ABSTRACT

JACOBSEN, THOMAS. Developing Tools and Techniques for Expanding the Synthetic Biology Toolkit (Under the direction of Dr. Chase L. Beisel and Dr. Gregory T. Reeves).

The field of synthetic biology has revolutionized biological research by making biology more predictable to engineer. By combining biology with fundamental engineering principles, synthetic biology has provided a platform to solve various challenges faced in society today. Synthetic biology has advanced due to the availability of gene regulatory tools. While these tools have been developed for eukaryotic systems, their development has been outpaced compared to unicellular systems. This thesis helps expand upon the tools and techniques available for synthetic biology, with a focus on eukaryotic systems.

First, a set of self-cleaving ribozymes, tools previously used to silence gene expression, was engineered to modulate gene expression. These ribozymes were flanked with upstream competing sequences designed to alter the ribozyme's secondary structure, thus potentially resulting in complete or partial inability of the ribozyme to self-cleave. As a proof-of-concept, these tools were implemented in two eukaryotic systems: HEK293T cells and *Drosophila* embryos. In both systems, these ribozymes modulated gene expression, with each competing sequence associated with varying extents of gene repression. Additionally, we correlated the empirical data with RNA folding algorithms and observed a lack of correlation. While this tool is currently not predictive, a set of ribozymes to modulate gene expression in various model systems are now available.

Next, we developed a technique to determine the PAMs of clustered regularly interspaced short palindromic repeats (CRISPR)-CRISPR associated (Cas) nucleases. Based on a cell-free transcription-translation (TXTL) system, this screen was shown to rapidly decipher protospacer-adjacent motifs (PAMs), sequences essential for the binding of Cas nucleases to target sequences of interest. As a proof-of-concept, the PAM of *Neisseria meningitidis* Cas9 was determined. While prior assays have been developed to determine the PAM sequences of Cas nucleases, this technique decreases the turnaround time for PAM determination, allowing the rapid expansion of the CRISPR toolkit.

Using this TXTL PAM screen, we discovered unreported non-canonical PAMs of the previously characterized *Acidaminococcus sp.* Cas12a (AsCas12a, also known as AsCpf1). While AsCas12a is associated with the canonical TTTV (V = A/C/G) PAM, this nuclease also recognizes CTTV, TCTV, and TTCV as non-canonical PAMs. The PAM screen revealed AsCas12a's ability to recognize GTTV and GCTV as non-canonical PAMs, while plasmid clearance in *E. coli*, DNA cleavage in TXTL, and indel formation in mammalian cells validated recognition of these PAMs. While this finding increases the targeting range of AsCas12a, it also increases the number of potential off-target sites within a genome.

In a separate study, the TXTL PAM screen was used to characterize the PAMs of previously unreported Cas12a nucleases, with Fn3Cas12a and PiCas12a sharing high sequence homology with the previously characterized FnCas12a and PdCas12a, respectively. While the characterized nucleases recognized T-rich PAMs commonly associated with Cas12a nucleases, PiCas12a recognized G-rich motifs. Mutating residues within PiCas12a transitioning PiCas12a to PdCas12a altered its PAM profile, thus increasing the targeting range of PiCas12a. These findings were validated using DNA cleavage in TXTL and indel formation in mammalian cells. Due to sequence homology between PiCas12a and PdCas12a, as well as the PAM flexibility of PiCas12a and its variants, PiCas12a provides a platform for CRISPR evolutionary studies and expands upon the CRISPR toolkit.

Overall, this thesis helps expand upon the synthetic biology toolkit. The ribozyme constructs have the potential to be applied in various biological studies, such as the tuning of gene networks. Our PAM screen reduces the turnaround time from Cas nuclease discovery to its application in biotechnologies, while the characterized nucleases increases the targeting space of Cas12a. Together, the tools and techniques developed from this thesis adds to the current synthetic biology toolkit and may offer insight into the evolution of CRISPR-Cas systems.

Developing Tools and Techniques for Expanding the Synthetic Biology Toolkit

by
Thomas Jacobsen

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Chemical and Biomolecular Engineering

Raleigh, North Carolina

2019

APPROVED BY:

Dr. Balaji Rao	Dr. Jeffrey Yoder
Dr. Chase Beisel Co-Committee Chair	Dr. Gregory Reeves Co-Committee Chair

DEDICATION

To my family, close friends, and mentors for their constant love and support.

BIOGRAPHY

Thomas Jacobsen was born in Hamamatsu, Japan in 1992. In 1997, he moved to Syracuse, NY with his younger sister where they were adopted by their aunt and uncle. He attended kindergarten and junior high school in the Cicero-North Syracuse school district. In 2006, his family moved to Watertown, NY and graduated from Watertown High School in 2010. Thomas chose to attend Clarkson University in Potsdam, NY where he majored in his chemical and biomolecular engineering. During his tenure, he worked in the labs of Selma Mededovic-Thagard and Paul Goulet at Clarkson University and Ernest Blatchley at Purdue University. After receiving his bachelor's, Thomas chose to pursue a PhD in chemical and biomolecular engineering at North Carolina State University under the direction of Dr. Chase Beisel and Dr. Gregory Reeves. After 5 years of research, he will (hopefully) have a PhD in Chemical and Biomolecular Engineering.

ACKNOWLEDGEMENTS

First, I would like to thank my PhD advisors, Dr. Chase Beisel and Dr. Greg Reeves, for all of their support throughout my graduate career. Their patience and encouragement throughout the years have been invaluable during my time here.

Also, I thank two former graduate students, Dr. Ashley Jermusyk and Dr. Michelle Luo, that I had the honor of learning under during the early years of my graduate career. Thank you for helping me acquire the skills and techniques I needed to become successful.

Moreover, I thank all of my colleagues, collaborators, and fellow Beisel and Reeves lab members for their technical guidance and expertise. I would especially like to thank Dr. Colin Maxwell, Dr. Chunyu Liao, Hadel Al Asafen, Gloria Yi, and collaborators from Benson Hill Biosystems for their significant contributions to my projects.

Furthermore, I would like to thank my family and close friends. They have loved me, supported me, and have always lent an ear whenever I needed it.

Finally, I would like to thank my fiancé Kimberley Knott for her endless love and support she has given me during my time here.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF DOCUMENTS

CHAPTER 1: An overview of gene regulatory tools in *Drosophila*

1.1 - INTRODUCTION

Synthetic biology aims to make biology more predictable and easier to understand by conceptualizing biological systems as a network comprised of interactions between various genetic components. Synthetic biologists modify, rewire, and/or engineer natural biological molecules and systems to produce novel products or components that have the potential to significantly impact various fields. Advancements in synthetic biology has been possible due to the wide array of genetic tools that are capable of regulating gene expression at various stages of its expression [1–4]. Subsequently, the development of these tools has led to the construction of various synthetic networks that were built upon fundamental engineering principles [5–7]. While a plethora of gene regulatory tools have been constructed in unicellular systems (i.e. bacteria and yeast), construction of these tools in more complex systems, such insect systems, have lagged behind.

In this chapter, the current technologies available to regulate gene expression in the model system *Drosophila* are discussed. This chapter serves as an addition to the introduction in **Chapter 2**, as the chapter lacked a discussion about gene regulatory tools in *Drosophila* although data associated with *Drosophila* were presented. These technologies include promoter swapping, orthogonal transcription factors, tetracycline-inducible systems, RNA interference (RNAi), and Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) technologies. Also described are potential genetic tools that have been used in other eukaryotic model systems and a brief discussion on their potential as gene regulatory tools in *Drosophila*. Finally, an outline of the following chapters in this thesis is provided.

1.2 - GENE REGULATORY TOOLS DEVELOPED IN *DROSOPHILA*

1.1.1 - Promoter swapping

A simple technique that has been used to regulate gene expression is the swapping of promoters with varying levels of gene expression activation. As expression levels across different promoters can vary widely, this method allows for increasing or decreasing levels of specific genes of interest [8]. This simply involves the replacement of promoters with

other synthetic promoters available in the synthetic biology toolkit or endogenous promoters (Figure 1.1A). While swapping promoters offers the user the ability to either increase or decrease gene expression, it does not allow for precise control of expression or the ability to further reduce or amplify expression levels beyond the available promoters in *Drosophila*.

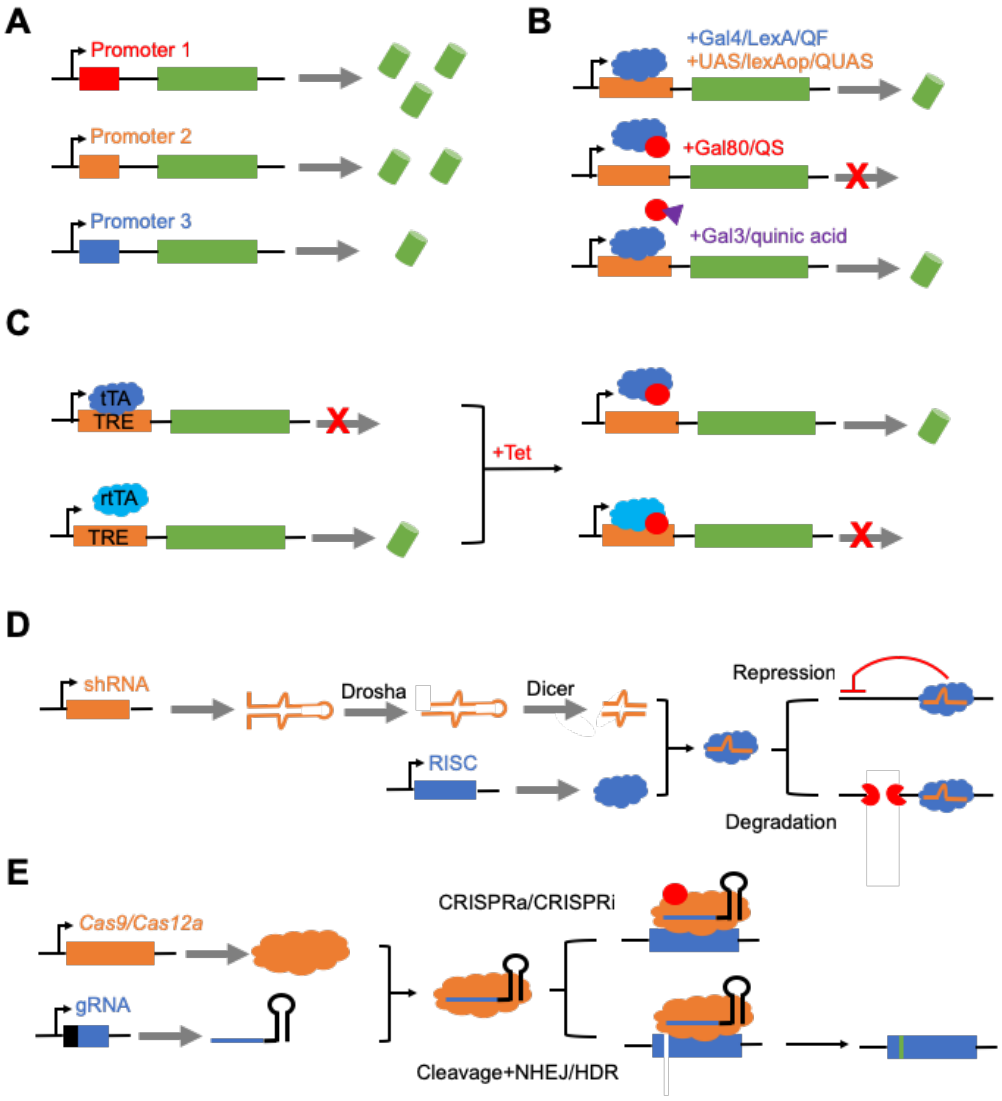


Figure 1.1: The current toolbox of gene regulatory tools adapted in *Drosophila*. (A) Various promoters (red/orange/blue) can express a gene (green) at varying levels. (B) Orthogonal transcription factors (blue) bind to consensus sequences (orange) to activate gene expression (green). These systems can be more flexible by expressing a repressor (red) to prevent gene expression, while a second repressor (purple) that represses the repressor can be introduced. (C) Tetracycline response elements (TREs) can be added upstream of a gene or interest (green). The tetracycline-dependent activator (tTA) or (reverse tTA (rtTA)) can bind (or not bind) to the TRE and inhibit (or activate) gene expression. Introduction of tetracycline will prevent (or allow) binding of tTA (or rtTA) to TRE sites, allowing (or preventing) gene expression. (D) The RNAi system involves the expression of a short RNA (orange) which is processed by proteins Drosha and Dicer, and then forms a ribonucleoprotein (RNP) complex with RISC (blue). This RNP then targets RNA transcripts to repress of cleave/degrade the transcript. (E) Cas9 or Cas12a (orange) and their gRNA (blue) can be expressed to form an RNP that has the ability to target DNA. If in its active form, Cas9/Cas12a can cleave DNA and form indels (green), which partially or completely inhibits gene expression. However, dCas9/dCas12a can be tagged with activating or inactivating domains (red) to increase or decrease gene expression.

1.1.2 - Orthogonal transcription factors

The Gal4-upstream activating sequence (UAS) system has been the primary method of regulating gene expression in *Drosophila*. Since its discovery as a viable genetic tool in *Drosophila*, thousands of fly lines have been developed containing this system (see https://bdsc.indiana.edu/stocks/uas/uas_nonrnai.html). The Gal4-UAS system involves the yeast Gal4 transcriptional activator binding to a consensus UAS site, which then allows for strong expression of the downstream gene of interest (Figure 1.1B). [9–11]. To offer more flexibility to this system, repressor Gal80 can be expressed to directly bind to Gal4 and prevent transcriptional activation [12–14], and Gal80 can be repressed by expression of Gal3 [15].

Similar to the Gal4-UAS system, other orthogonal transcription factors have been adapted in *Drosophila*. One of these include the LexA-lexA operon (lexAop) system [16]. Similar to the Gal4-UAS system, LexA binds specifically to consensus lexAop sites. However, unlike Gal4, LexA does not have the ability to drive gene expression after binding. Thus, to use this system for activating gene expression, the activation domain of Gal4 or VP16 must be tagged to LexA [17].

The most recent orthogonal transcription factor adapted in *Drosophila* is the QF-QUAS (Q) system [18]. Similar to Gal4-UAS and LexA-lexAop, this system involves the transcriptional activator QF binding to a consensus QUAS sequence, which then drives gene expression. Similar to Gal80 in the Gal4-UAS system, QF can be repressed by expressing its inhibitor, QS, while QS can be inhibited through supplementing the flies' food source with quinic acid [18]. Compared to the LexA-lexAop system, the Q system has the ability to drive gene expression without tagging the transcription factor with activating domains. Also, compared to the Gal4-UAS system, gene expression driven from the QF-QUAS system has shown to be less leaky [18].

Compared to swapping promoters, these systems allowed for greater control of gene expression levels as increasing or decreasing their respective binding sites can tune gene expression. Also, combining two of these systems can result in the ability to control gene

expression of two different transgenes (reviewed in [19]), which is useful for building synthetic gene circuits. While these systems allow for more control of gene expression, they require the addition of multiple binding sites and the creation of new fly lines when higher reduction levels of gene expression are needed.

1.1.3 - Tetracycline-inducible systems

The tetracycline-dependent transactivator (tTA) is a binary system that is able to regulate gene expression in various model organisms [20–25]. This system is composed of the bacterial tetracycline repressor (tetR) and the tetracycline response element (TRE) that it binds to (Figure 1.1C) [20]. The tagging of tetR with the VP16 transactivation domain can activate expression of a gene of interest in the absence of tetracycline, while the introduction of the drug inhibits gene expression [26]. To increase the flexibility of this system, random mutagenesis was performed to discover a reverse tetracycline-dependent transactivator (rtTA) that activates gene expression in the presence of tetracycline [25,27].

The tetracycline-inducible systems have allowed for the regulation of gene expression in *Drosophila* and had been especially useful for temporal-controlled gene expression that was lacking with the orthogonal transcription factor systems until the creation of a temperature-dependent expression of Gal4 [28]. Though tetracycline-inducible systems have allowed for gene regulation in *Drosophila*, it requires a drug for transactivation, thus introducing another variable for its use. Also, this tool cannot immediately be used for studies involving the early developmental stages due to the requirement of tetracycline consumption.

1.1.4 - RNAi

RNAi is a system that inhibits gene expression using double-stranded RNA to repress or cleave mRNA transcripts to prevent translation (Figure 1.1D) (reviewed in [29]). This process has been used to perform genetic screens across various model systems, including *Drosophila* [30–33]. Since then, many groups have harnessed this system as a genetic tool to target endogenous mRNA transcripts [34–36]. The first method of

introduction was through injection embryonic stages [37–39], but since has become more flexible when combined with the Gal4-UAS system for RNAi knockdown [40,41].

RNAi has been successfully used to regulate gene expression in *Drosoph*ila [42]. This system is particularly useful when studying gene function in early development of *Drosophila*, especially when Gal4-UAS lines are not available for specific genes of interest. While RNAi can be used as a tool to regulate gene expression, it has its limitations. First, the potential of off-target effects is an issue, especially when performing genetic screens. Also, as RNAi naturally occurs in *Drosophila*, introducing synthetic RNAi components to this model system may alter the endogenous biological machinery.

