DISSERTATION

SITE-SPECIFIC FUNCTION OF ENDONUCLEASE G AND CPS6 TO ENABLE VERTEBRATE FUNCTION IN AN INVERTEBRATE MODEL

Submitted by

Ryan Scott Czarny

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

Fall 2021

Doctoral Committee:

Advisor: P. Shing Ho

Laurie Stargell
Jeffrey Hansen
Lucas Argueso

ABSTRACT

SITE-SPECIFIC FUNCTION OF ENDONUCLEASE G AND CPS6 TO ENABLE VERTEBRATE FUNCTION IN AN INVERTEBRATE MODEL

The role of mitochondrial localized Endonuclease G (EndoG) remains relatively elusive. Studies have shown that EndoG has implications in mitochondrial DNA copy number, nuclear DNA cleavage during apoptosis, and oncogenesis; however, the mechanisms and pathways have yet to be determined. Our initial work investigates the nuclease activity of EndoG as well as its binding preference for duplex DNA and Holliday Junctions. It appears that EndoG and its *C. elegans* homolog, CPS6, have slightly different functions in their in vivo systems, which has led us to query the structural modifications between the proteins. EndoG has been shown to have a preference for the 5-hydroxymethylcytosine (5hmC) epigenetic marker, an interesting feature due to the fact that invertebrate systems do not contain 5hmC in their epigenome. A key difference in the homologs arises in their DNA binding domain. The invertebrate model (CPS6) contains two additional amino acids within this region that potentially allow for an alpha helix not seen in the vertebrate model to form. This helix repositions a cysteine pointed away from the active site in CPS6, which could have consequences with regards to function. Our work investigates the addition/removal of this helix from the vertebrate and inveterate system to elucidate its role. In conjunction with the primary DNA binding site, there is a second site next to and orthogonal to the first that differs in two prolines. We investigate the role of this secondary binding site as well as the importance of the invertebrate prolines. Overall, we propose a model to determine the role of EndoG in vivo utilizing the suite of protein mutations characterized herewithin.

ACKNOWLEDGEMENTS

My initial intrigue for science came from my high school chemistry teacher, Ms. Cantrell. She was incredibly passionate about science, and chemistry specifically, and didn't let the moodiness of a bunch of high school children dampen that. I fell in love with the subject and followed up by taking her AP Chemistry course the next year. I honestly couldn't get enough of the material. The next year, I took Ms. Whiteside's AP Biology and found another subject that fascinated me. Ultimately, these two courses are what influenced me to go to Colorado School of Mines and pursue my degree there. At Mines, Dr. Melissa Krebs was my research and academic advisor. Her guidance and mentoring have had a lasting impact throughout my academic career and my appreciation of her philosophies led me to choose the Ho lab for my PhD work.

I would like a small bit of time to thank some very special people that helped me in getting through my PhD graduate degree.

The first is Grace Heaslip. Grace has been an incredible mentor in life and in science. She has brainstormed countless experiments with me and truly helped shape me into the scientist I am today. Also, climbing. Grace introduced me to the sport during my first year of my graduate program and it has become a huge part of my life and method for relieving stress. Finally, Grace has been my biggest asset when it comes to my mental health. She has been there through the good times and the bad, helped me see things that I had chosen to ignore, and been the support through it all. I can't thank her enough.

Next, my roommates. Lindsay, Julie, and Wyatt are a goofy bunch of really smart scientists. The amount that we were able bounce ideas off each other at home, vent about our science, and be there to support with new directions was incredible. Outside of science, we also were able to

journey through the nature of Colorado and Wyoming and see beautiful places. I wouldn't have wanted to live with anyone else during my time in grad school.

I would also like to thank the whole Ho family. Shing, Margaret, Alex, and Ethan have been beyond kind to me and a great support throughout my education. They were the first to take me cross-country and alpine skiing, which was a life changing experience. All of them have also been so welcoming and open and have helped me immensely to have a good balance between friend and co-workers. Ethan was also a fantastic teacher and motivator with respect to learning to code for the lab. I had little to no previous coding/scripting experience before joining in the lab and Ethan helped in getting me started and addicted to the science.

Lastly, I would like to thank my partner, Spencer. Spenc has seen a pretty wide gambit of my crazy during my PhD and the stresses associated with it. I will never be able to thank him enough for the support and calming environment he provides that allows me to decompress after a day in the lab. I know that most of the time I don't make any sense when I am describing what I do in a day, but he always allows me to get it out and listens attentively.

TABLE OF CONTENTS

Chapter 1 – Introduction: Holliday Junctions and 5-hydroxymethylcytosine Roles in the cell and Potential Interactions with Endonuclease G and CPS6

1.1. Holliday Junction structure and function

Recombination events occur within a biological context to introduce variety within the genetic code or to correct errors that occur chemically, physically, or through replication[1–4]. Holliday Junctions (HJ) are DNA structures that form during the recombination process. Starting with a double-stranded break in a DNA sequences, the newly made single-stranded sequence crosses from one duplex to another in a process known as strand invasion[5,6]. Once strand invasion has displaced a portion of the doubled stranded sequence, the resulting superstructure takes the shape of a capitol H, HJ. The HJ is structurally dynamic and can switch between various exposed state with respect to the two different axis of the structure[7]. Resolvases can then interact with the DNA Junction to cleave across one of the two axis, resulting in the two different crossed over products[8].

1.2. 5hmC structure and function (and how/why it is vertebrate specific)

The transcriptional regulator methylcytosine, found throughout the genome and in mitochondrial DNA, (mtDNA) represents about 1% of all cytosines within the cell[9–11]. A small portion of those have been converted to 5-hydroxymethylcytosine (5hmC) as a byproduct of the de-methylation pathway[10–15]. Endonuclease G (EndoG), a mitochondrial localized nuclease, has been previously shown to have an increased preference for DNA cleavage when the 5hmC modification is present, introducing an interesting option for recognition through a hydrogen bond[5,8,16,17]. As hydrogen bonds form between two atoms that have a partial positive and partial

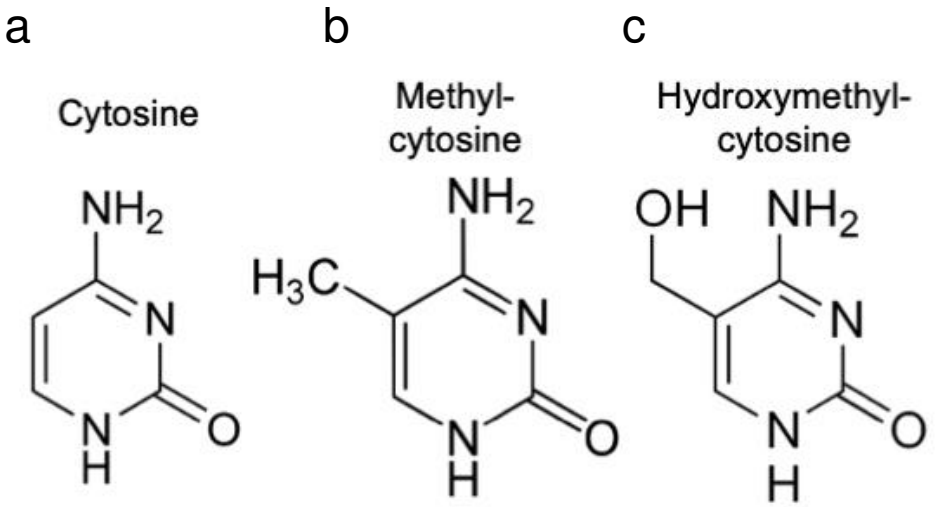

Figure 1.1. Cytosine and its epigenetic modifications of methylation and hydroxy methylation. a) Cytosine is one of the four DNA bases represented throughout life. b) When modified by a methyl group, methyl-cytosine acts as a transcription regulator. c) During the demethylation process in higher eukaryotes, an intermediate of hydroxymethyl-cytosine is formed and has been termed as an additional epigenetic marker.

negative charge respectively, the hydroxyl group of 5hmC provides the necessary electron in order to be an electron donor. It is believed the conserved cysteine forms a hydrogen bond between O-H···S-H from the 5hmC to the cysteine and an S-H···O-H from the cysteine to the phosphate backbone[17].

5hmC is a product of the ten-eleven translocation (TET) protein de-methylation process in vertebrate species[10,12,14,15,18]. Invertebrate models lack the TET protein, thus lower eukaryotes moderate their methylation through other pathways and lack the presence of the 5hmC in their epigenetic code[10,13–15]. Interestingly, 5hmC is highly enriched in recombination hotspots throughout vertebrate systems and is therefore thought to be involved in stabilizing the Holliday Junction structure[5,6,8,10].

1.3. EndoG's known functions and localizations

EndoG is a mitochondrially localized endonuclease, an enzyme that cleaves polynucleotide sequences at internal sites as opposed to at the ends of sequences[19–23]. EndoG resides within the intermembrane space (IMS) of the mitochondria, where little to nothing is known about its role[22,24–26]. Additionally, EndoG has been shown to lose its mitochondrial localization signal during apoptosis in a caspase-8 and Bid dependent manner. This triggers a migration to the nucleus to assist in genomic DNA degradation, though basal levels of EndoG have been detected in the nucleus constituatively[19,21,23,27–33]. *In vivo* knockdown of EndoG leads to an increase in genomic mutation rate, as well as an increase in oncogenesis from a failure to degrade DNA during apoptosis[29,33–36]. These findings suggest that EndoG might play a role in the Homologous Recombination pathway; a mechanism that is used to both encourage genetic diversity and correct errors in the DNA sequence[1,2,6–8,35,37,38]. Previous work in our lab has shown that EndoG has a high specificity for HJs as compared to duplex DNA[5,6,8,17]. Interestingly, the structure adopted by DNA

in HJs mimics that of DNA exiting a bound Histone in multiple positions, as can be seen in the 3D structure alignment of the Holliday Junction from PDB 2QNC and Nucleosomal DNA from PDB 5GSE (Figure 1)[39,40]. It is possible that due to *m*EndoG's (mouse) high specificity for this DNA structure, it could perform multiple functions simply based on this binding preference[5,17].

As previously mentioned, EndoG exists in the IMS, as opposed to the mitochondrial matrix (MM)[22,24–26]. This presents the potential for unique functions, given that the mtDNA exists within the MM[21,36,41–51]. Certain nucleases that exist in the IMS have been shown to process the mitochondrial mRNA or move back and forth across the membrane to cleave mtDNA. However, like EndoG, many of these proteins have no known mitochondrial function, [21,45,47,52]. Previously, it has been speculated that EndoG is essential for mtDNA replication[22,45,50,52,53]. mtDNA has a unique replication machinery that is different from the genomic DNA replication as it initiates off of a strand invasion and performs continuous replication in the positive and negative directions[31,45,48,51]. Though the exact role of EndoG in the process is unknown, cellular knockdowns have shown a decrease in mtDNA copy number, indicating that replication has been halted or slowed[22,31,50,52,53].

1.4. Similarities and differences between CPS6 and *m*EndoG

EndoG has been highly structurally conserved in eukaryotes with only slight variations[17,27]. The most prominent difference lies in its DNA binding domain observed in lower eukaryotes (Figure 2). For example, the *C. elegans* homolog CPS6 contains an alpha helix composed of five amino acids, whereas higher eukaryotes such as *Mus musculus* do not show the same higher order structure[27]. In looking at their sequence alignment, there are two amino acids that have been

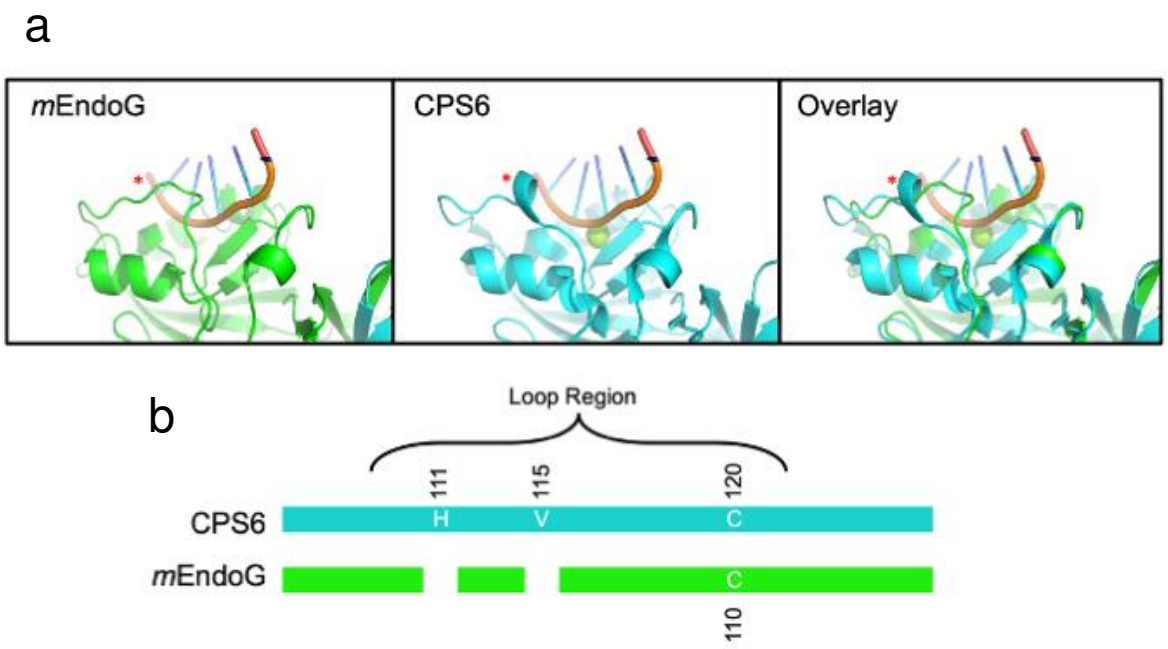

Figure 1.2. A-site for *m*EndoG and CPS6. *m*EndoG are almost structurally identical, with the exception of their A-site. a) *m*EndoG lacks an alpha helix seen in b) CPS6 that repositions the A-site loop and is believed to influence their functions. This helix difference is believed to be due to the removal of 2 amino acids when going from CPS6 to *m*EndoG.

deleted in moving from invertebrates to vertebrates. When the structures of the homologous proteins are overlaid, the start and stop of the DNA-interacting loop are anchored in the same location with the difference arising from the number of amino acids spanning the distance between them. Each turn of an alpha helix has a rise of 1.5 A and requires 3.6 amino acids to complete a turn. When comparing the structures of *m*EndoG and CPS6, it quickly becomes apparent that the two missing amino acids in *m*EndoG are at least partially responsible for the loss of the alpha helix structure, since the required sequence length needed to directly span the binding site is no longer present. Due to this alpha helix loss, a highly conserved cysteine changes position within the DNA-interacting loop that can be seen from the CPS6 co-crystal structure (PDB 5GKP)[17,27]. In invertebrate models, which contain the alpha helix, the cysteine is pointed away from the DNA sequence. In vertebrates that are missing the alpha helix, the cysteine is pointed toward the binding site. It is believed that this change in positioning is responsible for a loss in nuclease activity in invertebrates due to the cysteine pointed away from the bound DNA[17].

Little is known about CPS6's role in *C elegans*, similar to *m*EndoG[26,28,32,54]. However, CPS6$^{-/-}$ worms show a delay in cell death associated with worm development [32,55]. Since CPS6 appears to have reduced function and activity relative to *m*EndoG, it presents an interesting avenue allowing for a gain of protein function based on these amino acid deletions through protein evolution.

1.5. Appearance of B-site

Our lab has recently discovered what appears to be a secondary binding site, or the B site, directly orthogonal to the A-site based on crystallography data in which a DNA strand docked into roughly 10 amino acids downstream of the conserved cysteine (6JNU). At first glance, there doesn't appear to be an active site within this region and simply functions as a DNA binding site

and *m*EndoG appears to simply have charge interactions with both the phosphate backbone and individual nucleotides. In comparing *m*EndoG to CPS6, the B sites look structurally identical but differ in rigidity as CPS6 has two prolines where *m*EndoG has a glutamic acid and histidine, respectively. Using PyMol, we have docked a Holliday Junction into both the A and B sites and the structural fit is sound. We therefore predict that the A and B site work in tandem with respect to Holliday Junction binding and activity.

1.6. Nucleosome Structure Similarities to Holliday Junction

Previous work has looked at how *m*EndoG interactions with nucleosomal DNA and predicted that it cleaves in a nucleosome dependent manner[23]. Up to this point, the definite role of *m*EndoG in the nucleus during apoptosis has yet to be fully defined but holds potential as a nucleosome specific nuclease that assists with nuclear DNA degradation[21,22,24,31,49,56–58]. Given our labs work with Holliday Junctions, we looked at the structures of nucleosomal DNA and Junctions and super-imposed them with PyMol. When aligned, the DNA exiting the nucleosome matches that of the structure of the Junction. We thus predict that *m*EndoG is specifically binding at the nucleosome exit site and cleaving at that location. Interestingly, Histone H1 also binds the same site when incorporated into the nucleosome. Therefore, it would stand to reason that H1 would act as an *m*EndoG inhibitor with respect to nucleosomal DNA.

1.7. Research Goals

Through this work, we aim to tackle the following: How do the amino acid differences in the primary DNA binding site affect protein structure and function? What do these differences mean with respect to the conserved cysteine? How does the environment affect these proteins in structure and function? How do the structures of *m*EndoG and CPS6 change due to amino acid

addition and deletion? What is the importance of the B site and how do CPS6's proline influence its activity? How do the two sites communication and work in conjunction with one another?

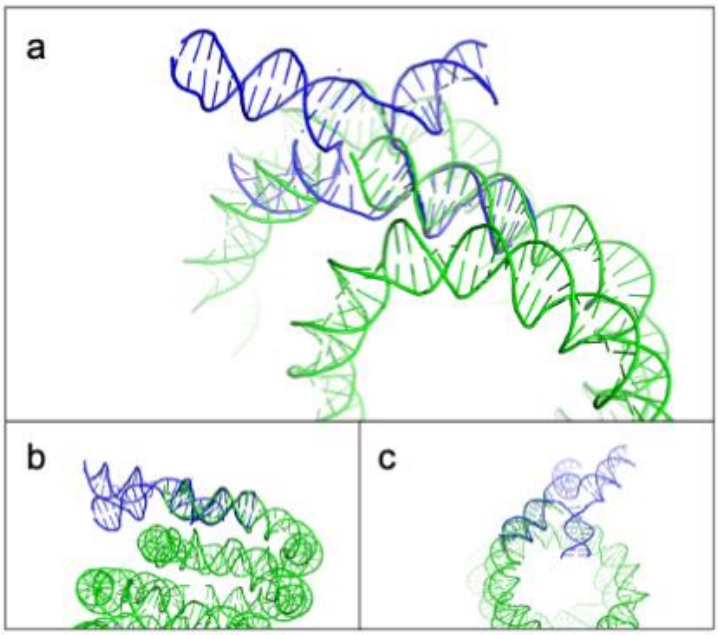

Figure 1.3. Comparing Nucleosomal DNA and Holliday Junction structure. The structures for nucleomsomal DNA and Junction DNA were super-imposed and aligned with PyMol. It becomes clear that the two structures share significant similarity, as can be seen throughout the variety of views (a being Top view, b Side view, and c Bottom view).

References

1. Nimonkar, A. V., Ozsoy, A. Z., Genschel, J., Modrich, P. & Kowalczykowski, S. C. Human exonuclease 1 and BLM helicase interact to resect DNA and initiate DNA repair. *Proc. Natl. Acad. Sci.* **105**, 16906–16911 (2008).
2. Takata, M. *et al.* Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells. *EMBO J.* **17**, 5497–508 (1998).
3. Constantinou, A. & West, S. C. Holliday junction branch migration and resolution assays. *Methods Mol. Biol.* **262**, 239–253 (2004).
4. Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J. & Stahl, F. W. The double-strand-break repair model for recombination. *Cell* **33**, 25–35 (1983).
5. Vander Zanden, C. M., Robertson, A. B. & Ho, S. P. Vertebrate Endonuclease G Preferentially Cleaves Holliday Junctions and Specifically Recognizes 5-Hydroxymethylcytosine. *Biophys. J.* **114**, 84a (2018).
6. Vander Zanden, C. M., Rowe, R. K., Broad, A. J., Robertson, A. B. & Ho, P. S. Effect of Hydroxymethylcytosine on the Structure and Stability of Holliday Junctions. *Biochemistry* **55**, 5781–5789 (2016).
7. Gibbs, D. R. & Dhakal, S. Single-Molecule Imaging Reveals Conformational Manipulation of Holliday Junction DNA by the Junction Processing Protein RuvA. *Biochemistry* **57**, 3616–3624 (2018).
8. Vander Zanden, C. M. 5-hydroxymethylcytosine and endonuclease G as regulators of homologous recombination. (2017).
9. Lee, W. *et al.* Mitochondrial DNA copy number is regulated by DNA methylation and

demethylation of POLGA in stem and cancer cells and their differentiated progeny. *Cell Death Dis.* **6**, (2015).

10. Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S. & Jacobsen, S. E. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol.* **12**, R54 (2011).

11. Song, C. X. & He, C. The hunt for 5-hydroxymethylcytosine: The sixth base. *Epigenomics* vol. 3 521–523 (2011).

12. Shi, D. Q., Ali, I., Tang, J. & Yang, W. C. New insights into 5hmC DNA modification: Generation, distribution and function. *Frontiers in Genetics* vol. 8 (2017).

13. Zhang, H. Y., Xiong, J., Qi, B. L., Feng, Y. Q. & Yuan, B. F. The existence of 5-hydroxymethylcytosine and 5-formylcytosine in both DNA and RNA in mammals. *Chem. Commun.* (2016) doi:10.1039/c5cc07354e.

14. Tahiliani, M. *et al.* Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science (80-.).* **324**, 930–935 (2009).

15. Delatte, B. *et al.* RNA biochemistry. Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. *Science* **351**, 282–5 (2016).

16. Robertson, A. B., Robertson, J., Fusser, M. & Klungland, A. Endonuclease G preferentially cleaves 5-hydroxymethylcytosine-modified DNA creating a substrate for recombination. *Nucleic Acids Res.* **42**, 13280–13293 (2014).

17. Zanden, C. M. V., Czarny, R. S., Ho, E. N., Robertson, A. B. & Ho, P. S. Structural adaptation of vertebrate endonuclease G for 5-hydroxymethylcytosine recognition and function. *Nucleic Acids Res.* **48**, 3962–3974 (2021).

18. Yamagata, K. & Kobayashi, A. The cysteine-rich domain of TET2 binds preferentially to

mono- and dimethylated histone H3K36. *J. Biochem.* **161**, 327–330 (2017).

19. van Loo, G. *et al.* Endonuclease G: a mitochondrial protein released in apoptosis and involved in caspase-independent DNA degradation. *Cell Death Differ.* **8**, 1136–1142 (2001).

20. Misic, V., El-Mogy, M. & Haj-Ahmad, Y. Role of endonuclease G in exogenous DNA stability in HeLa cells. *Biochem.* (2016) doi:10.1134/S0006297916020103.

21. Bruni, F., Lightowlers, R. N. & Chrzanowska-Lightowlers, Z. M. Human mitochondrial nucleases. *FEBS J.* **284**, 1767–1777 (2017).

22. Irvine, R. A. *et al.* Generation and Characterization of Endonuclease G Null Mice. *Mol. Cell. Biol.* **25**, 294–302 (2005).

23. Li, L. Y., Luo, X. & Wang, X. Endonuclease G is an apoptotic DNase when released from mitochondria. *Nature* **412**, 95–99 (2001).

24. Loll, B., Gebhardt, M., Wahle, E. & Meinhart, A. Crystal structure of the EndoG/EndoGI complex: mechanism of EndoG inhibition. *Nucleic Acids Res.* **37**, 7312–7320 (2009).

25. Uren, R. T. *et al.* Mitochondrial release of pro-apoptotic proteins: Electrostatic interactions can hold cytochrome c but not Smac/DIABLO to mitochondrial membranes. *J. Biol. Chem.* **280**, 2266–2274 (2005).

26. Zhou, Q. *et al.* Mitochondrial endonuclease G mediates breakdown of paternal mitochondria upon fertilization. *Science (80-.).* **353**, 394–399 (2016).

27. Lin, J. L. J., Wu, C. C., Yang, W. Z. & Yuan, H. S. Crystal structure of endonuclease G in complex with DNA reveals how it nonspecifically degrades DNA as a homodimer. *Nucleic Acids Res.* **44**, 10480–10490 (2016).

28. Parrish, J. Z., Yang, C., Shen, B. & Xue, D. CRN-1, a Caenorhabditis elegans FEN-1

homologue, cooperates with CPS-6/EndoG to promote apoptotic DNA degradation. *EMBO J.* **22**, 3451–60 (2003).

29. Zhdanov, D. D. *et al.* Apoptotic endonuclease EndoG induces alternative splicing of telomerase catalytic subunit hTERT and death of tumor cells. *Biochem. Suppl. Ser. B Biomed. Chem.* (2016) doi:10.1134/S1990750816040090.

30. Yang, S. *et al.* AKT2 blocks nucleus translocation of apoptosis-inducing factor (AIF) and endonuclease G (EndoG) while promoting caspase activation during cardiac ischemia. *Int. J. Mol. Sci.* (2017) doi:10.3390/ijms18030565.

31. Wiehe, R. S. *et al.* Endonuclease G promotes mitochondrial genome cleavage and replication. *Oncotarget* **9**, 18309–18326 (2018).

32. Parrish, J. *et al.* Mitochondrial endonuclease G is important for apoptosis in C. elegans. *Nature* **412**, 90–94 (2001).

33. Chen, C. J. *et al.* Ursolic Acid Induces Apoptotic Cell Death Through AIF and Endo G Release Through a Mitochondria-dependent Pathway in NCI-H292 Human Lung Cancer Cells In Vitro. *In Vivo (Brooklyn).* **33**, 383–391 (2019).

34. Ehrlich, M. & Wang, R. Y. 5-Methylcytosine in eukaryotic DNA. *Science* **212**, 1350–7 (1981).

35. Hoppe, M. M., Sundar, R., Tan, D. S. P. & Jeyasekharan, A. D. Biomarkers for Homologous Recombination Deficiency in Cancer. *JNCI J. Natl. Cancer Inst.* **110**, 704–713 (2018).

36. Saki, M. & Prakash, A. DNA damage related crosstalk between the nucleus and mitochondria. *Free Radical Biology and Medicine* vol. 107 216–227 (2017).

37. Lupski, J. R. Hotspots of homologous recombination in the human genome: not all

homologous sequences are equal. *Genome Biol.* **5**, 242 (2004).

38. Shah Punatar, R., Martin, M. J., Wyatt, H. D. M., Chan, Y. W. & West, S. C. Resolution of single and double Holliday junction recombination intermediates by GEN1. *Proc. Natl. Acad. Sci.* **114**, 443–450 (2017).

39. Biertümpfel, C., Yang, W. & Suck, D. Crystal structure of T4 endonuclease VII resolving a Holliday junction. *Nature* **449**, 616–620 (2007).

40. Kato, D. *et al.* Crystal structure of the overlapping dinucleosome composed of hexasome and octasome. *Science (80-.).* **356**, 205–208 (2017).

41. Kaufman, B. A. & Van Houten, B. POLB: A new role of DNA polymerase beta in mitochondrial base excision repair. *DNA Repair* (2017) doi:10.1016/j.dnarep.2017.11.002.

42. Wiley, C. D. *et al.* Mitochondrial dysfunction induces senescence with a distinct secretory phenotype. *Cell Metab.* (2016) doi:10.1016/j.cmet.2015.11.011.

43. Bailey, L. J. & Doherty, A. J. Mitochondrial DNA replication: A PrimPol perspective. *Biochemical Society Transactions* vol. 45 513–529 (2017).

44. Sharma, P. & Sampath, H. Mitochondrial DNA Integrity: Role in Health and Disease. *Cells* (2019) doi:10.3390/cells8020100.

45. Taanman, J. W. The mitochondrial genome: structure, transcription, translation and replication. *Biochim. Biophys. Acta* **1410**, 103–23 (1999).

46. Chen, X. J. Mechanism of Homologous Recombination and Implications for Aging-Related Deletions in Mitochondrial DNA. *Microbiol. Mol. Biol. Rev.* **77**, 476–496 (2013).

47. Thyagarajan, B., Padua, R. A. & Campbell, C. Mammalian mitochondria possess homologous DNA recombination activity. *J. Biol. Chem.* **271**, 27536–27543 (1996).

48. Côté, J. & Ruiz-Carrillo, A. Primers for mitochondrial DNA replication generated by

endonuclease G. *Science (80-.).* **261**, 765–769 (1993).

49. Moretton, A. *et al.* Selective mitochondrial DNA degradation following double-strand breaks. *PLoS One* **12**, (2017).

50. Yu, Z., O'Farrell, P. H., Yakubovich, N. & DeLuca, S. Z. The Mitochondrial DNA Polymerase Promotes Elimination of Paternal Mitochondrial Genomes. *Curr. Biol.* **27**, 1033–1039 (2017).

51. Wolf, D. P., Hayama, T. & Mitalipov, S. Mitochondrial genome inheritance and replacement in the human germline. *EMBO J.* **36**, 2177–2181 (2017).

52. Choi, Y. S. *et al.* Shot-gun proteomic analysis of mitochondrial D-loop DNA binding proteins: Identification of mitochondrial histones. *Mol. Biosyst.* **7**, 1523–1536 (2011).

53. Zhang, J. *et al.* Endonuclease G is required for early embryogenesis and normal apoptosis in mice. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 15782–15787 (2003).

54. Lin, J. L. J. *et al.* Oxidative Stress Impairs Cell Death by Repressing the Nuclease Activity of Mitochondrial Endonuclease G. *Cell Rep.* **16**, 279–287 (2016).

55. David, K. K., Sasaki, M., Yu, S. W., Dawson, T. M. & Dawson, V. L. EndoG is dispensable in embryogenesis and apoptosis. *Cell Death Differ.* **13**, 1147–1155 (2006).

