC core from a structural perspective, in terms of both direct and indirect readout implications. A hypothetical resolvase recognizing the junction core could distinguish the different cores, and hence, this would be a Tet-regulated control for sites of HR. In terms of indirect readout, the steric bulkiness of ^{5hm}C and ^{5m}C does impact the overall junction structure by opening the junction and relieving some strain on the contorted base pairs and backbone that kink to allow junction formation. This opening of the junction provides more space between the two duplex arms, possibly facilitating the ability of a protein to probe for specific interactions at that site.

Although the enthalpy−entropy compensation does not result in an overall more stable Holliday junction, it may affect the kinetics of junction migration, which in turn would affect the role of both the hydroxymethyl and methyl modifications on homologous recombination. Khuu *et al.*[52] proposed a model in which the sequence specificity of junction-cleaving proteins (resolvases) results from pausing migration at sequences that help stabilize the stacked-X junction structure. The kinetics of pausing, however, may not be reflected in the overall free energy of the stacked-X junction, but in the energetic barriers. The increased level of H-bonding interactions in both the ^{5hm}C and ^{5m}C junctions, thus, would provide such barriers, which may slow the migration of the junction away from the GCC core and provide sufficient time for a resolvase to indirectly recognize these modifications. It would be interesting to determine the effects of these epigenetic markers on the kinetics of junction migration and explore the concept of sequence-dependent pausing.

In conclusion, we see that the methyl substituent pushes the C·G base pair away from the junction crossover, resulting in a more open structure, as reflected in the larger J_{roll}. The hydroxymethyl has an even stronger effect. Given that select few sequences are capable of stabilizing a stacked-X junction[33], there is great potential for direct as well as indirect readout of these base modifications, which distort the stacked-X structure without dismantling it. In the context of the Khuu model[52], the H-bonds of ^{5hm}C could kinetically pause migration, as discussed, while the more open junction provides access for a protein to directly recognize the modified base, with the alternative rotamer allowing the hydroxymethyl group in the junction to be distinguished from the standard rotamer in a B-DNA duplex.

3.4.1 Funding

Research reported in this publication was supported by grants from the National Science Foundation (MCB-1515521) to P.S.H., a predoctoral fellowship from the National Institute of General Medical Sciences of the National Institutes of Health (F31GM113580) to C.M.V.Z., and a start-up grant from the 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.

REFERENCES

(1) Bestor, T. H. (2000) The DNA methyltransferases of mammals. *Hum. Mol. Genet. 9*, 2395–2402.

(2) Jaenisch, R., and Bird, A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet. 33*, 245–254.

(3) Goll, M. G., and Bestor, T. H. (2005) Eukaryotic Cytosine Methyltransferases. *Annu. Rev. Biochem. 74*, 481–514.

(4) Heithoff, D. M., Sinsheimer, R. L., Low, D. A., Mahan, M. J., Woude, M. van der, Braaten, B., Low, D., LeClerc, J. E., Li, B., Payne, W. L., Cebula, T. A., Braaten, B. A., Nou, X., Kaltenbach, L. S., Low, D. A., Conner, C. P., Heithoff, D. M., Julio, S. M., Sinsheimer, R. L., Mahan, M. J., Mahan, M. J., Slauch, J. M., Mekalanos, J. J., Roland, K. L., Martin, L. E., Esther, C. R., Spitznagel, J., Vescovi, E. G., Soncini, F. C., Groisman, E. A., Woude, M. van der, Hale, W. B., Low, D. A., Hale, W. B., Woude, M. W. van der, Low, D. A., Tavazoie, S., Church, G. M., Bandyopadhyay, R., Das, J., Julio, S. M., Conner, C. P., Heithoff, D. M., Mahan, M. J., Slauch, J. M., Silhavy, T., Smith, C. L., Cantor, C. R., Marinus, M. G., Poteete, A., and Arraj, J. A. (1999) An essential role for DNA adenine methylation in bacterial virulence. *Science 284*, 967–70.

(5) Ehrlich, M., Wilson, G. G., Kuo, K. C., and Gehrke, C. W. (1987) N4-methylcytosine as a minor base in bacterial DNA. *J. Bacteriol. 169*, 939–43.

(6) Vanyushin, B. F. (2005) Adenine Methylation in Eukaryotic DNA. *Mol. Biol. 39*, 473–481.

(7) 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.

(8) Penn, N. W., Suwalski, R., O'Riley, C., Bojanowski, K., and Yura, R. (1972) The presence of 5-hydroxymethylcytosine in animal deoxyribonucleic acid. *Biochem. J. 126*, 781–790.

(9) Kriaucionis, S., and Heintz, N. (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science 324*, 929–30.

(10) Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L. M., Liu, D. R., Aravind, L., and Rao, A. (2009) Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science 324*, 930–5.

(11) 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*, 1300–3.

(12) Jin, S.-G., Kadam, S., and Pfeifer, G. P. (2010) Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine. *Nucleic Acids Res. 38*, e125.

(13) Robertson, A. B., Dahl, J. A., Vågbø, C. B., Tripathi, P., Krokan, H. E., and Klungland, A. (2011) A novel method for the efficient and selective identification of 5-hydroxymethylcytosine in genomic DNA. *Nucleic Acids Res. 39*, e55.

(14) Song, C.-X., Szulwach, K. E., Fu, Y., Dai, Q., Yi, C., Li, X., Li, Y., Chen, C.-H., Zhang, W., Jian, X., Wang, J., Zhang, L., Looney, T. J., Zhang, B., Godley, L. A., Hicks, L. M., Lahn, B. T., Jin, P., and He, C. (2011) Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine. *Nat. Biotechnol. 29*, 68–72.

(15) Robertson, A. B., Dahl, J. A., Ougland, R., and Klungland, A. (2012) Pull-down of 5-hydroxymethylcytosine DNA using JBP1-coated magnetic beads. *Nat. Protoc. 7*, 340–50.

(16) Szwagierczak, A., Bultmann, S., Schmidt, C. S., Spada, F., and Leonhardt, H. (2010) Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic DNA. *Nucleic Acids Res. 38*, e181.

(17) Yu, M., Hon, G. C., Szulwach, K. E., Song, C.-X., Jin, P., Ren, B., and He, C. (2012) Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nat. Protoc. 7*, 2159–70.

(18) Reuters, T. Web of Science.

(19) Wen, L., and Tang, F. (2014) Genomic distribution and possible functions of DNA hydroxymethylation in the brain. *Genomics 104*, 341–346.

(20) Li, W., and Liu, M. (2011) Distribution of 5-hydroxymethylcytosine in different human tissues. *J. Nucleic Acids 2011*, 1–5.

(21) Ruzov, A., Tsenkina, Y., Serio, A., Dudnakova, T., Fletcher, J., Bai, Y., Chebotareva, T., Pells, S., Hannoun, Z., Sullivan, G., Chandran, S., Hay, D. C., Bradley, M., Wilmut, I., and De Sousa, P. (2011) Lineage-specific distribution of high levels of genomic 5-hydroxymethylcytosine in mammalian development. *Cell Res. 21*, 1332–1342.

(22) Irier, H., Street, R. C., Dave, R., Lin, L., Cai, C., Davis, T. H., Yao, B., Cheng, Y., and Jin, P. (2014) Environmental enrichment modulates 5-hydroxymethylcytosine dynamics in hippocampus. *Genomics 104*, 376–382.

(23) Putiri, E. L., Tiedemann, R. L., Choi, J.-H., and Robertson, K. D. (2014) Abstract 2319: Dynamics of TET methylcytosine dioxygenases in 5-methylcytosine and 5-hydroxymethylcytosine patterning in human cancer cells. *Cancer Res. 74*, 2319–2319.

(24) Pfeifer, G. P., Xiong, W., Hahn, M. a., and Jin, S. G. (2014) The role of 5-hydroxymethylcytosine in human cancer. *Cell Tissue Res. 356*, 631–641.

(25) Robertson, J., Robertson, A. B., and Klungland, A. (2011) The presence of 5-

hydroxymethylcytosine at the gene promoter and not in the gene body negatively regulates gene expression. *Biochem. Biophys. Res. Commun. 411*, 40–43.

(26) Stroud, H., Feng, S., Morey Kinney, S., 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.

(27) Wen, L., Li, X., Yan, L., Tan, Y., Li, R., Zhao, Y., Wang, Y., Xie, J., Zhang, Y., Song, C., Yu, M., Liu, X., Zhu, P., Li, X., Hou, Y., Guo, H., Wu, X., He, C., Li, R., Tang, F., and Qiao, J. (2014) Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain *15*, 1–17.

(28) Sun, W., Zang, L., Shu, Q., and Li, X. (2014) From development to diseases: The role of 5hmC in brain. *Genomics 104*, 347–351.

(29) Wang, T., Pan, Q., Lin, L., Szulwach, K. E., Song, C. X., He, C., Wu, H., Warren, S. T., Jin, P., Duan, R., and Li, X. (2012) Genome-wide DNA hydroxymethylation changes are associated with neurodevelopmental genes in the developing human cerebellum. *Hum. Mol. Genet. 21*, 5500–5510.

(30) Guo, J. U., Su, Y., Zhong, C., Ming, G., and Song, H. (2011) Hydroxylation of 5-Methylcytosine by TET1 Promotes Active DNA Demethylation in the Adult Brain. *Cell 145*, 423–434.

(31) El-Osta, A., and Wolffe, A. P. (2000) DNA methylation and histone deacetylation in the control of gene expression: basic biochemistry to human development and disease. *Gene Exp. 9*, 63–75.

(32) Holliday, R. (1964) A mechanism for gene conversion in fungi. *Genet. Res. 5*, 282–304.

(33) Hays, F. A., Teegarden, A., Jones, Z. J. R., 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*, 7157–62.

(34) Hays, F. A., Schirf, V., Ho, P. S., and Demeler, B. (2006) Solution formation of Holliday junctions in inverted-repeat DNA sequences. *Biochemistry 45*, 2467–2471.

(35) Lilley, D. M. J. (2000) Structures of helical junctions in nucleic acids. *Q. Rev. Biophys. 33*, 109–159.

(36) Roe, S. M., Barlow, T., Brown, T., Oram, M., Keeley, A., Tsaneva, I. R., and Pearl, L. H. (1998) Crystal Structure of an Octameric RuvA–Holliday Junction Complex. *Mol. Cell 2*, 361–372.

(37) Hadden, J. M., Déclais, A.-C., Carr, S. B., Lilley, D. M. J., and Phillips, S. E. V. (2007) The structural basis of Holliday junction resolution by T7 endonuclease I. *Nature 449*, 621–4.

(38) Biertümpfel, C., Yang, W., and Suck, D. (2007) Crystal structure of T4 endonuclease VII resolving a Holliday junction. *Nature 449*, 616–20.

(39) Eichman, B. F., Vargason, J. M., Mooers, B. H. M., and Ho, P. S. (2000) The Holliday junction in an inverted repeat DNA sequence: Sequence effects on the structure of four-way junctions. *Proc. Natl. Acad. Sci. 97*, 3971–3976.

(40) Minor, W., Cymborowski, M., Otwinowski, Z., and Chruszcz, M. (2006) HKL-3000: the integration of data reduction and structure solution--from diffraction images to an initial model in minutes. *Acta Crystallogr. D. Biol. Crystallogr. 62*, 859–66.

(41) Adams, P. D., Afonine, P. V, Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution.

Acta Crystallogr. D. Biol. Crystallogr. 66, 213–21.

(42) Lavery, R., Moakher, M., Maddocks, J. H., Petkeviciute, D., and Zakrzewska, K. (2009) Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res. 37*, 5917–29.

(43) Watson, J., Hays, F. A., and Ho, P. S. (2004) Definitions and analysis of DNA Holliday junction geometry. *Nucleic Acids Res. 32*, 3017–3027.

(44) Carter, M., and Ho, P. S. (2011) Assaying the Energies of Biological Halogen Bonds. *Cryst. Growth Des. 11*, 5087–5095.

(45) Vander Zanden, C. M., Carter, M., and Ho, P. S. (2013) Determining thermodynamic properties of molecular interactions from single crystal studies. *Methods 64*, 12–8.

(46) Frisch, M. J., Trucks, G. W., Schlegel, H. B., Scuseria, G. E., Robb, M. A., Cheeseman, J. R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G. A., Nakatsuji, H., Caricato, M., Li, X., Hratchian, H. P., Izmaylov, A. F., Bloino, J., Zheng, G., and Sonnenb, D. J. (2009) Gaussian 09, Revision A. 02, Gaussian. *Gaussian Inc.* Wallingford, CT.

(47) Renciuk, D., Blacque, O., Vorlickova, M., and Spingler, B. (2013) Crystal structures of B-DNA dodecamer containing the epigenetic modifications 5-hydroxymethylcytosine or 5-methylcytosine. *Nucleic Acids Res. 41*, 9891–900.