1.1.5 - CRISPR-Cas9/12a

A background of CRISPR-Cas systems can be seen in **Chapters 3, 4, and 5**. Briefly, CRISPR-Cas systems are adaptive immune systems that protect prokaryotes against mobile genetic elements [43–46]. In *Drosophila*, these systems have been harnessed as a genetic tool to produce transgenic fly lines containing sequence-specific mutations, insertions, or deletions [47,48]. This involves the expression of Cas9 and a guide RNA (gRNA) that serves to guide Cas9 to a sequence-specific site in the genome specified by the gRNA [45,46,49]. While Cas9 has been mainly used for site-directed mutagenesis in *Drosophila*, CRISPR interference (CRISPRi) and activation (CRISPRa) has been shown to regulate gene expression using the KRAB and VPR domains, respectively (Figure 1.1E) [50–53].

More recently, another Cas nuclease, Cas12a (also known as Cpf1), has been discovered [54] and used for genome-editing in *Drosophila* [55]. Compared to Cas9, Cas12a may be a better option for regulating gene expression due to its slightly smaller in size, its cleavage products forming staggered ends (compared to blunt ends produced by Cas9), and its ability to process its own CRISPR RNA (crRNA) [54]. The latter could potentially be advantageous for multiplexing crRNA and targeting multiple genes using a single CRISPR array. It was previously shown that expression of Cas12a in *Drosophila* was weak compared to Cas9 [55]. However, further work showed that the specific Cas12a

used in [55] was temperature-sensitive, and using a different variant of Cas12a resulted in greater cleavage activity [56].

The use of CRISPR technologies in *Drosophila* has allowed for simple and rapid reduction of gene expression through DNA cleavage and repair, resulting in partial or complete loss of gene function. Along with its ability to cleave DNA, CRISPR-Cas9 has shown the ability to form a synthetic transcription factor that can target virtually any locus, thus increasing or decreasing gene expression without altering the genome [50–53]. While this technology has revolutionized genome-editing and gene regulation in various model systems, it may be associated with off-target effects and has shown to reduce fitness of *Drosophila* [57].

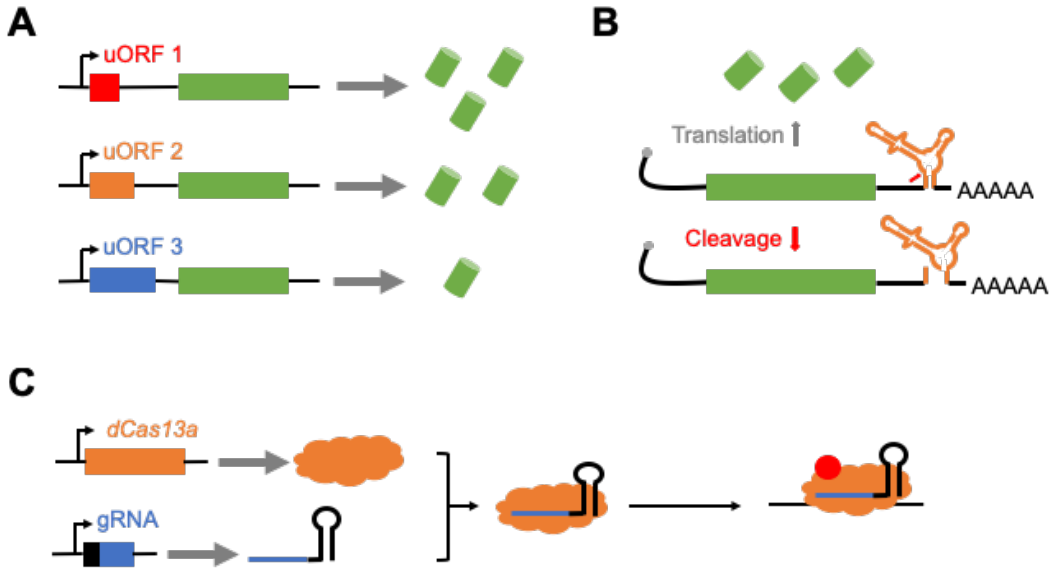


Figure 1.2: Potential tools to regulate gene expression in *Drosophila*. (A) Synthetic uORFs (red/orange/blue) can be placed in the 5'UTR of coding genes (green) to control its expression. The strength of each uORF is dependent on their sequence content and length and can be tuned to vary the reduction level of gene expression. (B) Self-cleaving ribozymes (orange) can be placed in various loci of a gene of interest (green), including within introns, 5'UTR, or 3'UTR (depicted). Due to their self-catalytic properties, these ribozymes will cleave the RNA transcript, which leads to degradation and inhibition of gene expression. (C) dCas13a (orange) can be expressed with its gRNA (blue) to form an RNP that targets an mRNA transcript of interest. An activating or inactivating domain (red) can be tagged to dCas13a to increase or decrease gene expression.

1.3 - POTENTIAL TOOLS TO REGULATE GENE EXPRESSION IN *DROSOPHILA*

1.2.1 - Short upstream open reading frames (uORFs)

Short uORFS are short sequences containing a start and termination codon (Figure 1.2A) that are occasionally found upstream of coding genes in various eukaryotic systems, including yeast, mice, and humans [58,59]. Studies have shown that synthetic uORFs can be used to regulate gene expression in model systems, and that the strength of reduction of gene expression was based on sequence length and content of the uORF [60,61]. These uORFs have also been discovered in *Drosophila* [62,63], and has been shown to regulate gene expression [64,65]. While uORFs seem promising as gene regulatory tools, preliminary work from a former graduate student showed results indicating that these tools may have weak regulatory ability in *Drosophila*.

1.2.2 - Self-cleaving ribozymes

Self-cleaving ribozymes are RNA sequences that contain self-enzymatic activity (reviewed in [66]). If placed in the untranslated regions or an intron of an RNA transcript, previous work has shown that this tool can prevent gene expression through self-cleavage and subsequent degradation (Figure 1.2B) [67–69]. While self-cleaving ribozymes have been used to regulate gene expression in bacteria, yeast, and mammalian cells [67–69], it has yet to be used in *Drosophila* as a tool to regulate gene expression. Work in **Chapter 2** of this thesis provide evidence of self-cleaving ribozymes as viable tool to regulate gene expression in mammalian cells and *Drosophila*, and potentially in other model systems.

1.2.3 - CRISPR-Cas13a

Recently, the CRISPR-Cas13a system was discovered [70,71]. Compared to Cas9 and Cas12a, Cas13a has the ability to bind and cleave RNA instead of DNA (Figure 1.2C). Similar to Cas9 and Cas12a, the catalytic domain of Cas13a was revealed by identifying conserved catalytic residues responsible for Cas13a's nuclease activity thus allowing for the generation of a catalytically-inactive version of Cas13a (dCas13a)

Using dCas13a in *Drosophila* and other multicellular systems could potentially lead to a genetic tool that regulates gene expression. While Cas13a has shown to cleave non-specific RNA in bacteria [70,71], mutating specific domains has inhibited this effect. However, when this nuclease was implemented in mammalian cells, non-specific RNA cleavage was not observed and showed enhanced RNA targeting specificity compared to RNAi [72]. Thus, this nuclease has the potential to be used as a tool for RNA cleavage a synthetic RNA-binding protein.

1.4 - CONCLUSIONS

This thesis helps expand the current toolbox of synthetic biology by the development of novel tools and techniques. To introduce a set of tools that can be used to regulate gene expression in mammalian and insect systems, a set of self-cleaving ribozymes was engineered with tunable upstream competing sequences, which is presented in **Chapter 2**. In **Chapter 3**, a novel technique to determine protospacer-adjacent motif (PAM) sequences of Cas nucleases was developed. Finally, in **Chapters 4 and 5**, the technique developed in **Chapter 3** was used to reveal previously unreported non-canonical PAMs of the *Acidaminococcus sp.* Cas12a nuclease and to characterize four previously unreported Cas12a nucleases, respectively.

1.5 - REFERENCES

[1] L. Guzman, D. Belin, M.J. Carson, J. Beckwith, Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter, J Biotechnol. 177 (1995) 4121–4130.

[2] L. Yen, J. Svendsen, J.-S. Lee, J.T. Gray, M. Magnier, T. Baba, R.J. D'Amato, R.C. Mulligan, Exogenous control of mammalian gene expression through modulation of RNA self-cleavage, Nature. 431 (2004) 471–476. doi:10.1038/nature02844.

[3] H.M. Salis, E.A. Mirsky, C.A. Voigt, Automated design of synthetic ribosome binding sites to control protein expression, Nat. Biotechnol. 27 (2009) 946–950.

[4] K.E. McGinness, T.A. Baker, R.T. Sauer, Engineering controllable protein degradation, Mol. Cell. 22 (2006) 701–707.

[5] T.S. Gardner, C.R. Cantor, J.J. Collins, Construction of a genetic toggle switch in Escherichia coli, Nature. 403 (2000) 339–342.

[6] M.B. Elowitz, S. Leibler, A synthetic oscillatory network of transcriptional regulators, Nature. 403 (2000) 335–338.

[7] C.C. Guet, M.B. Elowitz, W. Hsing, S. Leibler, Combinatorial synthesis of genetic networks, Science. 296 (2002) 1466–1470.

[8] J.Y. Qin, L. Zhang, K.L. Clift, I. Hulur, A.P. Xiang, B.Z. Ren, B.T. Lahn, Systematic comparison of constitutive promoters and the doxycycline-inducible promoter, PLoS One. 5 (2010) 3–6.

[9] D.A. Elliott, A.H. Brand, The gal4 system : a versatile system for the expression of genes, Methods Mol Biol. 420 (2008) 79–95.

[10] A.H. Brand, N. Perrimon, Targeted gene expression as a means of altering cell fates and generating dominant phenotypes, Development. 118 (1993) 401–415.

[11] S.E. McGuire, G. Roman, R.L. Davis, Gene expression systems in drosophila: a synthesis of time and space, Trends Genet. 20 (2004) 384–391.

[12] J. Ma, M. Ptashne, The carboxy-terminal 30 amino acids of GAL4 are recognized by GAL80, Cell. 50 (1987) 137–142.

[13] M. Carrozza, S. John, A. Sil, J. Hopper, J. Workman, Gal80 confers specificity on HAT complex interactions with activators, J. Biol. Chem. 277 (2002) 24648–24652.

[14] F. Janody, R. Sturny, V. Schaeffer, Y. Azou, N. Dostatni, Two distinct domains of

Bicoid mediate its transcriptional downregulation by the Torso pathway, Development. 128 (2001) 2281–2290.

[15] O. Egriboz, S. Goswami, X. Tao, K. Dotts, C. Schaeffer, V. Pilauri, J.E. Hopper, Self-association of the Gal4 inhibitor protein Gal80 is impaired by Gal3: Evidence for a new mechanism in the GAL gene switch, Mol. Cell. Biol. 33 (2013) 3667–3674.

[16] S.L. Lai, T. Lee, Genetic mosaic with dual binary transcriptional systems in Drosophila, Nat. Neurosci. 9 (2006) 703–709.

[17] S.J. Triezenberg, R.C. Kingsbury, S.L. Mcknight, Functional dissection of VP16, the trans-activator of herpes simplex virus immediate early gene expression, Genes Dev. 2 (1988) 718–729.

[18] C.J. Potter, B. Tasic, E. V. Russler, L. Liang, L. Luo, The Q system: A repressible binary system for transgene expression, lineage tracing, and mosaic analysis, Cell. 141 (2010) 536–548.

[19] A. Del Valle Rodríguez, D. Didiano, C. Desplan, Power tools for gene expression and clonal analysis in Drosophila, Nat. Methods. 9 (2012) 47–55.

[20] M. Gossen, H. Bujard, Tight control of gene expression in mammalian cells by tetracycline-responsive promoters, Proc Natl Acad Sci. 89 (1992) 5547–5551.

[21] P. Weinmann, M. Gossen, W. Hillen, H. Bujard, C. Gatz, A chimeric transactivator allows tetracycline-responsive gene expression in whole plants, Plant J. 5 (1994) 559–569.

[22] L. Hennighausen, R.J. Wall, U. Tillmann, M. Li, P.A. Furth, Conditional gene expression in secretory tissues and skin of transgenic mice using the MMTV-LTR and the tetracycline responsive system, J. Cell. Biochem. 59 (1995) 463–472.

[23] A. Kistnert, M. Gossentt, F. Zimmermannt, J. Jerecict, C. Ullmer, H. Lubbert, H. Bujardt1, Doxycycline-mediated quantitative and tissue-specific control of gene expression in transgenic mice, Proc Natl Acad Sci U S A. 93 (1996) 10933–10938.

[24] B. Bello, D. Resendez-Perez, W.J. Gehring, Spatial and temporal targeting of gene expression in drosophila by means of a tetracycline-dependent transactivator system, Development. 125 (1998) 2193–2202.

[25] M.J. Stebbins, S. Urlinger, G. Byrne, B. Bello, W. Hillen, J.C. Yin, Tetracycline-inducible systems for drosophila, Proc Natl Acad Sci U S A. 98 (2001) 10775–

10780.

[26] W. Hillen, A. Wissmann, Tet repressor-tet operator interaction, in: W. Saenger, U. Heinemann (Eds.), Protein-Nucleic Acid Interact., Macmillan Education UK, London, 1989: pp. 143–162.

[27] M. Gossen, S. Freundlieb, G. Bender, G. Müller, W. Hillen, H. Bujard, Transcriptional activation by tetracyclines in mammalian cells, Science. 268 (1995) 1766–1769.

[28] C. Grabher, J. Wittbrodt, Efficient activation of gene expression using a heat-shock inducible Gal4/Vp16-UAS system in medaka, BMC Biotechnol. 4 (2004) 1–6.

[29] R.C. Wilson, J.A. Doudna, Molecular mechanisms of RNA interference, Annu. Rev. Biophys. 42 (2013) 217–239.

[30] L.H. Hartwell, R.K. Mortimer, J. Culotti, M. Culotti, Genetic control of the cell division cycle in yeast: V. Genetic analysis of cdc mutants, Genetics. 74 (1973) 267–286.

[31] H.R. Horvitz, Genetic control of programmed cell death in the nematode Caenorhabditis elegans, Cancer Res. 59 (1999) 1701s–1706s.

[32] C. Nüsslein-Volhard, E. Wieschaus, Mutations affecting segment number and polarity in Drosophila, Nature. 287 (1980) 795–801.

[33] D. St Johnston, C. Nüsslein-Volhard, The origin of pattern and polarity in the Drosophila embryo, Cell. 68 (1992) 201–219.

[34] S. Grimm, The art and design of genetic screens: Mammalian culture cells, Nat. Rev. Genet. 5 (2004) 179–189.

[35] R. Bernards, T.R. Brummelkamp, R.L. Beijersbergen, shRNA libraries and their use in cancer genetics, Nat. Methods. 3 (2006) 701–706.

[36] M. Boutros, J. Ahringer, The art and design of genetic screens: RNA interference, Nat. Rev. Genet. 9 (2008) 554–566.

[37] J.R. Kennerdell, R.W. Carthew, Use of dsRNA-mediated genetic interference to demonstrate that frizzled and frizzled 2 act in the wingless pathway, Cell. 95 (1998) 1017–1026.

[38] L. Misquitta, B.M. Paterson, Targeted disruption of gene function in Drosophila by RNA interference (RNA-i): A role for nautilus in embryonic somatic muscle formation, Proc Natl Acad Sci. 96 (1999) 1451 –1456.

[39] R.W. Williams, G.M. Rubin, ARGONAUTE1 is required for efficient RNA interference in Drosophila embryos, Proc Natl Acad Sci. 99 (2002) 6889–6894.

[40] G. Dietzl, D. Chen, F. Schnorrer, K.C. Su, Y. Barinova, M. Fellner, B. Gasser, K. Kinsey, S. Oppel, S. Scheiblauer, A. Couto, V. Marra, K. Keleman, B.J. Dickson, A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila, Nature. 448 (2007) 151–156.

[41] J.Q. Ni, M. Markstein, R. Binari, B. Pfeiffer, L.P. Liu, C. Villalta, M. Booker, L. Perkins, N. Perrimon, Vector and parameters for targeted transgenic RNA interference in Drosophila melanogaster, Nat. Methods. 5 (2008) 49–51.

[42] J.Q. Ni, R. Zhou, B. Czech, L.P. Liu, L. Holderbaum, D. Yang-Zhou, H.S. Shim, R. Tao, D. Handler, P. Karpowicz, R. Binari, M. Booker, J. Brennecke, L.A. Perkins, G.J. Hannon, N. Perrimon, A genome-scale shRNA resource for transgenic RNAi in Drosophila, Nat. Methods. 8 (2011) 405–407.

[43] R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. Romero, P. Horvath, CRISPR provides acquired resistance against viruses in prokaryotes, Science. 315 (2007) 1709–1712.

[44] S.J. Brouns, M.M. Jore, M. Lundgren, E.R. Westra, R.J. Slijkhuis, A.P. Snijders, M.J. Dickman, K.S. Makarova, E.V. Koonin, J. van der Oost, Small CRISPR RNAs guide antiviral defense in prokaryotes, Science. 321 (2008) 960–964.

[45] C.R. Hale, P. Zhao, S. Olson, M.O. Duff, B.R. Graveley, L. Wells, R.M. Terns, M.P. Terns, RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex, Cell. 139 (2009) 945–956.

[46] J.E. Garneau, M.È. Dupuis, M. Villion, D.A. Romero, R. Barrangou, P. Boyaval, C. Fremaux, P. Horvath, A.H. Magadán, S. Moineau, The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA, Nature. 468 (2010) 67–71.

[47] A.R. Bassett, C. Tibbit, C.P. Ponting, J.L. Liu, Highly Efficient Targeted Mutagenesis of Drosophila with the CRISPR/Cas9 System, Cell Rep. (2013).

[48] S.J. Gratz, A.M. Cummings, J.N. Nguyen, D.C. Hamm, L.K. Donohue, M.M. Harrison, J. Wildonger, K.M. O'connor-Giles, Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease, Genetics. (2013).