56. Ishihara, Y. & Shimamoto, N. Involvement of endonuclease G in nucleosomal DNA fragmentation under sustained endogenous oxidative stress. *J. Biol. Chem.* **281**, 6726–6733 (2006).

57. Gregg, C., Kyryakov, P. & Titorenko, V. I. Purification of mitochondria from yeast cells. *J. Vis. Exp.* **30**, (2009).

58. Temme, C. *et al.* The Drosophila melanogaster gene cg4930 encodes a high affinity inhibitor for endonuclease G. *J. Biol. Chem.* **284**, 8337–8348 (2009).

Chapter 2 – Vertebrate Endonuclease G's Structural Adaption for 5-hydromethylcytosine Recognition*

2.1. Summary

The mitochondrial Endonuclease G (EndoG) is a protein whose function and role within the cell has remained relatively unknown. Preliminary evidence has indicated that EndoG is involved in homologous recombination and has shown recognition for the 5-hydroxymethylcytosine (5hmC) epigenetic marker. The recognition of 5hmC has been predicted to work in a hydrogen bond dependent manner through an interaction with a conserved cysteine. Our work will explore *m*EndoG's interaction with duplex and Holliday Junction DNA as well as its specificity for 5hmC in order to elucidate the protein's potential role in homologous recombination.

2.2. Introduction

Previous work in the field has established that EndoG is a mitochondrially localized proteins that migrates to the nucleus upon apoptotic signaling. However, little is known regarding the protein's function in either location throughout its lifetime. Adam Robertson et al. have shown that *m*EndoG taken from mice liver nuclear extracts promotes homologous recombination. They propose that *m*EndoG provides the cleavage needed in order for the nick to occur in the strand. The nicked strand would then be allowed to strand invade a homologous sequence to follow the homologous recombination pathway. Additionally, Robertson's group determined that sequences containing the 5hmC epigenetic modification showed an increase in their cleaved product, indicating that *m*EndoG specifically recognizes 5hmC. In this study, we sought to elucidate a preferential substrate for the nuclease as well as determine its mechanism for specifically targeting

*Adapted from published article "Structural adaption of vertebrate endonuclease G for 5-hydroxymethylcytosine recognition and function" by Vander Zanden; C.M., Czarny, R.S.; Ho, E.N.; Robertson, A.B.; Ho, P.S. (2020)

the 5-hydroxymethylcytosine epigenetic modification, as had been previously reported. Additionally, crystallographic data from our lab led us to explore the possibility of a secondary DNA-binding site. Taking these data into account, we propose potential roles for *m*EndoG in homologous recombination as well as a potential role in nuclear DNA cleavage during apoptosis.

2.3. Experimental Design

2.3.1. Cloning and Expressing *m*EndoG

One obstacle we had to overcome was expressing *m*EndoG, which is a toxic protein, in order to be used for in vitro assays. *m*EndoG was taken from a pMal plasmid and inserted into a pET-28a plasmid containing the *m*EndoG inhibitor from drosophila. *m*EndoG was inserted both up and downstream of the inhibitor (with both behind a T7 promotor) in order to determine which expression system would work best (Figure 2.1). After testing the sequences, it was ultimately determined that *m*EndoG followed by *m*EndoG inhibitor worked best. The plasmid was then transformed into BL21 Codon+ cells and we used a customized toxic expression protocol which included adding 0.1% w/v dextrose to the media and transferring the culture solution to 10 °C after reaching an OD600 of 0.5. Induction with IPTG was done at an OD600 between 0.6-0.7 and allowed to culture for 48 hours at 10 °C while shaking at 200 rpm. The cells were then spun down, pelleted, and stored at -20 °C until purification.

2.3.2. *m*EndoG Purification

Cell pellets were resuspended in EndoG Buffer A (50 mM Tris pH 8.0, 50 mM NaCl, 1 mM MgCl2, 750 uL/L BME, and 0.5% w/v glycerol) to about 25 mLs. 3 mLs of 0.5 M EDTA were added before sonicating the solution, with an additional 1-2 mLs added post sonication. The solution was then spun down at 15k rpm for 45 minutes and filtered through a 0.22 um filter. After filtering, the sample was loaded onto an MBPTrap Column with the running buffer as EndoG

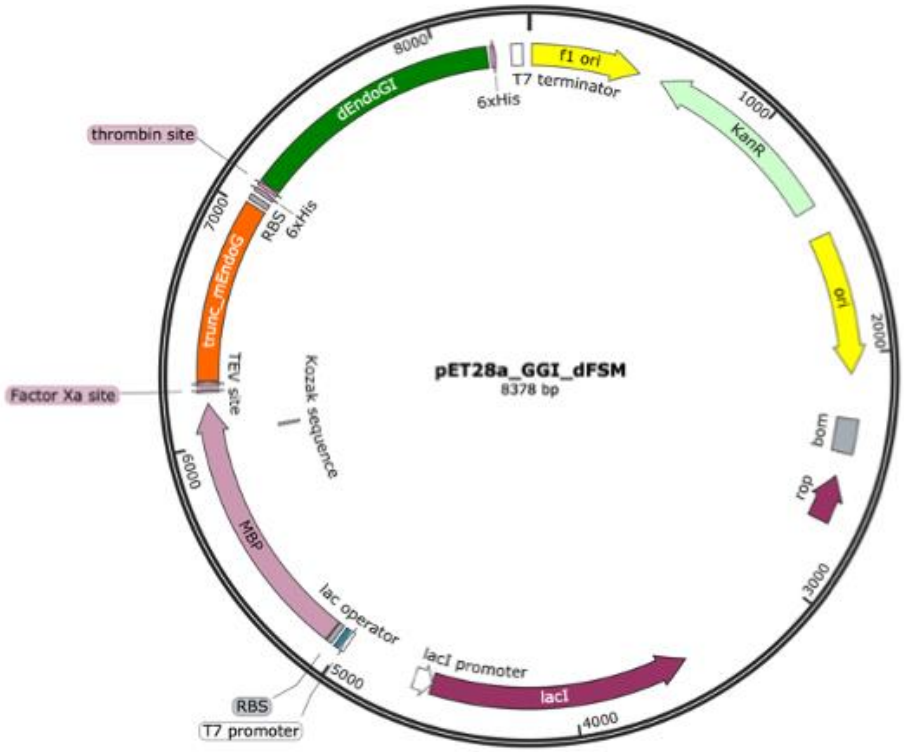

Figure 2.1. *m*EndoG with *d*EndoG Inhibitor expression system. A pET28a vector system was used and modified to include *m*EndoG with an MBP tag followed by a Ribosomal Binding Site and the *d*EndoG Inhibitor from the Drosophila system. Both proteins sat behind a T7 promotor system inducible with IPTG.

Buffer A and eluting buffer EndoG Buffer B (Buffer A + 20 mM maltose). The elution was then collected and nano-dropped in order to determine the concentration of protein, taking into account that *m*EndoG has an MBP tag attached. TEV protease was then added to cleave *m*EndoG from the MBP tag and allowed to react for 16 hours at 4 °C on a spinning plate. Finally, the sample was loaded onto a Heparin Column to bind *m*EndoG, as Heparin mimic the charge of DNA and therefore can be used to separate out DNA binding proteins. EndoG Buffer A was again used at the running buffer and EndoG Buffer C (Buffer A + 1 M NaCl) as the eluting buffer. Eluents were collected, concentrated down, and buffer exchanged into EndoG Buffer A before being stored at -80 °C until use.

2.3.3. Holliday Junction synthesis and purification

Holliday Junction was synthesized using a combination of 4 DNA oligos, in which two contained fluorescent labeling of Cy5 or FITC (see Figure 2.2). Construct 6 and 8 contain a 5hmC modified sequence, while Construct 5 and 7 do not. Equal molar ratios of each oligo, at a final concentration of 7 uM each, were added to a mix with 10X Junction Buffer (2X TBE, 0.1 M MgCl2, and 0.2 M NaCl at pH 7.8) and diluted with nano-pure water. The solution was then heated to 90 °C for 20 minutes under foil to protect from UV light. After 20 minutes, the heat block was turned off and allowed to cool to room temperature over a 2–3-hour period. Samples were then stored at -20 °C until ready for purification.

DNA Native gels were made by mixing 1.5 mL of 10X Junction Buffer, 6 mL of water, 7.5 mL of acrylamide, 75 uL of 10% APS, and 7.5 uL of TEMED. Once polymerized, gels were run in 1X Junction Buffer at 4 °C and 6 mA per gel for 1 hour. Samples were mixed with 40% sucrose at 1:10 uL sucrose to sample and then loaded into the gel. The gels were then run in a dark box for 3-4 hours to allow for Holliday Junction and duplex contaminant separation. After running, gels

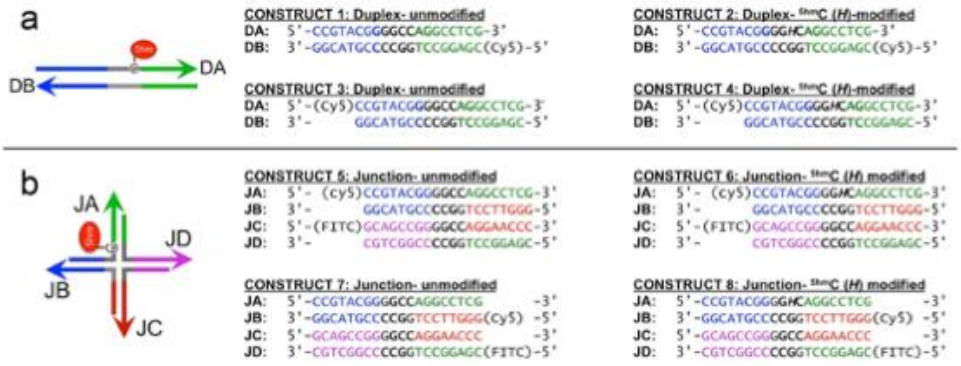

Figure 2.2. Duplex and Holliday Junction constructs used through the experimentation. a) Duplex constructs were designed to have the Cy5 label on one strand or the other as well as with and without the 5hmC modification. b) Holliday Junctions were made in the same manner with the addition for a FITC label in order to monitor cross-junction cleavage.

were imaged on a Typhoon system to precisely determine each DNA products location before being cut out of the gel with a razor blade. Gel fragments were then placed in a 0.75 mL eppi tube that had a hole punched in it with a 2 gauge needle, which sat within a 1.5 mL eppi tube. The tubes were placed in a centrifuge and spun at 14k rpm for 10 minutes to allow the fragments to pulverize through the punched hole. 1 mL of 1X Junction Buffer was then added to the pulverized material and allowed to sit at 4 °C overnight to pull the Holliday Junction into solution. The next morning, the samples were run through a 0.22 um filter and then concentrated using a 10kDa MWCO filter before being nano-dropped to determine their final concentration. Samples were then stored at -20 °C until use.

2.3.4. Nuclease Assay

For the Nuclease Assay, the reaction samples was prepared as follows: Holliday Junction was added to a final concentration of 112.5 nM (diluted in water), 11.25% v/v 10X Reaction Buffer (200 mM Tris pH 7.5, 40% glycerol, 1% Triton x-100, 0.7% v/v BME), and 74% v/v nano-pure water were mixed together. 10 uL of the reaction solution was placed in each well. 1.12 uL of 1 uM protein (diluted in EndoG Buffer A) was then added to start the reaction. The reaction was allowed to proceed at 37 °C in a thermocycler for 20-50 minutes (depending on the specific experiment) before 5 uL of Quenching Buffer (equal volumes of 0.5% SDS and Proteinase K at 1.2 mg/mL) was added and heated to 50 °C for 30 minutes. Samples were then stored at -20 °C until run on gel.

2.3.5. Denaturing Gel Assay

A BioRad Protean II xi gel assembly was used to make the denaturing gels. Additionally, a hot water circulator was setup using a water pump, hot plate, and thermometer to maintain a water temperature of 40 °C. A 16 x 20 x 0.15 cm gel was poured containing 19% acrylamide, 1%

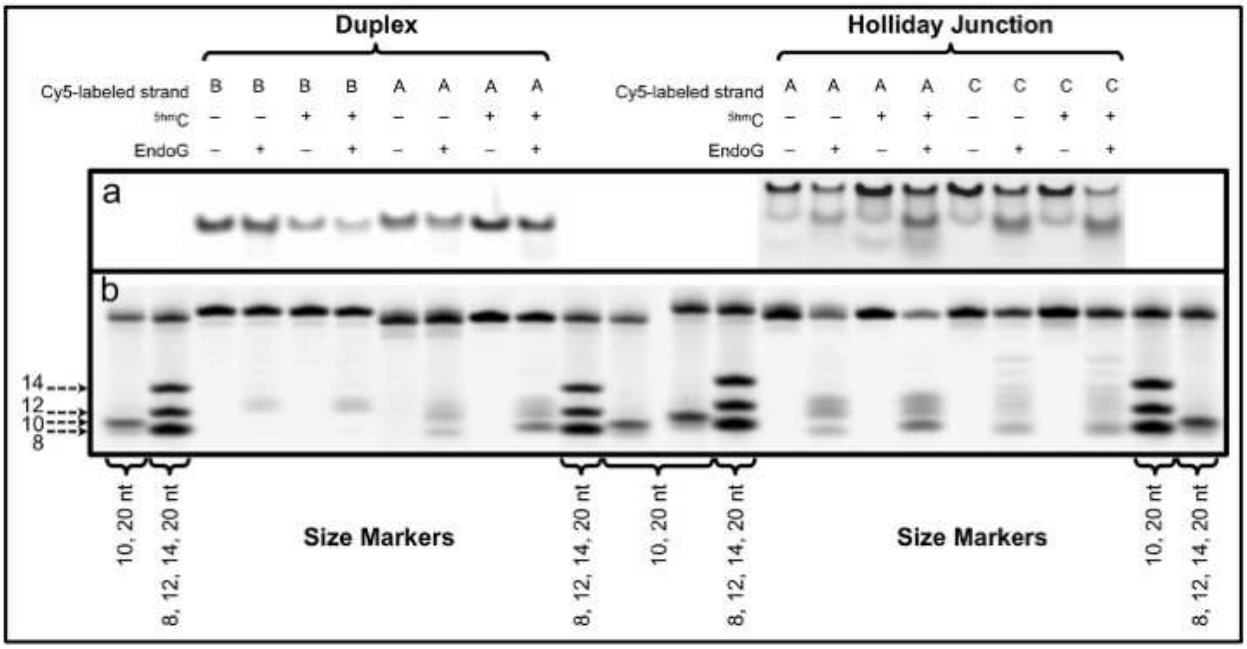

Figure 2.3. *m*EndoG nuclease native and denaturing gels for site specific cleavage. a) *m*EndoG was incubated with duplex and Holliday Junction DNA with and without the 5hmC modification in order to determine total cleavage percent of the substrate. b) The samples were then loaded into a denaturing gel to look for site specific cleavage and show a 1 nucleotide shift when 5hmC is present.

bisacrylamide, 7M urea, 12% formamide, 10% APS, and TEMED and allowed to polymerize for more than 1.5 hours. Gels were then loaded into the cassette with 0.5X TBE in the upper chamber and 1X TBE in the lower chamber. The gels were then run at 220-240 V for 1.5 hours with the 40 °C water circulating. DNA samples were prepared by adding 7.8 uL of the reaction solution, 3 uL of formamide mix (95% formamide and 10 mM EDTA), and 1.5 uL of 10X TBE and then heating to 95 °C for 10 minutes. Following heating, the samples were briefly spun down before 4 uL of 40% sucrose were added. Prior to loading, each well was rinsed of urea with 0.5X TBE using a syringe and then immediately loaded. The gels were then allowed to run in a dark box for 4-6 hours before being imaged on a Typhoon.

2.3.5. FRET Based Cleavage Assay

A FRET Based Cleavage assay was developed utilizing a PerkinElmer Victor3 and the non-canonical FRET pairs of Cy5 and FITC (see Figure 2.4 for spectral overlap). Fortunately, Cy5 does not have any absorbance at 485 nm, which is where FITC is optimized for excitation. Additionally, FITC does not have an emission at 650 nm, where the optimal Cy5 emission lies. Therefore, all emission in from Cy5 has to come from FRET pairing with FITC. Samples were prepared by adding 3.5 uL of 10X Reaction Buffer, 1.5 uL of EndoG Buffer A, and 20 uL of 50 nM Holliday Junction (diluted in nano-pure water). These 25 uL were placed in a Corning OptiPlate 384 and heated to 37 °C. 10 uL of 500 nM protein was then added to each of the wells and collection was started immediately at 37 °C. Data was then taken from the device and normalized for DNA concentration using the Cy5 signal before being curve fit in Kaleidograph.

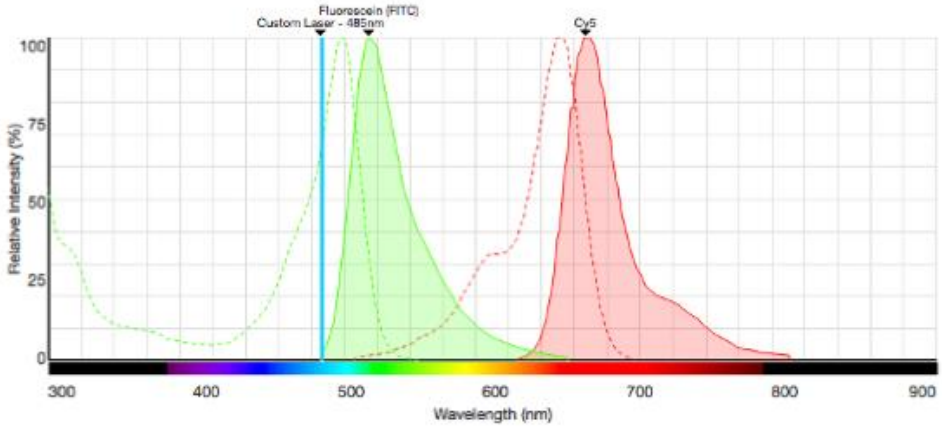

Figure 2.4. Non-traditional FRET pair of Cy5 and FITC. The Holliday Junction constructs contained both a FITC and Cy5 label. There two fluorophores are not traditional FRET pairs but show enough overlap in the FITC-emission and Cy5-excitation to properly FRET pair. No contamination occurs between the signals as the excitation energy for FITC does not excite Cy5 and the emission from Cy5 is out of the range of FITC emission.

2.4. Results

2.4.1. Binding to HJ over Duplex

We first set out to determine *m*EndoG's binding preference for duplex DNA. A catalytically inactive form of *m*EndoG was generated with the H138A mutant and DNA sequences were generated using the recognition sequence 5'-GGGGCCAG-3' reported by Robertson et al[1]. All DNA oligoes were a total of 20 nucleotides in length giving the duplex a total of 40 nucleotides. The DNA was titrated into H138A inactive *m*EndoG and run through gel electrophoretic mobility shift assay (EMSA). Intensities for each of the lanes were taken and the area for each band determined in order to calculate the binding curve for the protein to substrate. Based on this assay, the binding of H138A inactive *m*EndoG to duplex had a Kd > 50 uM. When 5hmC was introduced to the duplex sequence, no noticeable difference was found in binding preference. Initially, this brought on some confusion as the 5hmC modified duplex had been shown to have a 4-fold higher Km as compared to un-modified duplex. However, when taking this information together, it becomes clear that the binding of *m*EndoG to the nucleotide substrate does not directly influence that catalysis of the nuclease, and therefore accounts for the lack of binding specificity in the presence of 5hmC.

Next, we looked at binding preference with respect to Holliday Junction. Each Junction was comprised of 4 oligoes of 20 nucleotides each, forming the 4-armed Junction. The same recognition sequence present in the duplex DNA was also represented here in the crossing-over portion of the strands. An EMSA was also performed for H138A inactive *m*EndoG to both Holliday Junction and 5hmC modified Junction. Interestingly, both substrates have a multi-step binding model in which the first and third steps have a lower affinity of about 10 uM while the second and fourth have tighter affinities of about 0.3 and 1 uM (Figure 2.5). Yet again, 5hmC

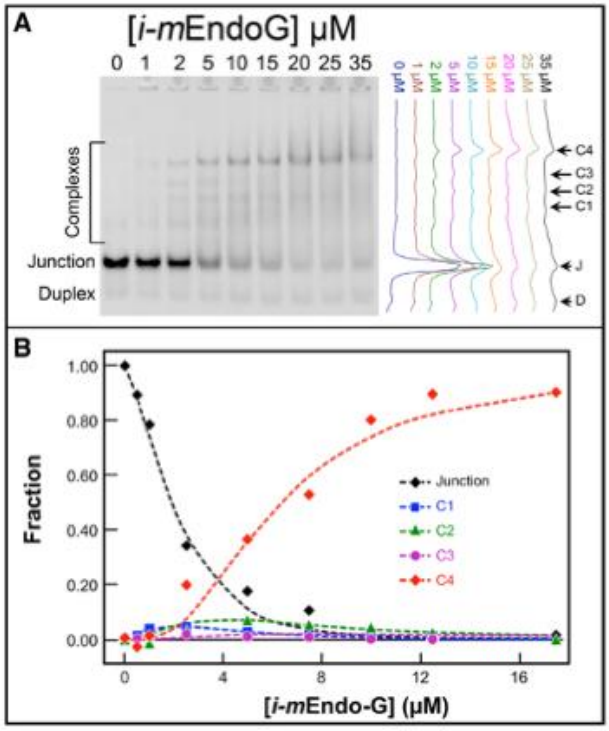

Figure 2.5. *m*EndoG to Holliday Junction binding with native gel. a) Representative gel with Holliday Junction and a titration series of inactive *m*EndoG (H96A). As the concentration of i*m*EndoG increases, 4 distinct binding events can be seen and are indicated by C1-C4. b) The fraction of DNA in each of the binding states was calculated and fit to a 4-step binding system. The Kds for each step indicate C1 at 10 uM, C2 at 1 uM, C3 at 10 uM, and C4 at 0.3 uM.

modified substrate showed no significant difference in binding preference with respect to unmodified Junctions.

Though 5hmC modified substrates did not increase the affinity for the protein, it is abundantly clear that *m*EndoG preferentially binds Holliday Junction about 50-fold more tightly as compared to duplex DNA.

2.4.2. 5hmC defines cutting efficiency and site specificity by active *m*EndoG

To further explore the disparities between the binding affinity and previously reported Km, we looked into the role of 5hmC plays in *m*EndoG's kinetic activity. An active form of *m*EndoG was expressed and purified in order to directly compare to the previously used nuclease, which had been extracted from mouse Liver Nuclear Extracts (LNE). *m*EndoG was incubated with duplex and Holliday Junction substrates, both with and without 5hmC modification, at 37 °C in order to determine cleavage of the substrates. Samples were loaded onto a native PAGE gel before being imaged to determine size through their migration by fluorescence labeling. Cleavage was determined by comparing the loss of substrate between a No Protein control and *m*EndoG sample. Simply by visual inspection, it is clear that *m*EndoG preferentially cleaved the Holliday Junction substrate and appears to generate a 20-base pair product, indicating that *m*EndoG could potentially function as a resolvase cleaving across the Junction (Figure 2.6).

More precise substrate cleavages were determined by measuring fluorescence intensities for each lane and comparing the loss of intensity for the substrate. Duplex DNA was shown to have a low amount of cleavage (at about 5-10%) and no discrimination between duplex and 5hmC modified duplex. However, *m*EndoG showed about a 5-fold increase in cleavage for Holliday Junction and 10-fold increase for 5hmC modified Holliday Junction (Figure 2.7). This

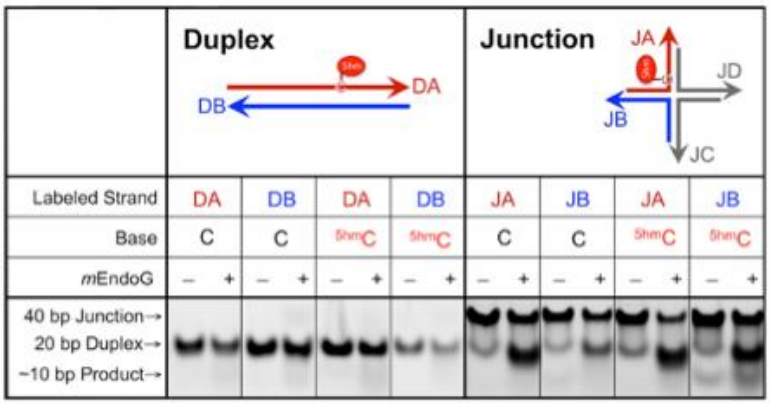

Figure 2.6. *m*EndoG substrate cleavage of duplex and junction DNA with and without 5hmC on native DNA PAGE gels. *m*EndoG was incubated with each DNA substrate and the products were run on a native DNA PAGE gel in order to determine cleavage of the product. Each lane was analyzed by intensity and the amount of cleavage was determined for each run.

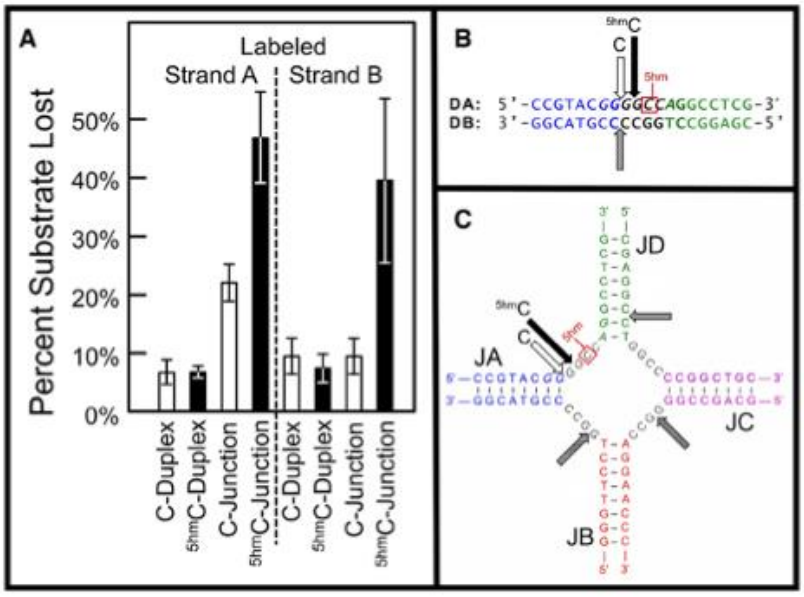

Figure 2.7. Analysis of *m*EndoG cleavage run on native DNA PAGE gels. a) Intensity data for each lane in the native gels was used to determine the percent of substrate that was cleaved during the reaction. Using the different variations of labeled strands, the data was broken down by cleavage between the A strand and the B strand in order to show specificity. b) Represents the cut sites for duplex DNA and the shift that occurs when 5hmC is present. c) Indicates the same but with junction and highlights were cleavage occurs as well as the 5hmC shift.

indicates that the catalytic preference increases from duplex < Holliday Junction < 5hmC modified Holliday Junction.

In order to validate that the cleavage specificity was not an artifact of a single time point, we developed a Förster Resonance Energy Transfer (FRET) assay in order to determine the rate of cleavage. For this experiment, the fluorescent labels on the Holliday Junction were used as a non-traditional FRET pair of Cy5-FICT on the Construct 7 and 8 Junctions. The Junction substrates were loaded into a 384 well plate and brought to 37 °C before *m*EndoG was added. Upon addition, the FRET signal was immediately read with continuous reads over about an hour time period. The loss of FRET signal indicated cleavage of the substrate by *m*EndoG as the two fluorescent labels were separated from each other through the cleavage event. Taking the data from the reads, kinetic traces were generated and fit to a pseudo-first order reaction to obtain the rate constants for cleavage. When comparing *m*EndoG's rate constants for Holliday Junction and 5hmC modified Junction, there is a 1.8-fold increase when 5hmC is present. This further validated what we had previously seen in the native PAGE assays.

Products from both the native PAGE and FRET assays were run on a denaturing PAGE gel in order to determine nucleotide resolution of the cleaved substrates. For the duplex DNA, the 20-nucleotide cleavage resulted in an 8, 12, and 13 nucleotide long products for the unmodified duplex while the 8 shifted to 9 nucleotides when 5hmC was present. Again, there were no change in overall cleavage percentage with the presence/absence of 5hmC but simply a change in cleavage site location. The same shift of 1 nucleotide towards the 5hmC modification also occurred for the Holliday Junction substrates. Taken together with the increase seem in the cleavage percentage and the rate constant, the shifted product indicates that 5hmC is responsible for recognition and

the catalytic preference increase. Overall, *m*EndoG has been shown to have a preference for Holliday Junction and specificity for the 5hmC modification.

2.4.3. Structure comparison to *m*EndoG homologs

Given this 5hmC specificity, we sought to crystalize the *m*EndoG structure in order to understand the interaction on a molecular level. A previous graduate student crystalized the H138A inactive *m*EndoG to a 2.1 A resolution, which produced a structure that was very similar to that of the C. elegans and Drosophila homologs (5gkp and 3ism respectively). One of the largest differences between the vertebrate and invertebrate models arises in the active sites of the protein in their secondary structure. All three of the proteins have a Mg+2 cation positioned within the active site in order to position the DNA for cleavage; however, the vertebrate model lacks an alpha helix that looks to interact with the DNA sequence and could potentially confer 5hmC specificity. The unwinding of this alpha helix with respect to the invertebrate models can be attributed to a loss in two amino acids within the DNA binding site, as the site is anchored on both sides and therefore lacks the primary sequence length necessary to form the secondary structure. This loss of helix repositions a highly conserved cysteine, which we believe is the primary factor involved in 5hmC recognition.