(48) Szulik, M. W., Pallan, P. S., Nocek, B., Voehler, M., Banerjee, S., Brooks, S., Joachimiak, A., Egli, M., Eichman, B. F., and Stone, M. P. (2015) Differential stabilities and sequence-dependent base pair opening dynamics of Watson-Crick base pairs with 5-hydroxymethylcytosine, 5-formylcytosine, or 5-carboxylcytosine. *Biochemistry 54*, 1294–305.

(49) Carter, M., Voth, A. R., Scholfield, M. R., Rummel, B., Sowers, L. C., and Ho, P. S. (2013) Enthalpy–Entropy Compensation in Biomolecular Halogen Bonds Measured in DNA Junctions. *Biochemistry 52*, 4891–4903.

(50) van Holde, K. E., Johnson, W. C., and Ho, P. S. (2006) Principles of Physical Biochemistry 2nd ed. Pearson Education Inc.

(51) Ho, P., Frederick, C., Quigley, G., Vandermarel, G., Vanboom, J., Wang, A., and Rich, A. (1985) GT wobble base-pairing in Z-DNA at 1.0-A atomic resolution - The crystal-structure of D(CGCGTG). *EMBO J. 4*, 3617–3623.

(52) Khuu, P. A., Voth, A. R., Hays, F. A., and Ho, P. S. (2006) The stacked-X DNA Holliday junction and protein recognition. *J. Mol. Recognit. 19*, 234–42.

CHAPTER 4

VERTEBRATE ENDONUCLEASE G PREFERENTIALLY CLEAVES HOLLIDAY JUNCTIONS AND HAS A DISTINCT STRUCTURE TO RECOGNIZE 5-HYDROXYMETHYLCYTOSINE

Endonuclease G (EndoG) is characterized as a non-specific DNA cutting protein that cleaves G/C rich sequences of DNA. EndoG has been implicated in a variety of cellular recombination functions, and was recently found to preferentially cleave 5-hydroxymethylcytosine (^{5hm}C) DNA in a sequence-specific context to promote recombination. ^{5hm}C is a recently recognized epigenetic marker that represents up to 1% percent of cytosines in the mammalian genome, and it is found in all vertebrate organisms. In this study, we demonstrate that the Holliday junction, the 4-stranded DNA intermediate in homologous recombination, is a preferred substrate for EndoG. EndoG cuts ^{5hm}C-modified Holliday junctions to produce unique cleavage products, suggesting ^{5hm}C is a marker in EndoG mediated recombination. Furthermore, we present the single-crystal structure of mouse EndoG and propose a mechanism for vertebrate EndoG recognition of ^{5hm}C. An α-helix seen at the DNA binding site of the homologous enzymes from Drosophila and *C. elegans* has unraveled into a long structured loop, allowing the side chains of Ala109 and Cys110 to enter the binding site and potentially recognize the ^{5hm}C modification. The unraveling of this helix is attributed to a two amino acid deletion near the binding site, which is conserved for all vertebrate EndoG sequences. Although EndoG is found in all eukaryotic species, we suggest EndoG has evolved to recognize ^{5hm}C in vertebrate species.

4.1 Introduction

Endonuclease G (EndoG) is a DNA-cutting enzyme first identified in mammals and named for its ability to cleave G rich tracts of DNA[1–3] It is a member of the ββα-Me finger nuclease family, and the catalytic site residues are conserved across eukaryotic species from yeast to humans[4], although the preference to cleave G-rich sequences is not entirely conserved[5]. The EndoG catalytic site requires a histidine as part of the conserved residues of the NHN motif more broadly found in other nucleases[6,7]. *In silico* modeling suggests EndoG structure is highly conserved across eukaryotic species[4], despite low sequence identity (only 45% sequence identity between mouse EndoG and *C. elegans* EndoG homologue CPS-6, for example). EndoG's proposed substrates have included double and single stranded DNA, RNA[5], RNA-DNA hybrids[8], damaged DNA[9], R-loops[10], and chromatin[11]. In this study we show that Holliday junctions are preferred over double stranded DNA as a substrate for EndoG, supporting a role for EndoG in homologous recombination. Furthermore, we present the single crystal structure of mouse EndoG and propose a recognition mechanism for 5-hydroxymethylcytosine (^{5hm}C) DNA that is specific to vertebrate EndoGs.

Mouse EndoG was successfully crystallized in 2009; however, no structure was reported[12]. In that same year, Loll *et al*. solved the structure of Drosophila EndoG bound to an inhibitor (EndoGI), and showed that EndoG crystallizes as a homodimer with its active sites on opposing faces of the complex[13]. As expected from the modeling, the DNA binding interface is positively charged to attract the DNA backbone, and the inhibitor works by binding that positive surface and preventing the protein from coordinating Mg^{2+}. In 2016, Lin *et al.* solved the structure of the *C. elegans* EndoG homologue CPS-6 bound to a single strand of poly-T DNA[14]. They found that EndoG primarily contacts the DNA backbone, explaining the non-specific

nature of EndoG cleavage activity. The cleavage mechanism works through a hydrolysis reaction, instigated by a bound water molecule that performs nucleophilic attack on a scissile phosphate, resulting in a single stranded cleavage product with a 5'-end phosphate and a 3'-end OH group.

EndoG has been implicated in normal mitochondrial and nuclear DNA processes, and in degradation of nuclear DNA during apoptosis[5,15,16]. In the mitochondria, EndoG facilitates DNA replication by cleaving RNA/DNA hybrids to generate RNA primers necessary for mtDNA replication[8]. Mitochondrial EndoG also cleaves R-loops, a necessary function because untimely R-loops can inhibit transcription and promote unwanted DNA replication[10]. EndoG prefers to cleave damaged DNA[9], which is needed for genomic maintenance in the highly oxidative mitochondrial environment. During apoptosis, mass quantities of EndoG are translocated from the mitochondria to the nucleus to degrade chromosomal DNA[15]. The proteins HSP70, AIF, and FEN-1 regulate EndoG's apoptotic activity[17], and reactive oxygen species are also regulators that decrease EndoG dimerization to reduce cleavage activity[18]. A proposed non-apoptotic function for nuclear EndoG is to degrade foreign viral DNA, as EndoG knockdown in cells causes proliferation defects and promotes replication of foreign viral DNA[19]. Finally, EndoG also promotes DNA recombination.

EndoG's role in recombination has been illustrated in a variety of cellular contexts. This includes promoting recombination in transfected plasmids[20], Herpes Simplex Virus[21], myeloid/lymphoid leukemia break point clusters[22], and immunoglobulins[23]. EndoG is hypothesized to promote genome maintenance through recombination, thus enabling viability of polyploidy cells[24]. In most cases, EndoG promotes recombination by non-specifically cleaving G/C rich sites found in these cellular systems.

^{5hm}C is a recently recognized[25] epigenetic DNA marker representing up to 1% of cytosines in the mammalian genome[26]. ^{5hm}C was previously understood to primarily function as a marker to defensively prevent degradation of bacteriophage DNA[27,28]. Interestingly, there is currently no evidence for ^{5hm}C DNA function in non-vertebrate eukaryotic organisms[29–31]. Mammalian ^{5hm}C has been implicated in many genomic functions including embryonic development[32–34], neuronal regulation[35–37], gene expression[33,38,39], 5-methylcytosine regulation[40], and recombination[33,41]. ^{5hm}C's role in recombination was first evidenced by high levels of ^{5hm}C in G/C rich genomic regions that are more prone to recombination[33]. In 2014, Robertson *et al.* showed that EndoG also promotes nuclear DNA recombination through specific recognition of ^{5hm}C DNA[41] and cleavage in the sequence 5'-GGGG^{5hm}CCAG-3' / 5'-CTGGCCCC-3' to generate double-strand breaks that promote strand exchange. Additionally, ^{5hm}C is able to stabilize Holliday junctions, 4-stranded DNA intermediates involved in homologous recombination. ^{5hm}C creates a structurally unique Holliday junction that could be recognized by a protein via either direct or indirect readout[42]. With strong evidence linking EndoG to recombination, we ask whether EndoG can bind and cleave Holliday junction substrates, and whether EndoG is able to recognize ^{5hm}C in a Holliday junction context. Additionally, we aimed to understand how mammalian EndoG specifically recognizes ^{5hm}C DNA.

In this study, we explore the binding of EndoG to Holliday junctions and test the effect of ^{5hm}C incorporation. Our results show that indeed junction DNA is a preferred substrate for EndoG, evidenced by higher activity and binding affinity for junctions over duplex DNA. We observed unique cleavage products when ^{5hm}C was incorporated into the junction, indicating EndoG does recognize ^{5hm}C in a junction context. Finally, we present the crystal structure of mouse EndoG and identify the unraveling of a helix (found in the fly and worm homologues) to

instead form a structured loop, which we believe to be the source of mammalian EndoG specificity for ^{5hm}C DNA. A two amino acid deletion in the sequence responsible for this structural perturbation is conserved for all known vertebrate EndoG sequences, and we propose that the recognition mechanism can be extended to describe all vertebrate organisms.

4.2 Methods

Oligonucleotide Design and Purification: Oligonucleotides were designed to incorporate the EndoG recognition sequence 5'-GGGG^{5hm}CCAG-3' / 5'-CTGGCCCC-3' into either a junction or duplex context (Table 4.1). The junction was designed with unique arms only capable of correctly base-pairing with the intended annealing partner. Each junction or duplex construct contains one copy of the hydroxymethylated GGGG^{5hm}CCAG sequence, and duplicate constructs were made containing the recognition sequence without hydroxymethylation (GGGGCCAG).

All oligonucleotides were purchased from Midland Certified Reagent Company with necessary ^{5hm}C or cy5 modifications included. The 5'-dimethoxytrityl (DMT) protecting group was left intact from the synthesis in order to facilitate hydrophobicity-based purification from prematurely truncated products. Oligonucleotides were suspended in ammonium hydroxide for 24 hours to remove the CPG solid support bead that remained from synthesis. Full-length products were isolated by HPLC purification on a C18 column, and the DMT group was removed by incubation in 3% acetic acid for 15 minutes. Finally a G-25 Sephadex size exclusion chromatography column was used for desalting. DNA constructs were annealed by combining 7μM of each required strand (final concentration of 7μM junction or 7μM duplex) in a solution containing 10mM $MgCl_2$, 0.4mM EDTA, 20mM boric acid, and 19mM Tris pH 8.3. The mixture was heated to 90°C for 20 minutes and allowed to slowly cool over 2 hours.

Table 4.1: Sequences of duplex and junction constructs. Unique base pairs are colored accordingly and the EndoG recognition sequence is underlined. Cy5-labeled fluorescent strands are marked with an asterisk (*).

Duplex – C	**Duplex – ^{5hm}C**
***5'-CGAGGCCTGGCCCCGTACGG-3'** **5'-CCGTACGGGGCCAGGCCTCG-3'**	***5'-CGAGGCCTGGCCCCGTACGG-3'** **5'-CCGTACGGGG^{5hm}CCAGGCCTCG-3'**

Junction – C	**Junction – ^{5hm}C**
5'-GGGTTCCTGGCCCCGTACGG-3' **5'-CCGTACGGGGCCAGGCCTCG-3'** **5'-CGAGGCCTGGCCCCGGCTGC-3'** ***5'-GCAGCCGGGGCCAGGAACCC-3'**	**5'-GGGTTCCTGGCCCCGTACGG-3'** **5'-CCGTACGGGG^{5hm}CCAGGCCTCG-3'** **5'-CGAGGCCTGGCCCCGGCTGC-3'** ***5'-GCAGCCGGGGCCAGGAACCC-3'**

Activity Assays: Activity assays were prepared in 25µL reactions containing 100nM duplex or junction, 1.613µg/mL WT EndoG, 0.1% Triton x-100, 4% glycerol, 20mM Tris pH 7.5, 2mM $MgCl_2$, and 10mM 2-mercapoethanol. All non-protein components were mixed and a 10µL sample was removed to quantify the original amount of DNA. Finally the protein was added and the cleavage reaction was allowed to proceed for 10 minutes at 37°C. A 10µL sample was combined with 2µL of stop solution (0.25% sodium dodecyl sulfate and 0.6mg/mL proteinase K), and the reaction was incubated for 30 minutes at 50°C. Cut and uncut samples were loaded onto native gels made with 15% polyacrylamide and containing 0.2X TBE and 10mM $MgCl_2$. Gels were run for 6 hours at 5mA and quantified with a Typhoon FLA 9500 gel imager (GE). ImageJ[43] was used to quantify gel bands, and EndoG activity was determined as percent of DNA substrate cut during the reaction.