[49] E.R. Westra, P.B.G. van Erp, T. Künne, S.P. Wong, R.H.J. Staals, C.L.C. Seegers, S. Bollen, M.M. Jore, E. Semenova, K. Severinov, W.M. de Vos, R.T. Dame, R. de Vries, S.J.J. Brouns, J. van der Oost, CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3, Mol. Cell. 46 (2012) 595–605.

[50] L.A. Gilbert, M.H. Larson, L. Morsut, Z. Liu, G.A. Brar, S.E. Torres, N. Stern-Ginossar, O. Brandman, E.H. Whitehead, J.A. Doudna, W.A. Lim, J.S. Weissman, L.S. Qi, CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes, Cell. 154 (2013) 442–451.

[51] M.L. Maeder, S.J. Linder, V.M. Cascio, Y. Fu, Q.H. Ho, J.K. Joung, CRISPR RNA-guided activation of endogenous human genes, Nat. Methods. 10 (2013) 977–979.

[52] L.S. Qi, M.H. Larson, L.A. Gilbert, J.A. Doudna, J.S. Weissman, A.P. Arkin, W.A. Lim, Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression, Cell. 152 (2013) 1173–1183.

[53] S. Konermann, M.D. Brigham, A.E. Trevino, J. Joung, O.O. Abudayyeh, C. Barcena, P.D. Hsu, N. Habib, J.S. Gootenberg, H. Nishimasu, O. Nureki, F. Zhang, Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex, Nature. 517 (2015) 583–588.

[54] B. Zetsche, J.S. Gootenberg, O.O. Abudayyeh, A. Regev, E. V Koonin, F. Zhang, I.M. Slaymaker, K.S. Makarova, P. Essletzbichler, S.E. Volz, J. Joung, J. Van Der Oost, Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system, Cell. 163 (2015) 759–771.

[55] F. Port, S.L. Bullock, Augmenting CRISPR applications in Drosophila with tRNA-flanked sgRNAs, Nat. Methods. 13 (2016) 852–854.

[56] M.A. Moreno-Mateos, J.P. Fernandez, R. Rouet, C.E. Vejnar, M.A. Lane, E. Mis, M.K. Khokha, J.A. Doudna, A.J. Giraldez, CRISPR-Cpf1 mediates efficient homology-directed repair and temperature-controlled genome editing, Nat. Commun. 8 (2017) 1–9.

[57] F. Port, H.-M. Chen, T. Lee, S.L. Bullock, Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in Drosophila, Proc. Natl. Acad. Sci. 111 (2014) E2967–E2976.

[58] C. Lawless, R.D. Pearson, J.N. Selley, J.B. Smirnova, C.M. Grant, M.P. Ashe, G.D. Pavitt, S.J. Hubbard, Upstream sequence elements direct post-transcriptional regulation of gene expression under stress conditions in yeast, BMC Genomics. 10 (2009) 1–20.

[59] S.E. Calvo, D.J. Pagliarini, V.K. Mootha, Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans, Proc Natl Acad Sci. 106 (2009) 7507–7512.

[60] P.P. Mueller, B.M. Jackson, P.F. Miller, A.G. Hinnebusch, The first and fourth upstream open reading frames in GCN4 mRNA have similar initiation efficiencies but respond differently in translational control to change in length and sequence., Mol. Cell. Biol. 8 (1988) 5439–5447.

[61] J.P. Ferreira, K.W. Overton, C.L. Wang, Tuning gene expression with synthetic upstream open reading frames, Proc. Natl. Acad. Sci. 110 (2013) 11284–11289.

[62] C.A. Hayden, G. Bosco, Comparative genomic analysis of novel conserved peptide upstream open reading frames in Drosophila melanogaster and other dipteran species, BMC Genomics. 9 (2008) 1–14.

[63] E. Ladoukakis, V. Pereira, E.G. Magny, A. Eyre-Walker, J.P. Couso, Hundreds of putatively functional small open reading frames in Drosophila, Genome Biol. 12 (2011) R118.

[64] J. Medenbach, M. Seiler, M.W. Hentze, Translational control via protein-regulated upstream open reading frames, Cell. 145 (2011) 902–913.

[65] S. Schleich, K. Strassburger, P.C. Janiesch, T. Koledachkina, K.K. Miller, K. Haneke, Y.-S. Cheng, K. Küchler, G. Stoecklin, K.E. Duncan, A.A. Teleman, DENR–MCT-1 promotes translation re-initiation downstream of uORFs to control tissue growth, Nature. 512 (2014) 208–212.

[66] A.R. Ferré-D'Amaré, W.G. Scott, Small Self-cleaving Ribozymes, Cold Spring Harb. Perspect. Biol. 2 (2010) a003574.

[67] K.H. Link, L. Guo, T.D. Ames, L. Yen, R.C. Mulligan, R.R. Breaker, Engineering high-speed allosteric hammerhead, Biol. Chem. 388 (2007) 779–786.

[68] M.N. Win, C.D. Smolke, A modular and extensible RNA-based gene-regulatory platform for engineering cellular function, Proc. Natl. Acad. Sci. 104 (2007) 14283–

14288.

[69] L. Yen, J. Svendsen, J.S. Lee, J.T. Gray, M. Magnier, T. Baba, R.J. D'Amato, R.C. Mulligan, Exogenous control of mammalian gene expression through modulation of RNA self-cleavage, Nature. 431 (2004) 471–476.

[70] O.O. Abudayyeh, J.S. Gootenberg, S. Konermann, J. Joung, I.M. Slaymaker, D.B.T. Cox, S. Shmakov, K.S. Makarova, E. Semenova, L. Minakhin, K. Severinov, A. Regev, E.S. Lander, E. V. Koonin, F. Zhang, C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector, Science. 353 (2016) aaf5573.

[71] A. East-Seletsky, M.R. O'Connell, S.C. Knight, D. Burstein, J.H.D. Cate, R. Tjian, J.A. Doudna, Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection, Nature. 538 (2016) 270–273.

[72] O.O. Abudayyeh, J.S. Gootenberg, P. Essletzbichler, S. Han, J. Joung, J.J. Belanto, V. Verdine, D.B.T. Cox, M.J. Kellner, A. Regev, E.S. Lander, D.F. Voytas, A.Y. Ting, F. Zhang, RNA targeting with CRISPR-Cas13, Nature. 550 (2017) 280–284.

CHAPTER 2: Tunable self-cleaving ribozymes for modulating gene expression in eukaryotic systems

Thomas Jacobsen, Gloria Yi, Hadel A. Asafen, Ashley A. Jermusyk, Chase L. Beisel, and Gregory T. Reeves

Original publication:

T. Jacobsen, G. Yi, H.A. Asafen, C.L. Beisel, G.T. Reeves, Tunable self-cleaving ribozymes for modulating gene expression in eukaryotic systems, Manuscript submitted for publication. (2019).

ABSTRACT

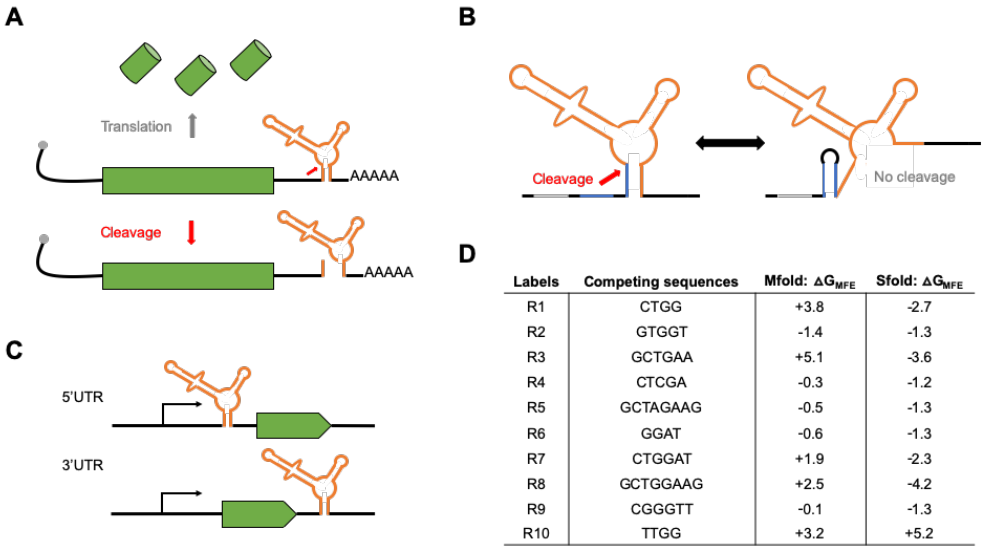
Advancements in the field of synthetic biology have been possible due to the development of genetic tools that are able to regulate gene expression. However, the current toolbox of gene regulatory tools for eukaryotic systems have been outpaced by those developed for simple, single-celled systems. Here, we engineered a set of gene regulatory tools by combining self-cleaving ribozymes with various upstream competing sequences that were designed to disrupt ribozyme self-cleavage. As a proof-of-concept, we were able to tune the expression of GFP in mammalian cells, and then showed the feasibility of these tools in *Drosophila* embryos. For each system, the fold-reduction of gene expression was influenced by the location of the self-cleaving ribozyme/upstream competing sequence (i.e. 5' untranslated region (UTR) vs. 3'UTR) and the competing sequence used. Together, this work provides a set of genetic tools that can be used to tune gene expression across various eukaryotic systems.

2.1 - INTRODUCTION

Synthetic biology is an interdisciplinary field that relies on biologists, engineers, mathematicians, and others to create novel biological systems by engineering and interchanging genetic parts derived from nature [1,2]. This has led to the advancements of various fields in medicine, molecular biology, and biotech. The ability to construct and analyze these systems has increased due to the availability of gene regulatory tools. Previous work has shown that these tools have the ability to regulate different steps of gene expression, including transcription [3], mRNA processing and stability [4], translation [5], and protein synthesis/stability [6]. This ability has been particularly useful in the construction of synthetic gene circuits, such as counting devices [7], patterning devices [8], toggle switches [9], and gene oscillators [10], as well as the production of novel drugs, therapeutics, and biofuels.

While gene regulatory tools have been developed for various model systems, the development of these tools in eukaryotic systems has been outpaced compared to those developed in single-celled systems like bacteria and yeast. Initially, the development of gene regulatory tools in eukaryotic systems had been focused on transcriptional control [1]. The tools to regulate transcription include the use of naturally-occurring (e.g. LacI, TetR, Gal4) and synthetic (e.g. zinc fingers, transcription activator-like effectors) transcription factors that have the ability to activate or inhibit gene expression [11–16]. Later, other methods of gene regulation have been developed to control translation (upstream open reading frames (uORFs), microRNAs, aptamers) and protein turnover [17–23]. More recently, clustered regularly interspaced short palindromic repeats (CRISPR) nucleases have been repurposed to act as synthetic transcription factors that have the ability to target virtually any gene of interest [24,25]. Even with these tools available, more powerful tools are needed to precisely control gene expression within eukaryotic systems.

One promising gene regulatory tool that has the potential to fine-tune gene expression are self-cleaving ribozymes, which are natural RNA structures that are able to catalyze their own cleavage [26]. When inserted into a transcript, these ribozymes reduce protein



Labels	Competing sequences	Mfold: ΔG_{MFE}	Sfold: ΔG_{MFE}
R1	CTGG	+3.8	-2.7
R2	GTGGT	-1.4	-1.3
R3	GCTGAA	+5.1	-3.6
R4	CTCGA	-0.3	-1.2
R5	GCTAGAAG	-0.5	-1.3
R6	GGAT	-0.6	-1.3
R7	CTGGAT	+1.9	-2.3
R8	GCTGGAAG	+2.5	-4.2
R9	CGGGTT	-0.1	-1.3
R10	TTGG	+3.2	+5.2

Figure 2.1: Gene regulatory tools based on self-cleaving ribozymes. (A) Inserting self-cleaving ribozymes in the 3' untranslated region (UTR) of a gene leads to cleavage (red arrow) and subsequent mRNA transcript destabilization/decay and inhibition of protein synthesis. (B) Conceptual design of tunable self-cleaving ribozymes. A competing sequence (blue) is placed directly upstream of the ribozyme (orange). Base-pairing of this sequence with a part of the ribozyme stem prevents ribozyme self-cleavage. The ribozyme is flanked by insulating sequences (gray) to aid in preventing base-pairing interactions between the ribozyme and other sequences in the 3'UTR. (C) Schematic of the constructs used to test the ribozyme constructs in mammalian cells and Drosophila. We placed the ribozyme (orange) either in the 5'UTR or 3'UTR of the reporter genes used (green). (D) List of the competing sequences used in this study, along with their labels used in Figures 2 and 4. Also listed are the free energy differences between the minimal free energy structures of ribozymes in a cleavable and non-cleavable conformation using each competing sequence derived Mfold and Sfold.

levels through self-cleavage and subsequent RNA degradation (Figure 2.1). Previous work has shown that inserting ribozymes in various loci of an mRNA transcript disrupts mRNA stability within bacteria, yeast, and mammalian cells [4,27,28]. Previous work in bacteria has also shown that the insertion of sequences flanking a ribozyme and ribosome binding site can alter the ribozyme's cleavage activity [29]. Here, we used Mfold to engineer a set of genetic tools based on self-cleaving ribozymes that can be used to regulate gene expression in eukaryotic systems. By combining ribozymes with upstream competing sequences that have the potential to base-pair with a major stem of the ribozyme and prevent ribozyme self-cleavage (Figure 2.1B), we show that gene expression can be tuned in two model systems. We initially show that these tools can tune expression of a fluorescent reporter in HEK293T cells, and then we implemented the ribozyme constructs in *Drosophila* embryos. While we observed that these tools were able to modulate gene expression in two model systems, there was a lack of correlation between RNA secondary structure prediction algorithms and the experimental data.

Together, these results show that self-cleaving ribozymes combined with upstream competing sequences can modulate gene expression in eukaryotic systems, and that other factors, besides ribozyme self-cleavage and base-pair interactions, influence gene expression.

2.2 - MATERIALS AND METHODS

2.2.1 - Strains, plasmids, oligonucleotides, and fly lines.

All strains, plasmids, oligos, gBlocks, and fly lines used in this work can be found in Supplementary Table 2.1. All PCR amplifications were performed using Q5 Hot Start High-Fidelity 2X Master Mix (NEB, Cat: M0494S) unless specified. All fly lines were generated using site-specific PhiC31-mediated insertion from Genetivision.

We used the pcDNA3.1$^{(+)}$ mammalian expression vector (Thermo Fisher, Cat: V79020) for expression of GFP in HEK293T cells. For this study we used the hammerhead self-cleaving ribozyme from *Schistosoma mansoni* as it has been associated with a high catalytic activity *in vitro* and *in vivo* [29,30]. We first built the active ribozyme constructs by first amplifying GFP and inserting it into the NotI and PstI sites of pCB1180. The inactive ribozyme constructs were built by creating a single point mutation that abolishes catalytic activity of the ribozyme [31]. Then, annealed/phosphorylated oligos containing the inactive and active ribozymes were inserted into the XhoI and NotI sites, located in the 5' untranslated region (UTR), to make pCB1134/1135. To insert these ribozyme/GFP sequences into pcDNA3.1$^{(+)}$, we PCR amplified the ribozyme-GFP sequence from pCB1134/1135 and inserted it into the HindIII and XbaI restriction sites in pcDNA3.1$^{(+)}$ to create pCB1136/1137. The upstream competing sequences were inserted into pCB1136/1137 by linearizing the plasmids with EcoRI and XhoI and then ligating with phosphorylated/annealed oligos containing the competing sequences of interest (Figure 2.1D). For insertion of the ribozyme/upstream competing sequences in the 3'UTR of GFP, the ribozyme and upstream competing sequences were PCR amplified from the previously built 5'UTR constructs and inserted into the XbaI site of pCB1133.

We used the pUAST-attB *Drosophila* expression vector (Drosophila Genomics Resource

Center, Cat: 1419) for creating the transgenic fly lines containing the *lacZ* reporter. To generate the ribozyme constructs, we first removed the UAS-hsp70 sequence using the HindIII and KpnI restriction sites and added the hunchback (hb) proximal enhancer (hbpe), the eve minimal promoter, and the *lacZ* reporter to create pCB1181. Expressing *lacZ* from the hbpe creates a well-established domain of *hb* to easily study the effects from the self-cleaving ribozymes [32–34]. For the insertion of the self-cleaving ribozymes into the 5'UTR and 3'UTR of *lacZ*, the StuI and KpnI restriction sites of pCB1181 were used, respectively. To insert the upstream competing sequences, both the EcoRI and AvrII sites were added upstream of the ribozyme sequence for ligation with phosphorylated/annealed oligos containing the competing sequences of interest.

2.2.2 - Predicting secondary structures of self-cleaving ribozymes/upstream competing sequences.

The online tools Mfold and Sfold were used to predict the minimal free energy (MFE) structures of the ribozymes lacking or containing an upstream competing sequence using the default settings [35,36]. We extracted the ΔG of the structures associated with the lowest free energy of a ribozyme in a cleavable and non-cleavable conformation. The ΔG of each upstream competing sequence was calculated as the difference between the ΔG of the cleaved and non-cleaved structures. See Supplementary Figure 2.1 for a representative secondary structure of ribozymes in a cleaving or non-cleaving conformation.

2.2.3 - Transient transfections of pcDNA3.1$^{(+)}$-ribozyme constructs.