2.4.4. Conserved Cysteine specificity to 5hmC

In both of the invertebrate models of the protein, the conserved cysteine is pointed in towards the proteins (away from the DNA binding pocket), which potentially forms a disulfide bond with C169 (in Drosophila) or a weak hydrogen bond (3.5 A) with Q159 (in C. elegans). However, in mouse EndoG, the conserved cysteine is positioned to point into the DNA pocket and presents the possibility for a hydrogen bond to form. Of note, the crystal structure determined in our lab actually showed the cysteine forming a disulfide bond to a neighboring *m*EndoG protein

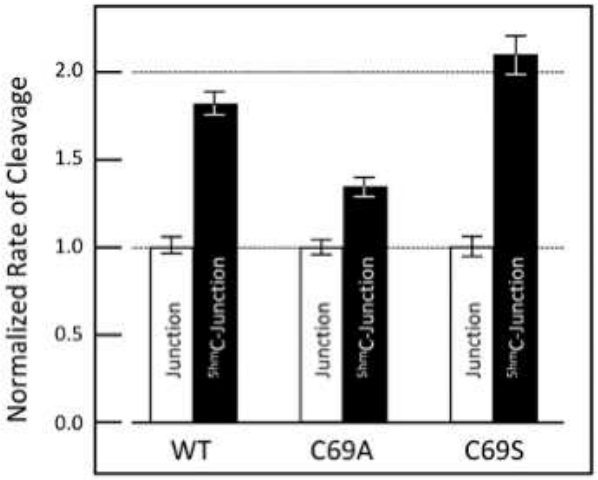

Figure 2.8. Cleavage rate constants for *m*EndoG and the conserved cysteine mutants. Using a Bulk FRET based assay, the rate constants were determined for *m*EndoG and the cysteine mutants for Junction and 5hmC modified junction. The rates were then normalized to each proteins cleavage of junction in order to determine the fold increase when 5hmC was present.

within the crystal. We believe, as has been indicated in many other crystal structures, that this interaction is simply opportunistic and not indicative of disulfide linking between *m*EndoG monomers nor is it responsible for the unwinding of the alpha helix seen in the invertebrate models.

We next modeled in the T5 DNA sequence from the C elegans structure (5gkp), replaced the nucleotides with the 5'-GGG5hmCC-3' sequence, and performed energy minimizations on the resulting structures[2]. Interestingly, two hydrogen bonds appear from the structure with one forming from the OH-SH of the hydroxyl group of 5hmC to the thiol of the conserved cysteine (C69) as well as another from the SH-OH of C69 to the phosphate backbone of the DNA sequence. When the 5hm modification is removed from the sequence and allowed to minimize again, a weaker hydrogen bond forms between the cytosine as CH-SH while the hydrogen bond to the backbone remains intact. Therefore, we believe that the conserved cysteine (C69) functions as a 5hmC recognizer in a hydrogen bond dependent manner. Additionally, this positioning of 5hmC to C69 would orient the catalytic site to be 1 nucleotide upstream of the 5hm modified base, consistent with the previous data we had collected.

In order to further validate these calculations, we generated mutants for C69 to an alanine (C69A) and to a serine (C69S) to determine the rate constants in comparison to wild type protein. Utilizing the FRET based assay, we determined that C69A has a 1.3-fold increase in rate constant from Holliday Junction to 5hmC modified Holliday Junction, 1.8-fold for wild type *m*EndoG, and 2.1-fold for C69S. Encouragingly, this follows the expected partner for hydrogen bonding with Ala-OH < Cys-OH < Ser-OH. Overall, this indicates that the C69 position is participating in a hydrogen bond with the 5hmC modification and potentially forming an additional hydrogen bond to the DNA backbone. We can thus conclude that *m*EndoG recognizes 5hmC through its conserved cysteine in a hydrogen bond dependent manner (Figure 2.8).

2.4.5. Bsite Appearing in structure

An additional feature that was present within our *m*EndoG crystal structure was the appearance of a secondary binding site, which we have termed the B-site. The B-site was determined by refinement data showing a double-helix DNA sequence that was not associated with the A-site (primary DNA binding site) and matched as an expected structure for the GC-rich oligo nucleotide that was added to the crystallization conditions. Interestingly, the B-site sits orthogonal to the A-site, which led us to consider how a Holliday Junction might position with both of the sites (Figure 2.9). Using the Holliday Junction structure from the T4 Endo VII structure, we modeled the DNA structure to align with both the DNA in the B-site found in our structure and the T5 sequence found in the C. elegans structure and were able to clearly predict the docking of *m*EndoG into the Holliday Junction. This positioning indicates that *m*EndoG will cleave the Holliday Junction 2-3 nucleotides from the center of the cross-over, which aligns with the denaturing migration assays previously discussed.

In analyzing the amino acid sequence within the B-site, it appears that all interactions with the DNA sequence occur from a charge-charge interaction with the DNA backbone, leading us to believe that the B-site does not have any sequence specificity. We then looked that the homolog C. elegans' B-site and discovered two Proline substitutions which re-orient the amino acids in a way that seems to preclude the potential for protein-DNA interactions. We therefore believe that this B-site is essential for Holliday Junction binding that may be lacking in the C. elegans homolog.

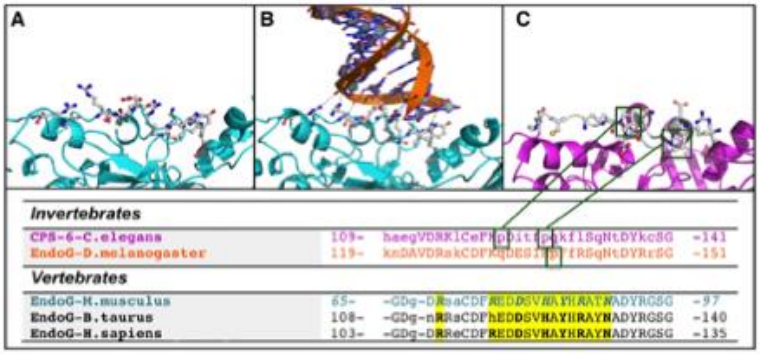

Figure 2.9. Inspection of the Bsite determined in 6NJU. Using the Bsite identified in the 6NJU structure, the Bsites of the homologous sequences of vertebrates and invertebrates were compared to determine variations. Invertebrates contain prolines in their Bsite while vertebrates do not. This could lead to potential differences in DNA interaction due to repositioning of the amino acid sequence.

2.5. Discussion

2.5.1. Specificity for 5hmC and Holliday Junction

Through this work, we have shown that Endonuclease G has a preference for Holliday Junction substrates and specificity for the 5hmC modification. *m*EndoG has a significant increase in binding affinity for 4-arm Holliday Junction as compared to duplex DNA; however, the 5hmC modification does not provide and assistance in term of binding. 5hmC recognition plays a role in catalytic rate and appears to help in aligning the substrate for cleavage through a hydrogen bond forming from C69 to the hydroxyl group of 5hmC. This conserved cysteine is positioned in a way that points it into the DNA binding pocket for mouse *m*EndoG while invertebrate models have the cysteine pointed away from the binding site and in towards the protein. The repositioning of the cysteine is believed to be due to an unwinding of an alpha helix that arises with the loss of two amino acids in the A-site DNA binding domain.

2.5.2. Relationship between A-site and B-site

Here we propose the presence of a B-site for *m*EndoG based on our crystallographic data. The B-site can be appreciated as a secondary binding site for DNA when considering the positioning and distance from the A-site. Orthogonally positioned with respect to one another, the two sites are very close in position, with a 13.5 A gap between them. We believe this to be responsible for a two-nucleotide bridge (each being ~7 A in length) that connects the two sites and encourages the idea that the two coordinate in order to bind to Holliday Junction, or Holliday Junction like structures.

2.5.3. Role of the *m*EndoG -5hmC Interaction

5-hydroxymethylcytosine has been implicated in promoting homologous recombination in acting as a signal for nuclease cleavage for eventual strand invasion by Robertson et al. We now

know that *m*EndoG specifically recognizes 5hmC and cleaves nucleotides within respect to 5hmC positioning. However, our data is not consistent with the idea of *m*EndoG functioning to cleave the duplex DNA, as it has a significant preference for Holliday Junction substrate in both binding and cleavage rate.

A potential function for the *m*EndoG-5hmC interaction would be as a Holliday Junction resolvase. Our data clearly showed the appearance of a 20-base pair product from the cleavage of our Holliday Junction substrate, which indicates that *m*EndoG has the capacity to cleave across the Junction. Data from our denaturing gels shows that *m*EndoG does have a slight increase in cleavage for one strand over another when 5hmC is present, which complicates *m*EndoG being a true resolvase. Additional work will be needed in order to definitively label *m*EndoG as a novel resolvase.

2.5.4. Potential of Junction as model for exiting nucleosome structure

An alternative to *m*EndoG's role in homologous recombination would be that the Holliday Junction could be acting as a representative substrate for DNA exiting a nucleosome or supercoiled duplex DNA[3]. Through these models, *m*EndoG would have the potential to interact with the supercoiled mitochondrial genome as well as nucleosome structures within the nucleus. We modeled Holliday Junction onto the structure of nucleosomal DNA which clearly displays that the two are incredibly similar in their conformations and would allow for *m*EndoG to dock to both in similar manners.

2.5.5. Function of Dimer *m*EndoG

Our crystal structure showed *m*EndoG as a dimer in which the active sites of each are on opposite sides of one another. This positioning could play a role in *m*EndoG's apoptotic pathway[2,4–8]. Upon apoptosis, the mitochondrial localization signal is removed and a mass

migration of *m*EndoG enters the nucleus. The supercoiled DNA in need of complete degradation would provide a prime binding site for the dimer to dock and efficiently cleave from both ends of the dimer. Additionally, there is still potential for the dimer to be involved in homologous recombination, but the details have yet to be defined.

References

1. Robertson, A. B., Robertson, J., Fusser, M. & Klungland, A. Endonuclease G preferentially cleaves 5-hydroxymethylcytosine-modified DNA creating a substrate for recombination. *Nucleic Acids Res.* **42**, 13280–13293 (2014).
2. Lin, J. L. J., Wu, C. C., Yang, W. Z. & Yuan, H. S. Crystal structure of endonuclease G in complex with DNA reveals how it nonspecifically degrades DNA as a homodimer. *Nucleic Acids Res.* **44**, 10480–10490 (2016).
3. Vander Zanden, C. M. 5-hydroxymethylcytosine and endonuclease G as regulators of homologous recombination. (2017).
4. Bruni, F., Lightowlers, R. N. & Chrzanowska-Lightowlers, Z. M. Human mitochondrial nucleases. *FEBS J.* **284**, 1767–1777 (2017).
5. Wiehe, R. S. *et al.* Endonuclease G promotes mitochondrial genome cleavage and replication. *Oncotarget* **9**, 18309–18326 (2018).
6. Ohsato, T. *et al.* Mammalian mitochondrial endonuclease G. *Eur. J. Biochem.* **269**, 5765–5770 (2002).
7. Loll, B., Gebhardt, M., Wahle, E. & Meinhart, A. Crystal structure of the EndoG/EndoGI complex: mechanism of EndoG inhibition. *Nucleic Acids Res.* **37**, 7312–7320 (2009).
8. Lin, J. L. J. *et al.* Oxidative Stress Impairs Cell Death by Repressing the Nuclease Activity of Mitochondrial Endonuclease G. *Cell Rep.* **16**, 279–287 (2016).

Chapter 3 – Site Specific Function Mapping of *m*EndoG and CPS6 to Holliday Junction DNA*

3.1. Summary

The strong relationship between a protein's function and its amino acid sequence and associated structure is a fundamental concept in biology. 5-Hydroxylmethylcytosine (^{5hm}C) is functionally significant as an epigenetic modification in vertebrate DNA but has no known function in invertebrates. Thus, it is surprising that mouse endonuclease G (*m*EndoG), which has been shown to promote ^{5hm}C-specific recombination events, has such high sequence and structural similarity to CED-3 protease suppressor-6 (CPS6), the ortholog from *Caenorhabditis elegans*, which shows no specificity for ^{5hm}C or function in recombination. In this study, we construct a series of chimeric enzymes in which sequences are swapped between *m*EndoG and CPS6 in order to determine the roles of individual amino acid deletions, insertions, and substitutions in defining the molecular properties of the vertebrate enzyme, in particular its efficiency in cleaving ^{5hm}C-modified DNAs in the context of four-stranded Holliday junctions. Our results that deleting two amino acids preceding the DNA binding site in CPS6 introduces both ^{5hm}C and junction recognition to this worm enzyme. In addition, this double mutation unwinds a single turn a-helix, supporting the model that links these vertebrate functions to this particular structural element. Using the DNA cleaving results from these and related chimeric constructs, we propose a model that couples this conserved DNA binding site (*A*-site) that is responsible for ^{5hm}C-recognition, with a second site (*B*-site) that binds an orthogonal arm of the DNA junction in *m*EndoG. This model identifies an acidic residue (aspartic acid 70) between the two sites as the important residue that couples ^{5hm}C to junction recognition. These studies on constructs of *m*EndoG indicate that the winding of this short helix requires not the insertion of the analogous amino acids from the CPS6

*Article in submission "Introducing Vertebrate Function into an Invertebrate Enzyme: Recapitulating Mouse Endonuclease G Recognition of 5-Hydroxymethylcytosine and Holliday Junctions in the Worm Ortholog" by Czarny, R.S.; Ho, E.N.; Ho, P.S. (2021)

sequence but breaking of a hydrogen bond from the *A*-site arginine to this coupling residue. Thus, vertebrate EndoG is evolutionary distinguished from its invertebrate orthologs primarily by the two amino acids that are deleted and a set of hydrogen bonds that help couple the *A*- and *B*-sites for DNA binding. The constructs created here that decouple these two sites can potentially serve as tools to help delineate the effects of ^{5hm}C from junction recognition in phenotypes of vertebrate organisms.

3.2. Introduction and Background

3.2.1. EndoG functional differences across species

Vertebrate and invertebrate organisms are readily distinguishable by their phylogenic features but cellularly distinctive functions can arise from subtle variations in biomolecular structures. One example is the recognition of 5-hydroxymethylated cytosine (^{5hm}C) DNA bases by the vertebrate endonuclease G (EndoG) enzyme [1–5]. The crystal structure of mouse EndoG (*m*EndoG) is nearly identical to that of the CED-3 protease suppressor-6 (CPS6) ortholog from *C. elegans*—a nonspecific and highly inefficient nuclease [6,7]. We had recently shown that *m*EndoG efficiently cleaves ^{5hm}C-modified DNAs in the context of the four-stranded Holliday junction [7]. In this study, we show that very simple amino acid deletions and/or substitutions to CPS6 introduce the vertebrate specific functions of *m*EndoG to the worm ortholog. In the process, we test structural models for how *m*EndoG recognizes ^{5hm}C and DNA junctions and how these functions are coupled.

3.2.2. 5-Hydroxymethylcytosine as an epigenetic marker

Modifications to cytosine bases are now recognized as essential epigenetic markers that affect the genetic functions of DNA [2–5]. It is well understood that 5-methylcytosine (^{5m}C) helps to regulate gene expression in eukaryotes [8]. The oxidation of the methyl substituent to form ^{5hm}C and subsequently to the aldehyde and carboxylic variants by the TET family of oxygenases was first

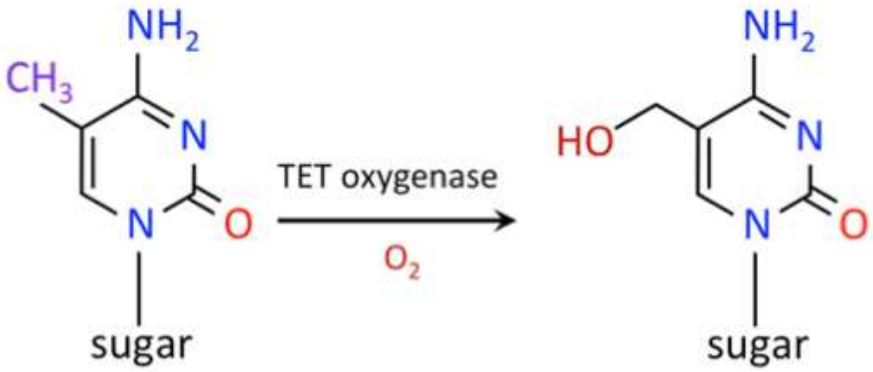

Figure 3.1. Structures of 5-methylcytosine (^{5m}C) and 5-hydroxymethylcytosine (^{5hm}C) resulting from oxidation of ^{5m}C.

seen as a mechanism to demethylate ^{5m}C (Figure 3.1) [4]. The potential biological functions of ^{5hm}C in vertebrate organisms have recently been expanded to include roles in organ development [1,2], gene regulation [9–12], DNA repair [8,10,13,14], and recombination events [7,15,16]. Robertson, *et* al demonstrated that ^{5hm}C promotes conservative homologous recombination in mice through the enzyme *m*EndoG [16], which was further shown to recognize the cognate sequence 5'-GGGGC^{5hm}***C***AG-3'.

3.2.3. Role of CPS6 in vivo

*m*EndoG has many homologs in other vertebrates and orthologs in invertebrate organisms [17–20]. This class of nuclearly encoded endonucleases is generally found in the mitochondria [21], but is present at basal levels in the nucleus [22], functioning in both organelles in cellular apoptosis [10,11,17,23–26]. The CPS6 ortholog in *C. elegans* is thought to be essential for programmed death of specific cells during development, but shows no DNA sequence or modified base specificity [13,17].

3.2.4. Previous work with *m*EndoG

In trying to understand the mechanism of *m*EndoG in the context of recombination, we had determined its single-crystal structure and found that it was very similar to the orthologs from *C. elegans* and from *Drosophila* [7]. A significant deviation from the invertebrate enzymes, however, was in the consensus DNA active site (labeled as the *A*-site), where a short a-helix (the a1-helix) in the worm and fly structures was seen to be unwound in *m*EndoG (Figure 3.2.a). The loss of the a1-helix repositioned a conserved Cys residue from the surface of the invertebrate enzyme into the DNA binding *A*-site of the mouse structure, allowing recognition of a ^{5hm}C base through an S–H···OH type hydrogen bond (H-bond). The other functionally important difference was the identification of a putative second DNA binding site (the *B*-site) present in the vertebrate, but not

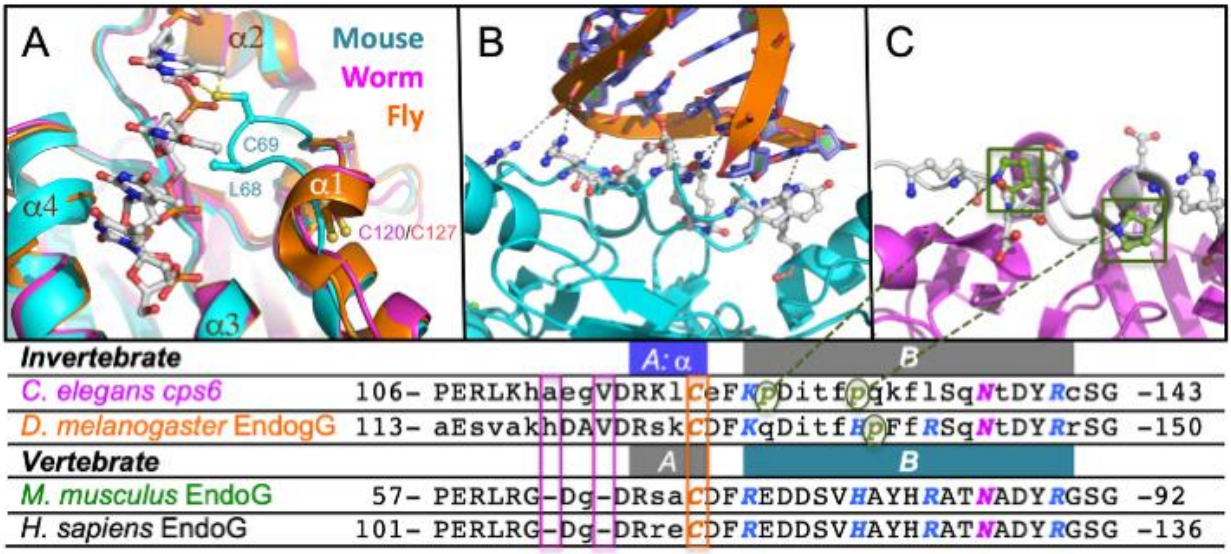

Figure 3.2. **Structure of *m*EndoG (cyan, 6NJU [7]) overlaid on homologs from *Drosophila* (orange, 3ISM [18]) and *C. elegans* CPS6 (magenta 5GKP [6]). a.** Structures with the T_5 substrate from *C. elegans* CPS6 placed in the DNA-binding *A*-sites. **b.** DNA binding *B*-site identified in the *m*EndoG structure (H-bonds between protein side chains and DNA shown as dotted lines). **c.** Analogous *B*-site in CPS6 (prolines highlighted in green carbons). Comparisons of sequences (bottom) highlight the conserved Cys (orange box) and deletions in the vertebrate sequences (magenta boxes). *A*- and *B*-sites are indicated above the sequences. Conserved amino acids seen to contact the DNA in the *B*-site are highlighted in bold italic fonts.

the invertebrate enzymes (Figure 3.2.b). We had proposed that this *B*-site positioned orthogonal to the *A*-site, is responsible for the recognition of four-armed Holliday junctions by *m*EndoG.

3.2.5. Research Goals

In the current study, we test the model that unwinding of the a1-helix and the ^{5hm}C recognition is a consequence of the deletion of two amino acids 5 and 9 residues *N*-terminal from this conserved Cys in the vertebrate *versus* the invertebrate sequences. In addition, we determine whether one or more Pro residues found in CPS6 is responsible for the inability of the invertebrate orthologs to recognize junction DNAs (Figure 3.2.c). Finally, our studies explore a possible mechanism by which the *A*- and *B*-sites communicate to couple ^{5hm}C to junction recognition in *m*EndoG.

3.3. Experimental Methods

3.3.1. FRET kinetic cleavage assays and Nuclease Assay

FRET kinetic cleavage assays and Nuclease assays were completed as previously described in Chapter 2[7].

3.3.2. Circular Dichroism

Proteins were buffer exchanged into 50 mM Tris and 150 NaClO4 at pH 7 over a Gel Filtration column and concentrated down. 200 uL of the buffer was loaded into a BioLogic MOS-500 detector to determine the High Voltage numbers and ensure a value between 300-400 mVolts in the 180-193 nm wavelengths. Triplicate runs from 185-265 nm were performed for buffer and the average was used to baseline subtract for the protein samples. All vials were rinsed with nano-pure water and acetone and allowed to sit on a vacuum for at least 10 minutes to remove any trace solutions. 200 uL of each protein solution was loaded at about 5 uM for its initial run. 100 uL of solution were read on a Cary 50 Bio UV-VIS Spectrometer to determine exact concentration based

on the 280 nm peak. If the protein had an absorbance read of greater than 1 AU on the BioLogic, the solution was diluted down to 2-2.5 uM and run again. These data were combined and the averages over multiple runs were determined. Values were then run through BeSTSEL in order to determine their structural composition.

3.3.3. Protein Mutagenesis and Purification

All primers for mutation were ordered from IDT. Solutions were made of 100 mM by suspension in 0.1X TE and then a 1:10 dilution into water. An InFusion HiFi PCR kit was used to amplify the product and a 1% agarose gel was run to confirm amplification and size. Product solutions were cleaned using the InFusion Clone Enhancer and transformed as described previously in Chapter 2[7]. All sequences were verified by GeneWiz and then transformed into BL21 Codon+ cells for expression and purified as described previously in Chapter 2[7].

3.4. Results

3.4.1. Outline of planned mutagenesis

A series of chimeric enzymes have been constructed in which sequences from *m*EndoG are swapped into the equivalent sequences in CPS6 and *vice versa* in order to test models for how the mouse enzyme has gained the abilities to recognize ^{5hm}C modified DNAs in the context of Holliday junctions. For this study, we have deleted amino acids of the DNA binding *A*-site of CPS6 and, in the complementary study, residues from CPS6 have been inserted into the analogous sites of *m*EndoG to determine the role of these specific amino acids in defining the helical structure and associated recognition of ^{5hm}C in the *A*-site. In addition, a stretch of 10 residues from the *m*EndoG *B*-site was spliced into the equivalent region in CPS6 in order to determine the role that this secondary site plays in junction recognition and binding. Proline residues conserved in the *B*-site

of CPS6 have then been introduced into *m*EndoG to test the hypothesis that these conformationally restricted amino acids disrupt the ability of this site to promote junction binding. Finally, we have mutated a Glu residue that is conserved in vertebrate systems to study its role in coupling the functions of the two DNA binding sites. The effects of these modifications to the functions of the chimeric enzymes were probed by monitoring the binding affinities and cutting efficiencies on unmodified and ^{5hm}C-modified DNA junction substrates—hallmarks of *m*EndoG.

3.4.2. a1-Helix and ^{5hm}C recognition

The alignment of sequences of EndoGs and its orthologs identified two amino acids (5 and 9 residues *N*-terminal from the conserved Cys) that are deleted in the vertebrate from the invertebrate enzymes. We had previously proposed that the loss of these two residues does not provide *m*EndoG with a sufficient number of amino acids to form the a1-helix in the *A*-site while still maintaining the conserved structural topologies that are common to all endonucleases of this type. As a result, the side chain of the conserved Cys residue (Cys 69 in *m*EndoG) is oriented into the DNA binding pocket to recognize a ^{5hm}C base through an S–H···OH H-bond. To test this model, we started by deleting one or both of these two amino acids from the CPS6 sequence and monitored the effects on its helical structure by circular dichroism (CD) spectroscopy and associated ability to cleave ^{5hm}C-modified junctions.

The CD spectrum of the CPS6 construct in which residues H111 and V115 were deleted (CPS6ΔH111ΔV115 construct) is consistent with a decrease in a-helical content of 1.8% (Figure 3.3). The complete unwinding of the a1-helix is expected to reduce the overall helical content of the (308 amino acid) protein by 1.6%. The differences in the CD spectra, therefore, support the model that the loss of two residues prior to the a1-helix results in the destabilization of this helix.

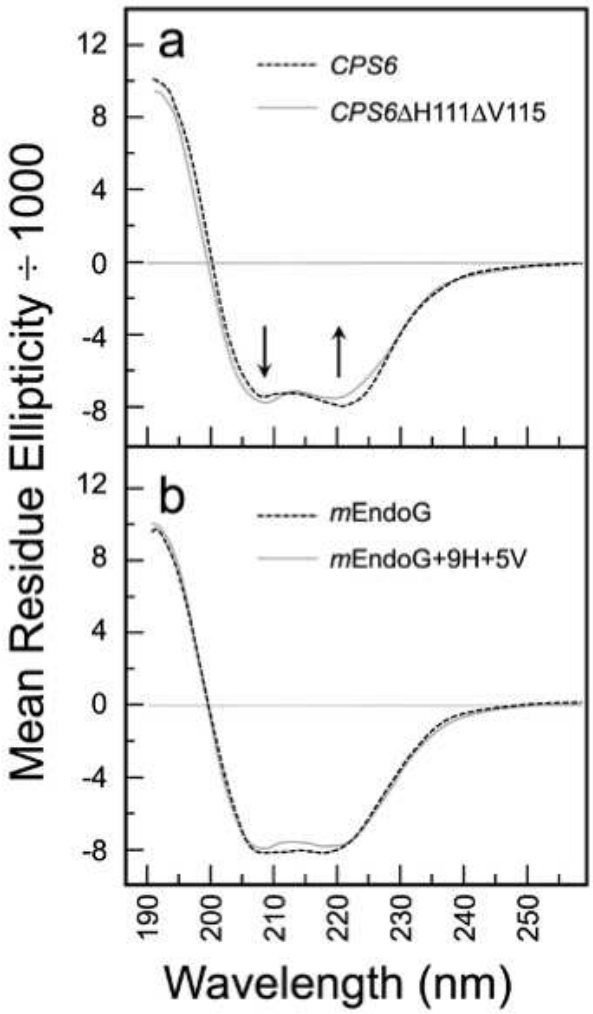

Figure 3.3. Circular Dichroism (CD) spectra of CPS6 and *m*EndoG constructs. a. The CD spectrum of wildtype CPS6 (dashed black line) is compared to its construct with residues H111 and V115 deleted (CPS6ΔH111ΔV115, solid grey line). BeSTSEL analysis of the secondary structures indicates that the helical content of CPS6ΔH111ΔV115 is 1.8% less than that of the wildtype. **b.** The CD spectrum of wildtype *m*EndoG (dashed black line) is compared to its construct with insertion of a His 9 residues and Val 5 residues *N*-terminal of the conserved Cys69 (*m*EndoG+H9+V5, solid grey line). BeSTSEL analysis of the secondary structures indicates that the helical content is nearly identical in both *m*EndoG constructs.

Wildtype CPS6 is a particularly poor nuclease compared to *m*EndoG, requiring hours to show significant cleavage of any DNA form, whether duplex or junction, ^{5hm}C modified or unmodified (Figure 3.4.a). The CPS6ΔH111ΔV115 double mutant was seen to cut unmodified and ^{5hm}C-modified DNA junctions at nearly the same levels as wildtype *m*EndoG, indicating that deleting these two amino acids increases the cutting efficiency and recapitulates the specificity for ^{5hm}C and DNA junctions of *m*EndoG in the invertebrate enzyme.

Constructs with single deletions at H111 and V115 (CPS6ΔH111 and CPS6ΔV115 constructs, respectively) all cleave the ^{5hm}C junctions. The specificity for the unmodified junction of these single deletion mutants, however, were variable, with the CPS6ΔV115 (having the deletion closer to the *A*-site) showing significantly reduced specificity and CPS6ΔH111 (deletion further from the *A*-site) showing no specificity for the unmodified substrate. These results suggest that the ability of the *A*-site to recognize ^{5hm}C affects recognition of the second arm of the DNA junction by the *B*-site. The potential coupling of the two DNA binding sites will be discussed in the next section.

The complement to the deletion mutation studies on CPS6 is to construct mutations of *m*EndoG with the comparable amino acids inserted at the equivalent sites (His at 9 residues and Val at 5 residues *N*-terminal of the conserved Cys). The prediction is that inserting one or both of these amino acids would provide the *m*EndoG sequence with sufficient length of peptide to fold the a1-helix and, in the process, displace Cys69 from the *A*-site and eliminate specificity for ^{5hm}C in the junction substrate. We found that the double insertion mutant (*m*EndoG+9H+5V) did not significantly increase the helical content relative to the wildtype enzyme (Figure 3.3.b), suggesting that the stability of the a1-helix is not determined explicitly by the length of available peptide sequence.