Electrophoretic Mobility Shift Assays: Samples contained constant 70nM duplex or 70nM junction with varying concentrations of inactive EndoG H138A (10µL total reaction volume). The reaction buffer included 0.1% Triton x-100, 4% glycerol, 20mM Tris pH 7.5, 10mM $MgCl_2$, and 10mM 2-mercaptolethanol. 10mM $MgCl_2$ was chosen in order to promote junction formation[44] and encourage binding of inactive EndoG H138A, which is more active at high concentrations of $MgCl_2$[6]. Components were mixed on ice and then allowed to bind at room temperature for 15 minutes. 15% polyacrylamide native gels containing 0.2X TBE and 10mM $MgCl_2$ were run at 35V for 6 hours and cy5-labeled DNA samples were detected with a Typhoon FLA 9500 gel imager (GE). The gel bands were quantified using ImageJ[43] and data was fit with a two-state non-cooperative binding model.

Protein Expression and Purification: Inactive EndoG H138A (without mitochondrial localization signal) was cloned into pMal-c2 vector and expressed as a maltose binding protein

(MBP) fusion. EndoG H138A pMal-c2 was expressed in BL21-CodonPlus(DE3)-RIPL strain (*E. coli* B F^- ompT hsdS(r_B^- m_B^-) dcm^+ Tet^r gal λ(DE3) endA Hte (Cam^R)). Cells were grown at 37°C to an OD_{600} of 0.5AU, then transferred to 10°C and allowed to continue growing to 0.7AU. Expression was induced with 1mM IPTG and cells continued to grow at 10°C for 48 hrs. Cells were harvested by centrifugation and suspended in purification buffer containing 50mM Tris pH 8.0, 50mM NaCl, 1mM $MgCl_2$, 5% glycerol and 10mM 2-mercaptoethanol. Cells were lysed by sonication, and the lysate was centrifuged at 17,000 RPM for 40 minutes at 4°C. The soluble cell lysate contained the majority of EndoG H138A-MBP and so only that portion was kept for purification. EndoG H138A-MBP fusion was purified by a gravity-fed amylose column and eluted with purification buffer containing 20mM maltose. The fusion protein was dialyzed into fresh purification buffer and cleaved overnight with Tobacco Etch Virus (TEV) protease (1mg TEV per 200mg fusion protein). Cleaved MBP was removed from the sample by purification with a heparin column which bound the positively charged DNA-binding patches on EndoG H138A. EndoG H138A was eluted from the heparin column in a gradient of purification buffer containing up to 0.5M NaCl. EndoG H138A was again purified on the amylose column, collecting the flow through (cleaved EndoG H138A) while the remaining EndoG H138A-MBP fusion contaminants bound to the column. Finally, the protein was concentrated and loaded onto a G-100 Sephadex column to remove high molecular weight contaminants. The final protein was concentrated to ~3mg/mL, quantified with ε= 34505 $M^{-1}cm^{-1}$, and stored at -80°C until used for crystallography or EMSA. Protein is estimated to be >95% pure (Fig. 4.4b).

Wild type mouse EndoG was expressed and purified as previously described[41]. In brief, His-tagged mouse EndoG lacking the mitochondrial localization signal was expressed on a pET28a vector in Rossetta(DE3) ((F− ompT $hsdS_B$ (R_B^- m_B^-) gal dcm λ (DE3 [lacI lacUV5-T7

gene 1 ind1 sam7 nin5]) pLysSRARE (Cam^R)) cells. The cells were grown at 37°C in Studier Autoinducing Media until they reached OD 0.8, then the culture was grown at 18°C for 16 hrs. The cells were harvested and stored in buffer containing 25mM HEPES pH 7.9, 500mM NaCl, 5mM Imidazole and 10% (v/v) glycerol. Prior to sonication, cells were treated with lysozyme and triton X-100, and the supernatant was collected from centrifugation of the cell lysate. His-tagged EndoG was purified with TALON resin (BD Biosciences) and eluted with imidazole buffer. The final protein was dialyzed into a final storage buffer containing 250 mM NaCl, 25 mM Hepes pH 7.9, 1 mM EDTA pH 8.0 and 50% (v/v) glycerol.

Crystallography and Structure Analysis: Crystals were grown at 16°C in a hanging drop vapor diffusion setup equilibrated against 450uL of mother liquor. The drop contained 1µL of 2.61mg/mL inactive EndoG H138A (stored in 50mM Tris pH 8.0, 50mM NaCl, 1mM $MgCl_2$, 5% (v/v) glycerol and 10mM 2-mercaptoethanol) and 1uL of mother liquor containing 25% (v/v) isopropanol, 0.1M HEPES pH 7.6, and 0.2M $MgCl_2$. Nucleation appeared after a few weeks and crystals reached full maturity after 3 months. The crystal was harvested under paraffin oil and flash frozen at 100K. Data were collected using a Rigaku Compact Home Lab equipped with a PILATUS detector and HLK3000[45] was used to index, integrate, and scale the data. The structure of dimeric *C. elegans* homologue CPS-6 (PDB 3S5B)[46] was used to phase the data by molecular replacement, and the model was subsequently refined using Phenix[47] Autobuild and Refine.

Minimization and Energy Calculations: Minimizations were performed using Amber12[48] and AmberTools13[49]. For CPS-6 minimizations, starting coordinates were obtained from the crystal structure of CPS-6 bound to poly-T DNA (PDB 5GKP), mutating the DNA sequence as necessary. Mouse EndoG starting coordinates were taken from the protein-only crystal structure,

and the DNA position was assumed from the CPS-6/poly-T cocrystal. Antechamber was used to calculate parameters for ^{5hm}C residues (derived from cif files, using the AM1-BCC charge model), and ff12SB was chosen as the force field for all topology file generation. The DNA was solvated in an octahedral box with TIP3P waters, and Mg^{2+} or Cl^- ions were added to neutralize the charge. Structures were minimized in three steps, allowing first the solvent and ions to minimize, then the hydrogens, and finally removing all constraints on the structure. ΔE was calculated for the energy of that DNA sequence complex relative to the energy of the minimized protein with poly-T.

4.3 Results and Discussion

The goal of this study is to learn about the substrate preferences of EndoG – specifically to determine if EndoG recognizes and cleaves Holliday junctions, and to learn how mammalian EndoG preferentially recognizes ^{5hm}C DNA. We determined the activity and binding affinity of mouse EndoG to junctions and duplexes containing the EndoG recognition sequence 5'-GGGG^{5hm}CCAG-3' / 5'-CTGGCCCC-3'. Non-hydroxymethylated DNA was also tested to determine if EndoG displayed combinatorial recognition of ^{5hm}C and junction structure, or if those recognitions play into different roles for the endonuclease. To learn about the mechanism of EndoG substrate specificity we determined the structure of mammalian EndoG. This structure was compared to those previously reported for the Drosophila and *C. elegans* (CPS-6), which are not expected to sense ^{5hm}C. Assuming that DNA binding is conserved, we were able to deduce a mechanism of sequence specific recognition of ^{5hm}C from contact points between the mammalian structure and the likely DNA binding site.

4.3.1 EndoG preferentially binds and cuts Holliday junction DNA

There is strong evidence for EndoG involvement in homologous recombination (HR); thus, it was logical to test the affinity and cutting specificity of EndoG for Holliday junctions, the 4-stranded DNA intermediate in HR. To promote EndoG binding, we designed duplex and junction constructs with the specific sequence GGGG^{5hm}CCAG, the sequence motif cleaved by wild type mouse EndoG[41] (Table 4.1). As a control to isolate the effects of ^{5hm}C on EndoG recognition, we tested this recognition sequence with the ^{5hm}C replaced by a canonical cytosine. Sequences for the junction constructs were designed with asymmetric ends that are only fully complementary in a four-stranded junction, thereby discouraging unwanted duplex formation. The GCC and G^{5hm}CC junction cores have both been illustrated as junction-stabilizing sequences[42,50]. The duplex and junction forms of the DNA were confirmed by electrophoresis using native gels run at various concentrations of polyacrylamide, comparing migration of each species against a duplex ladder of known strand lengths.

We first tested the activity of EndoG on each of the different duplex and junction DNA constructs, and found that EndoG cleaves junctions with a higher activity than it cleaves duplexes (Fig 4.1). EndoG cut 82 and 86% of the ^{5hm}C- and C-junction substrates respectively, while only cutting ~45% of the duplexes. To promote maximum cleaving efficiency, reactions were set up with excess protein, DNA substrate, and time.

The various DNA substrates also yielded different EndoG cleavage products (Fig. 4.2). No specific product was observed from interaction of EndoG with C- or ^{5hm}C-duplexes, which may simply be explained by not having enough duplex cleaved to observe accumulation of a specific product. C-junction was cleaved to produce a quickly migrating major product that, from alignment with the standard ladder, is likely a 10bp fragment representing one arm of the

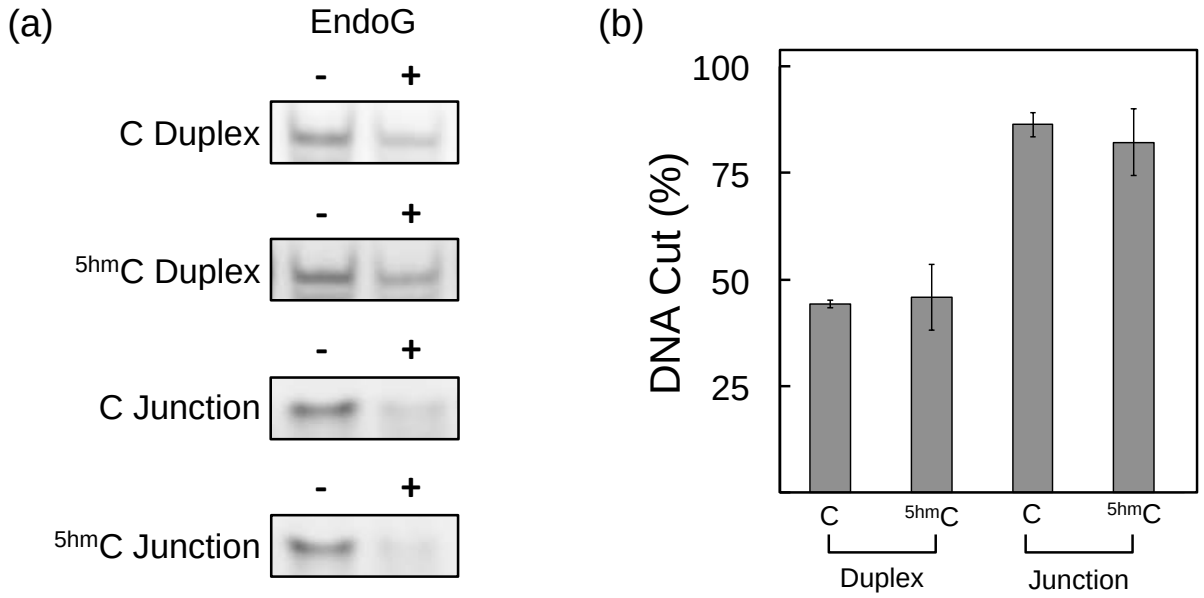


Figure 4.1: Activity of EndoG on junction and duplex DNA containing ^{5hm}C. **(a)** Gels showing activity of EndoG to 100nM duplex and junction constructs containing C or ^{5hm}C. The left lane (-EndoG) shows the original pre-cut DNA sample, and the right lane (+EndoG) is the remaining substrate after 10 minutes of cleavage with EndoG. EndoG cleaves all substrates, but junction substrates are the most depleted after the incubation period. **(b)** Quantification of activity assays. EndoG cuts 44 and 46% of C- and ^{5hm}C-duplexes respectively, while exhibiting increased cutting of C- and ^{5hm}C-junctions (86 and 82%). Errors bars represent the standard deviation of the measurements for 2 to 4 experimental replicates.

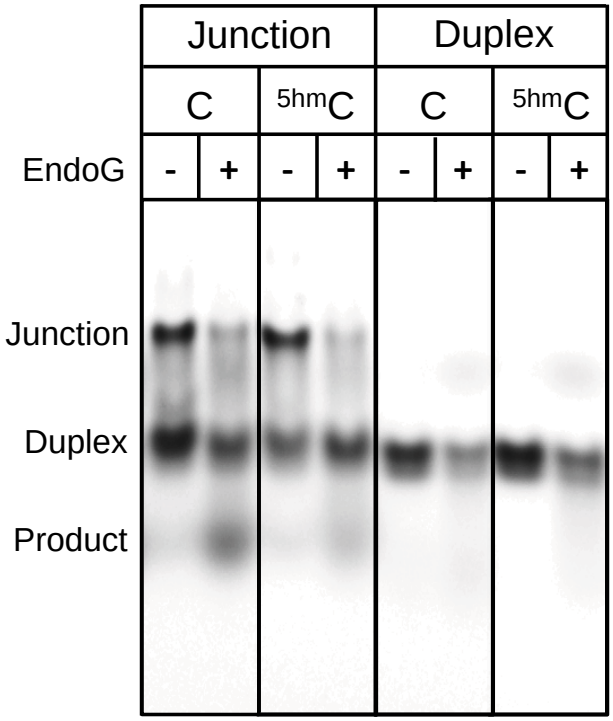


Figure 4.2: EndoG cleavage products. EndoG cleaves C-junction to produce a small product (likely 10bp fragment), while the primary product of ^{5hm}C-junction is a larger band co-migrating with duplex DNA. No specific product is observed from cleavage of C- and ^{5hm}C duplexes.