Transfection-grade DNA was prepared using the QIAGEN Plasmid Mini Kit (QIAGEN, Cat: 12125). One day prior to the transient transfections, HEK239T cells were seeded onto either 35mm or 24-well plates with complete media (Dulbecco's Modified Eagle Medium (Invitrogen, Cat: 11965-092) supplemented with 10% fetal bovine serum (Invitrogen, Cat: A3840001)). Each pcDNA3.1$^{(+)}$-ribozyme construct was transiently transfected using FuGeneHD (Promega, Cat: E2311). Cells were then incubated for 48 hours prior to preparing the cells for flow cytometry. See Supplementary Table 2.1 for details of the transient transfections performed using each plate format.

2.2.4 - Flow cytometry analysis of transfected HEK293T cells.

We trypsinized the transiently transfected HEK293T cells using trypsin-EDTA (Thermo Fisher, Cat: 25200056) and resuspended them in 500mL 1xPBS (Fisher Scientific, Cat: MT21040CV). The cells were analyzed for fluorescence using the Accuri C6 Flow Cytometer with CFlow plate sampler (Becton Dickinson). The events were gated based on the forward scatter and side scatter, with fluorescence measured in FL2-H, using the 533/30 filter, from at least 10,000 gated events. The fold-reduction of GFP was calculated as the ratio of the fluorescence values for the cells transfected with an inactive ribozyme with a specific competing sequence over that of an active ribozyme with the same competing sequence.

2.2.5 - Fluorescent in situ hybridization of Drosophila embryos.

All embryos were aged to 2-4 hours from laying and then fixed using 37% formaldehyde following standard protocols [37]. Fluorescent *in situ* hybridization (FISH) was combined with fluorescent immunostaining following standard protocols [37]. Briefly, fixed embryos were washed in 1xPBS buffer supplemented with 0.05% Tween-20, and then hybridized with a fluorescein (ftc)-conjugated anti-sense *lacZ* probe at 55°C. The embryos were washed and incubated with the rabbit anti-histone (Abcam, Cat: ab1791) (1:10,000 dilution) and goat anti-ftc (Rockland, Cat: 600-101-096) (1:5,000 dilution) primary antibodies overnight at 4°C. Embryos were then washed and incubated for 1.5 hours with fluorescent donkey anti-rabbit-546 (Invitrogen, Cat: A10040) (1:500 dilution) and donkey anti-goat-647 (Invitrogen, Cat: A21447) (1:500 dilution) secondary antibodies at room temperature. Finally, the embryos were washed and stored in 70% glycerol at -20°C prior to being imaged. All prepared embryos were imaged within two weeks of protocol completion.

2.2.6 – Imaging and analysis of embryos.

To reduce variability from the fluorescence measurements, the intensity output of the 488 nM laser was used for laser calibration prior to embryo imaging [38]. The calibration was performed by measuring the intensity of the 488 nM laser through the transmitted light channel giving us the output strength of the laser. This allowed us to compensate for

potential variability of laser strength between imaging sessions. The prepared embryos were mounted laterally using 70% glycerol using two pieces of double-sided tape. A Zeiss LSM 710 microscope was used to acquire 15-25 z-slices 45-60 μm apart at 40x magnification.

Using Fiji, the z-max intensity projection for each embryo was measured for its fluorescence intensity. The *hb* expression domain was used as the cutoff for signal (see Figure 2.3A), as the expression profile of *lacZ* should match the endogenous *hb* expression pattern due to expression from the hbpe. The fluorescent signal was obtained by measuring the intensity from the anterior pole to the edge of *hb* domain using the tools available in Fiji. After measuring signal, background noise was measured as the intensity outside of the *hb* expression pattern. The fold-reduction of *lacZ* was calculated as the ratio of the fluorescence values for the embryos with an inactive ribozyme with a specific competing sequence over that of an active ribozyme with the same competing sequence. Refer to Supplementary Document 2.1 for an in-depth protocol.

With the same embryos, the width of the *lacZ* gradient was compared with the active and inactive ribozyme constructs. For this analysis, we used a supervised MATLAB script to first locate/orient the embryo, and then shape the embryos' periphery boundary. We then measured the fluorescence of the embryo across the anterior-posterior axis (see supplementary material for MATLAB scripts). To measure the distance from the anterior pole to the boundary of the *lacZ* domain, we selected three points along the y-axis and extracted the width corresponding to 50% loss of the maximum intensity. We selected three different y-values to account for asymmetrical *lacZ* gradients (Supplementary Figure 2.3). The median of the three values was used to represent the measurement of the *lacZ* gradient.

2.3 - RESULTS

2.3.1 - Designing self-cleaving ribozymes containing tunable upstream competing sequences.

For this study, we used the hammerhead self-cleaving ribozyme as it has shown high

activity *in vitro* and *in vivo* [29,30]. Though these ribozyme constructs can be placed in various locations within a transcript, we chose to test two specific locations: the 5' and 3'UTR of the reporter genes tested (Figure 2.1C). The competing sequences were placed upstream of the ribozyme to ensure that transcription of the ribozyme before the competing sequence did not result in self-cleavage prior to the transcription of the competing sequence. Insulating sequences were flanked upstream of the ribozyme/competing sequence to limit ribozyme misfolding due to flanking sequences (Figure 2.1B). Finally, we designed the competing sequences using Mfold [35] to obtain a set of sequences that were associated with varying levels of predicted folded and misfolded ribozyme structures (Figure 2.1D). Each competing sequence varied in

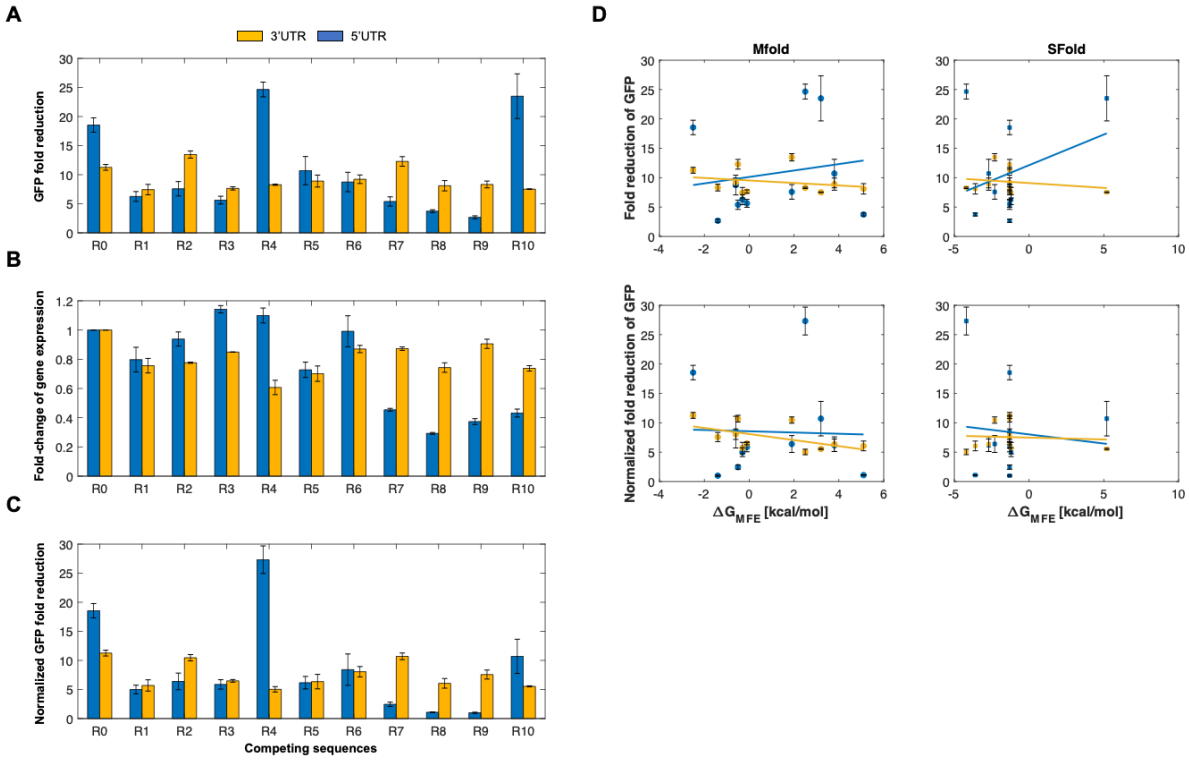


Figure 2.2: Self-cleaving ribozymes can tune gene expression in mammalian cells. (A) The average fold-reduction of GFP observed from the flow cytometry data for various competing sequences used in the 3'UTR (yellow) and 5'UTR (blue). The constructs were transiently transfected and incubated for 48 hours at 37°C. After incubation, the cells were trypsinized and resuspended in 1xPBS for flow cytometry analysis. (B) Percentage of constitutively-expressed *gfp* with a ribozyme/competing sequence located in the 3'UTR (yellow) or 5'UTR (blue) of the transcript. A value of one indicates no change. (C) Normalized average of GFP fold-reduction using the data from Figures 2.2A/B. This represents the loss of reporter gene expression only due to ribozyme activity. All error bars represent the standard deviation from at least three independent transfections. Note that R0 indicates a self-cleaving ribozyme lacking competing sequence. (D) Predicted relationship between the fold-reduction of GFP and the free energy difference between cleavable and non-cleavable ribozyme conformations. Plots in column one and two compare the fold-reduction levels with the free energies calculated from Mfold and Sfold, respectively. The first and second rows represent the raw fold-reduction data (Figure 2.2A) and the normalized fold-reduction data (Figure 2.2C), respectively.

sequence length and composition and were associated with different propensities to base-pair with the stem of the ribozyme. Finally, each competing sequence lacked a start codon to prevent premature translation initiation.

2.3.2 - Self-cleaving ribozymes combined with upstream competing sequences can tune gene expression in mammalian cells.

We first sought to test these ribozyme constructs in a mammalian system. To this end, we tested the ribozyme constructs in HEK293T cells. We inserted the self-cleaving ribozymes and 10 different upstream competing sequences in the 5' UTR or the 3'UTR of GFP to observe how various sequence configurations impacted reporter gene expression. For each ribozyme/competing sequence tested, we used an inactive ribozyme with the same competing sequence to act as a control. As the inactive and active ribozymes differ by a single point mutation [31], the overall structure of the ribozyme was preserved. After transiently transfecting these reporter constructs, the fluorescence of the cells was analyzed by flow cytometry analysis. We found that these ribozyme constructs were able to reduce expression of GFP in HEK293T cells, with fluorescence generally being associated in a bimodal distribution (untransfected cells and cells associated with varying GFP levels) (Supplementary Figure 2.2). When located in the 5'UTR, the ribozymes/upstream competing sequences generally resulted in greater range of fold-reduction levels compared to when located in the 3'UTR (Figure 2.2).

As the GFP fold-reduction levels between the 5' and 3'UTR constructs were variable, we wanted to assess the effect of competing sequence insertion on GFP expression. Due to prior work showing that the formation of secondary structures strongly effects transcript stability [39], we compared the fluorescence of the cells transiently transfected with ribozyme constructs containing an inactive ribozyme lacking an upstream competing sequence to that of inactive ribozymes containing an upstream competing sequence (Figure 2.2B). While the loss of GFP expression was fairly consistent for the constructs containing ribozymes with competing sequences in the 3'UTR (~20-40% loss of GFP expression), GFP expression loss was more noticeable when the ribozyme/competing sequences were placed in the 5'UTR. When placed in the 5'UTR, the loss of gene

expression ranged from negligible loss (e.g. R2, R6) to ~70% loss (e.g. R8) (Figure 2.2B). Interestingly, the insertion of some upstream competing sequences resulted in increased expression of GFP (e.g. R3, R4). We then accounted for the loss of gene expression due to the insertion of a competing sequence by normalizing the fold-reduction data from Figure 2.2A using the data from Figure 2.2B (Figure 2.2C). While this generally resulted in less fold-reduction of each construct, a wide dynamic range was generally maintained, from almost no fold-reduction to ~25-fold-reduction of gene expression.

After obtaining the experimental data, we then sought to gain insight into the relationship between the fold-reduction of gene expression and the predicted energies of misfolding. To this end, we compared the GFP fold-reduction levels with the predicted free energy differences obtained from Mfold. To obtain these values, the difference between the ΔG associated with the MFE structure of a ribozyme in a cleavable conformation and the ΔG associated with the MFE in a non-cleavable conformation was calculated (Supplementary Figure 2.1). While the experimental data from HEK293T cells showed a wide dynamic range of fold-reduction levels, there was a lack of correlation between the experimental data and predicted free energy differences (Figure 2.2D). We then sought to use a different RNA predictive folding algorithm to see if it could better correlate the fold-reduction of gene expression to predicted free energies. Thus, we used Sfold to compare MFE's to the GFP fold-reduction [36]. Similar to Mfold, there was a lack of correlation between the experimental fold-reductions to the predicted free energies (Figure 2.2D). The lack of a correlation indicates the presence of external factors that influence ribozyme self-cleavage, thus currently making this approach non-predictive. Even so, our data show that ribozyme/upstream competing sequences can be used to tune gene expression in mammalian cells.

2.3.3 - Self-cleaving ribozymes/upstream competing sequences can tune gene expression in Drosophila.

As a proof-of-concept, we next wanted to test these tools in a multicellular system. We chose to work with *Drosophila* embryos as we have previously used this model to study synthetic networks [40]. Thus, we generated transgenic fly lines carrying these ribozyme

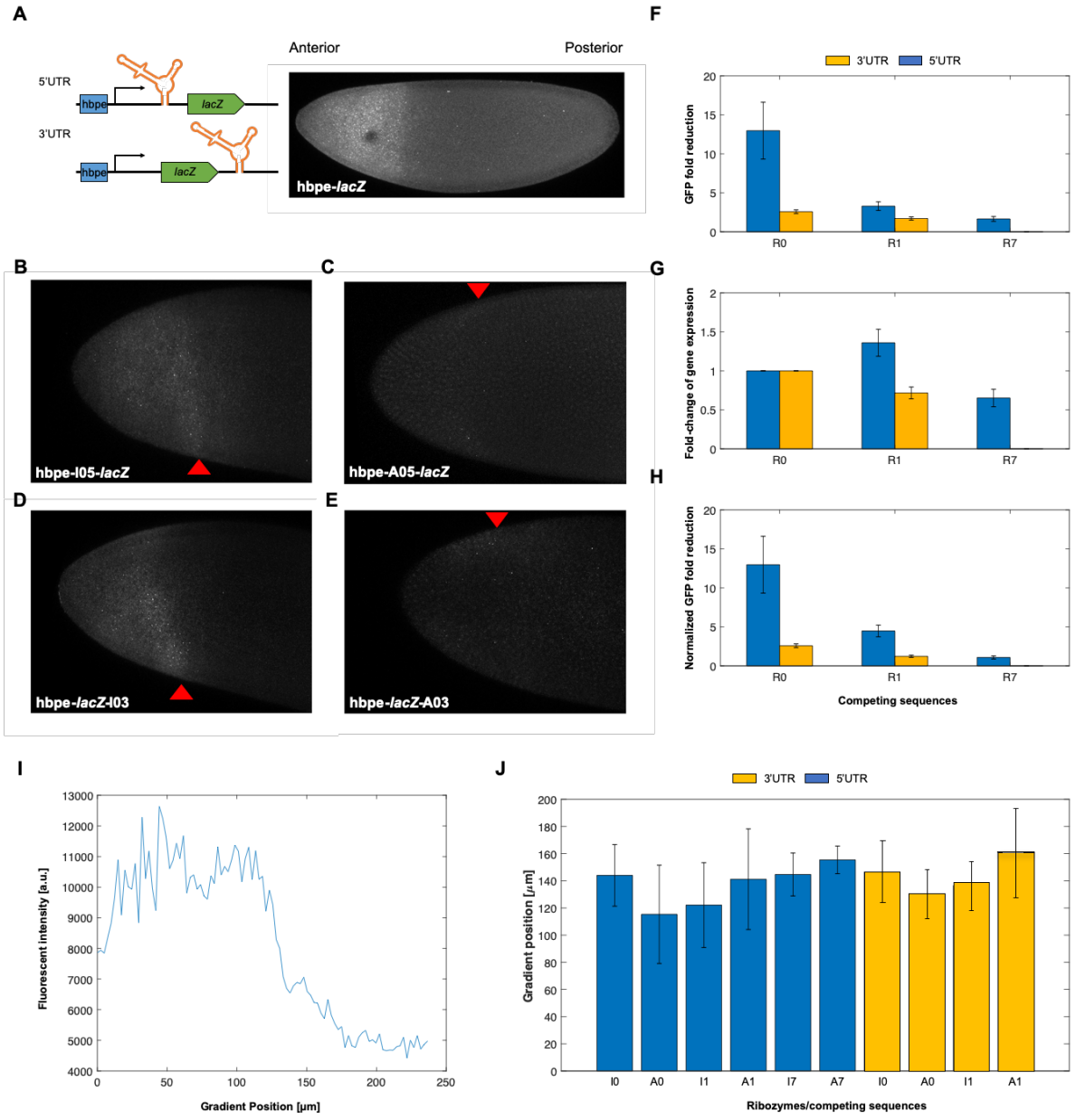


Figure 2.3: Self-cleaving ribozymes can tune gene expression in *Drosophila*. (A) Depiction of the ribozyme constructs and its expression domain in *Drosophila*. The domain of *lacZ* is similar to the endogenous *hunchback* (*hb*) gradient due to the hunchback proximal enhancer. During early development, *hb* is strongly expressed in the anterior of the embryo. (B-E) Images of *in situ* hybridized *Drosophila* embryos probed for *lacZ*. Each embryo imaged expresses *lacZ* under the control of the hbpe and contains an inactive (B/D) or active (C/E) ribozyme. Red triangles represent the width of the *lacZ* gradient. (F) The average fold-reduction of *lacZ* observed from the imaging data for various competing sequences. Embryos were collected from transgenic fly lines constitutively expressing *lacZ from* the hbpe containing a ribozyme sequence in the 3' UTR (yellow) or 5' UTR (blue) and prepared for image analysis. Images were acquired using a Zeiss LSM710 confocal microscope. (G) Percentage of *lacZ* expression loss due to effects other than ribozyme activity. A value of 1 indicates no change. (H) Normalized average fold-reduction of *lacZ* using the data from Figure 2.3F/G. This represents the loss of *lacZ* expression only due to ribozyme activity. All error bars represent the standard deviation from at least 10 embryos. Note that R0 indicates a self-cleaving ribozyme lacking a upstream competing sequence. Also note that fly lines containing the R7 competing sequence in the 3'UTR were not analyzed. (I) The fluorescent intensity at various positions of the embryo. A domain width of zero indicates the anterior pole and increasing values indicate a position closer to the posterior. (J) The average width of the *lacZ* domain for each ribozyme and competing sequence listed.

constructs. We first designed *Drosophila* expression vectors containing the *lacZ* reporter expressed from the hunchback proximal enhancer (hbpe). This enhancer results in an expression pattern similar to endogenous *hb*, which has a sharp boundary at roughly 50% anterior-to-posterior (AP) coordinate [32–34]. The hbpe drives expression with a boundary at roughly 33% AP coordinate (Figure 2.3A), which allowed us to quantitatively test these ribozymes *in vivo*. Similar to the work in HEK293T cells, each ribozyme/upstream competing sequence tested were compared to an inactive ribozyme containing the same competing sequence to act as a negative control. Embryos were first hybridized with an antisense *lacZ* probe, then imaged by confocal microscopy. We found that the insertion of ribozyme/competing sequences into a transcript expressing *lacZ* were able to tune *lacZ* expression levels in *Drosophila* embryos (Figure 2.3B-E). Unlike with the mammalian cell data, normalizing the fold-reduction data by accounting for the effects of inserting the upstream competing sequences on *lacZ* expression resulted in only in small changes to the measured dynamic range of fold-reduction values (Figure 2.3). While the fold-reduction values observed in *Drosophila* were generally less than those observed in HEK293T cells, the correlation of fold-reduction values, compared to the work in mammalian cells, remained the same (i.e. R0 > R1 > R7) and maintained a high dynamic range (~2-14 fold-reduction of *lacZ*).