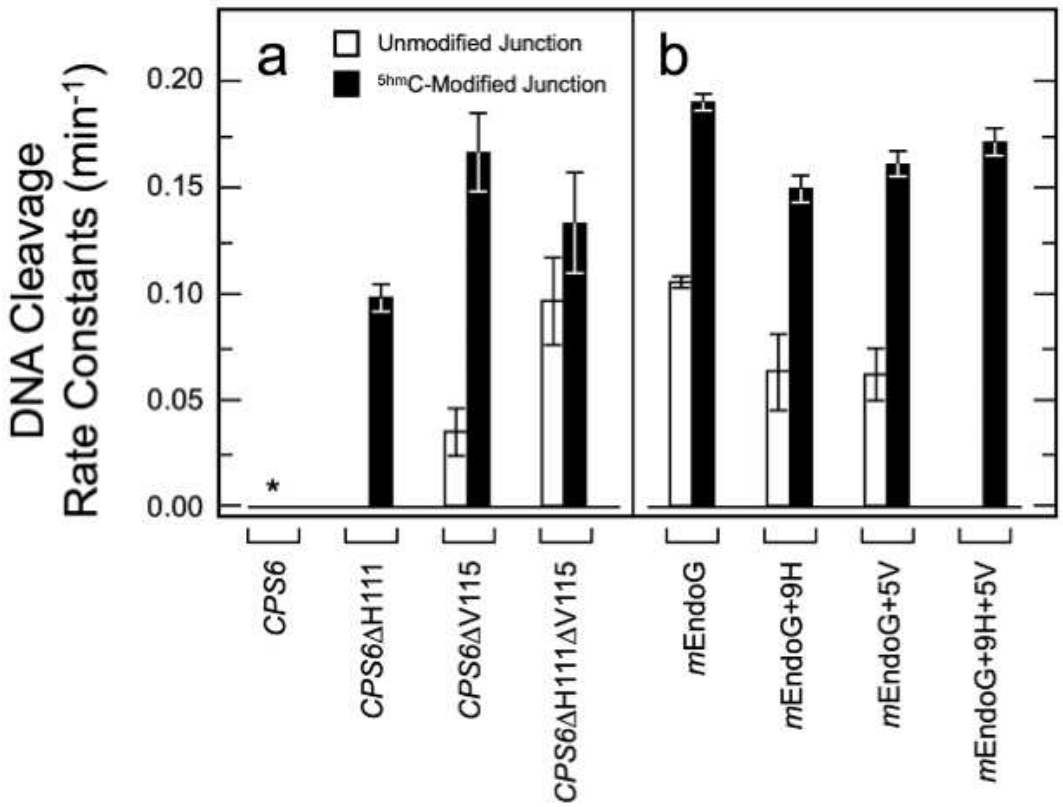

Figure 3.4. Rate constants for cleaving unmodified (open bars) and ^{5hm}C-modified (closed bars) DNA junctions. Error bars indicate the standard error of the mean values. **a.** Cleavage rates for CPS6 deletion constructs (CPS6 wildtype, CPS6ΔH111 with His111 deleted, CPS6ΔV115 with Val115 deleted, and CPS6ΔH111ΔV115 with both His111 and Val 115 deleted). The asterisk (*) for CPS6 indicates that the rate constants for the wildtype enzyme are on the order of hours^{-1} and, therefore, do not appear on this scale. **b.** Cleavage rates for *m*EndoG insertion constructs (*m*EndoG wildtype, *m*EndoG+9H with a His inserted 9 residues prior, *m*EndoG+5V with a Val inserted 5 residues prior, and *m*EndoG+9H+5V with His inserted 9 residues and Val inserted 5 residues prior to the conserved Cys69 residue).

All of the *m*EndoG constructs showed specificity for cleaving ^{5hm}C-modified DNA junctions. Both single insertion mutants showed reduced cutting of the unmodified junction relative to wildtype, while the double insertion mutant construct showed no cleavage of the unmodified DNA at all. When combined with results from the CPS6 deletion mutations, it appears that enzymes that have the a1-helix unwound due to deletions of His that is 9 residues and Val that is 5 residues from the conserved Cys (wildtype *m*EndoG and CPS6ΔH111ΔV115 double deletion) cut both junctions, with ~50% greater preference for ^{5hm}C-modified substrate. Constructs that deviate the most from native *m*EndoG (*m*EndoG+9H+5V and CPS6ΔH111) appear to cut ^{5hm}C-modified but not unmodified substrate, while those that are intermediate (*m*EndoG+9H, *m*EndoG+5V and CPS6ΔV115) are somewhere in between the two extremes in terms of preference for the modified *versus* unmodified junctions.

3.4.3. B-site and junction recognition

The crystal structure of inactive *m*EndoG showed duplex DNA bound to the *B*-site, leading to the model that this site provides an orthogonal surface to the *A*-site to accommodate junction binding by the vertebrate enzyme. To test this model, we swapped the 10 amino acid sequence from the *m*EndoG *B*-site into the comparable stretch of CPS6. As a result, the cutting efficiency of the chimeric CPS6<*B-EndoG* construct increased by greater than three-fold (from ~11% to ~33%) over the native invertebrate enzyme (Figure 3.5). There was, however, no significant difference in discrimination between the unmodified and ^{5hm}C-modified junction substrate.

In order to determine whether the entire *m*EndoG *B*-site or only the Pro residues in this site is required to introduce non-specific junction recognition, we constructed CPS6 mutants in which

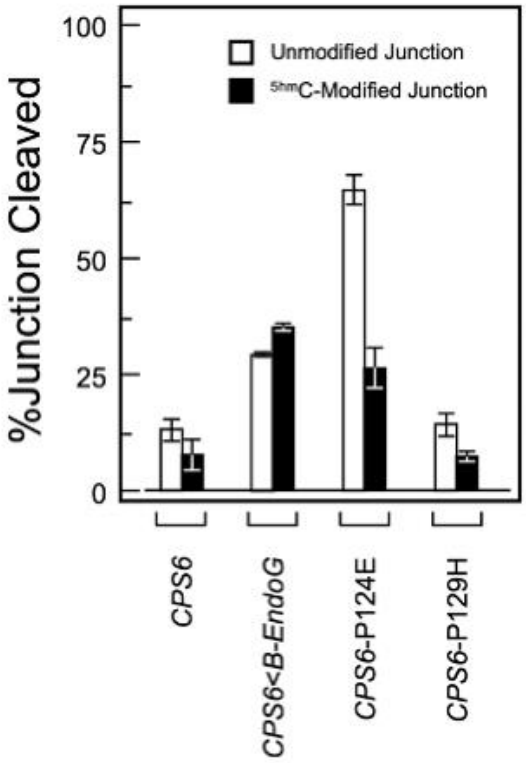

Figure 3.5. Cleavage of unmodified (open bars) and ^{5hm}C-modified (solid bars) DNA junction substrates by CPS6 constructs. The percent of substrate cleaved after 2 hours of incubation are compared for wildtype CPS6 to the chimeric enzyme with 10 amino acids from the *m*EndoG *B*-site swapped in, and to mutants in which the prolines at P124 or P129 have been mutated to those at the analogous mouse positions. Error bars are the standard errors of the mean.

Pro124 or Pro129 have been replaced by the analogous amino acids in the mouse sequence (CPS6-P124E and CPS6-P129H constructs, respectively). Interestingly, the construct with the mutation closest to, but still distinct from, the conserved *A*-site (CPS6-P124E) showed an even more dramatic effect on the efficiency in cutting the unmodified substrate than seen with CPS6<*B*-*EndoG*, with less of an effect on the ^{5hm}C-modified junction. The mutation further from the *A*-site was essentially the same as wildtype. Thus, the inability of the invertebrate enzyme to efficiently cleave DNA junctions can be attributed primarily to proline at the P124 position, which is proximal to the *A*-site.

3.4.4. Coupling between A- and B-sites

All the data from the CPS6 deletion and *m*EndoG mutants indicates that recognition the ^{5hm}C-modified base in the *A*-site is linked to junction binding at the *B*-site. Careful analysis of the crystal structures of *m*EndoG suggest that Asp70 could serve as the amino acid that couples the *A*- and *B*-sites of the mouse enzyme (Figure 3.6). In this model, the unwinding of the a1-helix in the invertebrate structures results in repositioning of the conserved Cys residue to recognize the ^{5hm}C base of the DNA in the *A*-site. In addition, the loss of this helix repositions a conserved arginine, allowing it to H-bond to Asp70 and the *A*-site to couple to the *B*-site through an additional H-bond to Arg77. The orientation of Arg117 in CPS6 does not allow H-bonding to the analogous Glu121, which further does not H-bond to Lys120.

To test this model for coupling the two DNA binding sites, we measured the rate constants for cutting unmodified and ^{5hm}C-modified junctions by constructs in which Asp70 in *m*EndoG or Glu121 in CPS6 have been mutated to Ala (Figure 3.7)—*A/B*-site decoupling mutants. The prediction is that these constructs will decouple the *A*- from the *B*-sites and, consequently, ^{5hm}C

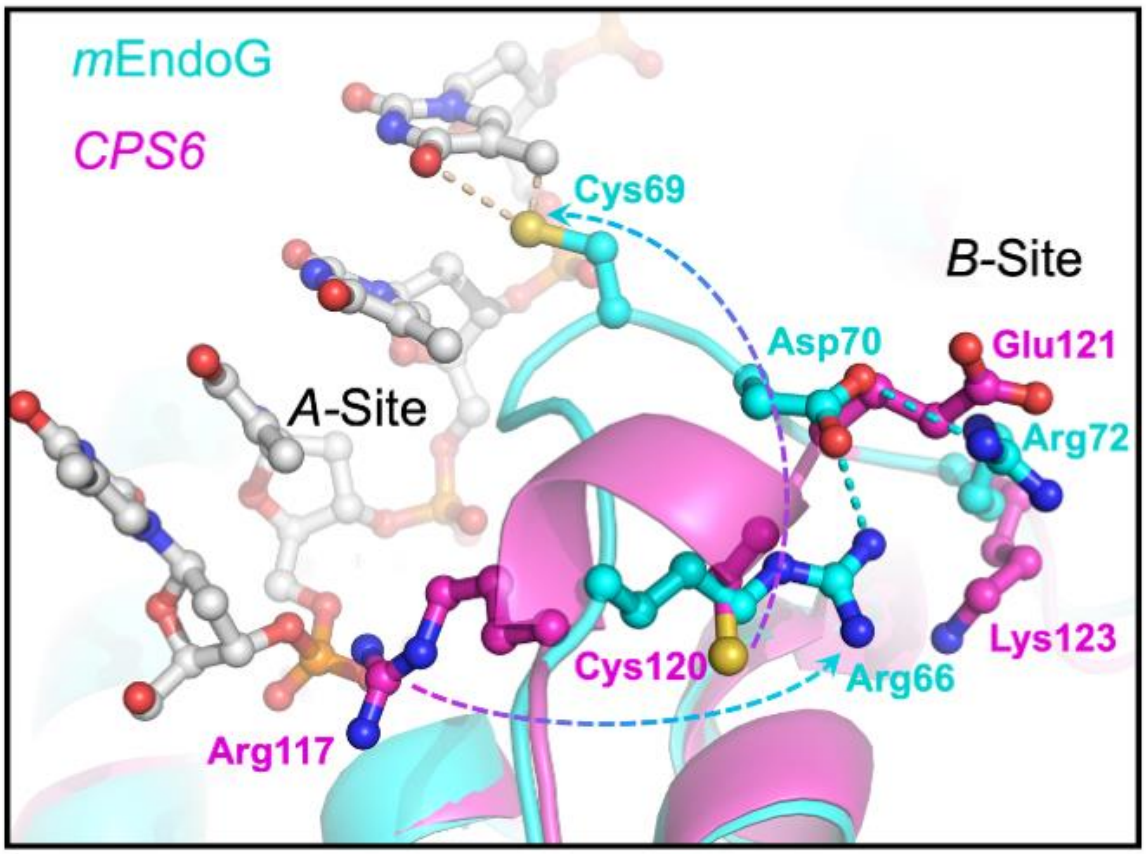

Figure 3.6. Potential coupling of DNA binding *A*-site and *B*-site in *m*EndoG (cyan) as a result of unwinding the a1-helix of CPS6 (magenta). The arrows indicate the repositioning of Arg 117 and Cys120 of CPS6 to those of Arg66 and Cys69 in *m*EndoG as a result of unwinding the a1-helix.

recognition from junction binding. The most dramatic reflection of the coupled sites can be seen in comparing the double insertion *m*EndoG+9H+5V constructs. Recall from Figure 3.4 that the *m*EndoG+9H+5V mutant was highly selective for the ^{5hm}C-modified over unmodified junction. The D70A variant (*m*EndoG+9H+5V-D70A) was entirely agnostic as to whether the substrate is ^{5hm}C-modified or not. In similar fashion, the CPS6ΔH111ΔV115 double mutant had shown significant (~50%) selectivity for the modified junction, while the E121A variant (the worm equivalent of the mouse D70A decoupling mutant) was indifferent to DNA modification.

There are some anomalies in these results. For one, the D70A mutant of *m*EndoG may have some residual selectivity for ^{5hm}C, but this difference is not definitive within the errors of the assay. More significantly, the E121A mutant of CPS6 shows a dramatic increase in selectivity for the modified over the unmodified substrate, while the wildtype enzyme shows no preference. It is not clear why this is the case, since this side chain sits entirely on a solvent exposed surface.

We further note that the overall rate constants for these *A/B*-site decoupling mutants of CPS6 are about 10-fold higher than the wildtype. It is not surprising that the cutting efficiency of the CPS6 mutants would increase, since replacing the strongly acidic E121 residue with a neutral Ala amino acid renders the surface less negative where the DNA junction would cross-over from the *A*- to the *B*-sites. It is not immediately obvious, however, why the decoupled *m*EndoG mutant behave similarly to wildtype CPS6 in terms of its lower DNA cutting efficiency and, within the errors of the measurement, loss of specificity for ^{5hm}C. The CD spectrum of the *m*EndoG-D70A construct showed an increase in helical content of ~1.7%. Thus, it appears disruption of the R66···D70 H-bond allows the loop of the *A*-site to adopt the a1-helix conformation seen in the invertebrate structures.

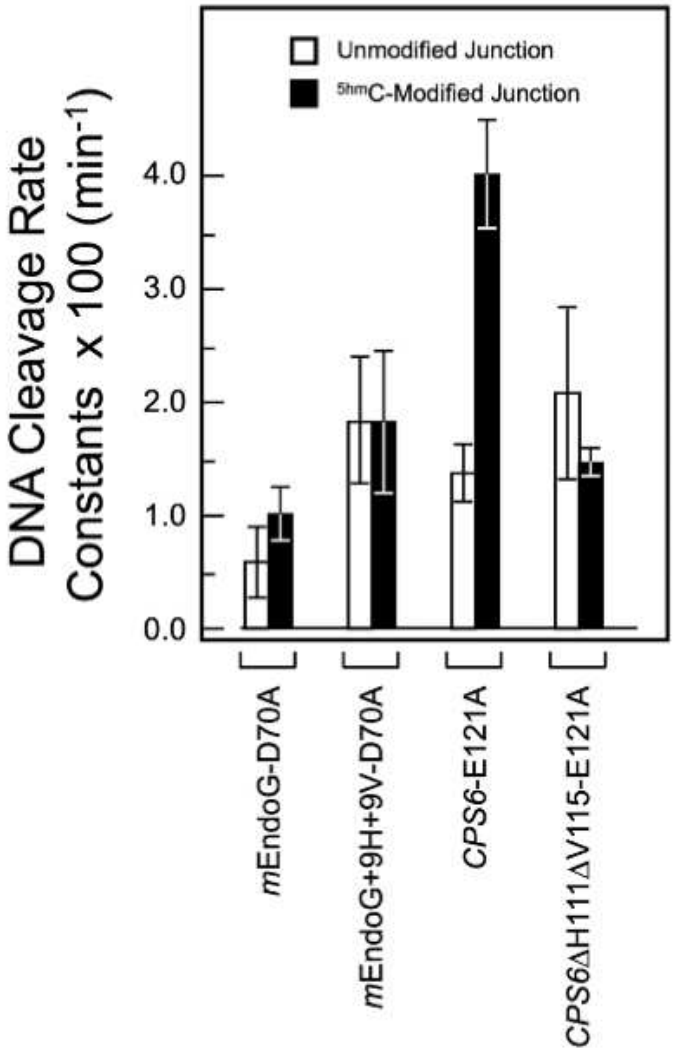

Figure 3.7. Rate constants for cleavage of unmodified (open bars) or 5hmC-modified (solid bars) DNA junctions by *A/B*-site decoupling mutants of *m*EndoG and CPS6. Error bars represent the standard errors of the means.

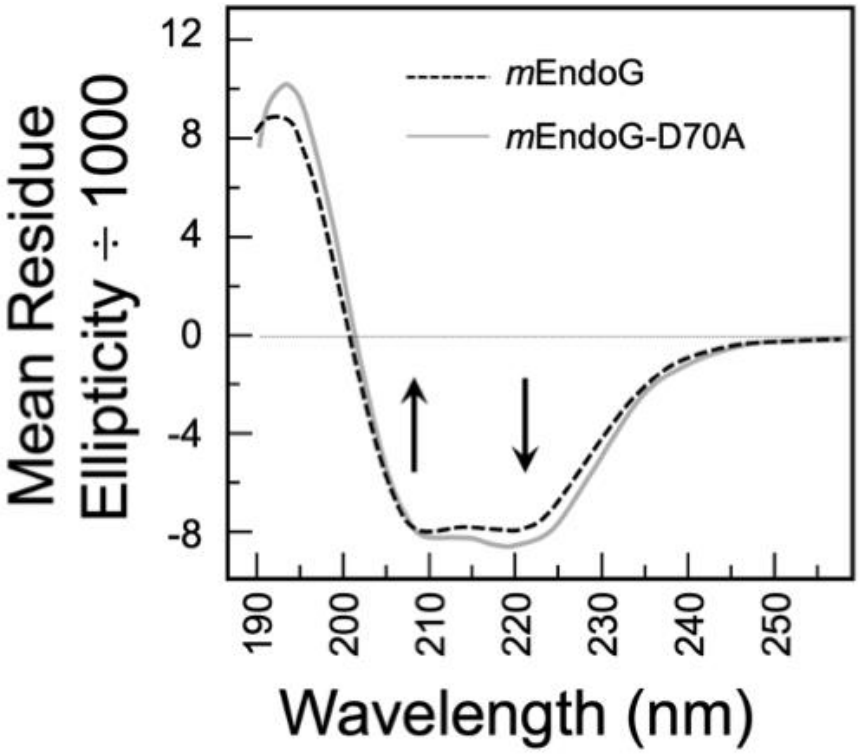

Figure 3.8. Comparison of CD spectra from wildtype *m*EndoG (dashed black curve) with the decoupled *m*EndoG-D70 mutant (solid grey curve). Arrows indicate the change in the relative ratio in the ellipticities at 208 nm *versus* 220 nm.

3.5. Discussion

In this study, we have shown that deletions of two amino acids in front of the active *A*-site of the worm CPS6 enzyme (the CPS6ΔH111ΔV115 double deletion mutant) significantly increases its catalytic activity as a nuclease and introduces specificity for the ^{5hm}C-modification in the context of DNA Holliday junctions, and is associated with a corresponding loss of the one turn of an a-helix. The immediate interpretation is that, as predicted by the original model, the reduced number of residues along this stretch of the vertebrate enzyme impairs the ability of the a1-helix to form in the *A*-site, which corresponds with the ability to recognize ^{5hm}C-modified DNAs. The corollary, however, is not true, in that simply inserting two amino acids into *m*EndoG (the *m*EndoG+9H5V construct) does not result in an increase in helical content and consequently the loss of ^{5hm}C recognition. The contradictory results between the double deletion and insertion mutants are rectified with the results from the *m*EndoG-D70A mutant, in which the activity of the mouse enzyme is reduced to that of CPS6 with a concomitant increase in helical content. In this case, however, forming the a1-helix was not a result of relaxing the spatial restriction for its formation, but in removing an H-bond that apparently constrains this region as a loop rather than a helix. Thus, results from these three constructs indicate that the elevated nuclease activity of and ^{5hm}C recognition by the vertebrate enzyme is directly associated with unwinding of the a1-helix. This potential helix is constrained to a loop conformation in the mouse enzyme by the limited number of amino acids along this edge of the *A*-site and, more significantly, by an H-bond anchors this loop to the carboxylate of an Asp residue.

Although we have shown that recognition of unmodified junctions is conferred by the *B*-site for DNA binding, we also see that the *A*- and *B*-sites are not entirely independent but are functionally coupled to each other through the Asp70 in *m*EndoG. We showed that swapping the

B-site sequences from *m*EndoG into CPS6 does increase the ability of the worm enzyme to bind junctions in a nonspecific manner. In addition, mutating the Pro residue in the *B*-site of CPS6 that is closest to the *A*-site also increased its efficiency for cleaving unmodified DNA. However, it is clear that more dramatic effects on both the modified and unmodified junctions are seen in the mutant constructs that affect the a1-helix or the coupling Asp/Glu residue that bridges the two sites. Consequently, we propose a model to help tie together the observed DNA cleaving efficiencies for the native CPS6 and *m*EndoG enzymes, along with the two constructs that show the most significant effects—CPS6ΔH111ΔV115, which essentially converts CPS6 to *m*EndoG, and *m*EndoG+5H+9V, which loses specificity for unmodified, but retains it for ^{5hm}C-modified substrates (Figure 3.9).

There are three primary concepts that define the model in Figure 3.9. First, the a1-helix inhibits recognition of ^{5hm}C by the *A*-site Cys and H-bonding of the *A*-site Arg to the coupling Asp/Glu. Second, the H-bond from the *A*-site Arg to the coupling Asp/Glu promotes H-bonds from the Asp/Glu to the nearby *B*-site Arg, allowing the *B*-site to rearrange and bind an unmodified arm of the junction. Finally, the efficiency in cleaving the modified *versus* unmodified junctions can be seen to be dependent on whether one or both DNA binding sites of the enzyme bind to their respective cognate features (^{5hm}C for the *A*-site and an orthogonal junction arm by the *B*-site). With these basic concepts in mind, we can start to rationalize the observed cutting efficiencies of the various constructs.

The wildtype CPS6 (Figure 3.9.a), with the a1-helix in the *A*-site, does not position its Cys120 residue to recognize ^{5hm}C nor Arg117 to H-bond with Glu121 (Figure 3.6). The lack of the H-bond to the coupling Glu121 inhibits the *B*-site from adopting a conformation to recognize the second arms of a junction and, consequently, CPS6 lacks specificity for both ^{5hm}C-modified and

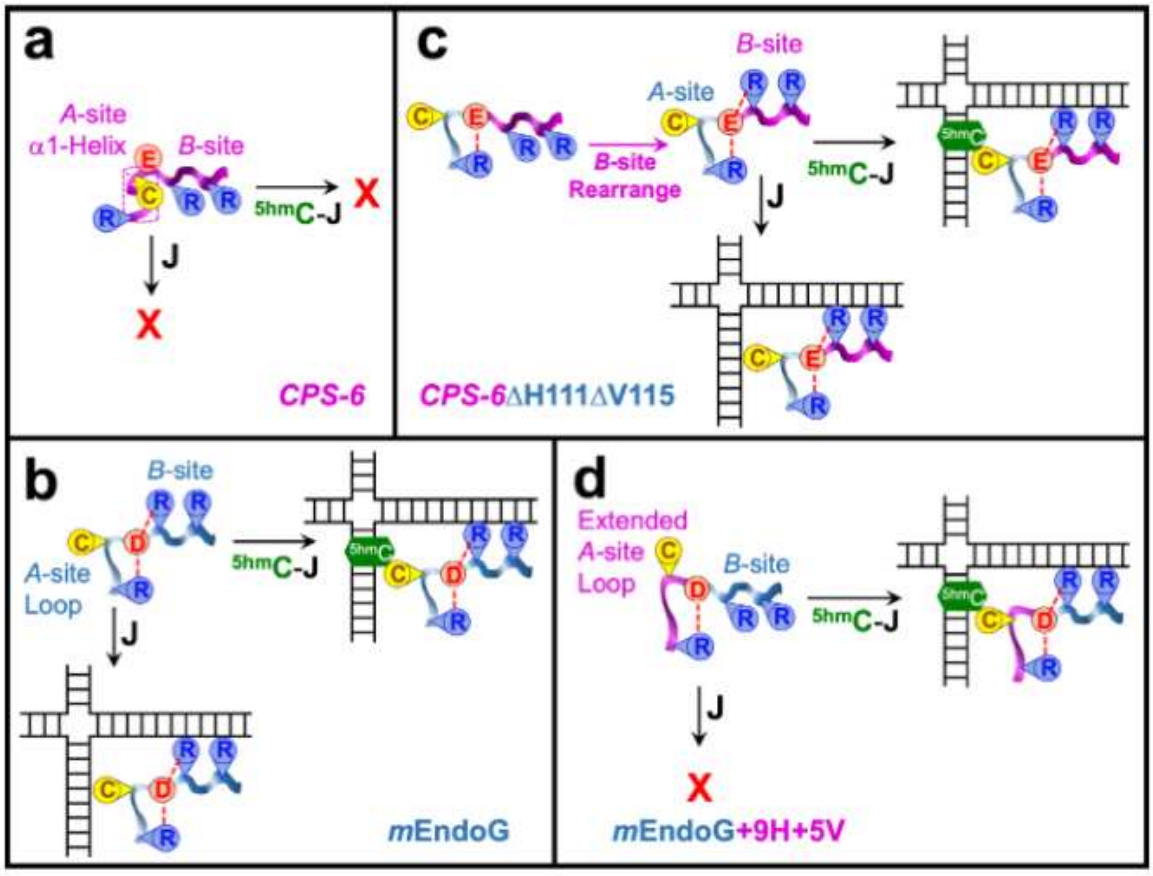

Figure 3.9. Model to couple the *A*- and *B*-sites for DNA binding in CPS6 and *m*EndoG constructs. Stretches of sequences associated with *m*EndoG are shown in blue ribbons while those associated with CPS6 are shown in red ribbons. The conserved Cys residues are shown as yellow circles and labeled "C", conserved Arg residues in blue circles and labeled "R", and the conserved acidic Asp (in *m*EndoG) or Glu (in CPS6) coupling residue in red circles and labeled "D" or "E". H-bonds from the R to D/E coupling residues are shown by red dashed lines. C residue recognizes ^{5hm}C when pointed to the left. The R in the *A*-site that H-bonds to the coupling D/E points to the right, while the Rs in the *B*-site that recognizes junction point up. The Unmodified junctions are designated "J" and shown as a four-armed schematic junction in its open X-form. The ^{5hm}C-modified junction is labeled "^{5hm}C-J" with the modified based indicated by the green hexagon. Non-binding/cleaving events are indicated by the red "X". **a.** Wildtype CPS6. **b**. Wildtype *m*EndoG. **c.** Double deletion mutant CPS6ΔH111ΔV115. **d.** Double insertion mutant *m*EndoG+9H+5V.

unmodified junction substrates. In wildtype *m*EndoG (Figure 3.9.b), the a1-helix is unwound to a loop, allowing Cys69 to recognize ^{5hm}C and Arg66 to H-bond to Asp70. The Arg66···Asp70 H-bond promotes the H-bond from Asp70 to Arg72, which arranges the *B*-site into the junction binding conformation. Thus, the *A*- and *B*-sites can each recognize their cognate DNA signals. The enzyme shows higher specificity for ^{5hm}C-modified junction over the unmodified substrate because the modified base provides an additional point of recognition.

The CPS6 double deletion mutant (CPS6ΔH111ΔV115, Figure 3.9.c) destabilizes the a1-helix, thereby positioning Cys120 in the *A*-site to recognize ^{5hm}C and Arg117 to H-bond to Glu121. This latter interaction further allows Glu121 to H-bond to a basic amino acid (in this case, Lys123) and, consequently, rearranging the *B*-site conformation to allow recognition of the second junction arm. The result is that the *A*- and *B*-sites are coupled, leading CPS6ΔH111ΔV115 to behave similarly to wildtype *m*EndoG. The CPS6 single mutations (CPS6ΔH111 and CPS6ΔV115) would fit into this general model if we project that these one amino acid substitutions have a destabilizing effect on the a1-helix, but does not entirely disrupt this helix in the uncomplexed enzyme. We can see, however, that once the enzyme binds a modified junction, the destabilized helix would unwind to allow Cys120 to recognize the ^{5hm}C base. The CPS6ΔV115 construct, which places the mutation closer to Arg117, may have some effect on the ability of this basic amino acid to H-bond to Glu121, which would account for its low but detectable level of specificity for the unmodified junction.

The two amino acids inserted into *m*EndoG double insertion mutant (*m*EndoG+9H+5V, Figure 3.9.d) does not allow folding of the a1-helix, likely because Arg66 remains H-bonded to Asp70, as discussed above. The extended length of the loop, however, could affect the conformation of the residues around the Cys69 and the Asp70, inhibiting the ability of Asp70 to form an H-bond to an Arg at the *B*-site. Binding and subsequent recognition of ^{5hm}C in a modified

junction would reestablish the H-bonded coupling between the sites, in much the same manner as proposed for the CPS6ΔH111 and CPS6ΔV115 constructs. Thus, the model proposed here to couple the two DNA binding sites of *m*EndoG provides a set of concepts that helps to rationalize the observed specificities for not only this vertebrate enzyme, but also for its various mutants and for the various forms of CPS6 constructed for this study.