4-armed junction. This product would be consistent with cuts at all strands of the DNA junction, making no distinction among the strand sequences.

Interestingly, the ^{5hm}C-junction substrate resulted in a primary cleavage product that co-migrates with 20 bp duplex DNA. This product requires only two cleavage events at opposite sides of the junction, suggesting EndoG acts as a resolvase for ^{5hm}C-junctions. A small amount of the 10bp product was also observed, similar to C-junction cleavage. It has previously been shown that G^{5hm}CC at the junction core is structurally and thermodynamically unique from GCC[42], so there may be direct or indirect effects of ^{5hm}C vs C on EndoG recognition of junctions. It should be noted that junction formation is not complete (despite careful sequence design) and duplex-like constructs still remain in the "junction only" substrate. Cleavage of these improper duplex-like constructs could obscure the determination of cleaved junction products.

To further isolate the driving factor for EndoG's higher activity with junctions, we tested the binding of inactive EndoG (H138A mutation)[4] to the same DNA constructs using electrophoretic mobility shift assays (EMSA). With inactive protein capable only of binding the DNA, we hoped to isolate the binding component of substrate activity and learn whether the junction preference is due to increased affinity or cleaving efficiency. Inactive EndoG H138A binds junction constructs with much higher affinity than it binds duplex constructs (Fig. 4.3a). EndoG H138A binds C-Junctions with a K_d of 1.8 ± 0.6 μM (Fig. 4.3b), while the K_d for C-Duplex is estimated to be greater than 100 μM. A three-armed junction (chicken foot) was also tested, and found to have an affinity similar to the duplex constructs (data not shown). Thus, EndoG recognizes a unique structural feature of the 4-armed junction involved in HR.

There was no difference in affinity between the ^{5hm}C and C DNA constructs. The K_d of 1.4 ± 0.4 μM for EndoG binding to ^{5hm}C-junction (Fig. 4.3c) is similar to that for C-junction

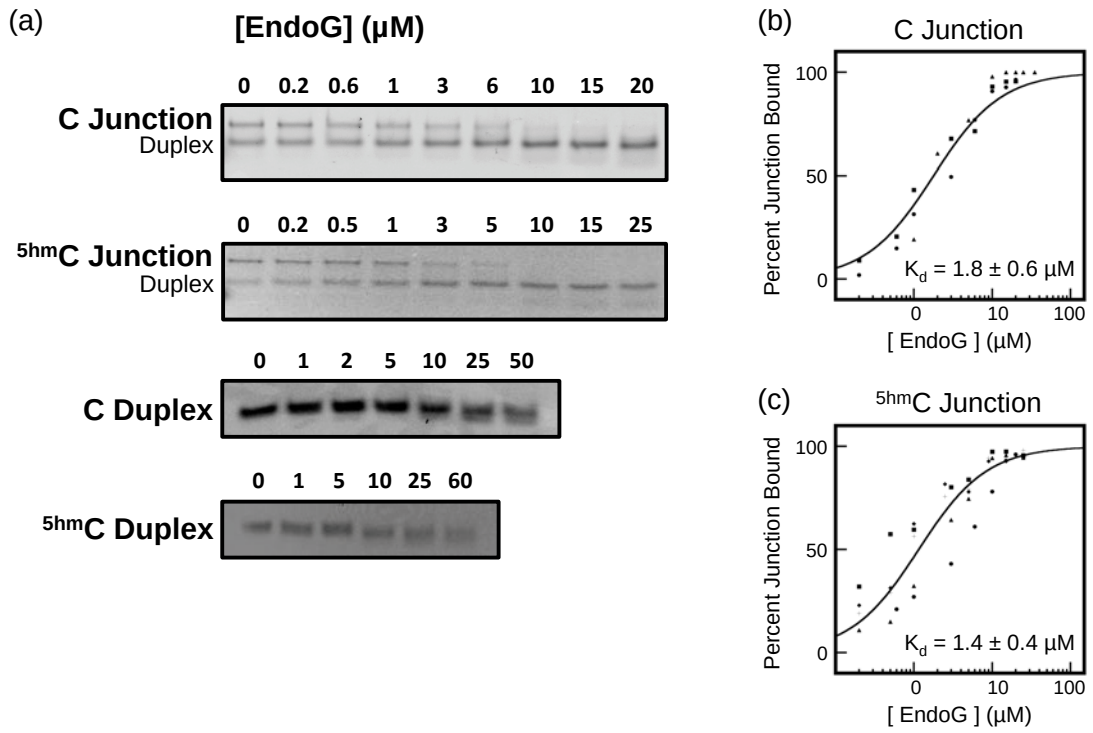


Figure 4.3: Binding of inactive EndoG-H138A to junction and duplex DNA containing ^{5hm}C. **(a)** EMSA showing binding of EndoG-H138A to 70nM duplex and junction constructs containing C or ^{5hm}C. For the both junction constructs, the junction (top band) shifts as it binds EndoG-H138A and is completely bound by ~15μM protein. The C- and ^{5hm}C-duplex constructs require much higher concentrations of EndoG-H138A to bind, displaying a much weaker binding affinity compared to the junction constructs. **(b)** Quantification of C-junction binding and fit of K_d. Data was fit with a two-state non-cooperative binding model. Binding experiment was repeated three times, each experiment represented with a different symbol shape. **(c)** Quantification of ^{5hm}C-junction binding and fit of K_d. Data was fit and quantified as for C-junction binding, but with five replicate data sets collected, each represented with a unique symbol.

(1.8 ± 0.6 μM). EndoG bound the ^{5hm}C-duplex with similar poor binding affinity (K_d > 100μM) as it did the C-duplex. Robertson *et al.* similarly reported an inability of inactive EndoG to distinguish between hydroxymethylated *versus* unmodified DNA[41]. EndoG's preference for ^{5hm}C is likely more related to catalytic activity than binding affinity, thus the H138A mutation may be responsible for the loss in ^{5hm}C recognition. Additionally, the preference for junction over duplex DNA is more dramatic (~50X K_d increase) than the reported ^{5hm}C sequence preference (~4X increase in K_m)[41].

4.3.2 Structure of mammalian EndoG confers sequence specificity

Mouse EndoG specifically recognizes ^{5hm}C DNA[41] and, within eukaryotes, ^{5hm}C DNA has only been detected in vertebrate species[30,31]. We have now solved the crystal structure of mouse EndoG (Table 4.2). A comparison of this mammalian enzyme to previous structures from Drosophila and *C. elegans* shows a structural perturbation to a potential mechanism for how the vertebrate protein preferentially recognizes ^{5hm}C.

The first barrier to crystallization was obtaining large quantities of pure EndoG. WT EndoG cannot be expressed at high concentrations because it degrades the host cell genome. The "inactive" H138A mutant still confers some activity, and it was still difficult to express the protein without causing toxicity to the cells. Sufficient expression of H138A inactive EndoG (Fig. 4.4a) was obtained with the gene inserted into a pMal-c2 plasmid, creating a fusion protein of EndoG-H138A-MBP (maltose binding protein). Our attempts at expressing EndoG-H138A alone were not fruitful, but the fusion construct allowed for cell viability while expressing large quantities of protein. Expression temperature was important, and the protein was best expressed at 10°C for 48hrs post-induction. The soluble fusion protein was then isolated on an amylose column, and after cleavage, EndoG-H138A was further purified by anion affinity, a re-pass

Table 4.2: Crystallographic parameters and refinement statistics for mouse EndoG H138A.

Mouse EndoG H138A	
Crystallographic Parameters	
Wavelength (Å)	1.5418
Resolution range (Å) *	36.50 - 2.55 (2.64 - 2.55)
Space group	P 3_1 2 1
Unit cell dimensions	
a=b (Å)	109.99
c (Å)	118.54
α = β (deg)	90
γ (deg)	120
Total reflections	2503882
Unique reflections *	27527 (2707)
Multiplicity*	17.3 (11.7)
Completeness (%) *	84.95 (78.33)
Mean I/σ(I) *	8.22 (0.83)
Wilson B-factor	22.72
R_{merge}	0.319
R_{pim}	0.105
R_{meas}	0.312
Refinement Statistics	
Reflections used for R-free	6.22%
R_{cryst} *	0.2061 (0.2554)
R_{free} *	0.2619 (0.3073)
Number non-hydrogen atoms	3932
RMSD for bond lengths (Å)	0.008
RMSD for bond angles (deg)	0.91
Average B-factor	26.97
Macromolecules	26.85
Solvent	28.31

**Values for the highest-resolution shell are given in parentheses*

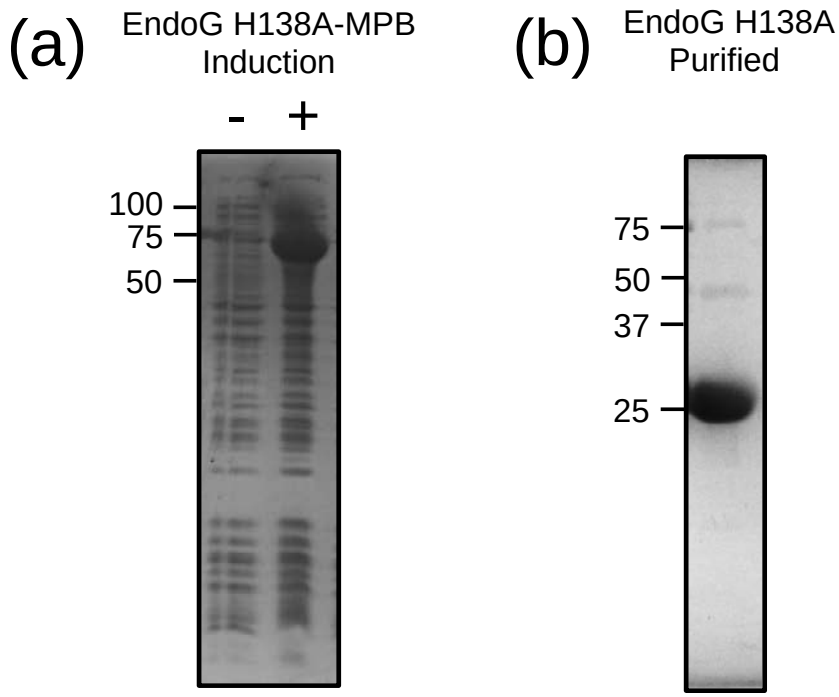


Figure 4.4: Expression and purification of EndoG-H138A-MBP. **(a)** SDS-PAGE of BL21 λDE3 expression cells pre (-) and post (+) induction. EndoG-H138A-MBP fusion has a molecular weight of 71kDa. **(b)** SDS-PAGE of final purified EndoG-H138A protein. 5μL of 3.65mg/mL protein was loaded onto the gel.

through the amylose column to remove uncleaved EndoG-H138A-MBP fusion proteins, and finally size exclusion. Purification occurred under reducing conditions and in the presence of 1mM $MgCl_2$ to bind and stabilize the protein's active site. Gel analysis of the final purified product is shown in Fig. 4.4b.

Mouse EndoG crystallized as a dimer with two copies of the protein in the asymmetric unit (ASU). The RMSD and TM score[51] of the two monomers are 0.33Å and 0.997, indicating both protein subunits are nearly identical. Mouse EndoG has a similar structure to that of the Drosophila[13] and *C. elegans*[14] homologues (Fig. 4.5a), although the two latter structures were more identical to each other than to the mouse (Table 4.3 reports RMSD and TM scores for all structure alignments). An N-terminal domain swap was also observed in the mouse structure, similar to that reported for the EndoG relative, human EXOG[52], but not observed in the Drosophila and *C. elegans* structures.

Similar to the other two species, a Mg^{2+} is bound in the active site, and the active site structure is largely conserved (Fig. 4.5b). Most of the residues contacting the DNA backbone in the *C. elegans* structure are conserved in the mouse structure (Fig. 4.5c). Although there is no DNA bound in the mouse structure, superposition of the mouse and worm structures reveals significant overlap for the placement of side chains that contact the phosphates. Thus, it can be assumed that the general binding of the DNA backbone and subsequent catalysis is identical between the mouse and *C. elegans* proteins. We assume conservation of DNA positioning and conformation, which allows us to simply dock the DNA from the worm structure into the binding site of the current mammalian structure for the following analyses.