Using the same images, we then compared the width of the *lacZ* domain along the anterior-posterior axis. We hypothesized that the embryos containing an active ribozyme construct would be associated with a reduced domain width as the expression of *lacZ* would be reduced at locations containing weak fluorescent intensity (i.e. distal to anterior pole). For each ribozyme construct, we observed that the differences in the *lacZ* domain width were small, but noticeable across all constructs. Interestingly, only the two strongest ribozymes (i.e. A0-5UTR, A0-3UTR) resulted in a noticeable *lacZ* gradient reduction (Figure 2.3J), though the average gradient width between active and inactive ribozymes were not statistically different. These results indicate that the *lacZ* domain width did not vary between active and inactive ribozymes regardless of location or competing sequence.

2.4 - DISCUSSION

In this work, we engineered a set of genetic tools that were able to modulate gene expression in HEK293T cells and *Drosophila*. At face value, inserting the ribozymes in the 5'UTR of the reporter genes yielded a greater range of fold-reduction levels compared to the 3'UTR. However, we observed that insertion of upstream competing sequences resulted in the inhibition of gene expression in the absence of ribozyme self-cleavage. This effect was greater when the ribozyme/competing sequence was located in the 5'UTR (Figure 2.2B). After normalizing the fold-reduction levels by accounting for the loss of gene expression, we observed that some ribozyme constructs (most notably the 5'UTR constructs) reduced gene expression more weakly compared to that data prior to normalization (Figure 2.2A/C). In general, the ribozymes/upstream competing sequences were observed to reduce gene expression more strongly in HEK293T cells compared to *Drosophila* (Figures 2.2 and 2.3), which has also been observed in recent work [41]. This difference could be due to different biological machinery between mammalian and insect models, different experimental assays, or the constructs themselves, as they contain different promoters and reporter genes. Even with the differences in fold-reduction levels between these model systems, these tools maintained a dynamic range of gene expression regulation (~1-25 in HEK293T cells and ~2-14 in *Drosophila*). While the experimental data did not show a high correlation with the RNA secondary structure predictions (Figure 2.3), we provide a set of gene regulatory tools based on empirical measurements between competing sequences and strength of gene reduction.

Prior to experimental work, we used Mfold [35,36] to design a set of competing sequences that were associated with a wide range of free energies (Figure 2.1D). When comparing these predicted free energies to the fold-reduction levels observed in our experimental data (Figures 2.2 and 2.3), we generally observed a weak correlation (Figure 2.3). This discrepancy could have been due to a variety of factors. For instance, the insulating sequences, used to prevent interactions between the ribozyme and flanking sequences, could have affected the ability of the competing sequences to base-pair with the ribozyme. While Mfold and Sfold predictions showed minimal interactions between the ribozyme and insulating sequences, the sequences flanking the insulating sequences could have

interacted with the competing sequence, ribozyme, and/or the insulating sequence. To prevent this phenomenon, the length and/or content of the insulator sequence could be altered. It is also possible that one or more of the competing and/or insulating sequences contain a target sequence for a native biological factor or pathway, such as an endogenous transcription factor, internal ribosome entry site (IRES), or RNAi. While the addition of a specific target sequence is unlikely, novel transcription factors, IRES', and non-coding RNAs continue to be discovered in eukaryotic systems, including *Drosophila* [42–48]. Finally, Mfold and/or Sfold may lack the ability to predict the fold-reduction of gene expression associated with the ribozyme constructs. Recent work has shown that hammerhead ribozymes are associated with varying cleavage activities across different model systems (e.g. mammalian vs. yeast) and experimental setups (e.g. *in vitro* vs. *in vivo*) [41], which show that cellular context is likely important for the observed activity. Another possibility is that Mfold and Sfold are not accurately capturing RNA folding. While algorithms, such as Mfold and Sfold, have the ability to predict RNA secondary structures, ribozymes can form complex 3D structures (e.g., pseudoknots) that cannot be predicted accurately. Due to these reasons, current predictive RNA folding algorithms may not be sufficient for accurate secondary structure predictions. Improvements on RNA structure prediction models will allow for a more accurate design of competing sequences.

Experimental data indicated that insertion of the upstream competing sequences generally inhibited gene expression when compared with the constructs lacking these sequences. This phenomenon could be due to various reasons. For one, the mRNA transcripts could have been subjected to the no-go decay pathway [49]. This mRNA surveillance pathway occurs when ribosomes have stalled during translation, resulting in cleavage and subsequent degradation. While some of the ribozyme constructs resulted in drastic reduction of gene expression from competing sequence insertion, the majority of these constructs had a small, but noticeable effect on gene expression (Figure 2.2B). Similar to the RNA folding algorithms, another possibility could be that one or more of the competing sequences was a target sequence for an endogenous biological factor or pathway. To prevent variation of gene expression when using these ribozyme constructs, longer insulating sequences can be flanked to both the 5' and 3' ends of the

ribozyme/upstream competing sequences. This could prevent interactions between the ribozyme or competing sequence with flanking sequences, resulting in fold-reduction levels only from ribozyme self-cleavage.

2.5 - CONCLUSIONS

We developed a set of tools that were able to tune gene expression in HEK293T cells and *Drosophila*. While the free energies obtained from the predictive RNA secondary structure tool did not correlate well with fold-reduction of gene expression, the competing sequences used in this work provides a set of genetic tools associated with a wide range of fold-reduction levels. Though tested in mammalian and insect systems, these tools should be applicable in other eukaryotic systems, such as *C. elegans*, zebrafish, and mice. Previous work has shown that self-cleaving ribozymes are found naturally in these organisms [50–52] and have been used for therapeutic applications [4,53]. These tools will be useful for studies involving synthetic biology, especially for the purposes of building and studying synthetic gene circuits.

2.6 - ACKNOWLEDGMENTS

The pUAST-attB plasmid was a gift from the Drosophila Genomics Resource Center, who are funded from the National Institutes of Health (2P40OD010949). This work was supported by the U.S. Department of Education [Graduate Assistance in Areas of National Need Biotechnology Fellowship (P200A140020)] and the National Science Foundation (MCB-1413044).

2.7 - REFERENCES

[1] F. Lienert, J.J. Lohmueller, A. Garg, P.A. Silver, Synthetic biology in mammalian cells: Next generation research tools and therapeutics, Nat. Rev. Mol. Cell Biol. 15 (2014) 95–107.

[2] V. Singh, Recent advancements in synthetic biology: Current status and challenges, Gene. 535 (2014) 1–11.

[3] L. Guzman, D. Belin, M.J. Carson, J. Beckwith, Tight regulation, modulation, and high-level expression by vectors containing the arabinose PBAD promoter, J Biotechnol. 177 (1995) 4121–4130.

[4] L. Yen, J. Svendsen, J.S. Lee, J.T. Gray, M. Magnier, T. Baba, R.J. D'Amato, R.C. Mulligan, Exogenous control of mammalian gene expression through modulation of RNA self-cleavage, Nature. 431 (2004) 471–476.

[5] H.M. Salis, E.A. Mirsky, C.A. Voigt, Automated design of synthetic ribosome binding sites to control protein expression, Nat. Biotechnol. 27 (2009) 946–950.

[6] K.E. McGinness, T.A. Baker, R.T. Sauer, Engineering controllable protein degradation, Mol. Cell. 22 (2006) 701–707.

[7] A.E. Friedland, T.K. Lu, X. Wang, D. Shi, G. Church, J.J. Collins, Synthetic gene networks that count, Science. 324 (2009) 1199–1202.

[8] S. Basu, Y. Gerchman, C.H. Collins, F.H. Arnold, R. Weiss, A synthetic multicellular system for programmed pattern formation, Nature. 434 (2005) 1130–1134.

[9] T.S. Gardner, C.R. Cantor, J.J. Collins, Construction of a genetic toggle switch in Escherichia coli, Nature. 403 (2000) 339–342.

[10] M.B. Elowitz, S. Leibler, A synthetic oscillatory network of transcriptional regulators, Nature. 403 (2000) 335–338.

[11] M. Brown, J. Figge, U. Hansen, C. Wright, K.-T. Jeang, G. Khoury, D.M. Livingston, T.M. Roberts, Lac repressor can regulate expression from a hybrid SV40 early promoter containing a lac operator in animal cells, Cell. 49 (1987) 603–612.

[12] M. Gossen, H. Bujard, Tight control of gene expression in mammalian cells by tetracycline-responsive promoters, Proc Natl Acad Sci. 89 (1992) 5547–5551.

[13] M.L. Maeder, S. Thibodeau-Beganny, A. Osiak, D.A. Wright, R.M. Anthony, M. Eichtinger, T. Jiang, J.E. Foley, R.J. Winfrey, J.A. Townsend, E. Unger-Wallace,

J.D. Sander, F. Müller-Lerch, F. Fu, J. Pearlberg, C. Göbel, J.P. Dassie, S.M. Pruett-Miller, M.H. Porteus, D.C. Sgroi, A.J. Iafrate, D. Dobbs, P.B. McCray Jr, T. Cathomen, D.F. Voytas, J.K. Joung, Rapid “open-source” engineering of customized zinc-finger nucleases for highly efficient gene modification, Mol. Cell. 31 (2008) 294–301.

[14] R. Morbitzer, P. Römer, J. Boch, T. Lahaye, Regulation of selected genome loci using de novo-engineered transcription activator-like effector (TALE)-type transcription factors, Proc. Natl. Acad. Sci. 107 (2010) 21617–21622.

[15] A. Garg, J.J. Lohmueller, P.A. Silver, T.Z. Armel, Engineering synthetic TAL effectors with orthogonal target sites, Nucleic Acids Res. 40 (2012) 7584–7595.

[16] H. Kakidani, M. Ptashne, GAL4 activates gene expression in mammalian cells, Cell. 52 (1988) 161–167.

[17] J. Medenbach, M. Seiler, M.W. Hentze, Translational control via protein-regulated upstream open reading frames, Cell. 145 (2011) 902–913.

[18] J.P. Ferreira, K.W. Overton, C.L. Wang, Tuning gene expression with synthetic upstream open reading frames, Proc. Natl. Acad. Sci. 110 (2013) 11284–11289.

[19] Z. Xie, L. Wroblewska, L. Prochazka, R. Weiss, Y. Benenson, Multi-input RNAi-based logic circuit for identification of specific cancer cells, Science. 333 (2011) 1307–1311.

[20] S. Ausländer, D. Ausländer, M. Müller, M. Wieland, M. Fussenegger, Programmable single-cell mammalian biocomputers, Nature. 487 (2012) 123–127.

[21] K. Nishimura, T. Fukagawa, H. Takisawa, T. Kakimoto, M. Kanemaki, An auxin-based degron system for the rapid depletion of proteins in nonplant cells, Nat. Methods. 6 (2009) 917–922.

[22] K.M. Bonger, L. Chen, C.W. Liu, T.J. Wandless, Small-molecule displacement of a cryptic degron causes conditional protein degradation, Nat. Chem. Biol. 7 (2011) 531–537.

[23] L.A. Banaszynski, L.-C. Chen, L.A. Maynard-Smith, A.G.L. Ooi, T.J. Wandless, A rapid, reversible, and tunable method to regulate protein function in living cells using synthetic small molecules, Cell. 126 (2006) 995–1004.

[24] L.S. Qi, M.H. Larson, L.A. Gilbert, J.A. Doudna, J.S. Weissman, A.P. Arkin, W.A.

Lim, Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression, Cell. 152 (2013) 1173–1183.

[25] L.A. Gilbert, M.H. Larson, L. Morsut, Z. Liu, G.A. Brar, S.E. Torres, N. Stern-Ginossar, O. Brandman, E.H. Whitehead, J.A. Doudna, W.A. Lim, J.S. Weissman, L.S. Qi, CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes, Cell. 154 (2013) 442–451.

[26] A.R. Ferré-D'Amaré, W.G. Scott, Small self-cleaving ribozymes, Cold Spring Harb. Perspect. Biol. 2 (2010) a003574.

[27] K.H. Link, L. Guo, T.D. Ames, L. Yen, R.C. Mulligan, R.R. Breaker, Engineering high-speed allosteric hammerhead, Biol. Chem. 388 (2007) 779–786.

[28] M.N. Win, C.D. Smolke, A modular and extensible RNA-based gene-regulatory platform for engineering cellular function, Proc. Natl. Acad. Sci. 104 (2007) 14283–14288.

[29] J.M. Carothers, J. a Goler, D. Juminaga, J.D. Keasling, Model-driven engineering of RNA devices to quantitatively program gene expression, Science. 334 (2011) 1716–1719.

[30] G. Ferbeyre, J.M. Smith, R. Cedergren, Schistosome satellite DNA encodes active hammerhead ribozymes, Mol. Cell. Biol. 18 (1998) 3880–3888.

[31] F. Seela, H. Debelak, N. Usman, A. Burgin, L. Beigelman, 1-Deazaadenosine: synthesis and activity of base-modified hammerhead ribozymes, Nucleic Acids Res. 26 (1998) 1010–1018.

[32] W. Driever, C. Nüsslein-Volhard, The bicoid protein is a positive regulator of hunchback transcription in the early Drosophila embryo, Nature. 337 (1989) 138–143.

[33] R. Lehmann, C. Nüsslein-Volhard, hunchback, a gene required for segmentation of an anterior and posterior region of the Drosophila embryo, Dev. Biol. 119 (1987) 402–417.

[34] M.W. Perry, J.P. Bothma, R.D. Luu, M. Levine, Precision of hunchback expression in the Drosophila embryo, Curr. Biol. 22 (2012) 2247–2252.

[35] M. Zuker, Mfold web server for nucleic acid folding and hybridization prediction, Nucleic Acids Res. 31 (2003) 3406–3415.

[36] Y. Ding, C.Y. Chan, C.E. Lawrence, Sfold web server for statistical folding and rational design of nucleic acids, Nucleic Acids Res. 32 (2004) 135–141.

[37] D. Kosman, C.M. Mizutani, D. Lemons, W.G. Cox, W. McGinnis, E. Bier, Multiplex detection of RNA expression in Drosophila embryos, Science. 305 (2004) 846.

[38] L.M. Liberman, G.T. Reeves, A. Stathopoulos, Quantitative imaging of the Dorsal nuclear gradient reveals limitations to threshold-dependent patterning in Drosophila, Proc. Natl. Acad. Sci. 106 (2009) 22317–22322.

[39] T.A. Carrier, J.D. Keasling, Controlling messenger RNA stability in bacteria: Strategies for engineering gene expression, Biotechnol. Prog. 13 (1997) 699–708.

[40] A.A. Jermusyk, N.P. Murphy, G.T. Reeves, Analyzing negative feedback using a synthetic gene network expressed in the Drosophila melanogaster embryo, BMC Syst. Biol. 10 (2016) 85.

[41] L.A. Wurmthaler, B. Klauser, J.S. Hartig, Highly motif- and organism-dependent effects of naturally occurring hammerhead ribozyme sequences on gene expression, RNA Biol. 15 (2018) 231–241.

[42] R.S. Young, A.C. Marques, C. Tibbit, W. Haerty, A.R. Bassett, J.L. Liu, C.P. Ponting, Identification and properties of 1,119 candidate LincRNA loci in the Drosophila melanogaster genome, Genome Biol. Evol. 4 (2012) 427–442.