We see from this study that the evolution of the invertebrate CPS6 enzyme to one that behaves similarly to vertebrate EndoG, with its enhanced efficiency and specificity for cleaving ^{5hm}C DNAs in the context of a DNA junction, requires only a simple set of two amino acid deletions that result in the unwinding of the single turn of an a-helix. There are other modifications seen in the *m*EndoG sequence that further enhances the discrimination for the ^{5hm}C-base, including loss of a Pro residue and coupling between the two DNA binding sites, which may help the vertebrate enzyme recognize and utilize this modified base more efficiently. Although at the amino acid sequence level the evolution of a protein from the non-discriminatory invertebrate form to one that provides functionality to this vertebrate specific epigenetic marker may in this case appear simple, it should be noted that these types of deletions at the gene level are not as simple, particularly if the rest of the protein must remain largely conserved. Most single amino acid mutations are deleterious to protein function and a double deletion should be that much more difficult from a genetic perspective [27]. For this class of endonucleases, however, we can see that a single substitution can in fact result in a mutant that is highly specific for ^{5hm}C (more so than even the wildtype mouse enzyme). The question then is why was it necessary to evolve a second deletion that, apparently, reduces specificity for the modified base? In order to address this question, it would be important to study the contributions of the two recognition features (^{5hm}C and Holliday junctions) separately. To do this, we would need to decouple the two functions, which we have

now done. We have created constructs that are specific for ^{5hm}C and constructs that are specific for unmodified junctions. These constructs are thus molecular tools that can potentially be used to determine how these two DNA features contribute to specific phenotypic functions in the respective vertebrate and invertebrate organisms.

References

1. Shi, D. Q., Ali, I., Tang, J. & Yang, W. C. New insights into 5hmC DNA modification: Generation, distribution and function. *Frontiers in Genetics* vol. 8 (2017).
2. Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S. & Jacobsen, S. E. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol.* **12**, R54 (2011).
3. Song, C. X. & He, C. The hunt for 5-hydroxymethylcytosine: The sixth base. *Epigenomics* vol. 3 521–523 (2011).
4. Tahiliani, M. *et al.* Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science (80-.).* **324**, 930–935 (2009).
5. Zhang, H. Y., Xiong, J., Qi, B. L., Feng, Y. Q. & Yuan, B. F. The existence of 5-hydroxymethylcytosine and 5-formylcytosine in both DNA and RNA in mammals. *Chem. Commun.* (2016) doi:10.1039/c5cc07354e.
6. Lin, J. L. J., Wu, C. C., Yang, W. Z. & Yuan, H. S. Crystal structure of endonuclease G in complex with DNA reveals how it nonspecifically degrades DNA as a homodimer. *Nucleic Acids Res.* **44**, 10480–10490 (2016).
7. Zanden, C. M. V., Czarny, R. S., Ho, E. N., Robertson, A. B. & Ho, P. S. Structural adaptation of vertebrate endonuclease G for 5-hydroxymethylcytosine recognition and function. *Nucleic Acids Res.* **48**, 3962–3974 (2021).
8. Ehrlich, M. & Wang, R. Y. 5-Methylcytosine in eukaryotic DNA. *Science* **212**, 1350–7 (1981).
9. Pardo, R. *et al.* EndoG Knockout Mice Show Increased Brown Adipocyte Recruitment in White Adipose Tissue and Improved Glucose Homeostasis. *Endocrinology* **157**, 3873–3887

(2016).

10. Wiehe, R. S. *et al.* Endonuclease G promotes mitochondrial genome cleavage and replication. *Oncotarget* **9**, 18309–18326 (2018).

11. Bruni, F., Lightowlers, R. N. & Chrzanowska-Lightowlers, Z. M. Human mitochondrial nucleases. *FEBS J.* **284**, 1767–1777 (2017).

12. Zhdanov, D. D. *et al.* Apoptotic endonuclease EndoG induces alternative splicing of telomerase catalytic subunit hTERT and death of tumor cells. *Biochem. Suppl. Ser. B Biomed. Chem.* (2016) doi:10.1134/S1990750816040090.

13. Parrish, J. Z., Yang, C., Shen, B. & Xue, D. CRN-1, a Caenorhabditis elegans FEN-1 homologue, cooperates with CPS-6/EndoG to promote apoptotic DNA degradation. *EMBO J.* **22**, 3451–60 (2003).

14. Reik, W. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature* **447**, 425–432 (2007).

15. Thyagarajan, B., Padua, R. A. & Campbell, C. Mammalian mitochondria possess homologous DNA recombination activity. *J. Biol. Chem.* **271**, 27536–27543 (1996).

16. Robertson, A. B., Robertson, J., Fusser, M. & Klungland, A. Endonuclease G preferentially cleaves 5-hydroxymethylcytosine-modified DNA creating a substrate for recombination. *Nucleic Acids Res.* **42**, 13280–13293 (2014).

17. Parrish, J. *et al.* Mitochondrial endonuclease G is important for apoptosis in C. elegans. *Nature* **412**, 90–94 (2001).

18. Loll, B., Gebhardt, M., Wahle, E. & Meinhart, A. Crystal structure of the EndoG/EndoGI complex: mechanism of EndoG inhibition. *Nucleic Acids Res.* **37**, 7312–7320 (2009).

19. Lin, J. L. J. *et al.* Structural insights into apoptotic DNA degradation by CED-3 protease

suppressor-6 (CPS-6) from Caenorhabditis elegans. *J. Biol. Chem.* **287**, 7110–7120 (2012).

20. Lin, J. L. J. *et al.* Oxidative Stress Impairs Cell Death by Repressing the Nuclease Activity of Mitochondrial Endonuclease G. *Cell Rep.* **16**, 279–287 (2016).

21. Li, L. Y., Luo, X. & Wang, X. Endonuclease G is an apoptotic DNase when released from mitochondria. *Nature* **412**, 95–99 (2001).

22. Irvine, R. A. *et al.* Generation and Characterization of Endonuclease G Null Mice. *Mol. Cell. Biol.* **25**, 294–302 (2005).

23. Zhang, J. *et al.* Endonuclease G is required for early embryogenesis and normal apoptosis in mice. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 15782–15787 (2003).

24. Ishihara, Y. & Shimamoto, N. Involvement of endonuclease G in nucleosomal DNA fragmentation under sustained endogenous oxidative stress. *J. Biol. Chem.* **281**, 6726–6733 (2006).

25. Moretton, A. *et al.* Selective mitochondrial DNA degradation following double-strand breaks. *PLoS One* **12**, (2017).

26. Ohsato, T. *et al.* Mammalian mitochondrial endonuclease G. *Eur. J. Biochem.* **269**, 5765–5770 (2002).

27. Simm, A. M., Baldwin, A. J., Busse, K. & Jones, D. D. Investigating protein structural plasticity by surveying the consequence of an amino acid deletion from TEM-1 β-lactamase. *FEBS Lett.* **581**, 3904–3908 (2007).

Chapter 4 – Computational Analysis of Raw DNA sequence for Z-DNA forming potential using Zhunt and Zmhunt*

4.1. Background and Introduction

4.1.1. Background on DNA Structures

The structure of DNA takes on different conformations throughout the genome and serves different roles in each state[1,2]. DNA available for transcription exists in B-form DNA, which is a right-handed helix[3]. Forming a tighter helix right-handed helix, A-form DNA is more compact and can be found in double stranded DNA or RNA as well as DNA-RNA complexes. Z-form DNA, however, takes on a left-handed helix structure that increases the number of base pairs per turn from 10 to 12 (with respect to B-DNA)[4,5]. This change in bases per turn results from a decrease in the helical rise from 2.3 A (B-DNA) to 1.9 A (Z-DNA) [4,5]. Negative supercoiling has been shown to be a major factor in the formation of Z-DNA. Z-DNA is thought to be involved in transcription regulation and genomic processes such as recombination, deletions, and translocation[6–8].

4.1.2. Antibody method of Z-DNA detection

The most common method for detecting Z-DNA has been through the use of antibodies, which have come under criticism. Concerns regarding this method have included: the antibodies could induce Z-DNA formation upon binding and thereby lead to false positive results; weak binding affinity between Z-DNA and the antibodies has led to a large underrepresentation of Z-DNA in a cellular context; and Z-DNA is thought to be a transient structure and therefore antibody binding is misleading and not representative[9,10].

4.1.3. Thermodynamics of Z-DNA Formation

Nucleotides (A and G – purines; T, C, and U – pyrimidines) can take on either an anti- or syn-conformation within a sequence. In the anti- conformation, the nucleotide base is positioned

*Adapted from published chapter "ZHUNT: Thermogenomic Analysis of Left-Handed Z-DNA Propensities in Genomes and Methylated Genomes" by Czarny, R.S.; Ho, P.S. (2021)

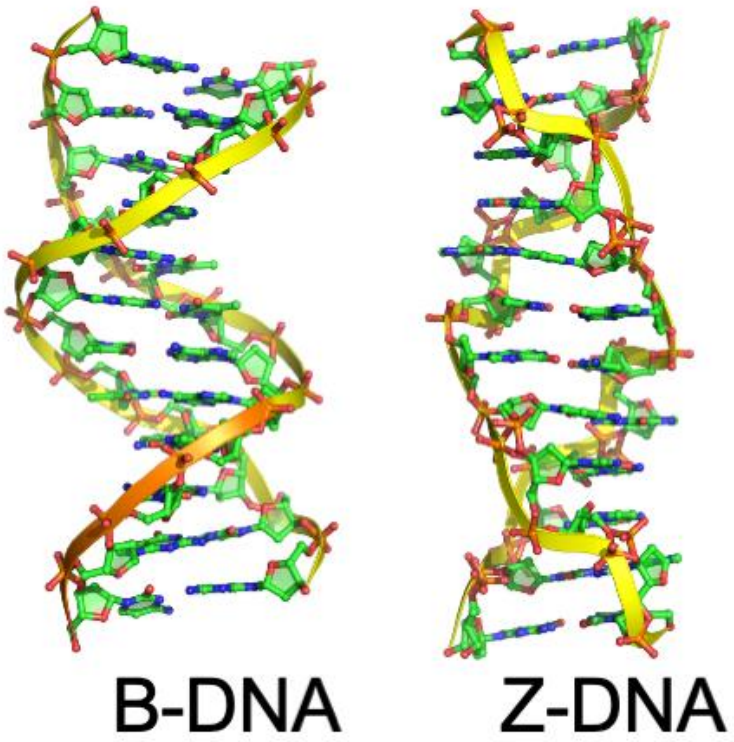

Figure 4.1. Representative structures of the right-handed helix of B-DNA and left-handed helix of Z-DNA. To the left is the canonical right-handed DNA that traditionally used for transcribing DNA. To the right is the left-handed Z-DNA structure, which often occurs due to supercoiling of the sequence. Both depict the major groove of the structures and it is easily seen that Z-DNA increases the distance between nucleotides and results in additional based needed for each turn of the helix.

directly above the backbone sugar while the syn- conformation represents the base positioned away from the sugar. Each base can take on either conformation; however, purines favor the anti- conformation while pyrimidines favor the syn- conformation[11]. Traditional computational methods for predicting Z-DNA in a sequence leveraged expected conformation positioning and calculated the number alternations between adjoining purines and pyrimidines, with the length of these alternations being directly proportional to the likelihood of Z-DNA formation. Since this is strictly based on the sequence and no other data, this method simply works as a rough estimate. To increase the accuracy of Z-DNA conformation prediction, rules were developed based off of synthetic nucleotides:

1) purine-pyrimidine alternation greatly increases Z-DNA formation[4,12–15]
2) purine-pyrimidine alternations are rated as dGC > dCA > dAT[4]
3) longer stretches of sequences increase propensity of formation[16]

Accounting for these ideas, a computational method was developed in 1986 which incorporated the free energy associated with the transition of dinucleotides from B-DNA to Z-DNA (Z-hunt)[17]. These transitional free energies have been determined through in vitro studies looking at negative supercoiling and have continued to be updated in the script as new work has been published. Within the script, the conformation is determined for each dinucleotide but finding the maximum energy difference in order to provide an overall energy minimization for Z-DNA formation. If the anti- syn- alternation is interrupted, a penalty is applied to the sequence; however, it does not negate the overall propensity of formation.

To get into the details for these calculations, Zhunt uses a number of different factors associated with DNA conformation. Negative Supercoils (NSCs) have been shown to induce the formation of Z-DNA with 2 turns of NSC resulting in 1 turn of Z-DNA, since the Z-DNA has a negative

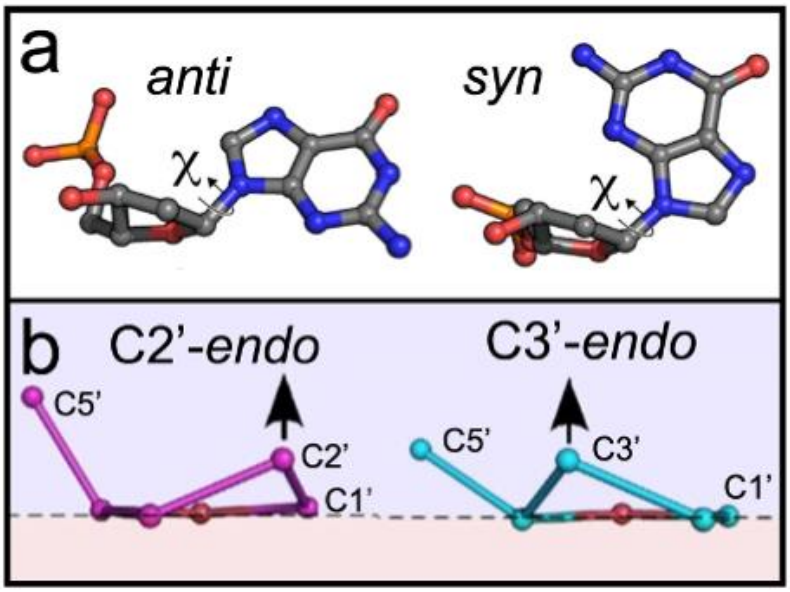

Figure 4.2. Anti- and syn- conformation of nucleotides. Comparing the anti and syn configuration guanine. Both purines and pyrimidines can exist in either the anti or syn configuration. In the syn configuration, the base is positioned over the sugar of the nucleic acid while the anti-configuration positions the base out away from the sugar. A) The guanine base is shown in its anti and syn conformations in which the base is rotated about the chi axis and B) the positioning of the C2 and C3 carbons within the sugar-phosphate backbone move with C2 positioned out of the plane for the anti conformation and the C3 positioned out of the plane for syn.

twist to it. To incorporate this into the calculation, the overall change in number of turns between structures for sequence that is completely B-DNA to Z-DNA is denoted by ΔLk. This change accounts for both the change in twist as well as the writhe in the NSC DNA. The Gibbs free energy is proportional to the square of the writhe, or the squared difference of the change in number of turns and the number of twists $(\Delta Lk - \Delta Tw)^2$. A fraction of Z-DNA formation is determined by using the standard Gibbs relationship and comparing the energies for Z-DNA vs B-DNA. Utilizing this, the fraction value can be used a probability of formation of Z-DNA for a given sequence. Once those calculations are complete for a given sequence, the program provides a Z-score, which represents the number of randomly generated sequences needed in order to obtain the same level of energy minimization as seen in the query.

4.1.4. Updates to Z-hunt

To expand on the functionality of Z-hunt, we incorporated energies associated with methylated cytosines, which had been thought to have roles in Z-DNA formation. The table index for the dinucleotides was updated to include methyl-cytosine and denoted as "M". Once these energies were added, the dictionary that classifies the input sequence and designates dinucleotides was expanded to allow for the modified nucleotide. Extensive testing was done on this new version of Z-hunt, which we have called Z-mhunt, to ensure that results directly matched for sequences run through Z-hunt and Z-mhunt. We then reached out to Wang et al regarding their recent publication identifying methylation sites using Bis-seq for the yeast genome[18]. Using yeast chromosome 1, we substituted each of the modified cytosines with an "M" in the sequence and ran it through Z-mhunt. Interestingly, the methylation sites change the overall profile of the Z-scores and thereby influences the formation potential of Z-DNA. We decided to look more deeply into the methylation

	AA	AT	AG	AC	AM	TA	TT	TG	TC	TM	GA	GT	GG	GC	GM	CA	CT	CG	CC	CM	MA	MT	MG	MC	MM
AS-AS	4.40	6.20	3.40	5.20	3.00	2.50	4.40	1.40	3.30	1.90	3.30	5.20	2.40	4.20	1.49	1.40	3.40	0.66	2.40	0.93	0.80	1.90	-0.71	0.93	0.90
SA-SA	4.40	2.50	3.30	1.40	0.80	6.20	4.40	5.20	3.40	1.90	3.40	1.40	2.40	0.66	-0.71	5.20	3.30	4.20	2.40	2.73	3.00	1.90	1.49	2.73	0.90
AS-SA	6.20	6.20	5.20	5.20	4.60	6.20	6.20	5.20	5.20	4.60	5.20	5.20	4.00	4.00	2.53	5.20	5.20	4.00	4.00	2.53	4.60	4.60	2.53	2.53	1.93
SA-AS	6.20	6.20	5.20	5.20	4.60	6.20	6.20	5.20	5.20	4.60	5.20	5.20	4.00	4.00	2.53	5.20	5.20	4.00	4.00	2.53	4.60	4.60	2.53	2.53	1.93

Figure 4.3. Free energy of transition of dinucleotides between conformations. The energy of transition between the anti and syn conformations for dinucleotide pairs has been calculated using molecular dynamics energy minimizations and updated with wet lab-based results. The energies were updated most recently to include methylation into the dinucleotide options.

sites and identified that 84.2% of the modified cytosines reside 225-275 base pairs upstream of Transcription start sites, which aligns with the yeast promotor site known to be -250 of the TSS. We then mapped the number of occurrences of methylation with respect to each TSS and were able to determine that 1 methylation is the most common, but sites of up to 7 methylations were seen.

Up until this point, the coding had been written and compiled in C and therefore required the user to have at least a basic understanding of Bash scripting in order to test their sequence. The first step was to develop a Python script that would compile Z-hunt locally for the user and then run with variable parameters set by the user within the Python script. Next, a Graphical User Interface (GUI) was developed so that the user would simply have to launch the script in the command line and then could either paste or load in a sequence file in order for it to run. Finally, the most recent version of the user experience involved using the Flask plugin for Python, which allows Python to work with and generate html files in order to create websites. Currently, the site is being served from a Mac Mini at zhunt.bmb.colostate.edu.

Since it was first developed, Z-hunt has come a long way in terms of processing power and computational time. The original code, written in Fortran-77, running on a VAX 11-780 computer took about 1 month to process the adenovirus genome at a length of about 37 kb. The current version of the program, running in C and using a Python wrapper, running on a Mac Mini can process yeast genome 1, at about 230 kb, in roughly 33 seconds. This has been tested for both Z-hunt and Z-mhunt with similar results.

Additional functionality was added to the results page, with code being available for viewing and use on the Code tab of the site. To use these scripts, a bit of preparatory work is needed: 1) a directory needs to be created for the Z-hunt data; 2) the raw data from Z-hunt needs

to be placed within the directory; 3) a python script (available on the site) needs to be run in order to generate output files that will be used for the graphics scripts; and 4) the R scripts used for graphing need to be downloaded and placed in the directory. Once the directory is setup, the R scripts provided can be run to generate a wide variety of graphs that can be used in data analysis. One final script allows the user to compare Z-hunt and Z-mhunt data as well. The sequence without the methylation sites needs to have been run in Z-hunt and the sequence containing methylation sites through Z-mhunt. After completion, the z-score and z-mscore can be placed in the same spreadsheet to be able to find the difference value between them (if methylation occupancy data is available, a rough estimate of a true value can be calculated by multiplying the difference by the percent occupancy). Running the final R script will then present these difference values. Of note, this script is not currently built into the website but has been discussed for future development.

4.2. Methods

4.2.1. Method for running Z-hunt

In order to run Z-hunt or Z-mhunt, the user navigates to zhunt.bmb.colostate.edu to find themselves on the home page. If the user wants to run Z-hunt, they upload their sequence file in either a ".fasta" or ".txt" format (any other format will be rejected by the website). Next, they enter their email so that we can keep track of usage and contact information should a run encounter issues. Finally, they select the Submit button and the script will run in the background. Once Z-hunt is finished (usually about 1-2 seconds for 8-10k base pairs, see Figure ___), the site will redirect to a run completion page. Here, the user can select to download the raw output from the run, download pre-defined graphs summarizing their data, or look at the data live within their browser.

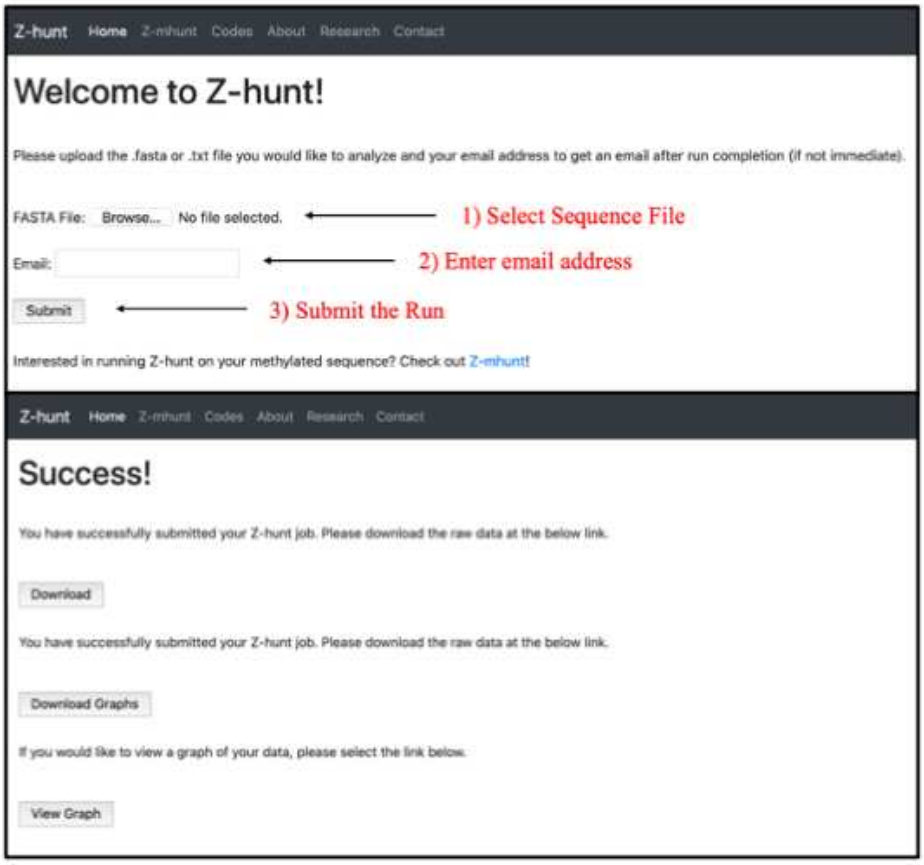

Figure 4.4. Z-hunt web interface for loading and finishing a run. Zhunt online Results page. Upon completion of the calculation, the site will redirect to the Results page. The user can then – 1) Download the raw Z-score information for their own analysis; 2) Download readymade graphs for the queried sequence; or 3) View the Z-score across the queried sequence within the web browser. Using the online generated graph, the user can directly print or download from that page. After the run, the results can be downloaded in their raw data format, downloaded in ready-made graphs, or viewed within the web browser.

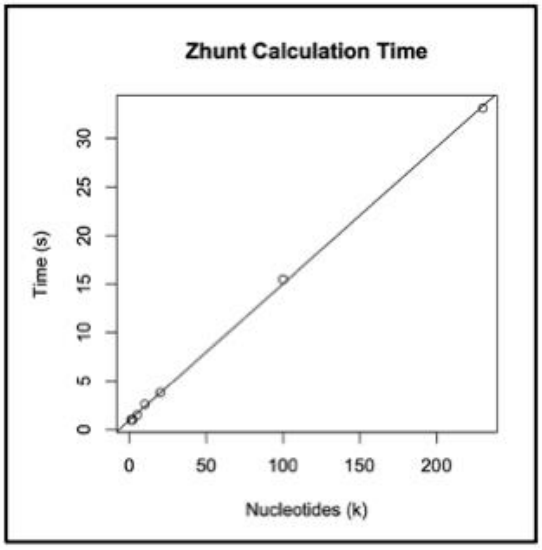

Figure 4.5. Computational time required to run Zhunt based on nucleotide length. Runtime performance for Zhunt online. Sequence of yeast chromosome 1 at varying sizes were tested on the Zhunt online server in order to determine calculation time. There is a linear relationship starting at less than 1 second for ≤ 2 kb and scaling to the full chromosome at 33 seconds for ~230 kb .

Within the raw data, a couple of different pieces of information are provided. In the first line, the input sequence file that was analyzed will be reported as well as the dinucleotide minimum and maximum lengths (the default value for the online portal is set to 6 and 12 respectively). There are then 4 columns which represent ΔLk, the slope at the first nucleotide positions, the Z-score, and the assigned anti or syn confirmations for the nucleotides within the queried sequence. The information from column 3 is what the site uses in order to generate the graph seen within the browser, with the x-axis being the sequence positions and the y-axis being the respective Z-score.

The method for running Z-mhunt is almost identical, with only a few modifications. When visiting the site, across the top toolbar is a link to Z-mhunt which will allow the user to run that version of the program. The setup for this page is exactly the same as Z-hunt. Before the user uploads their file, they must make changes to their sequence file in order to represent their methyl-cytosines with an "M" character (either upper or lower case will work). The sequence file can then be uploaded, and the run submitted and analyzed in the same manner as Z-hunt.

Should the user wish to do the analysis on their own, the codes for user to generate the graphs are available on the site to be copied and used locally. These graphing scripts were written in R and do not require the user to manipulate the data at all.

4.3. Future work and expected outcomes

4.3.1. Future Directions

The future focus of this program will be to collect Z-hunt runs from a large population of users in order to develop a "Big Data" repository. Utilizing the Z-scores from a wide array of species and portions of the genome, we will perform meta-analysis to localize regions in which Z-DNA are enriched throughout life. We will then map this back to species distinctions to correlate the protein and epi-genomic co-evolution to the presence of specific Z-DNA enriched sites.

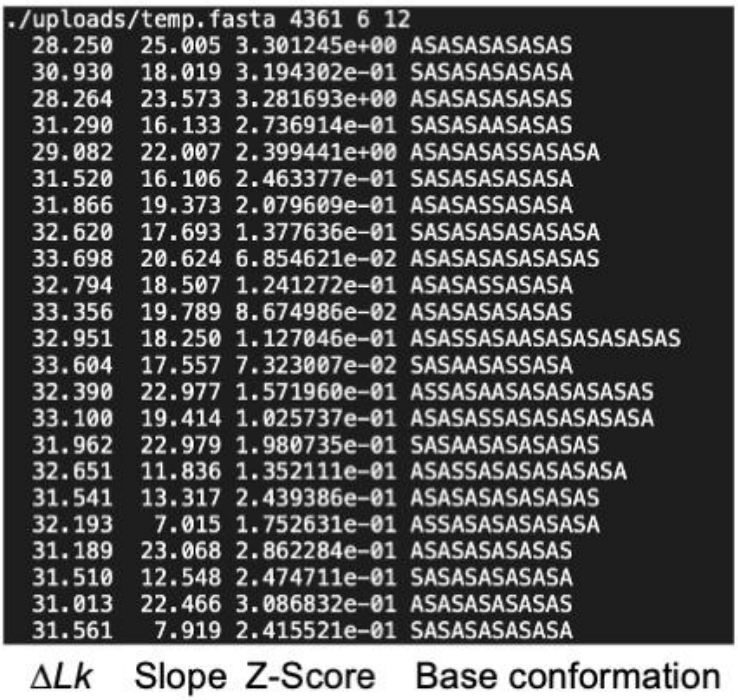

Figure 4.6. Representative output file from Z-hunt. Z-score Raw data file. Downloading the raw data file from the results page will provide a column separated file that can be opened and viewed in a text editor. The first line of the file contains the run information, with the name of the queried sequence, the number of nucleotides input, and the dinucleotides within the window (Zhunt online defaults these to 6 and 12 respectively). The first column represents the ΔLk value, the second represents the slope at the transition point, the third represents the Z-score for the window, and the fourth represents the assigned anti (A) or syn (S) conformation assigned to each nucleotide within the window.

Additional work will be done to expand the base energies included in the calculation table to encompass other epigenetic markers to more precisely predict the propensity of forming Z-DNA. Energies can initially be calculated using molecular dynamic simulations in their A-S

flipping and continually updated as additional in vitro work is published. We will also be developing a method to incorporate occupancy information determined in methylation assays. As of now, the program only runs in a way that assumes 100% occupancy and therefore shows an over representation of Z-DNA formation.

4.4. Developing skills

Skills for this project were collected as follows: Python was self-taught through the use of Code Academy and the book Python for Biologists by Martin Jones (lent by Sean Cascarina); R and C were self-taught through Code Academy and extensive online reading; and Flask and Docker (a container software used in developing the site) were learned from Harvard edX course covering Python Scripting and web hosting. Hosting and serving the site were done in collaboration with Ross Madden.

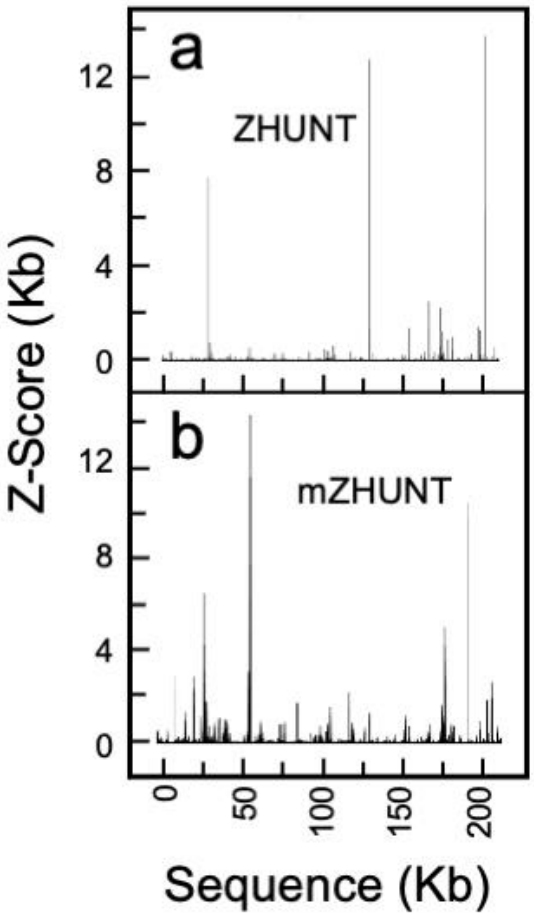

Figure 4.7. Representative structures of the right-handed helix of B-DNA and left-handed helix of Z-DNA. Comparing Zhunt and mZhunt calculations. Yeast chromosome 1 was used as a reference for testing the incorporation of methylation sites within the Zhunt calucation. A) The sequence was run through the traditional Zhunt or order to observe the Z-scores assigned before adding the methylation sites. Cytosines were then replaced with "M" to represent methyl-cytosine at locations determined by Fai Au et al. B) The updated, methylated sequence was then run through mZhunt in order to determine changes in Zscore taking into account the new energies.