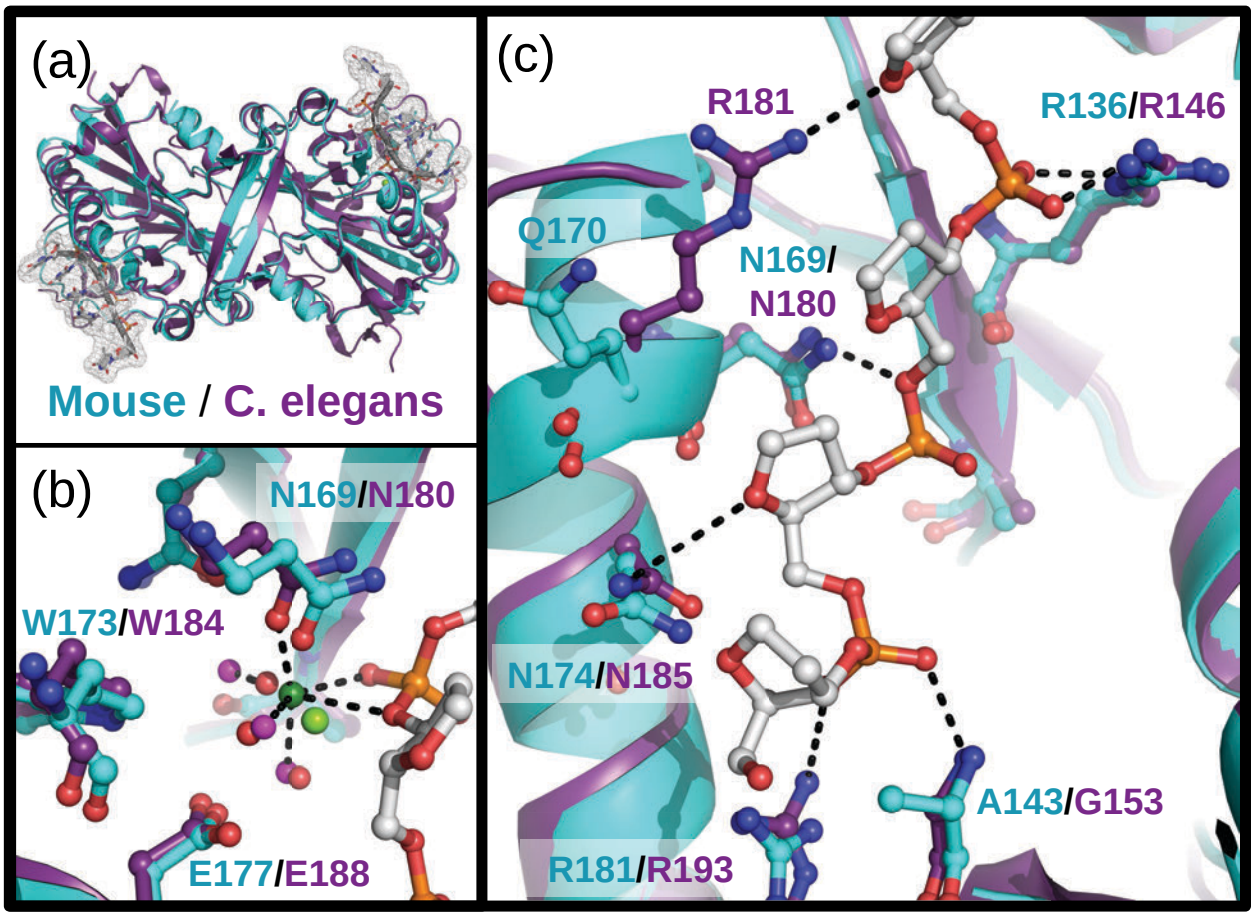


Figure 4.5: Mouse and *C. elegans* EndoG homologues have conserved structure and DNA positioning. **(a)** Overlay of mouse EndoG (cyan) and *C. elegans* CPS-6 (purple) ribbon structures reveals overall conserved folding. The bound DNA observed in the *C. elegans* structure (PDB 5GKP) is shown and outlined in white mesh, highlight the DNA binding sites on opposing sides of the EndoG dimer. **(b)** The active site is conserved between mouse and *C. elegans* EndoG homologues. Residues surrounding the bound Mg^{2+} are identical between the species, suggesting that catalytic mechanism and DNA cleavage site are conserved. The magnesium was observed in nearly identical positions for the mouse (lime green Mg^{2+}) and *C. elegans* (dark green Mg^{2+}) structures. Conserved waters (red-mouse, magenta-*C. elegans*) help to coordinate the magnesium in an octahedral geometry. **(c)** Mouse and *C. elegans* EndoG homologues have conserved contacts to the DNA backbone. Hydrogen bonds (black dashes) illustrate the contacts between *C. elegans* CPS-6 and the DNA backbone (bases omitted in the figure), thus positioning the DNA near the cleavage site. Nearly all DNA-positioning residues are conserved between the species.

Table 4.3: RMSD and template modeling (TM) scores for alignment of mouse EndoG monomers, and comparison of dimeric mouse EndoG with drosophila and *C. elegans* EndoG homologues.

	RMSD (Å)	*TM score*
Mouse Chain A + Chain B	0.33	0.997
Mouse + Drosophila	1.33	0.926
Mouse + *C. elegans*	1.37	0.912
C. elegans + Drosophila	0.97	0.976

The primary departure of mouse EndoG from the non-vertebrate structures is the unraveling of an α-helix at the DNA binding pocket (at residues D116 – A122 of the worm sequence, equivalent to D106 to F112 in mouse), resulting in a long structured loop. This structural perturbation in the mouse EndoG can be traced to the deletion of two residues between 102-103, and 105-106 in the mouse *versus* the non-mammalian sequences (Fig. 4.6). These deletions are conserved in vertebrate species sequences, while present in none of the non-vertebrate species. The key Cys residue is conserved among the three sequences, but the preceding deletions begin the structural change several amino acids prior. The α-helix observed in the non-mammalian structures is disrupted to create a long loop (Fig. 4.7a) that is stabilized by a salt bridge from Arg107 to Asp111, and an H-bond from Arg107 to the backbone of Asp111. We therefore propose that the unraveling of this α-helix provides a mechanism for vertebrate EndoG to recognize ^{5hm}C.

As a result of this conformational disruption of the α-helix, the nucleotides positioned +1 and +2 downstream of the cleavage site (Fig. 4.7b) come into close contact with the DNA at the major groove surface in the mouse EndoG-DNA docked complex. The side chains of residues Ala109 and Cys110 are well positioned to directly contact the nucleobases, thus effecting sequence specificity in the mouse enzyme—such contacts are not possible when these residues are in the α-helix of the fly and worm structures. This analysis is based on the observed DNA bound to *C. elegans* CPS-6, and slight positional adjustments are expected as DNA binds mouse EndoG. From this model, we expect that Cys110 interacts with the 5-position of a pyrimidine ring in the +2 nt position, optimal for detecting ^{5hm}C. The hydroxymethyl group would occupy an equivalent position to replace the methyl group of a thymine in this simple dock model, which is closely aligned to form an H-bond with the thiol of Cys110. This positioning of the ^{5hm}C is

Invertebrates			
NUC1-S.cerevisiae (yeast)	97-	PEsLaarnA-DRKnsfFKEDEvIp	-119
CPS-6-C.elegans (worm)	106-	PERLKhaegVDRKlCeFKpDitfp	-129
EndoG-D.melanogaster (fruit fly)	112-	aEsvaknDAVDRskCDFKqDESIH	-136
EndoG-A.gambiae (mosquito)	121-	PatvKhnDAVDRakCDFKpDESIH	-144
Vertebrates			
EndoG-D.rerio (zebrafish)	112-	aEtvt-Gss-DRKyCeFKEDESVH	-133
EndoG-X.laevis (frog)	95-	PdRLK-GsA-eRKdCeFqEDvSVH	-116
EndoG-M.musculus (mouse)	98-	PERLR-GDg-DRsaCDFrEDDSVH	-119
EndoG-B.taurus (cow)	103-	PEgLR-GDg-nRssCDFhEDDSVH	-124
EndoG-H.sapiens (human)	101-	PERLR-GDg-DRReCDFrEDDSVH	-122

Figure 4.6: Eukaryotic EndoG homologue sequence alignment. The Cys110 residue is conserved (highlighted), however vertebrate structure departure is likely due to two single amino acid deletions occurring before C110. The deletion mutations are conserved for vertebrate organisms.

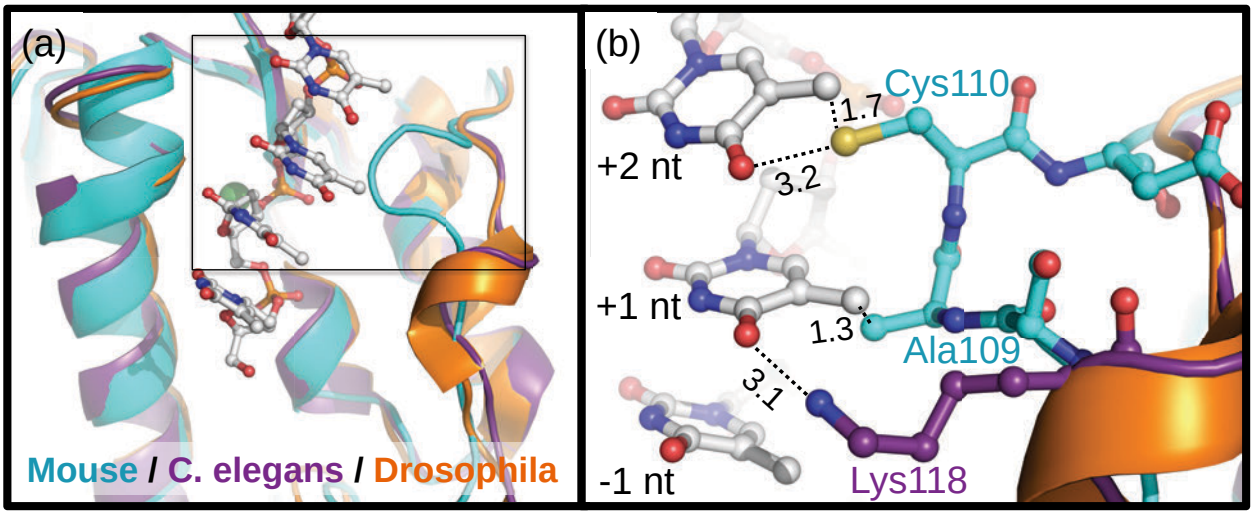


Figure 4.7: Specific contacts between mouse EndoG and nucleobases. **(a)** Overlay of published EndoG structures reveals mouse EndoG replaces the α-helix of the non-mammalian proteins with a loop shift, which brings the protein into contact with the nucleobases near the cleavage site. **(b)** Inset showing specific contacts between mouse EndoG and nucleobases. Approximated from the position of poly-T DNA bound to *C. elegans* CPS-6, mouse EndoG residues 109-110 are expected to contact the bases of nucleotides 1 and 2 positions downstream of the cleavage site. The side chain of Cys110 is expected to form an H-bond with the hydroxymethyl group of ^{5hm}C, thus conferring sequence specificity to the mammalian protein.

supported by the observation that mouse EndoG cleaves two nucleotides upstream of ^{5hm}C in the GG^GG^{5hm}CCAG sequence (^ denotes cleavage site reported by Robertson *et al.*)[41]. The backbone carbonyl or side chain of Ala109 is positioned to interact with the proceeding nucleobase, although the specific interaction is less obvious without knowing how a guanine base from the recognition sequence would be positioned.

It should be noted that the shift from helix to loop exposes Cys110 to the protein surface, rendering the vertebrate EndoG more susceptible to oxidation. We observed a disulfide bond between Cys110 of two neighboring proteins as a feature of crystal packing (Fig. 4.8). This opportunistic interaction was prompted by the close proximity of the Cys110 residues as the protein-protein lattice interface was formed. Although the disulfide bond may impact the observed orientation of the SH atoms in the Cys110 side chain, we believe that this is a result and not the cause of the disruption of the α-helix. Considering the two amino acid deletions, the remaining sequence is not long enough to form an alpha helix without unfolding other parts of the protein, thus the deletion is directly responsible for unraveling the helix. Additionally, other published EndoG crystallization conditions contained similar quantities of reducing agents[53], yet no disulfide bonds were observed in those structures. The crystallization space groups are all different for these structures, but some do pack along the DNA-binding interface. If the loop structures were identical and the conserved Cys residue was exposed, we would expect formation of disulfide bonds in the Drosophila and *C. elegans* structures.