[43] S. Inagaki, K. Numata, T. Kondo, M. Tomita, K. Yasuda, A. Kanai, Y. Kageyama, Identification and expression analysis of putative mRNA-like non-coding RNA in Drosophila, Genes to Cells. 10 (2005) 1163–1173.

[44] J.L. Tupy, A.M. Bailey, G. Dailey, M. Evans-Holm, C.W. Siebel, S. Misra, S.E. Celniker, G.M. Rubin, Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster, Proc. Natl. Acad. Sci. 102 (2005) 5495–5500.

[45] M.T. Marr II, J.A. D'Alessio, O. Puig, R. Tjian, IRES-mediated functional coupling of transcription and translation amplifies insulin receptor feedback, Genes Dev. 21 (2007) 175–183.

[46] D. Maier, A.C. Nagel, A. Preiss, Two isoforms of the Notch antagonist Hairless are produced by differential translation initiation, Proc. Natl. Acad. Sci. 99 (2002) 15480–15485.

[47] D.Y. Rhee, D.-Y. Cho, B. Zhai, M. Slattery, L. Ma, J. Mintseris, C.Y. Wong, K.P. White, S.E. Celniker, T.M. Przytycka, S.P. Gygi, R.A. Obar, S. Artavanis-Tsakonas, Transcription factor networks in Drosophila melanogaster, Cell Rep. 8 (2014) 2031–2043.

[48] C. Wang, F. Yeung, P.-C. Liu, R.M. Attar, J. Geng, L.W.K. Chung, M. Gottardis, C. Kao, Identification of a novel transcription factor, GAGATA-binding protein, involved in androgen-mediated expression of prostate-specific antigen, J. Biol. Chem. 278 (2003) 32423–32430.

[49] M.K. Doma, R. Parker, Endonucleolytic cleavage of eukaryotic mRNAs with stalls in translation elongation, Nature. 440 (2006) 561–564.

[50] C.-H.T. Webb, N.J. Riccitelli, D.J. Ruminski, A. Lupták, Widespread occurrence of self-cleaving ribozymes, Science. 326 (2009) 953.

[51] M. Martick, L.H. Horan, H.F. Noller, W.G. Scott, A discontinuous hammerhead ribozyme embedded in a mammalian messenger RNA, Nature. 454 (2008) 899–902.

[52] K. Salehi-Ashtiani, J.W. Szostak, In vitro evolution suggests multiple origins for the hammerhead ribozyme, Nature. 414 (2001) 82–84.

[53] B.E. Peace, J.B. Florer, D. Witte, Y. Smicun, I. Toudjarska, G. Wu, M.W. Kilpatrick, P. Tsipouras, R.J. Wenstrup, Endogenously expressed multimeric self-cleaving hammerhead ribozymes ablate mutant collagen in cellulo, Mol. Ther. 12 (2005) 128–136.

CHAPTER 3: A detailed cell-free transcription-translation (TXTL)-based assay to decipher CRISPR protospacer-adjacent motifs

Colin S. Maxwell*, **Thomas Jacobsen***, Ryan Marshall, Vincent Noireaux, and Chase L. Beisel

Original publication:

C.S. Maxwell*, **T. Jacobsen***, R. Marshall, V. Noireaux, C.L. Beisel, A detailed cell-free transcription-translation-based assay to decipher CRISPR protospacer-adjacent motifs, Methods. 143 (2018) 48-57.

ABSTRACT

The RNA-guided nucleases derived from the CRISPR-Cas systems in bacteria and archaea have found numerous applications in biotechnology, including genome editing, imaging, and gene regulation. However, the discovery of novel Cas nucleases has outpaced their characterization and subsequent exploitation. A key step in characterizing Cas nucleases is determining which protospacer-adjacent motif (PAM) sequences they recognize. Here, we detail an *in vitro* method based on an *E. coli* cell-free transcription-translation system (TXTL) to rapidly elucidate PAMs recognized by Cas nucleases. The method obviates the need for cloning Cas nucleases or gRNAs, does not require the purification of protein or RNA, and can be performed in less than a day. We improved on our previously published method by incorporating an internal GFP cleavage control to assess the extent of library cleavage as well as Sanger sequencing of the cleaved library to assess PAM depletion prior to next-generation sequencing. Furthermore, we detail the methods needed to construct all relevant DNA constructs, and how to troubleshoot the assay. We demonstrate the technique by determining PAM sequences recognized by the *Neisseria meningitidis* Cas9, and reveal subtle sequence requirements of this highly specific PAM. The overall method offers a rapid means to identify PAMs recognized by diverse CRISPR nucleases, with the potential to greatly accelerate our ability to characterize and harness novel CRISPR nucleases across their many uses.

3.1 - INTRODUCTION

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-CRISPR-associated (Cas) systems have become widespread biomolecular tools. In nature, these systems act as an adaptive immune system that protects prokaryotes from mobile genetic elements such as plasmids and bacteriophage [1–5]. In addition to their fascinating role in microbial ecology and evolution, these systems have proved to be remarkably useful, and have been repurposed for numerous applications such as gene regulation, imaging, and, notably, genome editing [6–9].

CRISPR-Cas systems are unique in that targeting them to cleave or bind to a new DNA sequence requires only the expression of a new non-coding guide RNA (gRNA). Upon expression, a Cas nuclease and its gRNA form a ribonucleoprotein complex (RNP). This RNP targets and cleaves near DNA sequences (called protospacers) that are complementary to the specificity-determining region (called the spacer) of the gRNA and are flanked by a protospacer-adjacent motif (PAM) (Figure 3.1A) [10]. Because the specificity of targeting derives from the spacer sequence of the gRNA base-pairing with the sequence of interest, re-engineering the target site simply involves the modification of the gRNA. In contrast, protein engineering was required to change the targeting specificity of other prior genome-editing technologies [11–13].

The vast majority of CRISPR-based technologies have utilized DNA-targeting Type II Cas9 nucleases [14,15] or Type V-A Cas12a (Cpf1) nucleases [16]. However, these two systems are a small subset of the stunningly diverse array of CRISPR-Cas systems found in nature. CRISPR-Cas systems are currently divided into two classes, six types, and over 30 subtypes [17], and are present in approximately 50% and 90% of bacteria and archaea, respectively [18]. The characterization of some of these systems has revealed nucleases containing the ability to degrade (rather than simply cleave) DNA, recognize diverse PAM sequences, recognize fewer off-target sequences, and efficiently process their own gRNA arrays for multiplexing in eukaryotic systems [19–23]. Furthermore, the recent discovery of the Type VI Cas nuclease, Cas13a, was found to target RNA instead of DNA and has opened new technological opportunities based

on RNA targets [19,24]. However, despite these advances, most CRISPR-Cas systems remain to be described and/or adapted for use in biotechnology.

Determining the PAM requirements of a CRISPR-Cas system is a critical step in its characterization. All characterized CRISPR-Cas systems, except for Type III, require a PAM for efficient cleavage of target DNA [14,25–27]. Conversely, identifying sequences flanked by PAMs is critical for predicting and minimizing off-target effects [28]. The PAM sequence required by a particular system varies widely in sequence length, content, and position relative to the protospacer. This variation occurs not only across higher-level (Class, Type) divisions of CRISPR-Cas systems, but also within subtypes among the species that the Cas nuclease was isolated from [23]. Furthermore, some Cas nucleases are able to utilize multiple PAM sequences with complicated relationships between the nucleotides required at different positions in the PAM [29]. Therefore, methods to characterize the PAM requirements of particular Cas nucleases play a vital role in studying their properties.

3.1.1 - Previous methods to characterize PAMs of CRISPR-Cas systems

Previous methods have been developed to identify PAMs of various Cas nucleases. These methods involve the original *in silico* method used to demonstrate the existence of PAMs [26], as well as the development of more recent high-throughput *in vivo* [16,30–34] and *in vitro* [16,35–38] assays. Here, we briefly describe each method. In-depth descriptions of each method have been provided elsewhere [10,39].

PAMs were originally identified using bioinformatics [26]. Once it was realized that CRISPR spacers were derived from mobile genetic elements, alignments of the spacers to plasmids and phage genomes revealed the existence of a motif that flanked the protospacers. This method continues to be used to identify PAMs. By aligning incorporated spacer sequences from a prokaryotic genome containing a CRISPR array to bacteriophage or plasmid sequences, PAMs of various CRISPR-Cas systems can be deciphered. However, this method is limited by the availability of identifying matching sequences; many spacers appear to recognize ‘dark matter’ [40–47]. Furthermore, the

relatively small number of spacers (tens to hundreds) that can be examined this way may not be sufficient to identify weakly-recognized PAMs. Finally, this method conflates the PAM requirements for acquiring a spacer and using a spacer for interference despite differences in these sequences [48]. These limitations have been overcome by assaying the ability of libraries of randomized sequences to act as PAMs.

The first *in vivo* PAM assay took advantage of the ability of CRISPR-Cas systems to cure bacteria of plasmids [16,30–33]. To assay for functional PAMs, *E. coli* cells are co-transformed with (1) a plasmid containing a randomized PAM library flanked by a unique protospacer and plasmids encoding (2) a Cas nuclease and (3) a gRNA targeting the protospacer. Cells are then selected for all three plasmids by recovering them on media containing appropriate antibiotics. Plasmids containing a PAM are cleaved and the cells that contain them cannot grow, while cells containing plasmids with non-PAMs are able to propagate. By comparing the frequency of a sequence in the library before and after selection, individual PAM sequences can be identified. While this assay has been used to successfully determine the PAMs of various Cas nucleases, it is limited by the requirement of high library coverage and the potential for cells containing mutated Cas or gRNA plasmids but functional PAMs to be amplified during cell outgrowth.

Soon after, another *in vivo* assay, named “PAM screen achieved by NOT-gate repression” (PAM-SCANR), was developed [34]. This assay was the first to involve the binding of catalytically inactive Cas nucleases to a target sequence containing a PAM. To assess the binding event, a genetic circuit involving the LacI repressor and GFP reporter was engineered to ensure a positive, fluorescent readout of PAMs. The cells expressing GFP were sorted through fluorescence-activated cell sorting (FACS), and with subsequent plasmid preparations, library prep, and a next- generation sequencing (NGS) run, functional PAMs were identified. While this assay reduces the limitations from the previous *in vivo* assay, PAM-SCANR is limited by the requirement of cloning in the Cas nuclease and its gRNA of interest into the PAM-SCANR plasmid system,

identifying and mutating the catalytic domains of the Cas nuclease of interest, as well as the requirement of high library coverage.

Previously developed *in vitro* assays are based on cleaving PAM library plasmids within *in vitro* conditions followed by either a positive [35,36,38] or negative [16,37] screen for cleavage. Negative screens incubate a PAM library with purified Cas nucleases and *in vitro* transcribed gRNAs. After an appropriate reaction time, plasmids containing PAMs are cleaved, while plasmids with non-PAMs remain intact. This depletion of PAMs is measured by preparing a high-throughput sequencing library using PCR, which does not amplify cleaved sequences. Positive screens proceed similarly, except that adapters are ligated to the free ends of cleaved sequences for NGS. While these assays give the user more control of reaction conditions and do not require high library coverage, they require the cloning and protein purification of each Cas nuclease in the system and *in vitro* transcription of gRNAs, limiting their throughput.

3.1.2 - TXTL-based PAM determination

While the previously developed systematic (as opposed to bioinformatic) PAM assays each have their own limitations, a common disadvantage is the time required to perform them. Be they transformations, cultures, or protein purification, all of the previously developed assays involve time-intensive protocols, requiring weeks to months to complete.

To contribute to the characterization of CRISPR-Cas systems, we developed a PAM assay based on an *E. coli* cell-free transcription-translation system (TXTL) [49,50]. All TXTL systems are capable of *in vitro* expression of RNA and proteins in a single reaction (Figure 3.1B). Our TXTL platform has the added advantage of containing the native transcriptional and translational machinery as well as all proteins (e.g. RNase III) found in *E. coli* which allows for flexible use of various expression constructs and proper prokaryotic RNA processing. DNA that encodes for non-coding RNA and proteins can be added to TXTL and expressed within a few hours [51]. Compared to the previous PAM assays, this system has two key benefits: (1) expression of the necessary protein

and RNA components for the PAM assay and cleavage by the RNP complex occur in the TXTL reaction, which removes the need for protein purification; and (2) linear DNA can be used to express the necessary components, which largely eliminates the need for cloning.

To perform the PAM characterization assay with TXTL, plasmid or linear DNA expressing the necessary Cas nucleases and gRNAs are added to the TXTL mix. RNA expression and protein translation by the TXTL mix result in the formation of the RNP complex (Figure 3.1C). If linear DNA is used, a RecBCD inhibitor must be added to protect the DNA from degradation [52]. The gRNA is designed to target a library of plasmids containing a conserved sequence flanked by a randomized set of potential PAM sequences. Depletion of PAM sequences from the library is measured by adding the adapters and indices necessary for high-throughput sequencing using PCR to both the cleaved library and to a control library expressing a non-targeting gRNA.

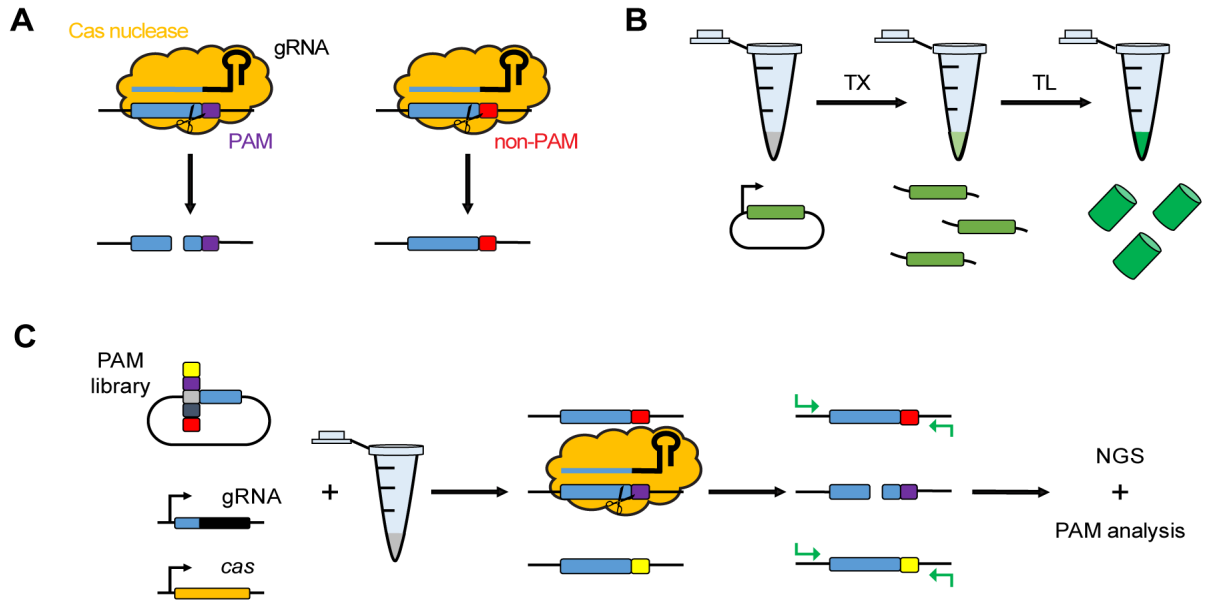


Figure 3.1: An overview of TXTL and its application for PAM determination. (A) Targeting by Cas nucleases. Targeting relies on a ribonucleoprotein complex composed of the Cas nuclease(s) and a guide RNA (gRNA) targeting DNA flanked by a PAM. Sequences lacking a PAM are ignored by the complex, even if the protospacer is perfectly complementary to the spacer portion of the gRNA. (B) Image showing the basic concept of TXTL. DNA encoding genes or non-coding RNA can be expressed in a single TXTL reaction. (C) The flow of a PAM characterization assay using TXTL. A plasmid library containing the randomized PAM sequence as well as DNA encoding a Cas nuclease and its gRNA can be added to a TXTL mix. After expression of the Cas nuclease and its gRNA, a ribonucleoprotein complex will be formed. Then the PAM libraries containing PAMs will be cleaved, while those containing a non-PAM PCR amplified with appropriate adapters and indices for analysis through an NGS platform. The PAM of the Cas nuclease will then be deciphered through a depletion analysis.

Here, we used this TXTL-based assay to rapidly characterize the PAM of a Cas9 originating from the bacterium *Neisseria meningitides* (NmeCas9). Our assay described here is improved relative to the assay described in Marshall and Maxwell et al. 2017 by the following: (1) using ten variable nucleotides flanking the protospacer in the randomized PAM library, (2) describing how to generate an internal control for Cas protein activity using a fluorescent reporter plasmid, and (3) describing how to use Sanger sequencing to verify Cas protein activity. The PAM of NmeCas9 was previously characterized [31,53,54]. We chose to demonstrate our assay using this system since it has an exceptionally long PAM (consensus NNNNGATT), which represents a 'worst-case scenario' for characterizing a new Cas nuclease. Because the PAM of NmeCas9 is known, we also showed additional control experiments that can be used to measure the kinetics of cleavage by the nuclease in TXTL, which is possible only when a PAM for the nuclease is known.