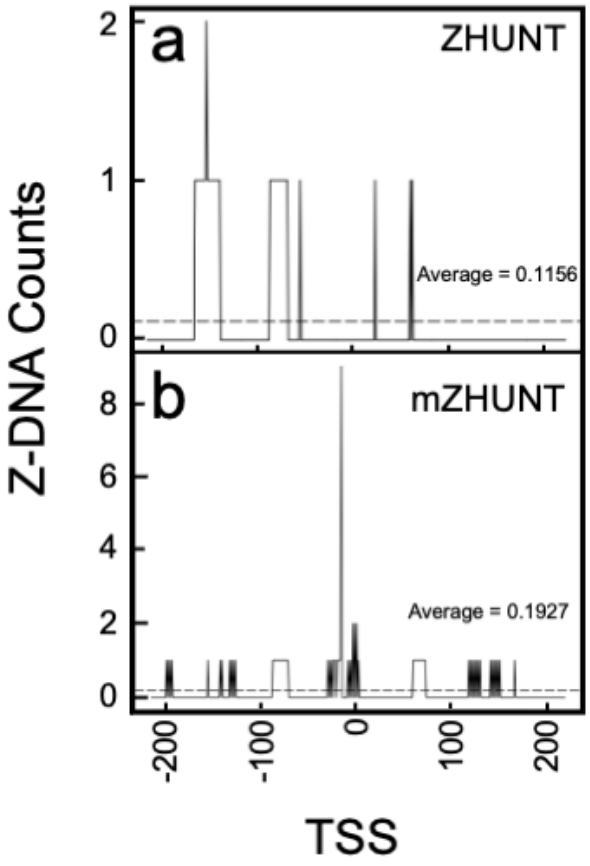

Figure 4.8. Comparing Z-hunt to mZhunt with respect to Transcription Start Sites. Comparison of Z-DNA regions relative to the transcription start site (TSS) from ZHUNT (a) and mZHUNT (b) analyses of chromosome 1 from S. cerevisiae.

References

1. Arnott, S. & Hukins, D. W. L. Optimised parameters for A-DNA and B-DNA. *Biochem. Biophys. Res. Commun.* **47**, 1504–1509 (1972).
2. Watson, J. D. & Crick, F. H. C. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738 (1953).
3. Dickerson, R. E. DNA structure from A to Z. *Methods Enzymol.* **211**, 67–111 (1992).
4. Rich, A., Nordheim, A. & Wang, A. H. J. The chemistry and biology of left-handed Z-DNA. *Annu. Rev. Biochem.* **VOL. 53**, 791–846 (1983).
5. Wang, A. H. J. *et al.* Molecular structure of a left-Handed double helical DNA fragment at atomic resolution. *Nature* **282**, 680–686 (1979).
6. Wang, J. C., Peck, L. J. & Becherer, K. DNA supercoiling and its effects on DNA structure and function. *Cold Spring Harb. Symp. Quant. Biol.* **47**, 85–91 (1982).
7. Nordheim, A. *et al.* Supercoiling and left-handed Z-DNA. *Cold Spring Harb. Symp. Quant. Biol.* **47**, 93–100 (1982).
8. Rich, A. & Zhang, S. Z-DNA: The long road to biological function. *Nature Reviews Genetics* vol. 4 566–572 (2003).
9. Lafer, E. M., Moller, A., Nordheim, A., Stollar, B. D. & Rich, A. Antibodies specific for left-handed Z-DNA. *Proc. Natl. Acad. Sci. U. S. A.* **78**, 3546–3550 (1981).
10. Pulleyblank, D. E., Haniford, D. B. & Morgan, A. R. A structural basis for S1 nuclease sensitivity of double-stranded DNA. *Cell* **42**, 271–280 (1985).
11. Thermogenomics: Thermodynamic-based approaches to genomic analyses of DNA structure | Elsevier Enhanced Reader. https://reader.elsevier.com/reader/sd/pii/S1046202308001539?token=699BD8070D37B7

A757335FC38B4A97FFD2628A2123CB66E79F97F9CC3D76CA7BC4CA6A57544A580E20B5CFB306093AF2.

12. Pohl, F. M., Jovin, T. M., Baehr, W. & Holbrook, J. J. Ethidium bromide as a cooperative effector of a DNA structure. *Proc. Natl. Acad. Sci. U. S. A.* **69**, 3805–3809 (1972).

13. Pohl, F. M. & Jovin, T. M. Salt-induced co-operative conformational change of a synthetic DNA: Equilibrium and kinetic studies with poly(dG-dC). *J. Mol. Biol.* **67**, 375–396 (1972).

14. Drew, H. R. & Dickerson, R. E. Structure of a B-DNA dodecamer. III. Geometry of hydration. *J. Mol. Biol.* **151**, 535–556 (1981).

15. Jovin, T. M. *et al.* Left-handed dna: From synthetic polymers to chromosomes. *J. Biomol. Struct. Dyn.* **1**, 21–57 (1983).

16. Peck, L. J. & Wang, J. C. Energetics of B-to-Z transition in DNA. *Proc. Natl. Acad. Sci. U. S. A.* **80**, 6206–6210 (1983).

17. Ho, P. S., Ellison, M. J., Quigley, G. J. & Rich, A. A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. *EMBO J.* **5**, 2737–2744 (1986).

18. Wang, Y. *et al.* Single-molecule long-read sequencing reveals the chromatin basis of gene expression. *Genome Res.* **29**, 1329–1342 (2019).

Chapter 5 – Conclusions and Future experiments/directions

5.1. Conclusions

5.1.1. *m*EndoG is a Holliday Junction Specific nuclease

We sought to identify *m*EndoG's interaction with Holliday Junction and duplex substrates. Our initial work using a gel-shift assay indicated that the binding to Holliday Junction was about 1-5 uM, while binding to duplex was > 50 uM and therefore indicating a > 10-fold preference for Holliday Junction substrate[1]. With the presence of 5hmC in either duplex or Holliday Junction, there was no noticeable change in bind. Later work further defined the overall binding through Fluorescence Polarization to be about 60 nM for Holliday Junction and maintaining that the duplex binding sits above 40 uM. Both versions of this assay confirm that *m*EndoG has a very strong binding preference for Holliday Junction with respect to duplex DNA[2].

5.1.2. *m*EndoG has 5hmC specificity in a hydrogen bond dependent manner

Though there is no difference in binding preference when 5hmC is present, the modification increases the rate of cleavage for the nucleotide substrates. Our early work indicated a 2-fold increase in cleaved product for both junction and duplex[1]. The amount of cleaved product increased from 20% to 40% for junction with the presence of 5hmC and from 5% to 10% for duplex with the presence of 5hmC. Additionally, our FRET based assay confirmed this 2-fold difference in cleavage for the Holliday Junction by showing a 1.8-fold increase in rate constant when 5hmC was present, thus indicating that the kinetic rate of cleavage for *m*EndoG is 5hmC specific[2].

Given the conserved cysteine that we have seen across species, we set out to determine if the 5hmC specificity occurred in a hydrogen bond dependent manner. Through the use of an

alanine and serine mutant at the C69 position, we determined the specificity for 5hmC to be 1.3-fold with respect to the alanine and 2.1-fold with respect to the serine. These data follow the traditional strength of hydrogen bonding in that CH-OH < SH-OH < OH-OH. Thus, we can conclude that the cysteine recognition of 5hmC occurs through the interaction of a hydrogen bond.

5.1.3. 2 amino acids in A-site contribute to formation of alpha helix between *m*EndoG and CPS6

In order to induce an alpha helix that is present in CPS6 and not in *m*EndoG, 2 amino acids were added in from the homologous positions in the *C. elegans* sequence. These additions saw a decrease in the rate constants for junction cleavage but almost complete recovery when 5hmC was present. To address this from the other direction, we deleted the respective amino acids from CPS6 and saw both an increase in the rate constants and 5hmC specificity. To validate these structure-based assertions, CD assays and MD calculations were done which both support the appearance of alpha helix, in the case of the *m*EndoG insertion mutant, and removal of the helix, in the case of the CPS6 deletion mutant[2].

5.1.4. B-site contributes to Holliday Junction function

Based on the crystal structure we had refined in our lab (6JNU), we next investigated the role of the B-site that sits orthogonal to the A-site in the *m*EndoG structure. By performing a domain swap, we inserted the mouse EndoG B-site into that of the *C. elegans* CPS6. Simply having the B-site mutant increased tightness of binding for CPS6 to Holliday Junction about 2-fold as well as increasing the rate at which it cleaves (though no 5hmC specificity arises). Given that there are two prolines in the CPS6 sequence at the 124 and 129 positions (with respect to C120), we mutated each to the homologous residue in the mouse structure to determine which most influences the loss of function within CPS6. Our data shows that the P124E increased both the binding

tightness and cleavage rate of CPS6 while P129H shows little to no difference compared to wild type CPS6. Thus, we can conclude that the P124 residue is responsible for the majority loss of function[2].

5.1.5. Communication Mutants

Since the presence of the 5hmC modification allowed for recovery of the rate constant for the *m*EndoG protein, we started to explore the idea of communication between the two DNA binding sites in order to determine if a compensation mechanism could be in place. Analyzing the structures indicated that the aspartic acid at the 70 position in *m*EndoG forms two hydrogen bonds, one to an arginine in each of the two binding sites. In generating the D70A mutant, we significantly decreased the overall cleavage of wild type *m*EndoG while simultaneously increasing that of the +9H+5V, the homolog E121A in CPS6, and E121A Bsite proteins. Of note, wild type *m*EndoG does not have the alpha helix present in its A-site while the other three proteins do, indicating that alanine mutation could remove the communication in *m*EndoG wild type while increasing A-site flexibility for the other three proteins, allowing for the increased cleavage[2].

5.1.6. Zhunt

Throughout the EndoG work, it became clear that nucleotide structure and modifications are increasingly important for our understanding of nucleases. We thus sought to update and expand upon a previously developed program in our lab that reads in raw DNA sequences and calculates the propensity of forming Z-DNA based on their conformational free energy. The energies for the traditional nucleotides were updated based on the most recently published in vitro work as well as adding in the methyl-cytosine DNA modification. The script was packaged into a Python wrapper and served on a port that is now available through the internet for open access.

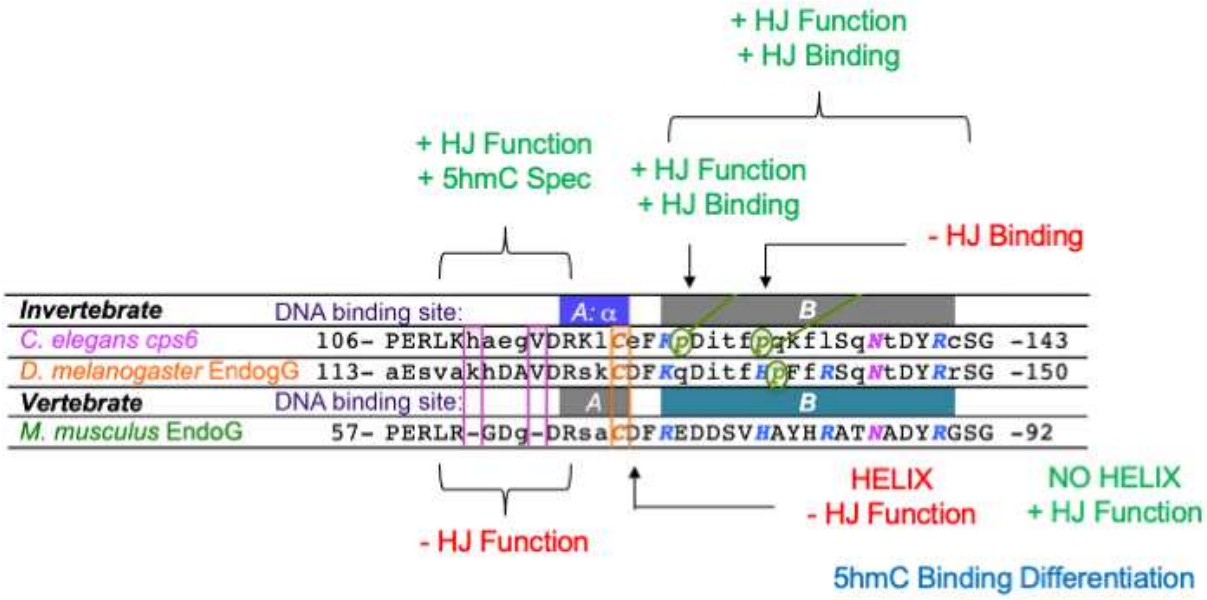

Figure 5.1. Outline of mutations and the respective changes in activity/function for *m*EndoG and CPS6.

5.2. Future Directions

5.2.1. Homologous Recombination on a Plasmid Level

Previous work has indicated that *m*EndoG is involved in the homologous recombination pathway[3]. Additionally, it has been shown that the knock-out of *m*EndoG reduces the overall copy number of mitochondrial plasmids within a cell[4,5]. In order to determine the role that *m*EndoG plays in homologous recombination (potentially both in the mitochondria and in the nucleus), an assay was developed that utilizes two sites of homology on a plasmid containing a kanamycin resistance and a plasmid containing an ampicillin resistance with GFP. Upon a homologous recombination event, the GFP is swapped between the two plasmids and can then be transformed into E. coli cells. Therefore, any cells that are expressing GFP upon induction and growing in kanamycin would be expected to have recombined. Preliminary evidence indicates that recombination events are occuring, though additional optimization must be done to ensure just recombination occurs and *m*EndoG is not allowed to proceed to plasmid cleavage (see Figure A.5.). Additional details regarding this assay and preliminary data can be found in the appendix.

5.2.2. Generating *C. elegans m*EndoG +/+ strains

As we have shown a significant difference in CPS6 binding and cleavage rate, it would be of interest to transfect *C. elegans* with the mouse homolog *m*EndoG[2]. Previous studies have generated CPS6 -/- worm strains and have noted a delay in apoptotic onset across cell types[3,6]. It would thus be hypothesized that *m*EndoG +/+ strains would restore, or potentially enhance, apoptotic onset. A simple return to normal cellular lifecycles would indicate that *m*EndoG /CPS6's role in *C. elegans* simply functions as a nuclease for end-of-cell-life processing and a potential increase could represent the rate constants seen in our studies. In vivo Bioscience strains, with the CPS6 -/- as the starting model, could be generated for mouse *m*EndoG, +9H+5V (which could

mimic most of CPS6 function), and mouse EndoG + TET3. In incorporation of TET3 would account for the lack of presence of 5hmC in *C. elegans* and present the opportunity for a whole new suite of pathways and reactions to occur[7].

5.2.3. Nucleosome and 5hmC Positioning

In addition to the previously mentioned homologous recombination assay, the function of EndoG in the nucleus during apoptosis (and potentially during other times) should be explored. The structure of DNA exiting a histone is almost identical to that of the Holliday Junction, which has led us to believe that *m*EndoG would have a preference for nucleosomal adjacent DNA[2]. Preliminary studies indicate that *m*EndoG cleaves nucleosomal DNA in a ladder-like pattern of about 150 base pairs each (when the DNA is in a repeat array format saturated with histone)[8]. Additionally, we investigated the same nuclease potential when histone H1 was present and saw a complete loss of cleavage (see Figure A.2.). This could indicate that *m*EndoG and histone H1 share a common binding site (be the exiting nucleotides from the histone). Further work would look to augment the linker region in the array repeat sequence to ensure that ample room is available for binding in the presence of histone H1 as well as incorporating 5hmC modifications within the sequence in order to specifically target cleavage events within respect to the nucleosome.

5.2.4. *m*EndoG + HJ crystal structure

The model we generated from the T4 Endo VII and *m*EndoG structures strongly indicate that the two interact in such a manner that two arms of a Holliday Junction fit into the A-site and B-site of *m*EndoG simultaneously[1]. To further prove this, co-crystals should be grown that contain *m*EndoG and Holliday Junction. Preliminary work has been done and trays setup, but have lacked any viable crystals. A new mutant of *m*EndoG was generated, Q129A, which has thus far been the

most inactive version of the proteins. Using Q170A incubated with Holliday Junction would provide the best chance for a co-crystal to form.

5.2.5. Mitochondrial DNA Isolation and Nuclease Digestion

Previously published work has identified a novel and simplified model for isolating and purifying mitochondrial DNA[9]. Since the role of EndoG in the mitochondria is mostly unknown at this point, in vitro nuclease assays utilizing mtDNA would be a great system for identifying function and activity, as compared to the other substrates already observed. The supercoiled native structure of mitochondrial DNA works as a direct structural homolog to the Holliday Junction and the dimeric state of *m*EndoG would allow for supercoiled DNA to sit within both monomers' active sites simultaneously for cleavage.

5.2.6. Mitochondrial Membrane Tracking

The exact location of EndoG within the mitochondria has yet to be determined. There is evidence that EndoG exists within the intermembrane space and potentially docks onto the inner membrane of the mitochondria[10]. It is possible that EndoG could undergo membrane flipping in order to access the mtDNA, which is located within the inner membrane. The location of EndoG could help elucidate its key function within the mitochondria as localization to the intermembrane space would indicate EndoG works mostly as an mt-mRNA regulator, while localization to the inner membrane space would implicate EndoG's involvement in mitochondrial DNA replication (as has been previously asserted)[4,11–14]. Preliminary data in our lab indicates that fluorescently labeled CPS6, with its mitochondrial localization signal (MLS) still attached, localizes to synthetically generated lipid membranes. Additional work needs to be done to test CPS6 without its MLS as well as optimizing the membrane composition to mimic that of the inner mitochondrial

membrane. Should EndoG show potential for membrane flipping, in vitro assays utilizing a D-loop substrate would help to further link the nuclease to mitochondrial DNA replication.

5.2.7. Nuc1, a stronger hydrogen bond

Of the homologs of *m*EndoG analyzed, the yeast Nuc1 is one of the only proteins for which the conserved cysteine is actually replaced with a serine[15]. Given our work in showing a slightly stronger 5hmC specificity with C111S and the recently published work indicating the potential presence of 5hmC in yeast, a new mechanism relating vertebrate function to yeast could be found. Our lab has started the process of isolating and cloning Nuc1 but have yet to express and characterize this protein.

5.2.8. Halogenated *m*EndoG

The catalytic site for *m*EndoG sites within the DNA binding A-site and utilizes a magnesium ion, which is positioned in a pocket and sits between the protein and DNA when bound. In the mouse sequence, there is a glutamic acid which stabilizes the position of the Mg+2 and is therefore positioned in the direction of the ion. We predicted that displacing the metal with a larger amino acid would disrupt the catalytic site and inhibit nuclease activity. Initially, we replaced the glutamic acid with a phenylalanine to completely fill the space with the mutated amino acid. It was immediately clear that the nuclease activity had been hindered as our cell pellet generated more protein than had yet to be collected from any of our preps. Further conformation came from a time-based nuclease assay when no detectable cleavage occurred throughout the two hours of reaction. Additional work and studies on X- *m*EndoG can be found in the appendix.

References

1. Zanden, C. M. V., Czarny, R. S., Ho, E. N., Robertson, A. B. & Ho, P. S. Structural adaptation of vertebrate endonuclease G for 5-hydroxymethylcytosine recognition and function. *Nucleic Acids Res.* **48**, 3962–3974 (2021).
2. Czarny, R. S. & Ho, P. S. Introducing Vertebrate Function into an Invertebrate Enzyme: Recapitulating Mouse Endonuclease G Recognition of 5-Hydroxymethylcytosine and Holliday Junctions in the Worm Ortholog. (2021).
3. David, K. K., Sasaki, M., Yu, S. W., Dawson, T. M. & Dawson, V. L. EndoG is dispensable in embryogenesis and apoptosis. *Cell Death Differ.* **13**, 1147–1155 (2006).
4. Wiehe, R. S. *et al.* Endonuclease G promotes mitochondrial genome cleavage and replication. *Oncotarget* **9**, 18309–18326 (2018).
5. Pardo, R. *et al.* EndoG Knockout Mice Show Increased Brown Adipocyte Recruitment in White Adipose Tissue and Improved Glucose Homeostasis. *Endocrinology* **157**, 3873–3887 (2016).
6. Parrish, J. *et al.* Mitochondrial endonuclease G is important for apoptosis in C. elegans. *Nature* **412**, 90–94 (2001).
7. Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S. & Jacobsen, S. E. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol.* **12**, R54 (2011).
8. Ishihara, Y. & Shimamoto, N. Involvement of endonuclease G in nucleosomal DNA fragmentation under sustained endogenous oxidative stress. *J. Biol. Chem.* **281**, 6726–6733 (2006).
9. Gregg, C., Kyryakov, P. & Titorenko, V. I. Purification of mitochondria from yeast cells.

J. Vis. Exp. **30**, (2009).

10. Ohsato, T. *et al.* Mammalian mitochondrial endonuclease G. *Eur. J. Biochem.* **269**, 5765–5770 (2002).
11. Choi, Y. S. *et al.* Shot-gun proteomic analysis of mitochondrial D-loop DNA binding proteins: Identification of mitochondrial histones. *Mol. Biosyst.* **7**, 1523–1536 (2011).
12. Zhang, J. *et al.* Endonuclease G is required for early embryogenesis and normal apoptosis in mice. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 15782–15787 (2003).
13. Parrish, J. Z., Yang, C., Shen, B. & Xue, D. CRN-1, a Caenorhabditis elegans FEN-1 homologue, cooperates with CPS-6/EndoG to promote apoptotic DNA degradation. *EMBO J.* **22**, 3451–60 (2003).
14. Chen, X. J. Mechanism of Homologous Recombination and Implications for Aging-Related Deletions in Mitochondrial DNA. *Microbiol. Mol. Biol. Rev.* **77**, 476–496 (2013).
15. Vincent, R. D., Hofmann, T. J. & Zassenhaus, H. P. Sequence and expression of NUC1, the gene encoding the mitochondrial nuclease in Saccharomyces cerevisiae. *Nucleic Acids Res.* **16**, 3297–3312 (1988).

Appendix A – Additional roles and functions of EndoG

A.1. Differential environmental preference of *m*EndoG and CPS6

A.1.1. Background and Preliminary Data

Since the conserved cysteine appears to be important for site-specific cleavage and the associated rate constant, we next looked at how pH changes would influence the activity[1,2]. Using a Buffer solution with either a pH of 7 or 8 and with and without the presence of BME, we tested the rate constants of both CPS6 and *m*EndoG. To be able to obtain values for CPS6, as they are below the limit of detection for the plate reader at our standard concentration, we bumped the amount of CPS6 to 1 uM for these experiments. Given that the –SH group becomes deprotonated for the cysteine amino acid above pH 8.3, it appears that *m*EndoG has an increase in activity when in the deprotonated state, while CPS6 prefers the protonated state. The preference of CPS6 for the lower pH of 7 might suggest CPS6 is involved in development, given that *C. elegans* cells have a pH of about 6.8 during their early stages. In both cases, the presence of BME exacerbated the effects of the pH change, indicating that protein reduction has opposite effects on the proteins.

Lin et al. have shown that CPS6 changes its oligomeric state based on oxidative stress on the system in the form of hydrogen peroxide, with a dimer being preferred in the natural state and the monomer appearing with H_2O_2[2]. In a similar fashion, we sought to determine if pH would influence the protein-protein interaction. The same solutions as described for the rate reactions were run on native protein gels showing that pH 8 solutions promote multimers for *m*EndoG (di-, tri-, and tetra-mers) while pH 7 promotes mono- and dimer formation. This points to a pH sensitive regulation in that the oligomeric state of the protein dictates its activity.

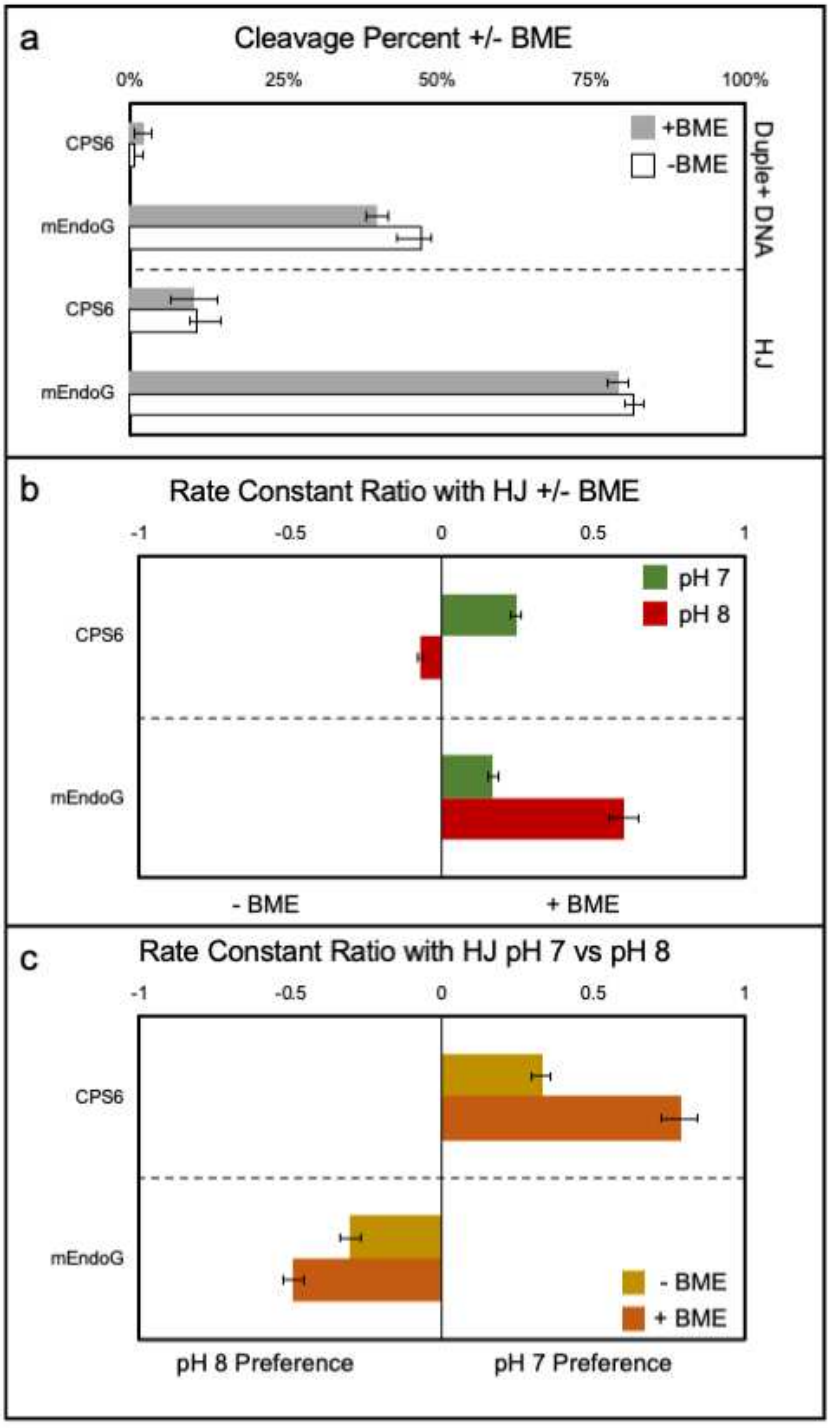

Figure A.1. Environmental influences of pH and BME on the cleavage rates of *m*EndoG and CPS6. a) *m*EndoG and CPS6 were both incubated at different pHs of 7.0 and 8.0, with and without BME present to determine their total cleavage of Junction. b) Rates were then determined by Bulk FRET and normalized to indicate a preference with or without BME. c) Additionally, the rates were normalized with respect to pH to show preference for neutral or slightly basic conditions.

A.2. Nucleosome Digest

A.2.1. Background and Preliminary Data

The potential pH dependency of activity made us consider the nucleus as well. Most mammalian nuclei sit in a pH range of 7.3-7.8, which would place *m*EndoG in a good environment for activity. Knowing this, we tested *m*EndoG's activity with respect to nucleosomes. Using repeat sequence nucleosome binding arrays with histones, *m*EndoG was incubated with the sequence and native agarose gels were run to resolve the products. After analysis with ImageJ, bands of ~150 base pairs can be seen in a stepwise manner, validating what had been seen previously. To compliment this data, the structures of Holliday Junctions and DNA exiting a nucleosome were compared and found to have high structural similarity. Overall, this adds to the idea that *m*EndoG's function is highly imparted by the structural conformation of the DNA to which it is binding.

We next looked at incorporating Histone H1 into the nucleosome arrays, as the H1 binding site is the same as that predicted for *m*EndoG. Traditional nucleosome nuclease assays using MNase have been shown to be severely, though not completely, inhibited in the presence of H1 as H1 assists in wrapping a portion of the linker DNA. Arrays were generated in a similar manner with the addition of H1 and then incubated with *m*EndoG to allow for cleavage activity. Even after a period of 80 minutes at 37 °C, there was no nuclease activity seen in the presence of H1. Though promising in these initial studies, additional work will be needed to confirm *m*EndoG's cleavage location at the H1 binding site.

A.2.2. Future Experiments

In order to explore where *m*EndoG is binding and cleaving the nucleosomal array, a variety of arrays can be generated. Additionally, the beforementioned and following assays could be done in tandem with the MNase protein to more clearly distinguish between the two nucleases.

A.2.2.1. Increased Linker Length

By increasing the linker sequence between nucleosomes on the array in the presence and absence of H1, there is potential that the larger exposed sequence would allow for *m*EndoG to bind and cleave throughout the linker region and indicate a nonspecific cleavage with respect to the nucleosome exit site. However, if cleavage continued to be inhibited with the increased length in the presence of H1, this would strongly indicate a binding site specific cleavage for *m*EndoG and the nucleosome.