4.3.3 Mouse EndoG is computationally predicted to favorably bind ^{5hm}C

In order to further understand mouse EndoG specificity for ^{5hm}C, we performed a molecular mechanics geometry optimization of the mammalian and the *C. elegans* enzymes in complex with sequence variations to the DNA from the worm complex. The structures of

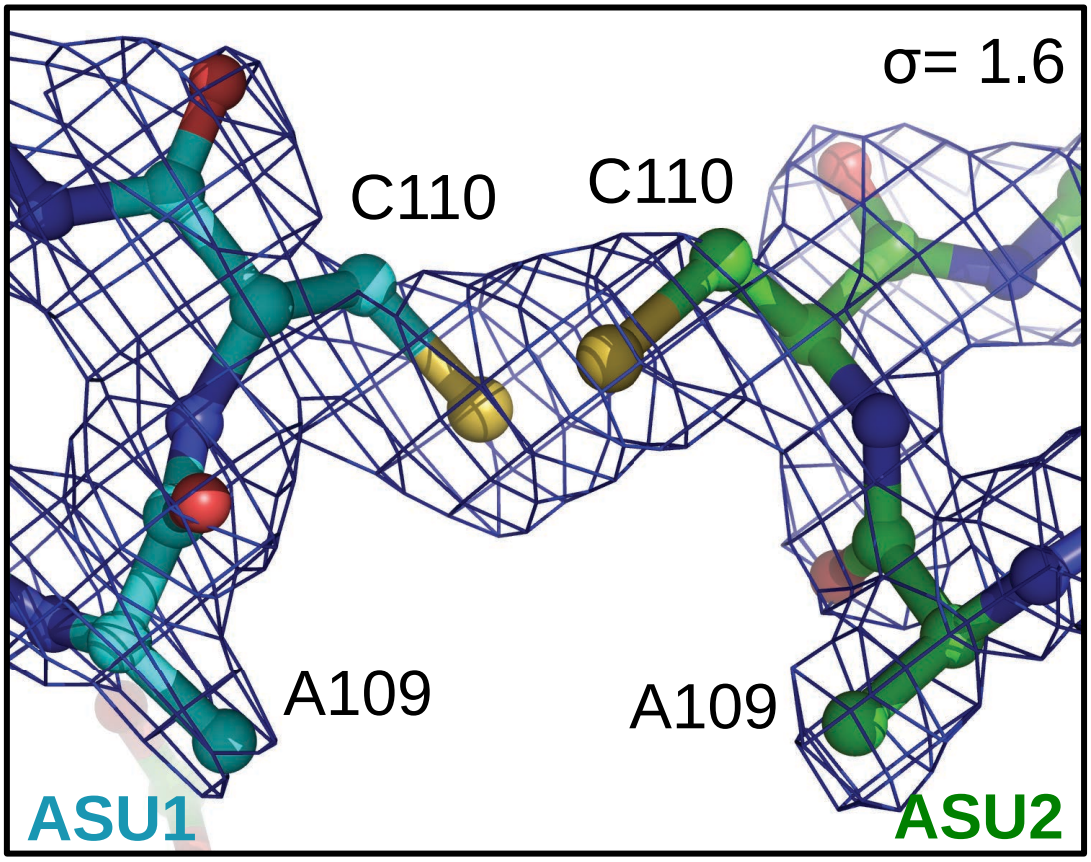


Figure 4.8: Electron density indicates presence of C110-C110 disulfide bond. $2F_o$-F_c map is calculated at 1.6(and drawn around C110 residues of two neighboring asymmetric units (ASU's). A disulfide bond is observed on both C110 residues in the ASU (one on each protein monomer) covalently bonding each ASU to its neighbor.

protein-DNA complexes were built assuming conservation of the DNA backbone position and mutating only the nucleotide base. Initial structures were solvated and minimized to calculate the ideal structure and energy for each complex.

The first observation from the minimized structures of the DNA complexes with mouse EndoG (Fig. 4.9a) is that the thiol group of Cys110 is H-bonded to the phosphate of the +2 nt nucleotide. In terms of the DNA, the +2 nt position showed the largest degree of sequence dependent conformational variability, as the nucleotide base shifts in order to accommodate an H-bond to the sulfur of Cys110. It is clear that of these H-bonds, the O—H···S interaction of ^{5hm}C would be the most energetically favorable. The cytosine base lacks a bulky group at the 5-position and rotates trying to accommodate a C—H···S interaction to the Cys110 thiol. These energy-minimized structures reflect EndoG bound to single-stranded DNA, and the nucleotide bases are predicted to be much more constrained when engaged in base pairing in a double-stranded or junction DNA context.

The energies associated with the geometry-optimized structures (ΔE_{min}) reflect the stabilization of each EndoG-DNA complex relative to the minimized energy of EndoG-polyT. For mouse EndoG, polyC is the preferred binding nucleotide (compared to polyT and polyG, Table 4.4). ^{5hm}C at the +2 nt position significantly improves energetic stability, making the hypothesized recognition sequence (GGG^{5hm}CC) much more favorable than a GGGCC sequence.

The number of atoms in each structure and the position of interacting waters influence the system's total minimization energy. To isolate the ^{5hm}C···Cys110 interaction energy from other interactions, we calculated the energy of the minimized CCC^{5hm}CC structure with the hydroxyl group pointed below the pyrimidine ring (able to H-bond with Cys110) and above the ring (no H-bonding potential). The Cys110 interaction added an additional 120 kcal/mol of increased

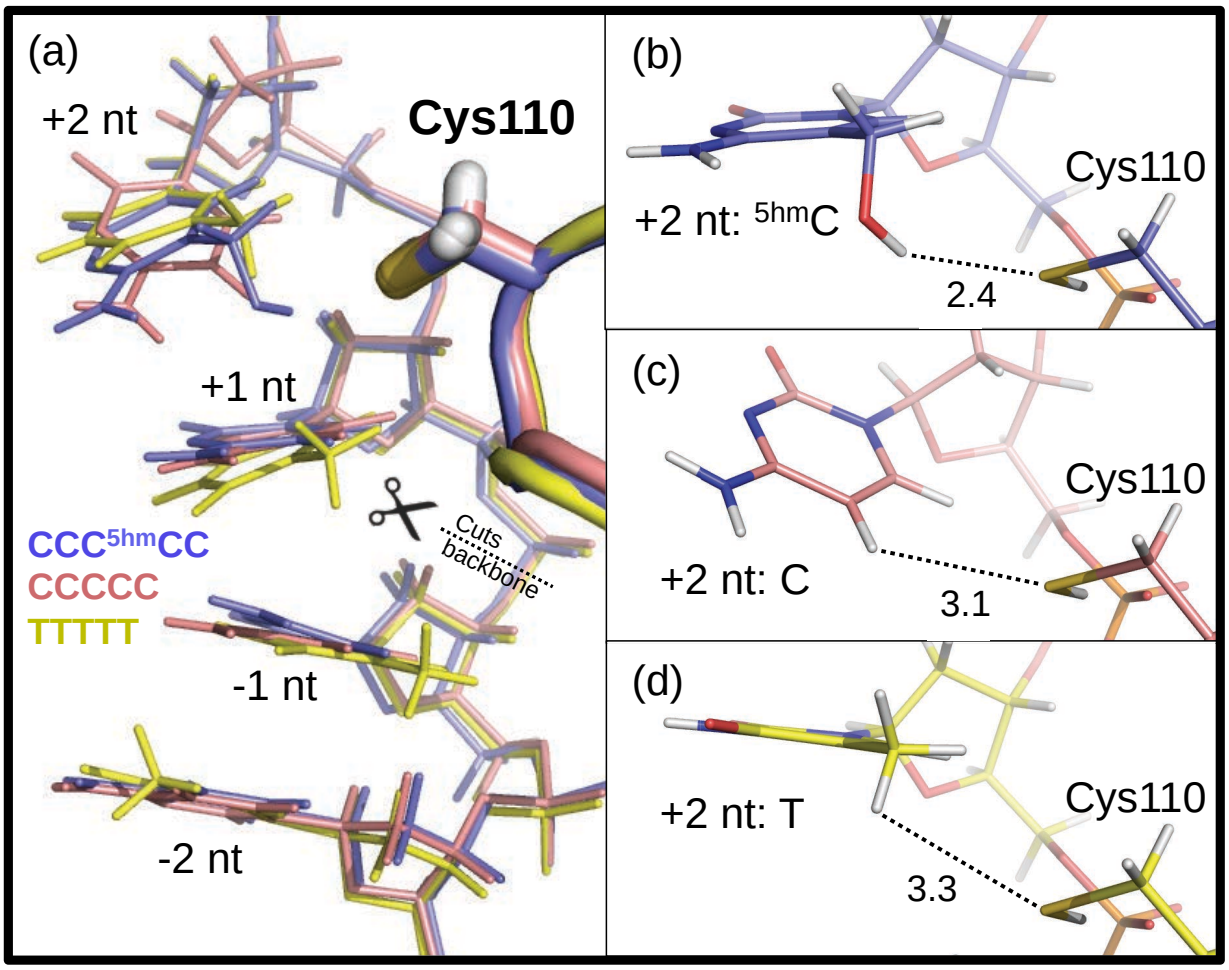


Figure 4.9: Minimized structures of mouse EndoG bound to DNA sequences. **(a)** Overlay of minimized structures for mouse EndoG bound to CCC^{5hm}CC, CCCCC, and TTTTT. The minimized position of the +2 nucleotide is highly dependent on the H-bonding potential to the sulfur on the Cys 110 residue. **(b-d)** ^{5hm}C, C, and T interactions with Cys 110. ^{5hm}C at the +2 nt position minimizes to form an H-bond between the hydrogen on the hydroxyl and the sulfur of the Cys 110 side chain. The cytosine ring does not easily accommodate an H-bond to the Cys, but the methyl of the thymine forms a weak H-bond with Cys 110.

Table 4.4: Minimization energy of mouse EndoG or *C. elegans* CPS-6 bound to different DNA sequences. ΔE is calculated relative to the minimized energy of EndoG (mouse or *C. elegans*) bound to poly-T DNA. The hydroxyl of ^{5hm}C can point in two directions – below the plane of the pyrimidine ring (H-bonding capabilities to the sulfur), or above the ring (pointed away from the sulfur).

DNA Sequence	ΔE_{min} (kcal/mol)	
	Mouse EndoG	CPS-6 (*C. elegans*)
TTTTT	0	0
CCCCC	-670	-1640
CCC^{5hm}CC (H-bonding)	-1330	-1610
CCC^{5hm}CC (no H-bond)	-1210	n.d.
GGGGG	90	-1570
GGGCC	-840	n.d.
GGG^{5hm}CC (H-bonding)	-1260	-1630

stability, attributing 60 kcal/mol per interaction since the system contains an EndoG dimer. *C. elegans* CPS-6 displayed similar ΔE_{min} for any C/G rich sequence bound. All tested sequences were much more stable than the polyT complex, but no specific ^{5hm}C preference was observed.

4.4 Conclusions

In this study we aimed to understand more about the DNA-binding behavior of EndoG. EndoG has proposed roles in DNA recombination[20–24,41], and we learned that 4-stranded junction DNA involved in HR is a substrate for EndoG cleavage. EndoG showed increased activity and affinity for junctions, meaning junction DNA is the preferred substrate in recombination contexts where it is available.

EndoG must be recognizing a unique structural feature of Holliday junctions. The 20-mer oligonucleotides used for these experiments are not long enough to reach both active sites on the EndoG dimer, so the preferential junction binding is not due to combined affinity of contacting both active sites. It is most likely that EndoG is specifically recognizing the junction crossover via the junction core. DNA junctions can take either an open-X (cruciform-like) or stacked-X (compacted) structure depending on the core DNA sequence and local divalent cation concentration[44]. In high salt the GCC and G^{5hm}CC cores form the rigid and kinetically trapped stacked-X junction[42]. This structure may incur better binding because it prevents the junction from migrating, thus providing a stable substrate for EndoG binding. Additionally the entropic penalty of binding is reduced as the substrate flexibility was already quite limited. In low salt conditions (<1-2mM), all junctions adopt the flexible open-X structure that accommodates junction migration events in HR. This structure is not thermodynamically or kinetically stable and has only been crystallized in the presence of junction-binding proteins. In this case, EndoG may induce an open-X structure on the junction as it binds. The open-X junction may bind with

higher affinity because the DNA strands are already curved to accommodate the shape of the Endo DNA binding site (Fig. 4.10). The junction structure has an inherent increased ability for single strand separation, thus EndoG might grab onto a single strand more easily.