3.2 - MATERIALS AND METHODS

Here, we describe the detailed methods to perform a PAM determination assay for any Cas nuclease of interest. All DNA used in this procedure should be suspended in molecular biology-grade, nuclease-free H_2O (e.g. ThermoFisher, 10977-015) and not an elution buffer provided in DNA preparation kits unless specified. All plasmid DNA should be prepared using a midiprep (not a miniprep) kit, eluting in nuclease-free H_2O by following the protocol associated with the midiprep kit. We have found that preparing plasmid DNA using midiprep leads to more consistent experimental results. All oligos, plasmid DNA, and gene fragment sequences are reported in the appendix section.

3.2.1 - Creating DNA required for PAM assay

Materials

- Plasmid CSM160
- Gene fragment CSM-GB089
- Oligos CSMpr1308-1311
- Oligos Chi6.FWD/REV
- NEBuilder HiFi DNA Assembly Cloning Kit (NEB, E5520S)

- Midiprep kit (e.g. Genesee Scientific, 11-550)
- Q5 Hot Start High-Fidelity DNA Polymerase (NEB, M0493S)
- DpnI (20,000 units/ml; NEB, R0176)
- LB media (e.g. ThermoFisher, 10855-001)
- LB + chloramphenicol agar plates (e.g. bioWORLD, 30627011-1)
- LB supplemented with 25µg/ml chloramphenicol
- Chloramphenicol (e.g. MilliporeSigma, R4408)
- DNA purification kit (e.g. Genesee Scientific, 11-302)
- NEB 5-alpha electrocompetent cells (NEB, C2989K)
- DNA ladder (e.g. NEB, N3200S)

Optional materials

- Oligos T7.FWD/REV (optional for expressing Cas nucleases using linear DNA)
- Oligos to add the PAM-library protospacer flanked by a known PAM in the reporter plasmid
- Q5 Site-Directed Mutagenesis Kit (NEB, E0554S) to add the PAM-library protospacer flanked by a known PAM in the reporter plasmid
- Plasmid P70a-deGFP (Addgene, #40019) to create a reporter plasmid for rate-of-cleavage control reactions

Table 3.1: Thermocycler program used to anneal oligos to create Chi6 DNA.

Temperature (˚C)	Time (s)
95	300
Reduce temperature by 1˚C	15
Repeat 79x	
4-10	∞

Annealing Chi6 DNA

To successfully express linear DNA in TXTL, inhibitors of the RecBCD complex such as GamS or DNA containing Chi sites must be added [52]. The latter is particularly useful because it is inexpensive and easily obtained. For the experiments described below, we added Chi6 DNA whenever linear DNA was added to the reaction. To make a 100µM stock of Chi6 DNA, 50µl of 100µM Chi6.FWD and Chi6.REV oligos suspended in

nuclease-free H_2O were mixed in a PCR tube and annealed in a thermocycler using the program in Table 3.1.

Creating DNA to express Cas nucleases

To prepare DNA for expressing your Cas nuclease of interest, three options are available.

First, the Cas nuclease can be expressed from a bacterial expression vector with a constitutive or inducible promoter (pBAD18, pET, etc.). Use standard cloning techniques to generate these constructs. Prepare the plasmid DNA using a midiprep kit. Measure the concentration of the plasmid using standard techniques. If the concentration of the DNA is less than 20nM, the plasmid can be concentrated using a DNA purification kit.

Second, the Cas nuclease can be expressed from linear DNA encoding the Cas nuclease under the control of a T7 promoter through gene synthesis or through PCR amplification of the gene with an extended 5' primer. A T7 promoter is used because: (1) it is short enough to allow for straightforward PCR amplification, and (2) the T7 polymerase can transcribe mRNA efficiently from linear DNA. Cas nucleases could likely be expressed from linear DNA using a constitutive σ^{70} promoter, but we have not investigated this possibility.

Expressing Cas nucleases from linear DNA is useful for at least two reasons. (1) We have found that some Cas nucleases are almost impossible to clone into vectors containing a prokaryotic promoter suitable for expression in TXTL due to cell toxicity. However, we have been able to readily clone these proteins into vectors that lack promoters. (2) This method can also be used to prepare DNA for expressing the Cas nuclease in TXTL directly from genomic DNA containing the *cas* gene-of-interest.

To generate linear DNA with the Cas nuclease being expressed from a T7 promoter from either genomic DNA encoding the *cas* gene or from a plasmid with the *cas* gene lacking a promoter, design primers as shown in the appendix (see TJpr371/372). These

primers add the T7 promoter, Shine-Dalgarno sequence, and a T7 terminator to the 5' and 3' ends of the *cas* gene. These primers will need to be adapted by altering their binding sequence for each Cas nuclease of interest. Use these primers to perform PCR using Taq or the Q5 polymerase. Verify that the PCR yields a single bright band at the expected molecular weight using gel electrophoresis using a small sample of the PCR reaction. Purify the DNA from the remaining PCR reaction using a DNA purification kit and elute in nuclease-free H_2O.

To synthesize linear DNA expressing the Cas nuclease, design the gene fragment to include: (1) a binding site for CSMpr1105, (2) the T7 promoter, (3) a strong ribosomal binding site, (4) the sequence of the cas gene, and (5) a terminator, and (6) a binding site for CSMpr1106. Lyophilized gene fragments can be resuspended in nuclease-free H_2O to 20-40nM and added directly to TXTL. If a large number of reactions are needed, use CSMpr1105/1106 to PCR amplify the gene fragment using either Taq or Q5 polymerase. See the sequence "Cas-GB" in the appendix for a gene fragment expressing FnCas12a.

Note that if the Cas nuclease is expressed from a T7 promoter, the plasmid P70a-T7RNAP must be added to the TXTL mixture to a final concentration of 0.2nM as described below for expression from linear DNA. If expressed from an arabinose-inducible promoter, arabinose must be added to the TXTL mix to 20mM. If the Cas nuclease is repressed by LacI (as is the case for pET vectors), IPTG must be added to the TXTL mix to 0.5mM.

Creating DNA to express gRNAs

Similar to the Cas nucleases, gRNAs can be expressed from either plasmid or linear DNA. We generally express gRNAs from linear DNA, so we only describe that method in detail here. Otherwise, clone a spacer similar to the one in CSM-GB191 into a suitable gRNA expression vector of interest and prepare the plasmid DNA using a midiprep kit as above.

To express the gRNA, order a gene fragment (IDT, Eurofins, etc.) containing the following required elements: (1) a binding site for CSMpr1105, (2) a constitutive promoter (e.g. sigma 70 promoter), (2) a sequence expressing the appropriate ncRNA needed to target the Cas nuclease to the appropriate sequence in the randomized PAM library, (4) a terminator, (5) a binding site for CSMpr1106. An annotated sequence (CSM-GB191 and CSM-GB019) is provided in the appendix that was used to target Cas9 and FnCas12a, which use a 5' and 3' PAM, respectively. If it is unknown whether the Cas protein of interest recognizes a 5' or 3' PAM, gRNA sequences targeting both a 3' and 5' PAM should be designed.

Resuspend the gene fragment in nuclease-free H20 and normalize to 20nM. If a large number of reactions are needed, use CSMpr1105/1105 to amplify the gene fragment using either Taq or Q5 polymerase, then use a DNA purification kit to resuspend the amplified DNA in nuclease-free H_2O.

Along with the gRNA targeting the PAM library, another gRNA construct should be ordered as a negative control that does not target any sequence in the PAM library or the reporter construct. An annotated sequence (CSM-GB019) expressing a non-targeting SpyCas9 sgRNA is provided in the appendix.

Note that, if required, a tracrRNA can also be expressed in TXTL using a gene fragment with the same elements as those described above.

Creating a reporter plasmid

If a PAM that is recognized by a Cas nuclease is known (e.g. if at least one PAM can be identified bioinformatically), it is helpful to create a reporter plasmid to monitor the rate of cleavage by the Cas nuclease and to optimize expression conditions. Cleavage anywhere in a reporter plasmid in TXTL leads to quenching of the reporter due to RecBCD-mediated degradation [52]. By inserting the PAM library protospacer flanked by a known PAM upstream of the promoter driving the expression of GFP, the efficiency and rate of cleavage by the Cas nuclease can be determined (Figure 3.2A). Note that

the reporter plasmid can still be used even if the Chi6 RecBCD inhibitor is added because the inhibitor does not completely block the activity for RecBCD, and is only active for approximately 5 h [51,52]. Indeed, Figure 3.2A was generated using a gRNA and Cas nuclease expressed from linear DNA. Although it is likely that the reporter gene activity lags behind the actual cleavage of DNA (e.g. library cleavage), it still provides an estimate of cleavage time, especially for Cas nucleases that cleave DNA slowly in TXTL. In the method below, a reporter plasmid is added to each reaction in order to monitor the activity of the Cas nuclease as it cleaves the randomized PAM library.

We have used the Q5 Site-Directed Mutagenesis Kit to insert a protospacer and PAM recognized by the Cas nuclease of interest into the plasmid P70a-deGFP to create reporter plasmids. An example annotated plasmid sequence (pTJ247) that we created to monitor cleavage of NmeCas9 is provided in the appendix. To create similar plasmids, use the Q5 Site-Directed Mutagenesis Kit to insert the PAM-library protospacer sequence of the appropriate length (labelled “mut protospacer” in pTJ247) at position 247 in the attached P70a-deGFP plasmid map flanked by a PAM that can be recognized by the Cas nuclease. The primers used to create pTJ247 from p70a-deGFP using the Q5 Site-Directed Mutagenesis Kit are listed in the appendix (TJpr373/374).

Table 3.2: Components and thermocycler program used for PCR amplification using Q5 Hot Start High-Fidelity DNA Polymerase. An in-depth protocol is packaged with the polymerase.

Component	Volume (µl)	Temperature (˚C)	Time (s)
Nuclease-free H_2O	32.5	98	30
5x Q5 reaction buffer	10	98	10
10mM dNTPs	1	T_{anneal}	30
10µM FWD primer	2.5	72	t_{ext}
10µM REV primer	2.5	repeat 25x	
Template DNA	1	72	120
Q5 Hot Start High-Fidelity DNA Polymerase	0.5	4-10	∞

Creating the randomized PAM library

To create a randomized PAM library, we used the NEBuilder HiFi DNA Assembly Cloning Kit. We used restriction-enzyme free cloning to prevent any nucleotide biases in the randomized region that would result from cleavage by a restriction enzyme.

Note that this is the only step in our protocol that requires cloning, and that it only needs to be performed once in order to assay a large number of Cas nucleases. Purified 10N library plasmids are available from the authors on request, contingent on availability.

1. Set up two separate PCRs using Q5 Hot Start High-Fidelity DNA Polymerase (Example reaction conditions in Table 3.2).
    - Reaction 1: Template is 1ng of CSM160, primers CSMpr1308/1309, T_{anneal} = 72˚C, t_{ext} = 41s, 1.64kB expected product size.
    - Reaction 2: 10ng of CSM-GB089, primers CSMpr1310/1311, T_{anneal} = 72˚C, t_{ext} = 12s, 0.48kB expected product size.
2. Confirm that each reaction has a product of the appropriate size by running the PCR products on a 1% agarose gel with a DNA ladder.
3. Add 1µl of DpnI to the PCR product from Reaction 1, gently mix, and incubate at 37˚C for 1hr to digest the plasmid DNA.
4. Use the protocol from the NEBuilder HiFi DNA Assembly Cloning Kit to assemble the PCR products. For a negative control reaction, omit the PCR product generated from Reaction 2 to estimate the background.
5. Purify the two assembled reactions using a DNA purification kit and elute in as small of a volume of Tris-HCl pH 8 as possible. Transform 1µL of the purified reaction into 50µL of NEB 5-alpha electrocompetent cells using the associated protocol. Recover the cells according to the associated protocol.
6. After recovery of the cells, make 1:10 serial dilutions of the two reactions and plate on LB + Chloramphenicol agar to estimate the number of transformants.
7. Incubate the plates at 37˚C and back-dilute the remaining recovered cells (1:50) in LB + Chloramphenicol media and grow overnight at 37˚C.

8. The total transformants from the assembly should be approximately 1-2 million and a ~30,000:1 plasmid to background ratio is expected.
9. Make glycerol stocks of the library as desired.
10. Midiprep the remaining overnight culture for subsequent use in the PAM library cleavage assay described below.

Table 3.3: Recipe for TXTL master mix for PAM determination assay for Cas nucleases expressed from pET vectors. The master mix does not include DNA encoding the Cas nuclease and the gRNA. Note that if the PAM of the Cas nuclease of interest is unknown or if no reporter plasmid was constructed during step 2.1.4, then the reporter plasmid does not need to be added. Note depending on how the Cas nucleases are expressed, the IPTG and/or the p70a-T7RNAP may be omitted.

Component	Volume (µl)	Final Conc.	Stock Conc.
myTXTL extract	75		
Randomized PAM library	1	0.5nM	50nM
Reporter plasmid from step 2.1.4	1	5nM	50nM
p70a-T7RNAP	1	0.2nM	20nM
IPTG	1	0.5mM	50mM
Chi6 DNA	2	2µM	100µM

*Final concentration accounts for the extra volume from the addition of the DNA constructs encoding the Cas nuclease and the gRNA.

Table 3.4: An example set of reactions to characterize the PAM requirements of four Cas nucleases prepared using the master mix in Table 3.3. Each Cas nuclease is incubated with the PAM library and either a targeting or a non-targeting gRNA.

Reaction condition	DNA 1	DNA 2
1	Cas nuclease A	Targeting gRNA A
2	Cas nuclease A	Non-targeting gRNA
3	Cas nuclease B	Targeting gRNA B
4	Cas nuclease B	Non-targeting gRNA
5	Cas nuclease C	Targeting gRNA C
6	Cas nuclease C	Non-targeting gRNA
7	Cas nuclease D	Targeting gRNA D
8	Cas nuclease D	Non-targeting gRNA

3.2.2 - PAM library cleavage in TXTL

This section describes how to use TXTL to determine the PAM requirements of Cas nucleases. We assume that four Cas nucleases will be assayed and that each Cas nuclease is expressed from a pET vector. If other expression vectors are used, then the master mix should be adjusted appropriately to ensure that the *cas* gene is transcribed

(e.g. by adding arabinose if it is expressed from a pBAD vector). Note that if fewer than four Cas nucleases are being assayed, the master mix assembled in Step 3 can be re-frozen and thawed one time without loss of activity.

We assume that a reporter plasmid was constructed in Section 4.2.1 in order to provide an internal fluorometric control for Cas nuclease cleavage, and we assume that each Cas nuclease being assayed recognizes a protospacer of the same length and can recognize the PAM cloned into the reporter plasmid. This is true when assaying closely related Cas homologs. However, if the Cas nucleases being assayed do not recognize the same PAM, a different reporter plasmid should be added to each reaction expressing that Cas nuclease, and the TXTL recipe should be adjusted accordingly.

If a reporter plasmid is not used either because no PAM is known that is recognized by the Cas nuclease, or for some other reason, then the assay need not be carried out in a plate reader. We also assume that the Cas nuclease is active at 30˚C, and that 3nM of DNA expressing the Cas nuclease is sufficient for expression. See Section 4.2.4 for troubleshooting Cas nuclease cleavage in TXTL.

Materials

- myTXTL *In vitro* TX-TL protein expression kit [55] (Arbor Biosciences, 507096)
- 50nM randomized PAM library
- 30nM pET vector expressing the Cas nuclease of interest
- 20nM DNA expressing a gRNA targeting the protospacer in the PAM library
- 20nM DNA expressing a gRNA not targeting the protospacer in the PAM library
- 100 µM Chi6 DNA
- 96 well V-bottom plates (e.g. MilliporeSigma, CLS3357)
- Cap mat (e.g. MilliporeSigma, CLS3080)
- 20nM p70a-T7RNAP
- 50nM reporter plasmid
- 50mM IPTG (e.g. ThermoFisher, R1171)
- Fluorescent plate reader capable of measuring GFP (e.g. BioTek H1)

Protocol

1. Thaw a tube of myTXTL on ice. Ensure that all the liquid in the tube is at the bottom of the tube by briefly centrifuging if necessary.
2. Add the contents of Table 3 into a thawed TXTL tube to assemble the master mix. Mix by gently vortexing.
3. Aliquot 9.6μL of the master mix to eight separate Eppendorf tubes for each of the eight reactions described in Table 3.4.
4. Add 1.2μL each of the gRNA and Cas nuclease DNA as shown in Table 3.4 to each reaction. Mix each reaction by pipetting. This brings each reaction to a final volume of 12μL.
5. Carefully load two 5μL sub-reactions per reaction in the bottom of a 96 well V-bottom plate **(A)**.
6. Seal the plate using a cap map. Ensure that no well has a loose seal, which would lead to evaporation of the reactions or potentially damaging the plate reader.
7. Load the plate into a plate reader pre-warmed to 30˚C and measure *gfp* fluorescence (Ex 485nm, Em 528nM) kinetically every 10 minutes.
8. Incubate for up to 16h in the plate reader. **(B)**
9. Remove and pool the 5μL sub-reactions for each condition, and store the samples at -20˚C until ready for NGS library preparation.

(A) Differences in the geometry of the TXTL droplet at the bottom of the plate seem to have a large effect on kinetics of the TXTL reaction. Care should be taken to load the reactions onto the plate uniformly.

(B) Shorter incubation times can be used, although we have found that extended incubation does not affect the PAMs identified by the assay. The internal fluorescent control is helpful to decide how long to incubate the reaction if a shorter incubation time is desired. When the fluorescence in the wells containing the targeting gRNA stops increasing, but the fluorescence in the wells containing the non-targeting gRNA continues to increase, cleavage is complete and the TXTL reaction can be stopped.