A.2.2.2. EndoG Recognition Sequence in Linker

Using the recognition sequence determined by Robertson et al., the 5'-GGGGCCAG-3' would be inserted into the linker region in a shifting window manner (with the 5' G starting at the nucleosome exit and moving the cassette stepwise down the linker)[3]. This system would allow us to parse out EndoG's binding and cleavage preference between the structure of the exiting nucleosome and its recognition sequence. There also is the possibility for multiple cleavage events, for which the small nucleotide sequence between the linker start and the cut site in the recognition sequence would need to be searched for in analysis.

A.2.2.3. 5hmC Modification in Linker

Similar to the recognition sequence, the 5hmC modification could be added to the linker region in tandem with the recognition sequence. Using this system, we would be able to elucidate if binding is occurring with the recognition sequence, as we would be able to observe an increase in cleavage products as indicated by our previous studies. Ultimately, this study could provide interesting in vivo roles with regards to EndoG's activity and binding.

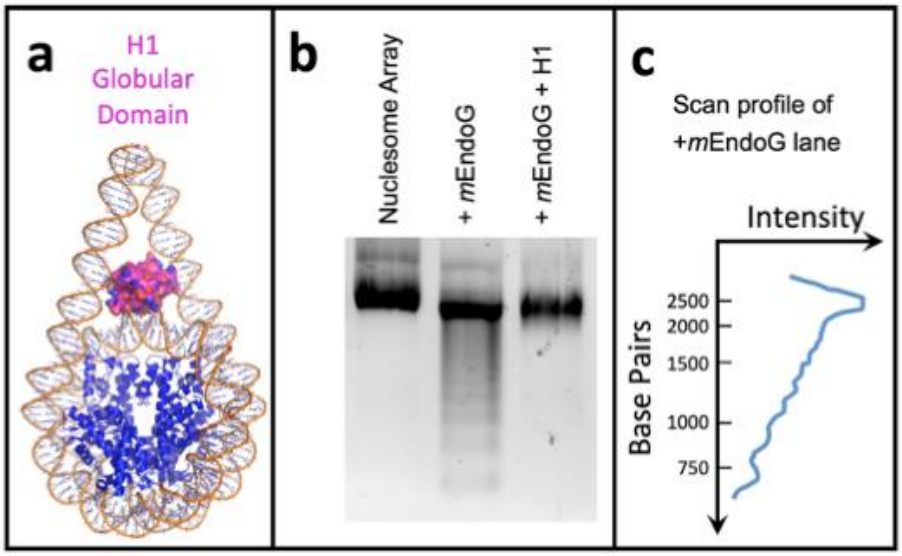

Figure A.2. *m*EndoG cuts nucleosome arrays specifically and is inhibited by the presence of H1. a) A nucleosome with extended DNA arms was modeled with *m*EndoG and shows potential binding consistent with the A and B-sites in *m*EndoG. b) *m*EndoG was incubated with nucleosome arrays with and without Histone H1 present and shows strong inhibition by H1. c) Lane 2 was then plotted using ImageJ to determine the band sizes of the nucleosome cleavage and indicate 150 base pair steps between each band.

A.3. Binding and Cleavage of variety of nucleic substrates

A.3.1. Background and Preliminary Data

Previously, we have reported that the binding of *m*EndoG to duplex strands to be > 75 uM[1]. However, this was done with Electrophoretic mobility shift assay (EMSA) due to restriction of resources1. With Fluorescence Polarization, the binding of *m*EndoG to duplex DNA (with and without the 5hmC modification) sits at 12-18 uM. The KD was also determined for Holliday Junctions (with and without the 5hmC modification) to be about 60 nM. With the consideration that *m*EndoG has a preference for the structure seen in the Holliday Junction, we also determined the KD for a 3-arm Junction to be about 112 nM. Given that mitochondrial replication initiates within a D-loop, a structure that is homologous to a 3-arm Junction, *m*EndoG's tight binding to this substrate could elucidate potential roles in the pathway. In addition to the potential D-loop interaction, we wanted to determine if *m*EndoG could act as an mRNA regulator and found the binding to mRNA is 330 nM, not quite as tight as for Holliday Junctions but significantly stronger than interactions with duplex DNA.

In all cases, the presence of the 5hmC modification did not present an increase in binding for *m*EndoG WT, indicating a process other than binding is responsible for the differences in rate constants. To elucidate this, the Km was determined for *m*EndoG for the Holliday Junction at about 2 uM and for the 5hmC-modified Holliday Junction at about 900 nM, which directly reflects the 2-fold difference seen in the rate constants for these substrates. These results indicate that *m*EndoG binds Holliday Junctions with a high specificity but needs positioning imparted by either the B-site or conserved cysteine, or combination of the two, in order to increase its catalytic rate.

Additionally, we looked at the cleavage of mRNA by both *m*EndoG and CPS6. When we allowed the reaction to proceed for 1 hour at 37 °C with 1 uM of protein and 500 nM RNA,

*m*EndoG showed cleavage of about 30% of the substrate while CPS6 had no detectable cleavage. This is interesting as it presents another potential difference in roles between the two proteins in different species.

A.3.2. Future Experiments

Based on the data we collected, *m*EndoG has a binding affinity for 3-arm Junctions that is almost as tight as that of Holliday Junctions. Interestingly, 3-arm Junctions mimic the structure of D-loops seen in the mitochondria, which are used in mtDNA replication. EndoG has previously be implicated in mtDNA copy number and replication and yet its function has remained a mystery. Using the idea of the 3-arm Junction, a synthetic D-loop system could be generated in which the embedded strand contained a Cy5 label and one of the encapsulating strands contained a FITC (alternating the strands in order to determine if there is a binding or cleavage preference). Should site specific cleavage be shown, there would be strong evidence that EndoG assists in D-loop cleavage and promotes strand replication.

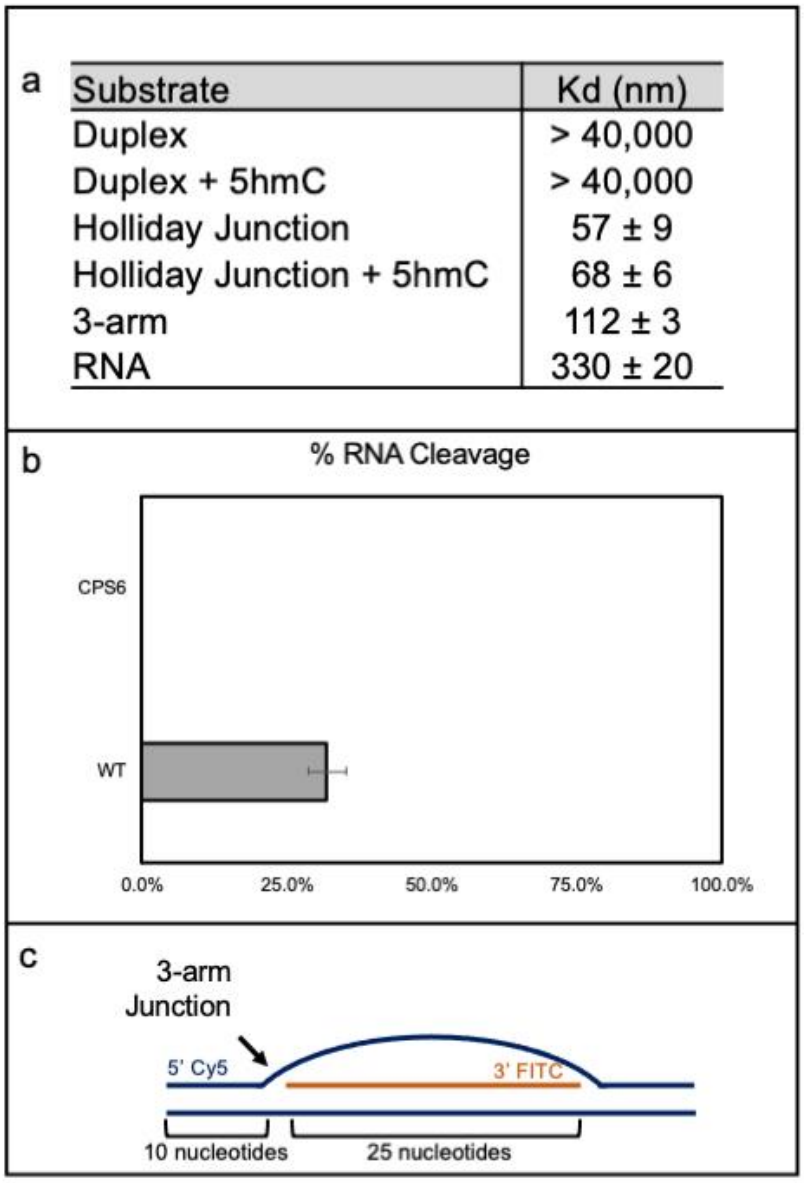

Figure A.3. *m*EndoG shows binding preference for Holliday Junction and strong affinity for 3-arm Junctions and mRNA. a) Fluorescence Polarization was used to determine *m*EndoG's binding to duplex and Holliday Junction DNA with and without the 5hmC modification as well as 3-arm Junction and RNA. b) The total cleavage of the RNA was determined for both CPS6 and *m*EndoG and only showed activity in the later. c) We next propose an in vitro D-loop system with respect to the 3-arm Junction to determine *m*EndoG's interaction with the substrate.

A.4. EndoGI Binding

A.4.1. Introduction and Background

Previous work done by Wahle et al indicated that the binding of EndoG to its inhibitor in drosophila was within the picomolar range for the interaction[4]. Given that the inhibitor was used in the expression and purification process for our system, we decided to test the affinity to the mouse version of EndoG.

Based on the co-crystal structure (PDB 3ISM) of Drosophila EndoG and its inhibitor (EndoGI) by Loll et al., EndoGI docks into a dimeric form of EndoG and binds in the A-site of each monomeric EndoG[5]. The binding of these proteins has been reported to be incredibly tight, in the pM range, which lead to concern regarding the bindings we observed for the DNA substrates[4]. Should binding of the inhibitor be that tight, there is a very low chance Holliday Junctions in the 60 nM range, or even less so duplex in the 13 uM range, would be able to compete for binding and thereby render the EndoG enzyme ineffective. The binding of *m*EndoG to EndoGI was thus determined by EMSA and determined to be about 25 uM. Inhibition at this level would allow for *m*EndoG to preferentially bind Holliday Junctions while more closely monitoring its binding to duplex DNA.

A.4.2. Experimental Methods

A.4.2.1. EndoG Inhibitor Purification

The EndoG inhibitor (EndoGI) was expressed in a pET plasmid system and contains a His tag on its C-terminus. Cultures were grown to an OD600 of 0.6 at 37 °C before being induced with 0.1% v/v of 1M IPTG and then cultured overnight at room temperature. The cell cultures were spun down at 7k rpm at 4 °C for 45 minutes before the cell pellet was removed and stored at -20 °C until purification. Pellets were thawed in EndoG Buffer I (EndoG Buffer A + 10 mM imidazole)

and sonicated before an additional spin at 15k rpm at 4 °C for 40 minutes. Supernatant was filtered through a 0.22 um filter before being loaded onto a HisTrap 5mL column. Protein was eluted from the column using EndoG Buffer E (EndoG Buffer A + 500 mM imidazole). EndoGI was then concentrated in a 10kDa MWCO filter and buffer exchanged into EndoG Buffer A.

A.4.2.2. EndoG(I) Fluorescent Labeling

Throughout testing, both *m*EndoG and EndoGI were fluorescently labeled using the ATTO488 maleimide method. Briefly, the protein was degassed at 4 °C and 0.85 atm for 20 minutes. 100X molar excess TCEP was added and allowed to sit at room temperature for 20 minutes. 20X molar excess maleimide ATTO488 was added and placed at 4 °C overnight. Labeling solution was buffer exchanged either through a 10kDa MWCO concentrator or over a GF column to remove TCEP and unlabeled dye and concentrated down.

Gel Shift Assay

A.4.2.3. FP (Straight Binding and Competitive Binding)

A solution for about 30 nM ATTO488 (~300 nM inhibitor) was made in buffer solution. 15 uL was added to 24 wells with 15 uL of a dilution series of protein starting at 241 uM and doing a 0.75-fold dilution across the 24 wells. A Victor 3 plate reader was used to obtain the FP signal for the ATTO 488 dye. The shift in signal is relatively low compared to other FP experiments shown and that is due to low labeling efficiency as well as binding weight (as the inhibitor has a mass of about 44 kDa while *m*EndoG has a mass of about 26 kDa in its monomeric state).

A.4.3. Results

Based on the well-defined binding data collected for *m*EndoG to a variety of nucleic acid substrates using FP, we first sought to generate fluorescently labeled EndoGI. The labeling method used showed to be about 10% effective. Though not great, 10% labeling efficiency has been shown

to be valid in FP experiments (the assumed concentration and Kd simply have to be taken into account to avoid substrate saturation). The EndoGI concentration was held constant throughout the series with a serial dilution taking place with the *m*EndoG. After multiple runs, there were no shifts seen from the control of no *m*EndoG added all the way up to 10 uM *m*EndoG (FigureA.4). We were initially concerned given the previous claims of EndoGI binding the Drosophila homolog in the pM range. Given how they did their experiments, we next pre-incubated *m*EndoG with Holliday Junction (at a concentration above its Kd of ~60 nM) and titrated EndoGI to look for HJ displacement. However, no change in signal was detected again up to 10 uM (data not shown).

Taking a different approach, we next decided to do an Electrophoretic mobility shift assay (EMSA) in which the *m*EndoG concentration was held constant and EndoGI was titrated in by serial dilution. The samples were then run in a Protein Native Gel in order to maintain the potential binding of the proteins. Binding approaching the Kd (50% of *m*EndoG bound) is predicted at ~60 uM and was limited by expression system concentrations (Figure A.4). Again, this discrepancy to the previously published findings led us to further investigate the inhibitor. To ensure proper folding of the inhibitor, we performed Circular Dichroism experiments. The results were then compiled and compared to the predicted signal based on the structure found in PDB 3ISM. Our inhibitor protein appears to be properly folded and aligns with the predicted structure signal.

Overall, this has some interesting implications. It is possible that the Drosophila homolog truly has a pM affinity for the inhibitor and that mouse EndoG is significantly weaker in binding. The variations in amino acids could contribute to the difference in affinity and could point to *m*EndoG evolving to more readily release its inhibitor and preferentially bind to Holliday Junction and HJ-like structures.

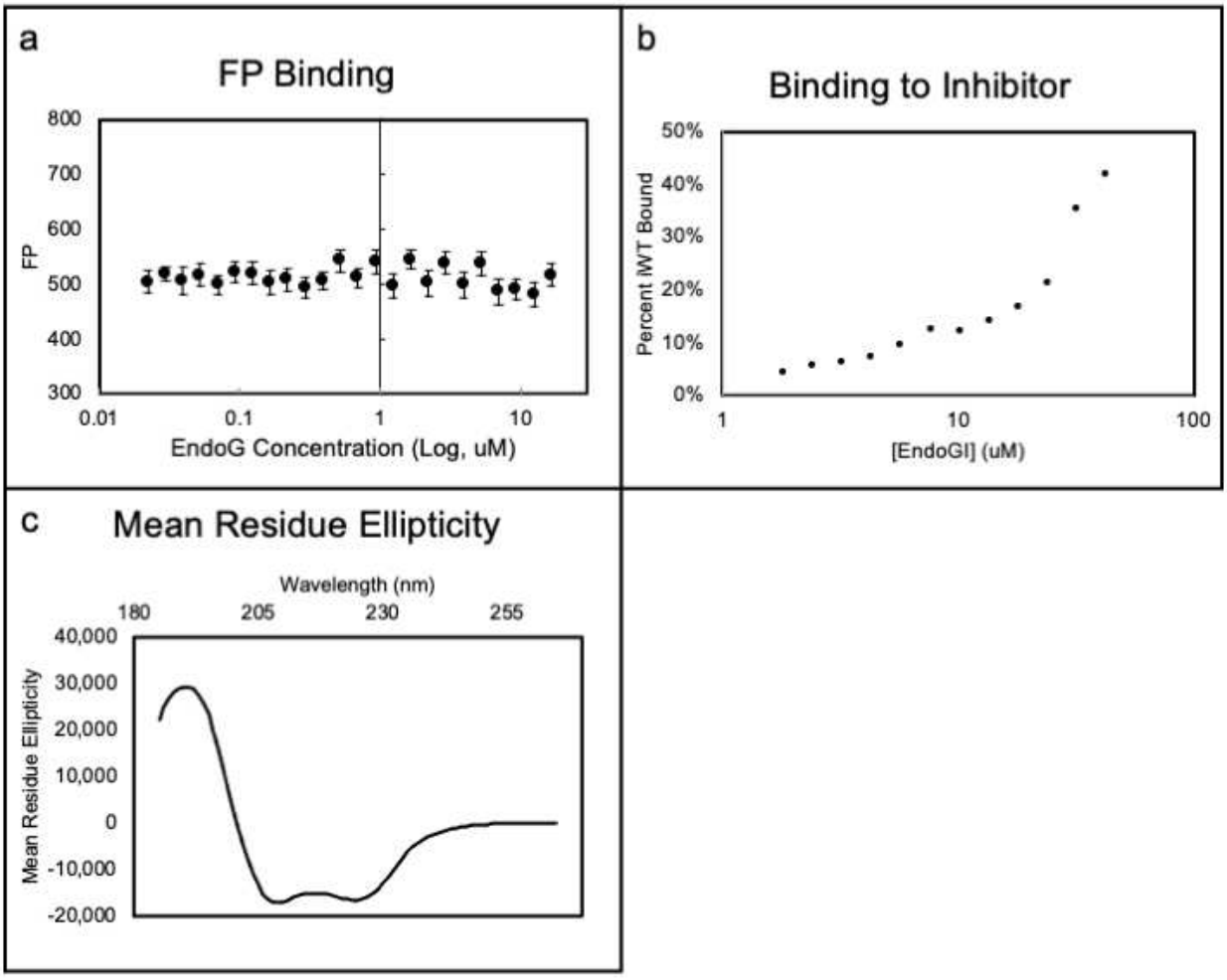

Figure A.4. Determining the binding of *m*EndoG Inhibitor to *m*EndoG. a) Fluorescence Polarization was used to determine binding up to 20 uM *m*EndoG with not binding detected. b) An EMSA was used and predicts a binding Kd of ~60 nM but was limited due to expression system setup. c) The CD spectrum was taken of the inhibitor to ensure proper folding for these experiments.

A.5. In vitro Homologous Recombination of Plasmids

A.5.1. Introduction and Background

Work done by our collaborator Robertson et al. implicated *m*EndoG as having a role in the homologous recombination pathway[3]. Based on our results, we have revised their original model, in which Endo is responsible for generating the nicked strand, as it has been shown that HR most often occurs through double-stranded break repair pathways, in order for strand invasion to occur, so that *m*EndoG participates later in the process by cleaving and resolving the 4-stranded crossing[6]. To test this potential role, we wanted to develop an in vitro system that would utilize full plasmids as a method to mimic an in vivo mitochondrial DNA system. Stark et al. previously developed a similar study in which they created two plasmids with two regions of homology[7]. Their study involved introducing a resolvase that cleaves sequences and assists with reannealing with a homologous sequence, in their case the other plasmids to generate an expanded plasmid. Taking inspiration from this, we generated a protocol to create a two-plasmid system that simulated the crossing-over seen in homologous recombination. We then introduced *m*EndoG to the plasmids in order to observe potential resolvase activity.

A.5.2. Experimental Methods

A.5.2.1. Generating and Purifying HR Plasmids

The first plasmid used for *m*EndoG expression in a pET28A vector was used as the first sequence and labeled pET through the experiment. The second plasmid, obtained from Amanda Koch, was pET21A containing a super-folded GFP, labeled pGFP. Both plasmids share 3 sites of complete homology at their f1 ori, ori, and lacI sequences. Typically, about 500 base pairs are needed for sufficient recombination to occur and each of these sites provides the needed length. 6.66 uL of HJ Buffer (as previously described) was mixed with 5 uL each of pET and pGFP at 20

ng/uL. The solution was then heated to 93.5 °C for 80 minutes and allowed to slowly cool back down to room temperature. 6X DNA Loading Dye was added to the reaction sample and loaded into a 1% agarose gel containing 0.01% SBRY safe. Gels were run at 100 V for 45 minutes and imaged. The gels were then transferred to a UV light benchtop and the combined plasmids were extracted with a razor blade. A small hole was poked into the bottom of a 0.75 mL Eppendorf tube with a 22-gauge needle and the gel extract was placed inside. The tube was then inserted into a 1.5 mL Eppendorf tube and spun at 14k rpm for 10 minutes to pulverize the gel. Material was then resuspended in 1 mL of HJ buffer and stored at 4 °C overnight. The mixture was then filtered through a 0.22 um filter and concentrated down using a 10 MWKO filter.

A.5.2.2. Recombination Assay

Diluted *m*EndoG protein to 400 nM in EndoG RXN Buffer (previously described in nuclease assay). Added 5 uL of EndoG or RXN Buffer (for no protein control) and 2 uL of DNA Ligase to 10 uL of DB solution (at the highest concentration possible to complete the reaction). Let the reaction proceed at 37 °C for 25 minutes. Quenched the reaction with Q Buffer (previously described in nuclease assay) and increase the temperature to 50 °C for 30 minutes. All samples were loaded onto a 1% agarose gel and imaged for their product strands.

A.5.2.3. Recombination Products Transforming Cells

To test the recombination event, cells were transformed with the reaction samples. 4-6 uL of ligated solution was added to 90 uL of Codon+ BL21 cells and incubated on ice for 20 minutes. Cells were then heat shocked at 42 °C for 35 seconds and filled to 1 mL with LB before being placed in a shaking incubator at 37 °C for 1 hour. The solutions were then spun down at 4k rpm for 5 minutes and 900 uL of supernatant was removed. A pipet tip was used to gently resuspend the cells in solution and plated on LB with Kanamycin and Chloramphenicol by bead spreading.

A.5.2.4. Inducing GFP Expression

Individual colonies were then selected and grown in 4 mL of LB with Kan/Chlor to an OD600 of ~0.6. 1 mL of the sample was taken and placed in 4 mL of LB with Amp/Chlor to check for multiple plasmid transformation. 0.1% v/v 1M IPTG was added to each of the Kan/Chlor tubes and allowed to express for 30 minutes. The tubes were then placed under a UV light and analyzed for GFP expression. All tubes expressing GFP were checked against their Amp/Chlor tube to ensure no growth occurred in the presence of Ampicillin. Any cultures that were positive for GFP and negative for growth in the presence of Amp were added to overnight cultures and miniprepped to obtain the resulting plasmid. The miniprepped plasmid was then run on an agarose gel against the original pET, pGFP, 2-plasmid heated product, and the recombination assay product to compare plasmid size changes throughout the experiment. Finally, the sequence was sent to the Harvard MGH CCIB DNA Core for full plasmid sequencing.

A.5.3. Preliminary Results

The first big step was generating and purifying the artificial Homologous Recombination plasmid. After doing a wide range of buffers, heating temperatures, and lengths of incubation time, we finally were able to obtain the parameters used in our methods. Initially, we used the raw material from these reactions to test *m*EndoG but the high level of contamination from the non-annealed plasmids presented a lot of difficulty in determining the final products. Ultimately, we started to gel purify our reaction samples in order to have a cleaner sample to incubate with *m*EndoG. Additional optimization occurred late when the reaction with *m*EndoG took place. We were seeing the cleaved bands appear on the gel, but they seemed to have difficulty in annealing to make the final plasmid. DNA Ligase was then added to the mix and showed the promise of the recombined plasmid length appearing on the gel (Figure A.5).

The remaining sample that wasn't used to run on the gel was then transformed into Codon+ BL21 cells and plated on Kan/Chlor resistant plates. Chloramphenicol was used as a resistance for the Codon+ cells and Kanamycin for the pET plasmid. If homologous recombination occurred, the GFP sequence from the pGFP strand should be found in the pET plasmid, which contains the Kan resistance. Only cells that grew on the Kan/Chlor plates were used for further testing. Individual colonies were selected from the plates, inoculated into LB media with Kan and Chlor, and allowed to grow to an OD600 of ~0.6. 1 mM (final concentration) IPTG was then added, and the cultures were checked under UV light for GFP expression. Cultures that were positive for GFP then had 1 mL of the system removed and placed in a Chlor/Amp solution and allowed up to 4 hours to grow in order to check for remaining pGFP vector present that might represent a double transformation of contaminating plasmids. Negative grow cultures were then identified back to their GFP expressing cultures and the expressing systems were then used to inoculate overnight cultures for plasmid purification by Miniprep. Using primers for the GFP and Kanamycin sequences, the plasmids were sent off for sequencing to confirm the presence of both components.

A.5.4. Future Directions

Future work on this project will look to optimize the steps from cell transformation onward to plasmid sequencing. Initial runs have been completed that showed signs of GFP expression in colonies plated on a Kan/Chlor LB plate and induced with IPTG. Additionally, optimization regarding the Miniprep collection and sequencing of the plasmids is needed. Though a good first step, the primer sequencing will not provide the complete picture and full plasmid sequencing will be needed. Due to the fact that the expression system is in Codon+ cells, the plasmid will need to be separated from the Codon+ plasmid before sending to the Harvard Sequencing Core as to prevent noise between the two sequences.

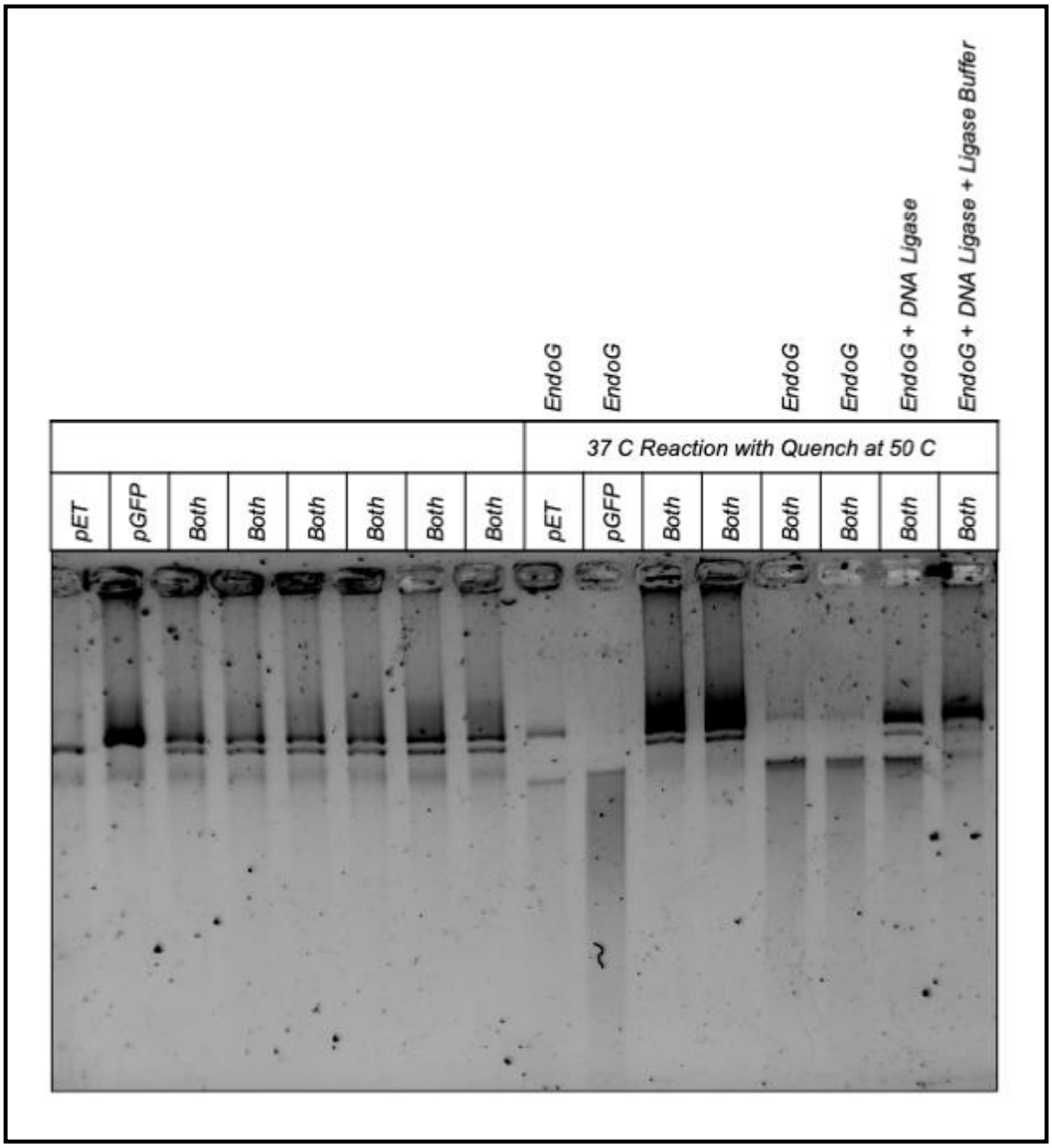

Figure A.5. In vitro Homologous Recombination induced by *m*EndoG. pET and pGFP were used to generate an artificial Homologous Recombination event and introduce to *m*EndoG, *m*EndoG with DNA Ligase, and *m*EndoG DNA Ligase and DNA Ligase Buffer. Extracts from the latter two were later transformed into Codon+ BL21 cells to check for GFP expression.

A.6. AUC data to show dominant dimer state

A.6.1. Experimental Methods

Holliday Junction and *m*EndoG protein were buffer exchanged in 0.05 M Tris-base, 1 M NaCl, 1 mM MgCl2, 5% v/v glycerol, 0.014 uM BME. Sample concentrations were adjusted to an absorbance of 0.6 measured at 280nm. Loaded 450 ul of buffer into the reference sector and 400 ul of the sample into the sample sector of a 2-sector aluminum AUC cell. Samples were run at 40k RPM at 20 °C and a wavelength of 280nm. A total of 400 scans were taken throughout the run. Data was analyzed using UltraScan using the Van Holde-Wieschets Analysis module to subtract for the effects of diffusion.