The structure of vertebrate EndoG reveals a loop shift that provides contact to the bound nucleobases and promotes sequence specificity. Mouse EndoG specifically recognizes ^{5hm}C DNA, and that specificity is likely a common trait in all vertebrates. Although most ^{5hm}C research has been conducted in mammalian contexts, the modified cytosine is found in all vertebrate genomes while absent from invertebrate eukaryotes. We found a two amino acid deletion conserved among vertebrate EndoG sequences, and we expect this deletion is responsible for a structural change that promotes ^{5hm}C sequence specificity in all vertebrates.

EndoG has many proposed roles in the cell, from apoptosis, regulation of mitochondrial gene expression, nuclear and mitochondrial DNA recombination, and others. It is unclear whether EndoG's ^{5hm}C specificity and junction preference are related activities, although we did observe different cleavage products between ^{5hm}C-junctions and C-junctions. Thus it is likely that EndoG recognizes ^{5hm}C in both duplex and junction contexts, perhaps while executing different cellular functions. To separate EndoG's many functions, its activity must be described in correlation with ^{5hm}C levels in nuclear and mitochondrial DNA, changes in cellular oxidation, transcriptional activity, recombination, and G/C rich genomic sequences.

A persisting question is how does EndoG bind junction DNA? EndoG also has affinity for other non-canonical nucleic acid structures including R-loops, RNA, and single stranded substrates, and so perhaps these recognition mechanisms are related. It is also curious that

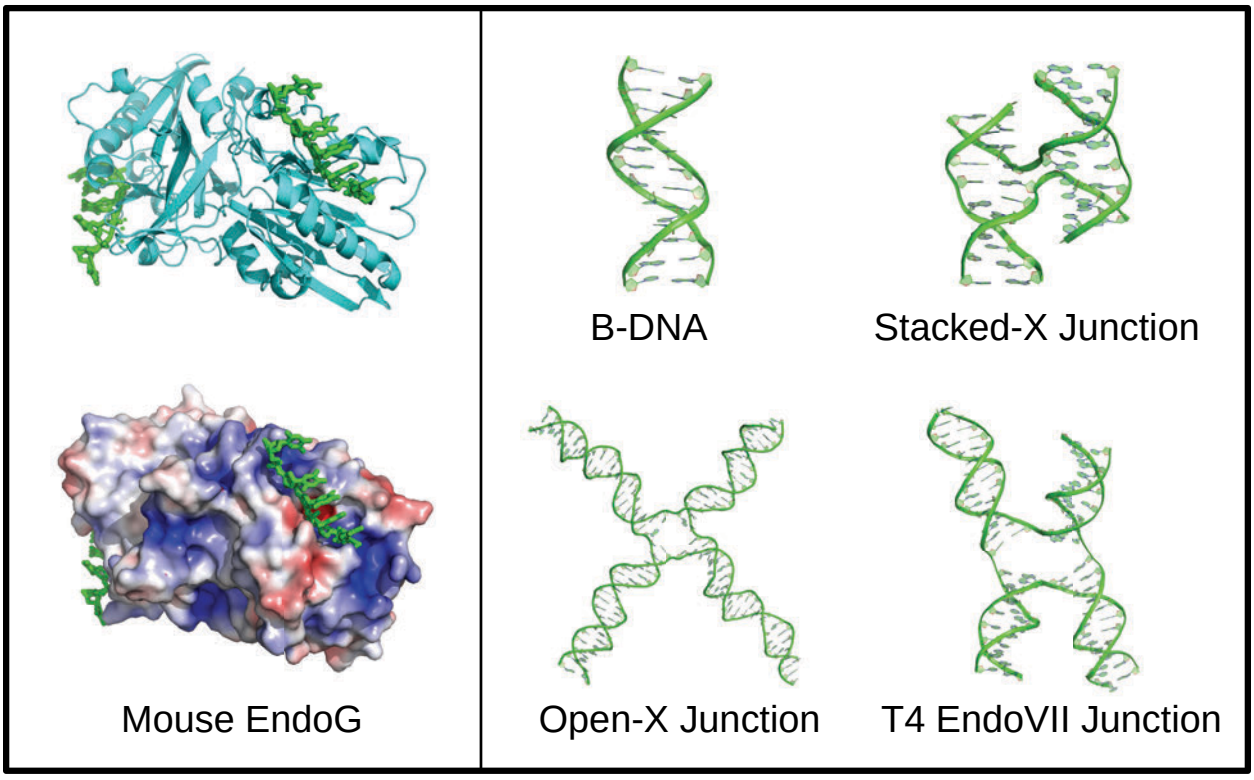


Figure 4.10: Structures of EndoG and potential DNA substrates. EndoG is shown as a ribbon structure and with a mapped electrostatic surface potential, and observed bound DNA is shown in green. The positive DNA binding patch is curved around the surface of the protein, suggesting EndoG may accommodate a curved DNA substrate. Structures of B-DNA (4C64), stacked-X junction (1P4Y), open-X junction (5J0N), and distorted junction bound to T4 Endonuclease VII (2QNC) are shown.

EndoG active sites are on opposing sides of the dimer, yet the literature suggests these active sites function completely independently of each other. In the context of EndoG's recombination function, it seems plausible that both active sites could be implicated to anchor two nearby strands of DNA. This hypothesis requires further exploration with a variety of longer canonical and non-canonical DNA substrates. Finally, more research is needed to understand EndoG's preference to cleave G/C rich DNA sequences and whether that sequence preference is solely related to the observed Cys110 and Ala109 contacts. Two other possible scenarios are that G/C rich sequences promote non-canonical DNA structures preferentially recognized by EndoG, or G/C rich sequences are recognized through an indirect readout effect in canonical duplex DNA.

In conclusion, we have discovered the Holliday junction is a preferred substrate for EndoG. Furthermore, ^{5hm}C in the junction regulates the cleavage outcome, suggesting ^{5hm}C is a regulatory marker in recombination. These results propose a role for EndoG as a resolvase in recombination. Furthermore, we have determined mouse EndoG has a unique structure that likely confers sequence specificity to ^{5hm}C DNA. The structural change is explained by a vertebrate departure from the original eukaryotic sequence, and the structure we observe is likely conserved in all vertebrate species.

4.4.1 Funding

Research reported in this publication was supported by grants from the National Science Foundation (MCB-1515521) to PSH and a pre-doctoral fellowship from the National Institute of General Medical Sciences of the National Institutes of Health (F31GM113580) to CMVZ. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation or the National Institutes of Health.

REFERENCES

(1) Cote, J., Renaud, J., and Ruiz-Carrillo, A. (1989) Recognition of (dG)n. (dC)n Sequences by Endonuclease G *264*, 3301–3310.

(2) Cummings, O. W., King, C., Holden, J. A., and Low, L. (1987) Purification and Characterization of the Potent Endonuclease in Extracts of Bovine Heart Mitochondria. *J. Biol. Chem. 262*, 2005–2015.

(3) Ruiz-carrilio, A., and Renaud, J. (1987) Endonuclease G: a (dG)n* (dC)n-specific DNase from higher eukaryotes *6*, 401–407.

(4) Schäfer, 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*, 217–28.

(5) Low, R. L. (2003) Mitochondrial Endonuclease G function in apoptosis and mtDNA metabolism: a historical perspective. *Mitochondrion 2*, 225–236.

(6) 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*, 1–14.

(7) Kieper, J., Lauber, C., Gimadutdinow, O., Urbańska, 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.

(8) Côté, J., and Ruiz-Carrillo, A. (1993) Primers for Mitochondrial DNA Replication Generated by Endonuclease G. *Science (80-.). 261*, 765–769.

(9) Ikeda, S., and Ozaki, K. (1997) Action of Mitochondrial Endonuclease G on DNA Damaged byl-Ascorbic Acid, Peplomycin, andcis-Diamminedichloroplatinum (II). *Biochem. Biophys. Res. Commun. 235*, 291–294.

(10) Ohsato, T., Ishihara, N., Muta, T., Umeda, S., Ikeda, S., Mihara, K., Hamasaki, N., and Kang, D. (2002) Mammalian mitochondrial endonuclease G. Digestion of R-loops and localization in intermembrane space. *Eur. J. Biochem. 269*, 5765–70.

(11) 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*, 48404–9.

(12) Yoon, S. M., Song, H. N., Yang, J. H., Lim, M. Y., Chung, Y. J., Ryu, S. E., and Woo, E. J. (2009) Purification, crystallization and data collection of the apoptotic nuclease endonuclease G. *Acta Crystallogr. Sect. F. Struct. Biol. Cryst. Commun. 65*, 504–7.

(13) 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*, 7312–20.

(14) Lin, J. L. J., 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.

(15) Parrish, J., Li, L., Klotz, K., and Ledwich, D. (2001) Mitochondrial endonuclease G is important for apoptosis in C . elegans *412*, 0–4.

(16) 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.

(17) Kalinowska, M., Garncarz, W., Pietrowska, M., Garrard, W. T., and Widlak, P. (2005) Regulation of the human apoptotic DNase/RNase Endonuclease G: involvement of Hsp70 and ATP. *Apoptosis 10*, 821–830.

(18) Lin, J. L. J., Nakagawa, A., Skeen-Gaar, R., Yang, W.-Z., Zhao, P., Zhang, Z., Ge, X., Mitani, S., Xue, D., and Yuan, H. S. (2016) Oxidative Stress Impairs Cell Death by Repressing the Nuclease Activity of Mitochondrial Endonuclease G. *Cell Rep.*

(19) Misic, V., El-Mogy, M., and Haj-Ahmad, Y. (2015) Endonuclease G depletion may improve efficiency of first generation adenovirus vector DNA replication in HeLa cells.pdf. *Cell. Mol. Biol. 61*, 1–8.

(20) 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. 50*, 252–261.

(21) 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. 103*, 8995–9000.

(22) Gole, B., Baumann, C., Mian, E., Ireno, C. I., and Wiesmüller, L. (2014) Endonuclease G initiates DNA rearrangements at the MLL breakpoint cluster upon replication stress. *Oncogene*.

(23) 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.

(24) Büttner, S., Carmona-Gutierrez, D., Vitale, I., Castedo, M., Ruli, D., Eisenberg, T., Kroemer, G., and Madeo, F. (2007) Depletion of endonuclease G selectively kills polyploid cells. *Cell Cycle 6*, 1072–1076.

(25) Kriaucionis, S., and Heintz, N. (2009) The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science 324*, 929–30.

(26) Szwagierczak, A., Bultmann, S., Schmidt, C. S., Spada, F., and Leonhardt, H. (2010) Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic DNA. *Nucleic Acids Res. 38*, e181.

(27) Kornberg, S. R., Zimmerman, S. B., and Kornberg, A. (1961) Glucosylation of Deoxyribonucleic Acid by Enzymes from Bacteriophage-infected Escherichia coli. *J. Biol. Chem. 236*, 1487–1493.

(28) Warren, R. A. J. (1980) Modified Bases in Bacteriophage DNAs. *Annu. Rev. Microbiol. 34*, 137–158.

(29) Delatte, B., Wang, F., Ngoc, L. V., Collignon, E., Bonvin, E., Deplus, R., Calonne, E., Hassabi, B., Putmans, P., Awe, S., Wetzel, C., Kreher, J., Soin, R., Creppe, C., Limbach, P. A., Gueydan, C., Kruys, V., Brehm, A., Minakhina, S., Defrance, M., Steward, R., and Fuks, F. (2016) Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. *Science (80-.). 351*, 282–285.

(30) Diotel, N., Yohann, M., Coumailleau, P., Gueguen, M.-M., Serandour, A. A., Salbert, G., and Olivier, K. (2017) 5-Hydroxymethylcytosine Marks Postmitotic Neural Cells in the Adult and Developing Vertebrate Central Nervous System. *J. Comp. Neurol. 525*, 478–497.

(31) Raddatz, G., Guzzardo, P. M., Olova, N., Rosado, M., and Rampp, M. (2013) Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proc. Natl. Acad. Sci. 110*, 8627–8631.

(32) Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L. M., Liu, D. R., Aravind, L., and Rao, A. (2009) Conversion of 5-methylcytosine to 5-

hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science 324*, 930–5.

(33) Stroud, H., Feng, S., Morey Kinney, S., 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.

(34) Ruzov, A., Tsenkina, Y., Serio, A., Dudnakova, T., Fletcher, J., Bai, Y., Chebotareva, T., Pells, S., Hannoun, Z., Sullivan, G., Chandran, S., Hay, D. C., Bradley, M., Wilmut, I., and De Sousa, P. (2011) Lineage-specific distribution of high levels of genomic 5-hydroxymethylcytosine in mammalian development. *Cell Res. 21*, 1332–1342.

(35) Wen, L., and Tang, F. (2014) Genomic distribution and possible functions of DNA hydroxymethylation in the brain. *Genomics 104*, 341–346.