A.6.2. Results

The first scan that was run was *m*EndoG on its own in order to determine how the protein natively formed in solution without DNA present. In alignment with data collected from native proteins and gel filtration columns, *m*EndoG formed a dimer in solution. We next introduced Holliday Junction to the solution and ran the AUC at the same parameters. It can be seen that the dimer binds to the Holliday Junction until it is completely saturated, further proving the model that had be predicted. This system provides an interesting mechanism for nuclease activity to occur as the two binding and active sites are oriented on opposite sides of the dimer and allow for multiple strands, or a super-coiled strand, to be cleaved simultaneously. Though in vivo roles have remained elusive, this dimeric binding assists in narrowing the wide array of potential functions.

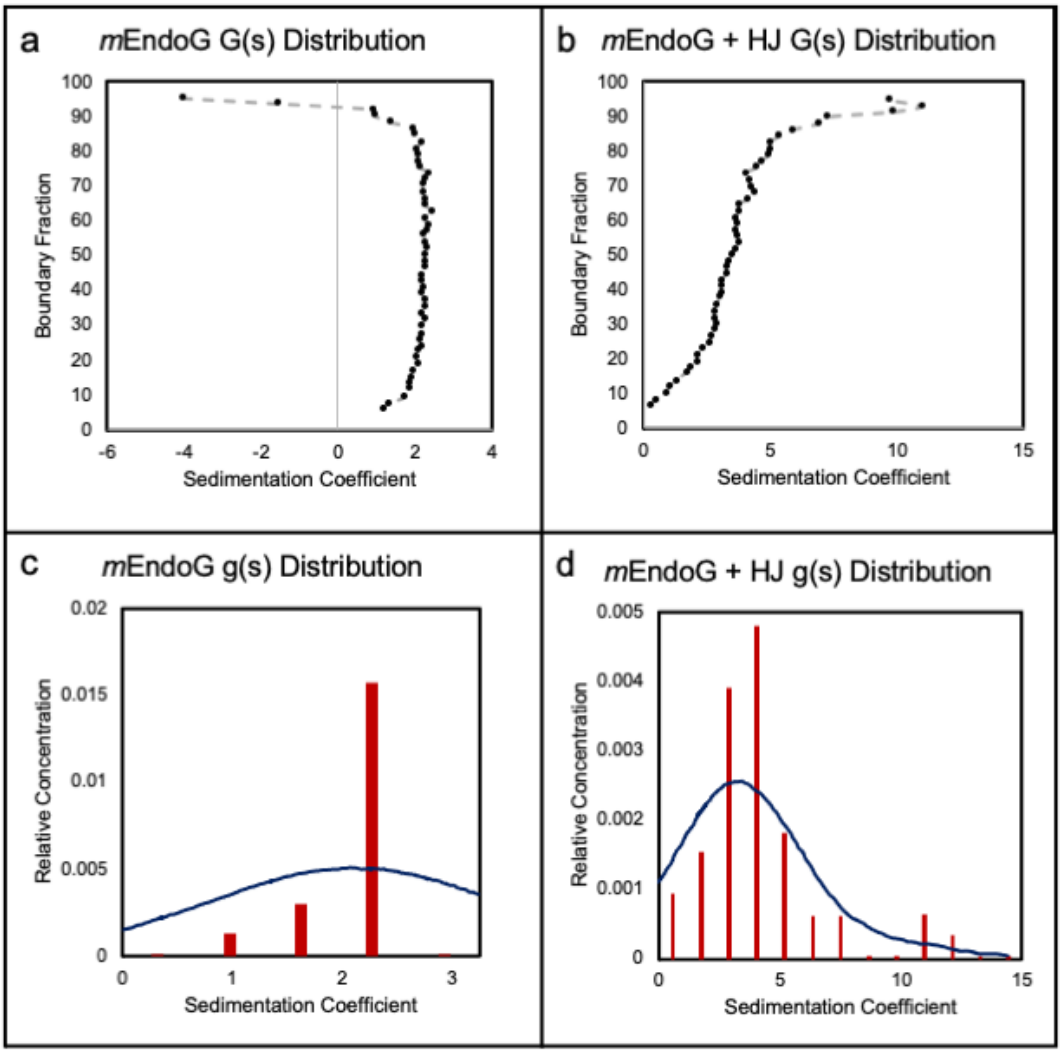

Figure A.6. AUC experiments indicate *m*EndoG exists in a dimeric state and binds HJ in dimer as well. a&c) Represent *m*EndoG protein alone and indicate a Sedimentation Coefficient close to 2 for the protein. b&d) Adding Junction shifts that value close to 4 and represents the dimeric form of *m*EndoG binding the Junction.

A.7. X-bonded *m*EndoG

A.7.1. Introduction and Background

Since we had successfully killed nuclease activity, we next sought to recover catalytic function with the incorporation of a Halogen Bond (X-Bond). The Halogen Bond is a subclass of non-classical noncovalent (NC-NC) interactions that are defined by the sigma-hole, and interaction that is electrostatic in nature. The basis for sigma-hole theory arises from electrons from covalent bonds entering the sigma molecular orbital. This depopulation of the outer electron shell generates an electropositive patch which can serve in a similar manner to that of a Hydrogen bond. Additionally, an electronegative waist is generated around the halogen atom allowed it to function as an X-bond donator and Hydrogen bond acceptor. The overall strength of the X-bond is dependent on the size of the sigma-hole and follows the trend of increased polarizability with the increasing size of the atom (I > Br > Cl >> F). Generally, X-bonds can be defined by their electrostatic potential but other factors including dispersion, steric repulsion, polarization, and charge transfer also contribute to the overall interactions.

Our lab has a sizable background with regards to halogen bonds and their incorporation into biological systems. Work done in our lab by Crystal Vander Zanden found that incorporation of a halogenated nucleotides could assist in stabilizing Holliday Junction formation. An array of different halogens were placed on the 2 position of a deoxi-uracil's sugar at a position 2 nucleotides from the center of the 10mer oligo sequence. It was found that the halogen formed a Halogen Bond to the backbone of the adjacent strand and stabilized the crossing-over needed to form the 4 stranded junction.

Additionally, in collaboration with Ryan Mell's group at OSU, others in the lab have incorporated halogens into protein systems with the use of non-canonical amino acids. Work done

by Matt Scholfield was one of the first cases studying Halogen Bonds in a protein system, where he specifically used the protein T4 Lysozyme. By targeting a tyrosine site within the sequence that forms an internal hydrogen bond, Matt mutated generated a mutant for the TAG stop codon, for which he added a tRNA synthetase for TAG that translates the codon to a ncAA. By supplementing the media with synthetically generated halogenated phenylalanine, the protein incorporated the ncAA specifically at the mutated position. Through thermal melting studies, Matt showed that the halogenated group, when solvent exposed, destabilized the protein but was able to rescue most of the lost stability when positioned to form a halogen bond. Though the halogen bond formed, a decrease in melting temperature of about 1° to 3° C indicates that a halogen bond cannot completely replace a hydrogen bond.

Due to the necessity of the hydrogen bond for stability, Anna-Carin Carlsson constructed a halogenated tyrosine system, in which the halogen resides on the *meta* position. Thermal stability assays showed an increase in stability with melting temperatures raising by 1° to 2° C as well as an increase in activity by about 15% at higher temperatures. Interestingly, the chlorinated tyrosine provided a higher contribution to stability than was originally expected, based on the previously mentioned DNA studies. Looking closely at the structure, it was observed that the OH group of the tyrosine could spin about its carbon-oxygen bond, thereby changing the position of the hydrogen. Quantum mechanical analysis was performed with the hydrogen positioned towards, orthogonal, and away from the halogen and showed there was a hydrogen bond forming between the OH group and the electropositive waist of the halogen, which resulted in additional polarization and increase in the sigma hole size. This new interaction was thus labeled as a Hydrogen Bond enhanced Halogen Bond (HBeXB).

Since T4 Lysozyme is considered a “rock” in terms of stability, Rhea Kay Rowe Hartje sought to insert an ncAA into the GCN4 coiled-coil system. Previous work had determined that the dimeric GCN4 formed a trimer in the presence of benzene when mutated to N16A. Rhea Kay determined that the N16 could be replaced with a tyrosine in a way that the tyrosine ring would sit in the same position as the trimeric inducing benzene. She first generated the N16A sequence and confirmed the trimeric formation through DSC and crystallographic analysis. Based on the crystal structure, there was potential for a halogen bond to form between the modified amino acid and the backbone of an adjoining peptide. Utilizing the same system as Matt Scholfield and Anna-Carin Carlsson, a halogenated tyrosine was inserted at the N16 position for one of the peptides and observed in its binding to two N16A peptides. Based on the CD data, analyzed for the helicity corresponding to the coiled-coil structure, it was shown that the heterotrimer forms very tightly (in the sub-nanomolar range) with a homotrimer of the halogenated peptides forming in a x1000 weaker manner (when looking at the tetra-fluoroiodo-tyrosine mutant). Thus, the specific binding of the halogenated peptide provides a target recognition element for BXBs.

Given our previous work with *m*EndoG, we then aimed to use the halogen bond as an alternative to the catalytic magnesium ion in the A-site. In looking at the active site for *m*EndoG, we determined E136 as an amino acid used in coordinating the magnesium ion and situated in a way that a halogenated tyrosine could fill the same space. To include the halogenated tyrosine, we needed to utilize the same TAG system for non-canonical incorporate the amino acid into the protein.

A.7.2. Preliminary Results

A mutant sequence for *m*EndoG was generated with E136 swapped for the TAG codon. Cells were transformed with the stop codon *m*EndoG plasmid as well as a plasmid containing a stop

codon tRNA synthetase that was originally obtained from Ryan Mehl's group (pDule, with the Tetracycline resistance swapped with Ampicillin resistance). First, we tested the expression of the TAG – *m*EndoG to ensure the expression of the truncated form of the protein. Expressions were run at 10 °C, Room, Temperature, and 37 °C for up to 48 hours with time points taken throughout. Based on these expressions, the TAG – *m*EndoG system successfully generated the truncated form of the protein up to about 4.5 hours. At 20 hours of expression, the products all dropped off and the cell lysates were lower across all temperatures, indicating that degradation and cell death could be induced by truncated *m*EndoG misfolding and/or aggregation.

Having confirmed the expression of the truncated form, we next sought to incorporate a non-canonical amino acid (ncAA) at our TAG site. Right before IPTG induction, 1 mM (final concentration) of Cl-3-tyrosine was added to the media so that the tRNA synthetase could bind the substrate. Again, we used the 3 different temperatures for up to 48 hours after induction and ran time points on an SDS Page gel.

The protein was purified by MBP column, followed by Tev Cleavage overnight, Heparin column, and finally over a GF column (which suspended it back into EndoG Buffer A). We first wanted to determine if the protein was catalytically active. A time-based nuclease assay was setup with a positive control of WT *m*EndoG and a negative control of E136F. The reaction was incubated with Holliday Junction DNA and run over a two-hour period with samples quenched in 0.5% SDS and Proteinase K throughout. The samples were then run on a native DNA and imaged. *m*EndoG cleaved the Holliday Junction and E136F showed no cleavage, as expected; however, Cl-*m*EndoG also showed no cleavage throughout the 2 hours of the reaction.

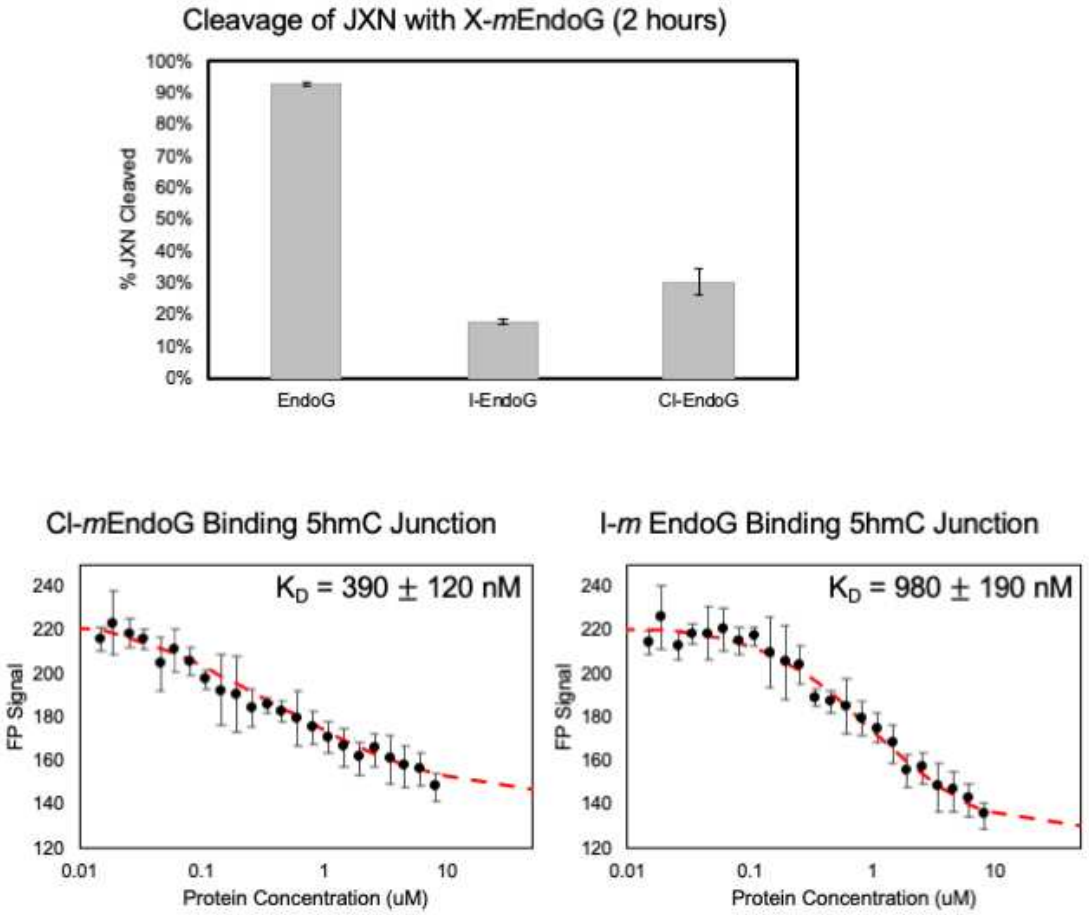

Figure A.7. Cleavage and binding of the Halogenated *m*EndoG mutants. A nuclease assay was performed in triplicate to determine the amount of cleavage of Holliday Junction done by Wild Type *m*EndoG with the Halogenated mutants. Additionally, the binding of the Halogenated mutants to junction was determined through FP.

We next thought to increase the size of the halogen substituent and incorporated an iodinated tyrosine at the E136 position. The same method for purification was used as above. The iodinated protein was then run in a cleavage assay for a total length of 2 hours and showed preliminary cleavage. Based on these data, we checked the chlorinated yet again to ensure that everything was run according to the same method. Everything together indicated that chlorinated as allowed for cleavage to occur (and indicating that protein was probably not added to the same during the first runs).

Since both of the proteins were able to cleave (albeit at a low level as compared to wild type *m*EndoG), we next decided to explore the pH effect on the activity of the halogenated nucleases. It is possible that the halogen is providing the energy needed for catalysis to occur, but there is also a possibility that the hydroxyl is functioning in a similar manner as that seen in topoisomerases. By deproteinating the hydroxyl group, a tyrosine is able to participate in an SN2 reaction with the substrate and might be used as a method for cleavage in X-*m*EndoG. In halogenating the tyrosine, the pKa of the side chain decreases to about 8.4, incredibly close to the pH of 8.0 used in the reactions. We thus modulated the pH of reaction between pH 6 and 9 in half-step intervals in order to determine a change in activity. Based on this model, if the hydroxyl group is responsible for the activity, the cleavage should be low up to and around the pKa and then increase dramatically at pH 9. However, if the halogen is responsible for the activity, the cleavage should follow the opposite trend and increase dramatically as it approaches pH 6.

Future work will explore – hydroxyl verses halogen cleavage for the Halogen mutants, magnesium independent cleavage, and modeling the system to calculate the energies associated with halogen-dependent cleavage.

A.8. Plasmid Digest Assay

A.8.1. Introduction and Background

Our work with the Homologous Recombination assay described before led us to consider if *m*EndoG was specifically cleaving those plasmids due to sequences it recognized or if structural components were influencing the cut site. The reason being, we have seen that *m*EndoG has a binding and catalytic preference for Holliday Junction substrates and the structure of the junction can be closely mimicked by supercoiled DNA plasmids. To test this, we took a number of sequenced plasmids in our lab and incubated them with *m*EndoG and known single cut site Restriction enzymes in order to map the specific cleavage site.

A.8.2. Experimental Methods

5 uL of 1 uM *m*EndoG was incubated with 20 ng/uL of plasmid types of pDule, pET-28a, and pMal for 15 minutes at 37 °C. The reaction was then quenched with 5 uL of Quenching Buffer before being heated to 50 °C for 30 minutes. Each solution was purified with a Monarch PCR Cleanup kit and the products were suspended in 15 uL of elution buffer. The solutions were split to have 4 uL in 3 wells, either 1 uL of EcoRI, HindIII, or both restriction enzyme, 1 uL of CutSmart, and then filled to 10 uL with DI water. The restriction enzyme was allowed to incubate at 37 °C for 1 hour. All samples were then placed in a 1% agarose gel with SYBR Safe and run at 100 V for 30 minutes before imaging.

A.8.3. Results

After imaging the gel, the sizes of each product were measured against the ladder and used to back-calculate the cleavage site by *m*EndoG. A 200-nucleotide window was narrowed down (due to resolution limits of the gel) and the sequence was scanned for the *m*EndoG recognition sequence. Interestingly, one such site was found within the 200 nucleotides. Though not definitive at this point, this represents strong preliminary evidence that *m*EndoG specifically cleaves at the

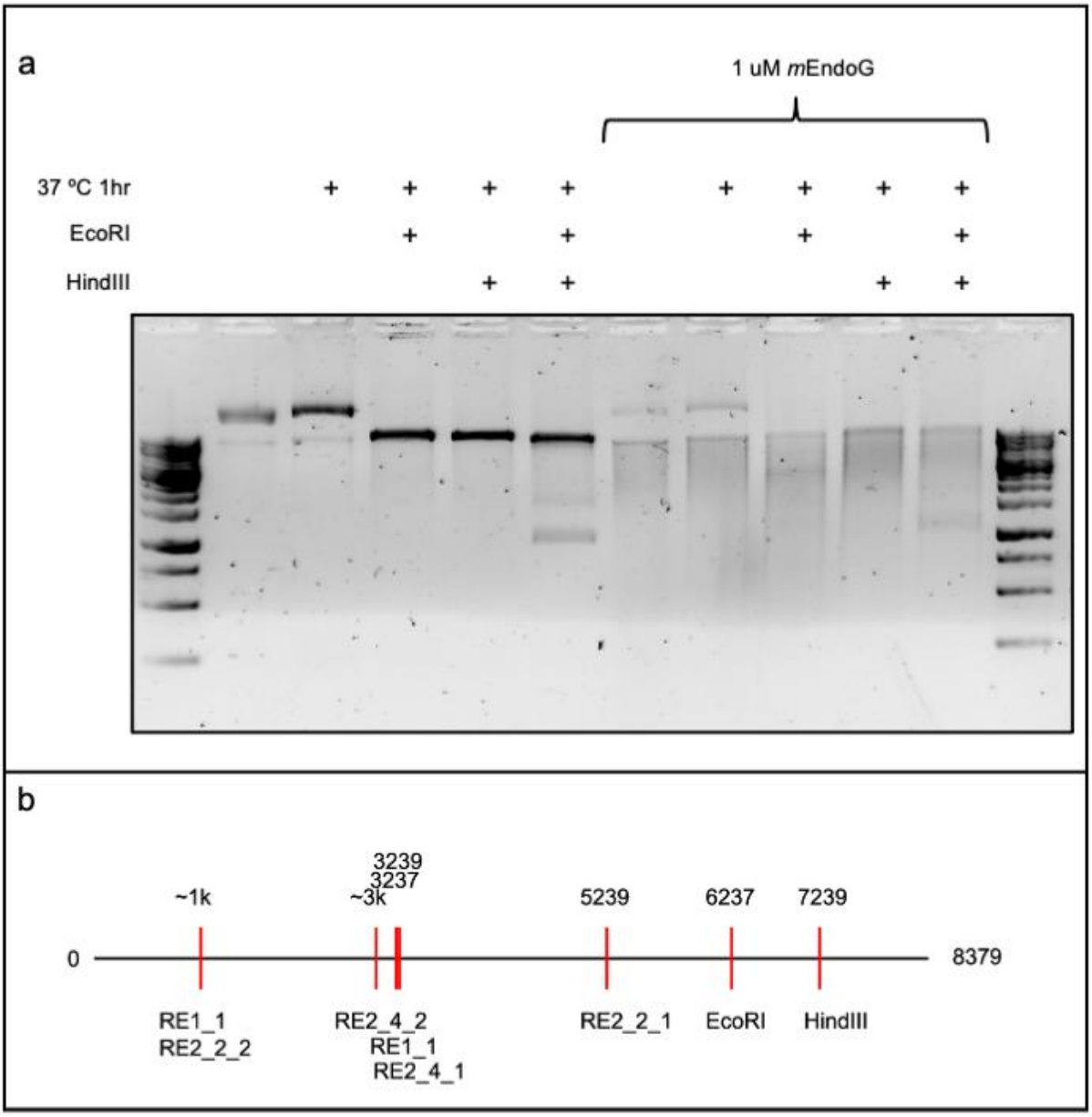

Figure A.8. pET Plasmid Digest assay to determine *m*EndoG cleavage location. Location mapped was performed on each of the cut sites with respect to EndoG, EcoRI, and HindIII. The locations were back calculated with respect to the known cut sites and the most common cut site was determined at about 3200 nucleotides in the sequence.

recognition site in plasmids as identified by Robertson et al.[3]. This is not to say that *m*EndoG can function as a restriction enzyme, as we have seen additional non-specific cleavage events occur throughout all of our studies, but does indicate that there is a strong preference for the recognition sequence.

A.9. EndoG Binding to Unmodified and Modified Junction

A.9.1. Introduction and Background

Throughout our studies, we tested the binding of our EndoG variants on both unmodified and modified junction. This became increasingly important as we discovered the differences in catalysis and we wanted to ensure that it truly was the activity of the enzyme that was changed and not a massive impact on the binding.

A.9.2. Experimental Methods

Fluorescence Polarization was carried out using a plate reader and the Cy5 label contained on duplex and Holliday Junction DNA constructs with and without the 5hmC modification. 15 uL of 50 nM DNA was mixed with 20 uL of a serial dilution of the protein diluted with the buffer in which the protein was purified. The 15 uL of DNA was loaded first and place in the plate reader to equilibrate at 37 °C for about 5 minutes. The protein was then added, placed in the plate reader, and allowed to sit for about 30 minutes. A Cy5-FP protocol was used to determine the change in signal over time. Points were then plotted and fit to a binding curve in KaleidaGraph.

A.9.3. Results

Overall, the difference in cleavage between *m*EndoG and CPS6 could be identified by the 2 to 6-fold difference seen in binding to junction and 5hmC junction respectively. Additionally, we saw a massive difference in binding for the proline mutants within the CPS6 B-site, further validating that P124 is responsible for the B-site repositioning to interaction with the arm of the

junction. Finally, the full B-site domain swap showed that 5hmC modified junction increased by 2-fold when simply swapped in, further indicating that the two sites communicate with one another.

a.

Mouse EndoG Proteins

	Holliday Junction (nM)	Holliday Junction + 5hmC (nM)
EndoG	57 ± 9	68 ± 6
EndoG_D111A	94 ± 8	41 ± 3
EndoG+102H105V_D111A	89 ± 17	48 ± 14

b.

C. elegans Proteins

	Holliday Junction (nM)	Holliday Junction + 5hmC (nM)
CPS6	112 ± 5	360 ± 60
CPS6+ EndoG Bsite	142 ± 17	155 ± 20
CPS6_P114E	850 ± 120	266 ± 32
CPS6_P119H	> 1700	> 1700
CPS6_E111A	59 ± 7	83 ± 24
CPS6+ EndoG Bsite_E111A	50 ± 21	111 ± 45

Figure A.9. Binding of Junction and 5hmC modified Junction to *m*EndoG and CPS6 and their respective mutants. a) Table for *m*EndoG and its mutant proteins to determine binding preference (if any) between junction and 5hmC modified junction. b) Contains the same type of data but pertains to CPS6 and its mutants.

Appendix B – Coding used in Zhunt

B.1. Python Script to run server for Zhunt

```
import pandas as pd
import numpy as np
import os
import csv

# updated on 200527 to have everything work
for documents in os.listdir("."):
    if documents.endswith(".fasta" or ".txt"):
        seq_col=[]
        working_doc=open(documents)
        for lines in working_doc:
            for nt in lines:
                if nt in ["A","T","G","C","M","a","t","g","c","m"]:
                    seq_col.append(nt+",")
        seq_col=np.vstack((seq_col))
        #print(seq_col)
    elif "Z-SCORE" in documents:
        zscore=open(documents)
        df=[]
        for lines in zscore:
            if "/" in lines:
                lines=""
            else:
                lines=lines.replace("     "," ")
                lines=lines.replace("    "," ")
                lines=lines.replace("  "," ")
                lines=lines.replace(" ",",")
                lines=lines.replace("\n","")
                lines=lines[1:]

                df.append(lines)
        df=np.asarray(df)
        df = np.array([l.split(',') for l in df])
        df=np.vstack((df[:,2]))

        if "Z-SCORE.txt" in documents:
            documents=documents.rstrip(".Z-SCORE.txt")
        else:
            documents=documents.rstrip(".Z-SCORE")

all_data=np.hstack((seq_col,df))
```

```
seq_dict={}

df = all_data

nts, zscore = map(list, zip(*df))


for i in range(len(nts)):
    if nts[i] not in seq_dict:
        seq_dict[nts[i]]=[float(zscore[i]),1]
    else:
        seq_dict[nts[i]][0] += float(zscore[i])
        seq_dict[nts[i]][1] += 1

output_file=open(documents+"_output.csv","w+")

writer=csv.writer(output_file)
writer.writerow(["nts","z-score sum","count"])
for key,value in seq_dict.items():
    writer.writerow([key.rstrip(","),value[0],value[1]])


    np.savetxt(documents+"_sequence&zscore.csv",np.array(all_data),fmt="%s")
```

B.2. R-script to generate summaries per nucleotide of the Zhunt data

```
library(ggplot2)
df <- read.csv("pBR322.fasta_output.csv")
nts <- df$nts
zscore <- df$z.score.sum
count <- df$count

nts=factor(nts)
count=as.numeric(count)

pdf("NTS_&_Sum_Zscore.pdf")
ggplot(df, aes(x = nts)) +
geom_col(aes(y = zscore), size = 1, color = "darkblue", fill = "white") +
geom_point(aes(y = 100*count), size = 1.5, color="red")+
scale_y_continuous(sec.axis=sec_axis(~./100,name="Counts"),name="Z-score")
```

B.3. R-script to present the Z-score across the sequence

```
df <- read.csv("pBR322.fasta_sequence&zscore.csv", header=FALSE)

pdf("Z-score_across_Sequence.pdf")
```

```
barplot(df$V2, ylab="Z-score",xlab="Sequence", main="Z-score Across Sequence")
```

B.4. R-script that looks at the differences in Z-hunt and Z-mhunt

```
library(ggplot2)
mainDir <- getwd()
subDir <- "pdfs"
dir.create(file.path(mainDir, subDir), showWarnings = FALSE)

df <- read.csv("pBR322.fasta_sequence&zscore.csv")
zscore <- df$zscore
methyl <- df$methyl
diff <- df$diff
nts <- df$nts
nts=factor(nts)

pdf("./pdfs/methyl_zscore.pdf")
barplot(methyl, main="Methylation Z-score")
dev.off()

pdf("./pdfs/zscore.pdf")
barplot(zscore, main="Z-score")
dev.off()

pdf("./pdfs/diff.pdf")
barplot(diff, main="Difference in Z-score")
dev.off()

pdf("./pdfs/combined_zscores.pdf")
ggplot(df, aes(x = nts)) +
geom_col(aes(y = methyl), size = 0.5, color = "darkblue", alpha=0.1, fill = "white") +
geom_col(aes(y = zscore), size = 0.5, color="red", alpha=0.5)+
scale_y_continuous(sec.axis=sec_axis(~.,name="Z-score"),name="Methylation Z-
score")
```

References

1. Zanden, C. M. V., Czarny, R. S., Ho, E. N., Robertson, A. B. & Ho, P. S. Structural adaptation of vertebrate endonuclease G for 5-hydroxymethylcytosine recognition and function. *Nucleic Acids Res.* **48**, 3962–3974 (2021).
2. Lin, J. L. J. *et al.* Structural insights into apoptotic DNA degradation by CED-3 protease suppressor-6 (CPS-6) from Caenorhabditis elegans. *J. Biol. Chem.* **287**, 7110–7120 (2012).
3. Robertson, A. B., Robertson, J., Fusser, M. & Klungland, A. Endonuclease G preferentially cleaves 5-hydroxymethylcytosine-modified DNA creating a substrate for recombination. *Nucleic Acids Res.* **42**, 13280–13293 (2014).
4. Temme, C. *et al.* The Drosophila melanogaster gene cg4930 encodes a high affinity inhibitor for endonuclease G. *J. Biol. Chem.* **284**, 8337–8348 (2009).
5. Loll, B., Gebhardt, M., Wahle, E. & Meinhart, A. Crystal structure of the EndoG/EndoGI complex: mechanism of EndoG inhibition. *Nucleic Acids Res.* **37**, 7312–7320 (2009).
6. Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J. & Stahl, F. W. The double-strand-break repair model for recombination. *Cell* **33**, 25–35 (1983).
7. Olorunniji, F. J. & Stark, W. M. The catalytic residues of Tn3 resolvase. *Nucleic Acids Res.* **37**, 7590–7602 (2009).