(36) Putiri, E. L., Tiedemann, R. L., Thompson, J. J., Liu, C., Ho, T., Choi, J.-H., and Robertson, K. D. (2014) Distinct and overlapping control of 5-methylcytosine and 5-hydroxymethylcytosine by the TET proteins in human cancer cells *15*, 1–20.

(37) Wang, T., Pan, Q., Lin, L., Szulwach, K. E., Song, C. X., He, C., Wu, H., Warren, S. T., Jin, P., Duan, R., and Li, X. (2012) Genome-wide DNA hydroxymethylation changes are associated with neurodevelopmental genes in the developing human cerebellum. *Hum. Mol. Genet. 21*, 5500–5510.

(38) Robertson, J., Robertson, A. B., and Klungland, A. (2011) The presence of 5-hydroxymethylcytosine at the gene promoter and not in the gene body negatively regulates gene expression. *Biochem. Biophys. Res. Commun. 411*, 40–43.

(39) Wen, L., Li, X., Yan, L., Tan, Y., Li, R., Zhao, Y., Wang, Y., Xie, J., Zhang, Y., Song, C., Yu, M., Liu, X., Zhu, P., Li, X., Hou, Y., Guo, H., Wu, X., He, C., Li, R., Tang, F., and Qiao, J. (2014) Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base

resolution in the human brain *15*, 1–17.

(40) Guo, J. U., Su, Y., Zhong, C., Ming, G., and Song, H. (2011) Hydroxylation of 5-Methylcytosine by TET1 Promotes Active DNA Demethylation in the Adult Brain. *Cell 145*, 423–434.

(41) 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.

(42) 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.

(43) Schneider, C. A., Rasband, W. S., and Eliceiri, K. W. (2012) NIH Image to ImageJ : 25 years of image analysis. *Nat. Methods 9*, 671–675.

(44) Hays, F. A., Schirf, V., Ho, P. S., and Demeler, B. (2006) Solution formation of Holliday junctions in inverted-repeat DNA sequences. *Biochemistry 45*, 2467–2471.

(45) Minor, W., Cymborowski, M., Otwinowski, Z., and Chruszcz, M. (2006) HKL-3000: the integration of data reduction and structure solution--from diffraction images to an initial model in minutes. *Acta Crystallogr. D. Biol. Crystallogr. 62*, 859–66.

(46) Lin, J. L. J., Nakagawa, A., Lin, C. L., Hsiao, Y.-Y., Yang, W.-Z., Wang, Y.-T., Doudeva, L. G., Skeen-Gaar, R. R., Xue, D., and Yuan, H. S. (2012) Structural Insights into Apoptotic DNA Degradation by CED-3 Protease Suppressor-6 (CPS-6) from Caenorhabditis elegans. *J. Biol. Chem. 287*, 7110–7120.

(47) Adams, P. D., Afonine, P. V, Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L.-W., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W.,

Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C., and Zwart, P. H. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D. Biol. Crystallogr. 66*, 213–21.

(48) Case, D. A., Darden, T. A., T.E. Cheatham, I., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Walker, R. C., Zhang, W., Merz, K. M., Roberts, B., Hayik, S., Roitberg, A., Seabra, G., Swails, J., Götz, A. W., Kolossváry, I., Wong, K. F., Paesani, F., Vanicek, J., Wolf, R. M., Liu, J., Wu, X., Brozell, S. R., Steinbrecher, T., Gohlke, H., Cai, Q., Ye, X., Wang, J., Hsieh, M.-J., Cui, G., Roe, D. R., Mathews, D. H., Seetin, M. G., Salomon-Ferrer, R., Sagui, C., Babin, V., Luchko, T., Gusarov, S., Kovalenko, A., and Kollman, P. A. (2012) AMBER 12. University of California, San Francisco.

(49) D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R., Luo, R.C. Walker, W. Zhang, K.M. Merz, B. Roberts, S. Hayik, A. Roitberg, G. S., J. Swails, A.W. Götz, I. Kolossváry, K.F. Wong, F. Paesani, J. Vanicek, R.M. Wolf, J. L., X. Wu, S.R. Brozell, T. Steinbrecher, H. Gohlke, Q. Cai, X. Ye, J. Wang, M.-J. Hsieh, G., Cui, D.R. Roe, D.H. Mathews, M.G. Seetin, R. Salomon-Ferrer, C. Sagui, V. Babin, T., and Luchko, S. Gusarov, A. Kovalenko, and P. A. K. (2012) AMBER 13. University of California, San Francisco.

(50) Hays, F. A., Teegarden, A., Jones, Z. J. R., 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*, 7157–62.

(51) Zhang, Y., and Skolnick, J. (2005) TM-align: A protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res. 33*, 2302–2309.

(52) Szymanski, M. R., Yu, W., Gmyrek, A. M., White, M. A., Molineux, I. J., Lee, J. C., and

Yin, Y. W. (2017) A domain in human EXOG converts apoptotic endonuclease to DNA-repair exonuclease. *Nat. Commun. 8*.

(53) Lin, J. L. J., Nakagawa, A., Lin, C. L., Hsiao, Y.-Y., Yang, W.-Z., Wang, Y.-T., Doudeva, L. G., Skeen-Gaar, R. R., Xue, D., and Yuan, H. S. (2012) Structural insights into apoptotic DNA degradation by CED-3 protease suppressor-6 (CPS-6) from Caenorhabditis elegans. *J. Biol. Chem. 287*, 7110–20.

CHAPTER 5

CONCLUSION AND DISCUSSION

In this study, we sought to learn how 5-hydroxymethylcytosine (^{5hm}C) promotes recombination via interaction with Endonuclease G (EndoG). This project is foundational biochemistry that informs about the natural processes that regulate recombination in the cell, with eventual applications towards developing technology for disease treatment and research. ^{5hm}C is an important regulatory marker for recombination, and this dissertation presents key aspects towards understanding the biochemical mechanism of this regulation.

I have determined the structural and thermodynamic consequences of the ^{5hm}C marker in Holliday junctions, exploring ^{5hm}C's potential for specific recognition by proteins. Furthermore I discuss a conserved sequence and structure in vertebrate EndoG that provides an explanation for how EndoG specifically recognizes ^{5hm}C DNA. Lastly, I found that EndoG prefers Holliday junctions to duplex DNA as a substrate. This implies EndoG has a potential function as a resolvase in addition to promoting recombination via introducing double strand breaks. EndoG recognizes ^{5hm}C in the junction context as well as the duplex context, and so we expect EndoG and ^{5hm}C may have multiple roles in the recombination machinery.

Single non-covalent interaction energies can be isolated from crystallographic structures to compare the stability of Holliday junctions in solution

The Holliday junction is an optimal system for isolating and comparing single bond energies from crystallographic and differential scanning calorimetry data. We constructed

identical junctions differing by only one changed nucleotide, and we used crystallography to validate that the changed nucleotide only impacted the structure at the junction core. Differential scanning calorimetry (DSC) was used to obtain melting energies for each entire molecule, and subsequent subtractions allowed the removal of all identical energies between the junctions. The remaining energy reflected only the difference between the original and changed nucleotide, and was thus a description of the relative stability of each nucleotide for the junction. Energies determined from single crystals matched those obtained by DSC, confirming that the crystallographic structure was representative of solution state junction analytes. In chapter 2 we validated this method of obtaining specific bond energies through combined crystallography and DSC on a junction system designed to study the structure and energetic consequences of halogen bonds. In chapter 3 we applied the validated method to learn about how ^{5hm}C impacts the structure and stability of Holliday junctions relative to canonical junctions stabilized by cytosine.

^{5hm}C stabilizes Holliday junctions and causes structural and thermodynamic changes that provide possibilities for direct or indirect readout mechanisms

^{5hm}C is able to stabilize Holliday junctions and is structurally unique from C-stabilized junctions. In chapter 3 we described the structure of the ^{5hm}C junction and find that the hydroxyl replaces the typical cytosine amine donor in the H-bond to the phosphate backbone. The ^{5hm}C junction is -1.5 kcal/mol more enthalpically stable than the C junction, but suffers an entropic penalty, which cumulates in the same G of melting. The ^{5hm}C hydroxyl typically favors an out-of-plane rotamer conformation relative to the nucleotide ring, but adopts a dominant planar geometry in the junction to accommodate the H-bond.

^{5hm}C also induces a global change to the junction structure and dynamics. The measurements J_{roll} and J_{twist} describe the relative angles between the two duplex arms of the junction, and we observed an increase in those measurements for the ^{5hm}C junction. The two junction arms are being pushed apart, which exposes the junction core for protein recognition. The residues involved in the ^{5hm}C···phosphate H-bond display more conformational entropy (measured by crystallographic B-factors) compared to the canonical C···phosphate H-bond. The decreased entropy measured by DSC likely indicates constrained solvent entropy associated with the modified base. Overall, the conformational entropy and change in junction shape provides a mechanism for indirect readout of the ^{5hm}C by a protein, perhaps EndoG.

Endonuclease G preferentially cleaves Holliday junctions

In chapter 4, we have revealed the first evidence of EndoG having high affinity for Holliday junctions. We discover that, in fact, EndoG has a higher activity on junction DNA than duplex DNA with a 40% increase in cleaving efficiency. This finding is congruous with the multiple recombination contexts that employ EndoG, and implies a role for EndoG as a resolvase. We further probed the interaction with inactive EndoG to isolate the binding affinity from the cutting efficiency, and found that inactive EndoG binds junctions with a K_d of approximately 1.6μM, compared to the duplex binding K_d > 100μM. EndoG specifically produces unique cleavage products from ^{5hm}C-junctions. EndoG cleaves ^{5hm}C junction to resolve the junction into two duplexes, while C junction is cleaved into smaller fragments. Previous descriptions of EndoG included cleavage of ^{5hm}C and G/C rich duplex DNA, but we have expanded that to include ^{5hm}C- and C-junction DNA as substrates.

Vertebrate EndoG has a unique structure to recognize ^{5hm}C

We have solved the single-crystal structure of EndoG from mouse and found a helix to loop conversion that could provide a mechanism for recognition of ^{5hm}C. This structural perturbation (relative to the invertebrate EndoG homologues) is a consequence of a two amino acid deletion in the sequence that is conserved for all vertebrate EndoG sequences. Although mouse EndoG crystallized without DNA, the structure reveals all catalytic and DNA backbone-binding residues are conserved with the published *C. elegans* structure that does contain DNA. This implies the mouse and *C. elegans* proteins bind DNA in the same position, and we can model the DNA positioning based on the *C. elegans* structure. In the mouse EndoG structure, Cys 110 and Ala 109 are both positioned to contact the bases of the bound nucleotides, and therefore confer an ability to sense and select the DNA sequence bound. Particularly, the Cys 110 residue is perfectly positioned to form an H-bond with the 5-position of a pyrimidine nucleotide bound n+2 downstream from the cleavage site. This implies a perfect opportunity for EndoG to select for ^{5hm}C in the binding site. As this two amino acid deletion is conserved for all vertebrates, we expect they all contain this shifted loop and are able to sense the sequence of the bound DNA. ^{5hm}C is only present in vertebrate DNA, and we propose that vertebrate EndoG coevolved to recognize ^{5hm}C as it appeared.

In this work we have begun to uncover the mechanism of how ^{5hm}C and EndoG interact to regulate recombination, however, some questions persist. Future studies will investigate whether ^{5hm}C may provide an augmented kinetic energy barrier in the model for recombination described by sequenced-based pausing of a migrating Holliday junction. We have described the thermodynamics of static stacked-X ^{5hm}C-junctions, but it would be interesting to investigate the

kinetics of conversion between stacked-X and open-X junctions containing ^{5hm}C. Furthermore, we will continue to learn about the binding relationship between EndoG and ^{5hm}C in junction and duplex contexts. We have uncovered the mechanism by which vertebrate EndoG is able to sense the sequence of its substrate, and the next step is to further probe how EndoG interacts with a variety of sequences. The expectation is that mouse EndoG most favorably interacts with ^{5hm}C via the conserved contacting Cys110, but perhaps other nucleotides are also tolerated in this position, despite a thermodynamic penalty. EndoG has long been known to preferentially cleave G/C rich sequences, and the reasons for this are still unknown. It may be an indirect readout feature of the DNA, or an increased prevalence of ^{5hm}C in G/C rich genomic regions. Future studies will also clarify the reason for EndoG's preference to bind junction DNA. Our current hypothesis is that junction DNA more readily accommodates the curved positive binding surface of EndoG, but this has not been confirmed. In conclusion, this work has begun to answer many questions about how ^{5hm}C and EndoG interact to promote recombination. The hope for the future is to continue building the foundational understanding of recombination so it can be controlled to develop better disease treatments and biotechnology for research.