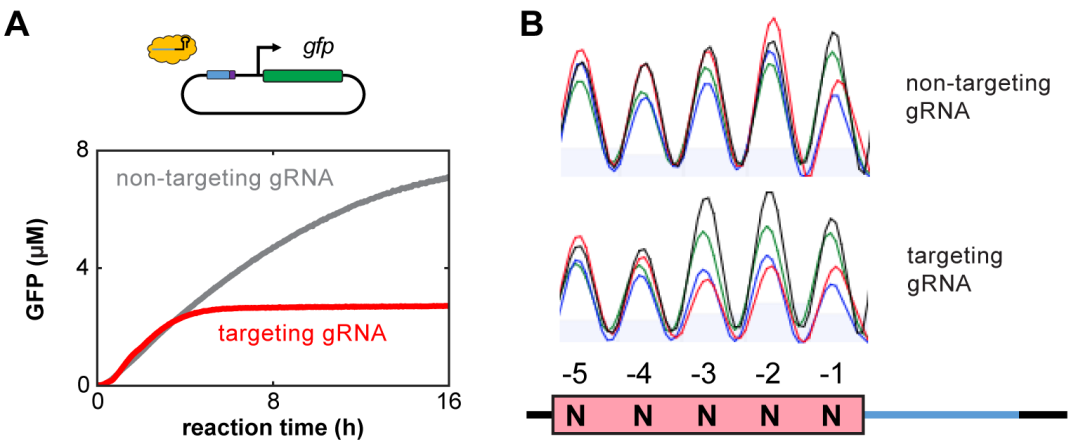


Figure 3.2. Data analysis for assessing cleavage by a Cas nuclease in TXTL. (A) A representative GFP cleavage analysis tracking fluorescence of GFP over time is shown. The red and gray lines indicate TXTL reactions containing a Cas nuclease with a gRNA targeting the GFP reporter or a non-targeting gRNA, respectively. At earlier times, GFP fluorescence is similar for both reactions. After a few hours, the reaction containing the GFP-targeting gRNA cleaves the reporter plasmid, and GFP production ceased, while in the reaction containing the non-targeting gRNA GFP continues being produced for $16 h. (B) An example comparison between two trace files obtained from Sanger sequencing. The images show chromatograms obtained by PCR amplifying directly from two TXTL reactions. Each reaction was conducted with the FnCas12a nuclease, a 5-nt randomized PAM library, and either the non-targeting (top) or targeting (bottom) gRNA. Cleavage with the targeting gRNA shows noticeable depletion of nucleotides C and T in the 2 and 3 positions relative to cleavage with the non-targeting gRNA. The inconsistent nt abundance in the non-targeting gRNA sample can be attributed to varying frequencies of each sequence in the PAM library.

3.2.3 - Assessing cleavage of the PAM library

Prior to the NGS library preparation, we recommend Sanger sequencing to evaluate Cas nuclease cleavage (Figure 3.2B). This is a quick and inexpensive way to verify that the Cas nuclease was expressed and is active in TXTL. It also provides an estimate of what PAMs are recognized by the Cas nuclease. Note that this provides another method to assess Cas nuclease activity in TXTL if a reporter plasmid cannot be constructed. If the Cas nuclease was expressed and active in TXTL, then a depletion of particular nucleotides will be observed in the reaction expressing the targeting gRNA relative to the reaction expressing the non-targeting gRNA.

Materials

- Q5 Hot Start High-Fidelity DNA Polymerase
- 100bp ladder (e.g. NEB, N3231S)
- Oligos TJpr416/417
- DNA purification kit

Protocol

1. Make a 1:10 dilution of each TXTL reaction in nuclease-free H_2O.
2. Set up a PCR reaction for each TXTL reaction using the standard Q5 Polymerase protocol (see Table 3.2 for an example) using primers TLpr416/TLpr417, T_{anneal} = 67˚C, t_{ext} = 20s, using the 1:10 diluted TXTL reaction as the template DNA, and supplemented with 4% DMSO (e.g. 1µl for a 25µl reaction).
3. Run the PCR product(s) on a 1% agarose gel by mixing 2µl 6x loading dye, 5µl nuclease-free H_2O, and 5µl PCR product on parafilm and loading 10µl on the gel to ensure amplification (~500bp product).
4. Sanger sequence the resulting amplicons using TJpr416 or TJpr417 as the sequencing primer.
5. Align the trace files (.ab1) from the Sanger sequencing of the TXTL reactions containing the targeting and non-targeting gRNA with the PAM library using Benchling or similar software.
6. Examine the aligned trace files by zooming in to the randomized sequence flanking the protospacer. Successful cleavage leads to a less random distribution of some nucleotides in the randomized area in the targeting gRNA control. An example trace file of a library cleaved by a Cas12a homolog can be seen in Figure 3.2.

3.2.4 - Troubleshooting Cas nuclease cleavage in TXTL

If no cleavage is observed using either the reporter plasmid or Sanger sequencing, first check that all of the expression constructs are correct. If the expression constructs are correct, there are a few parameters that can be tuned to improve the expression of the Cas nuclease. Troubleshooting is greatly facilitated by having a fluorometric readout.

- Double or quadruple the concentration of the DNA expressing the Cas nuclease.
- Double or quadruple the concentration of the DNA expressing the gRNA.
- Try increasing the temperature of the reaction to 37˚C or 42˚C.

- Some enzymes require supplementation with cations such as Mg^{2+} or Ca^{2+}. Try supplementing these trace elements to the TXTL reaction.

Out of the ~20 Cas nucleases we have assayed with this technique, we have only found two with which we were unable to observe DNA cleavage in TXTL and that did not respond to the optimization techniques above. It could be that these Cas nucleases cannot cleave DNA, or it could be that this method is broadly, but not universally applicable.

3.2.5 - NGS library preparation

This section describes the preparation of the NGS libraries from the cleaved PAM libraries. Oligos RL133/134 add Illumina Nextera adapters to the PAM library adjacent to the area of the randomized PAM sequence. These primers were previously used by Leenay and co-workers [34]. This method could be improved by incorporating random barcodes in these primers to enable the identification of PCR duplicates, but we have not found this to be necessary to yield accurate results.

Materials

- TXTL reaction(s) of interest
- Q5 Hot Start High-Fidelity DNA Polymerase
- Oligos RL133/134/135
- Oligos TJpr418/419/422/423
- Nextera indexing primers
- AMPureXP beads (Fisher Scientific, NC9933872)
- 100bp DNA ladder
- Magnetic rack (e.g. Thermofisher, AM10027)

Optional Materials

- Additional Illumina index oligos if analyzing more than two reaction conditions.

Protocol: Add Nextera adapters to the cleaved PAM libraries and controls.

1. Set up 50µl PCR(s) using the standard Q5 Polymerase protocol (see Table 3.2 for an example) using primers RL133/RL134, T_{anneal} = 62.9˚C, t_{ext} = 7s, using the 1:10 diluted TXTL sample as the template DNA, and supplemented with 2µl DMSO.
2. On parafilm, mix 2µl 6x loading dye and 10µl PCR product and load onto a 1% agarose gel.
3. Run the gel with 10µl 100bp ladder at 100V until dye travels ~80% of the gel.
4. Image the gel with a transilluminator. You should see a strong 200bp band.
5. Clean up the PCR product(s):
    i. Vortex the AMPureXP beads.
    ii. Mix 40µL beads with the PCR product(s) and mix well by pipetting up and down.
    iii. Incubate the mixture(s) at room temperature for 5min.
    iv. Place the PCR tube(s) onto a magnetic rack and wait 2min or until the supernatant is clear.
    v. With the tube(s) remaining on the magnetic rack, use a pipette to remove the clear supernatant.
    vi. Wash the beads by adding 200µl of fresh 80% ethanol to the tube(s) and wait 30s. Remove the supernatant and repeat the wash step.
    vii. After repeating wash step, remove the supernatant and let the sample(s) air dry for 10min.
    viii. Remove the PCR tube rack from the magnetic rack, and elute the DNA from the beads by adding 45µl of 10mM Tris-HCl pH 8 to the tube(s) and mixing well.
    ix. Incubate the sample(s) for 2min, then place the PCR tube(s) back onto the magnetic tube rack and allow the beads to separate from the liquid for additional 2min.
    x. Transfer 40µl of the supernatant to a new tube ensuring that no beads are transferred.

xi. Nanodrop the sample(s). The average concentration should be ~12ng/µl.

Table 3.5: Components and thermocycler program used for the final PCR amplification of the PAM library using Q5 Hot Start High-Fidelity DNA Polymerase. Note the cycle number is reduced from 25 to 8 to reduce errors during amplification.

Component	Volume (µl)	Temperature (˚C)	Time (s)
Nuclease-free H_2O	28.5	98	30
5x Q5 reaction buffer	10	98	10
10mM dNTPs	1	67	20
10µM FWD Nextera index (i5)	2.5	72	10
10µM REV Nextera index (i7)	2.5	repeat 8x	
purified PCR product	5	72	120
Q5 Hot Start High-Fidelity DNA Polymerase	0.5	4-10	∞

Protocol: Add Illumina indices and binding domains

With the Nextera adapters on the PAM library, Illumina indices and binding domains can be attached using PCR. To run more than one library on a single sequencing lane, a unique pair of primers must be used for each sample in order to attach a unique pair of i5 and i7 indices to each sample. A list of the Nextera indices and the oligos that they are embedded within are provided by Illumina in their technical support document "Illumina Adapter sequences" under the file heading "Illumina Nextera Adapters." The two pairs of indexing oligos from this table that we used to elucidate the NmeCas9 PAM are provided in the appendix (RL135 (i5) and TJpr422/423 (i7)).

1. Add indices to the library using PCR using the recipe and reaction conditions in Table 3.5. Use a unique pair of i5 and i7 primers for each reaction.
2. Clean the second PCR product(s) using the same protocol as in the first cleanup, except for the following changes:
    i. Mix 56µl beads with the PCR product(s) at step (2)
    ii. Elute with 25µl 10mM Tris-HCl and transfer 20µl to a new tube at steps (8-9)
3. Measure the final sample concentration(s) using a Nanodrop or Qubit. The average concentration(s) should be ~40ng/µl.

4. Based on the measured concentrations, run ~100ng of total DNA on a 1% gel (2µL 6x loading dye, 100ng PCR product, up to 10µL nuclease-free H_2O) with 10µl 100bp ladder. Run at 100V until dye travels ~80% of gel **(C)**.
5. Image the gel. The band should be at the 300bp marker, and the intensity should be comparable to the band corresponding to the 100ng marker.
6. Prepare the sample(s) to the concentrations required for sequencing (e.g. NCSU Sequencing Facility requires 10nM in 20µL). ***Note:*** To sequence low-complexity libraries, a high-diversity library must generally be spiked in during sequencing. Our sequencing libraries were spiked with 15% PhiX libraries before sequencing.

(C) For a more accurate measurement of DNA concentration, a Qubit (e.g. ThermoFisher, A30225) can also be used if available.

3.2.6 - Counting PAMs and calculating depletion

We have created a simple Python script to extract the NGS reads from PAM-SCANR libraries [34] and libraries constructed according to the method outlined above. The script looks for reads that contain the sequence flanking the variable nucleotides in the PAM library and restricts the counts to reads that have a threshold quality score at each of the variable nucleotides. The code is available as a public repository (see appendix). In addition, we have included a copy of the script in the appendix of this protocol. To use this script, follow the instructions in the readme file contained in the repository.

To calculate the depletion of each sequence in the PAM library for a Cas nuclease, we first restricted ourselves to measuring sequences that had at least 50 reads in the non-targeting control. We then normalized the counts of each sequence by the total number of reads in each library. We then calculated the quotient of counts for each sequence in the targeting library and the non-targeting library. Example R code is included in the repository along with sample output of the counting script and sample plots that can be used to represent the output of the script. PAM depletion data can also be represented using a PAM wheel as described in [10].

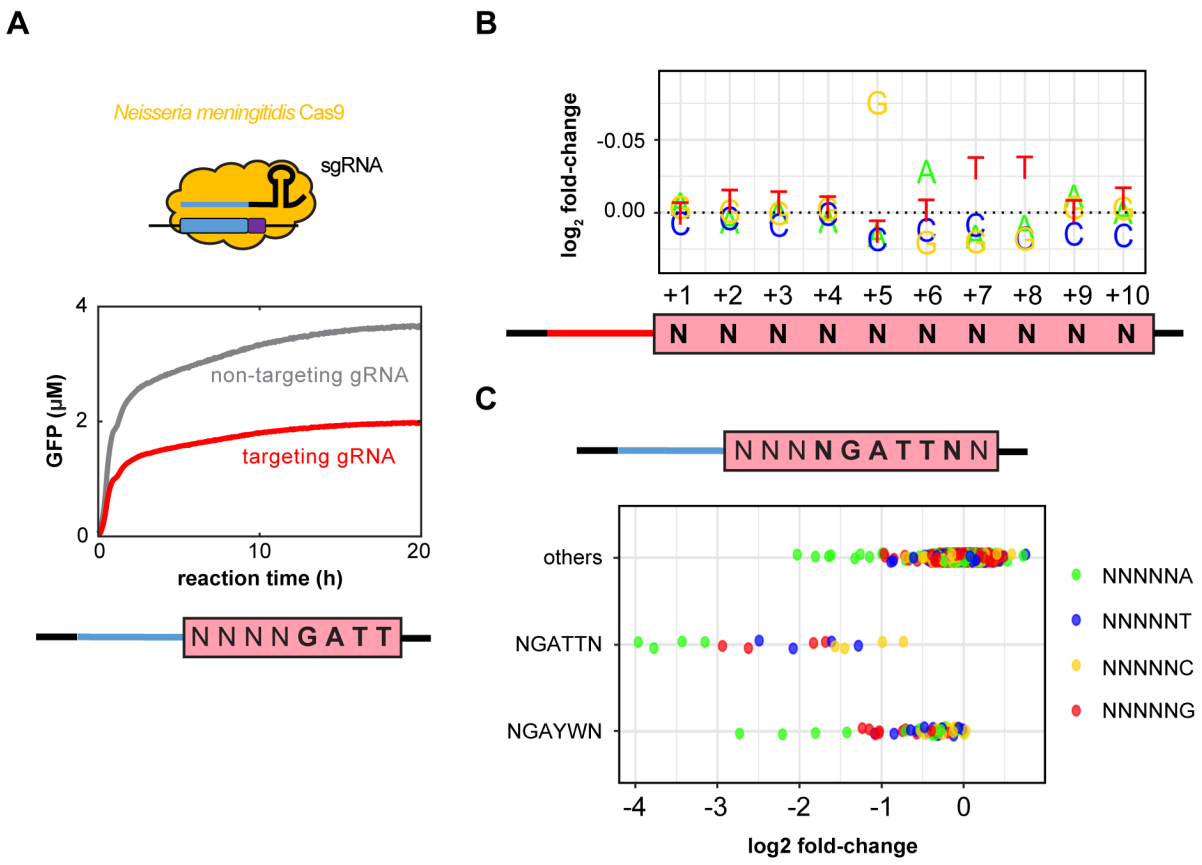


Figure 3.3: Data analysis for the TXTL-based PAM assay performed on the Cas9 from Neisseria meningitides (NmeCas9) expressed from linear DNA. (A) Fluorescence data from the GFP cleavage assay of NmeCas9 using an NNNNGATT PAM. The red and gray lines indicate TXTL reactions containing NmeCas9 with a gRNA targeting the reporter plasmid pTJ247 or a non-targeting gRNA, respectively. Both reactions had similar GFP expression rates, but after $1 h, the reaction containing the targeting gRNA began to show a decrease in GFP production. (B) Plot showing the log2-fold change in the nucleotide frequency across the 10-nucleotide PAM library. The depletion shows the strong NNNNGATTNN with a slight bias of D in the 9th and 10th positions. Note the inverted y-axis. (C) The average fold-change of 10-mers for selected PAM motifs in the 5th-10th position is shown. While a NNNNGATTNN PAM was strong, having an A in the 9th position increases cleavage efficiency compared to the other nucleotides, and having a T in the 10th position slightly increased cleavage for NNNNGATTNN PAMs.

3.3 - RESULTS AND DISCUSSION

To demonstrate our assay, we determined the PAM requirements of NmeCas9 expressed from linear DNA. We obtained linear DNA expressing NmeCas9 by adding a T7 promoter and terminator to the NmeCas9 sequence on the plasmid DS-Nmcas. The PAM for this particular Cas9 has previously been characterized [31,53,54]. These studies found a consensus PAM for NmeCas9 as NNNNGATT. We created a reporter plasmid for NmeCas9 by placing the, NNNNGATT PAM adjacent to the PAM library protospacer (see pTJ247 for NNNN sequence). Other PAMs with similar cleavage efficiency could have been used [29].

The GFP cleavage data from NmeCas9 (Figure 3.3) shows that the NmeCas9 RNP formed and began cleavage of the reporter plasmid by ~1hr. In contrast to previous Cas nucleases we have characterized [51], NmeCas9 did not completely cleave the reporter plasmid; fluorescence continued to increase even in the targeting gRNA condition throughout the experiment. Accordingly, we did not observe any obvious differences between the targeting and non-targeting gRNA conditions by Sanger sequencing. While an incubation time of ~10hrs would most likely have been sufficient, we opted for a longer reaction time to ensure adequate cleavage of the PAM library.

We counted the reads mapping to each of the ten variable nucleotides in the PAM library using the script 'count_pams.py' with the flags '-n 10' and '-q 10'. The former flag tells the script to look for ten variable nucleotides. The second relaxes the phred score needed at each of the ten variable nucleotides from the default 30 to 10. We found that by relaxing the quality score we were able to increase the signal of the canonical NmeCas9 PAM because otherwise a large number of reads were rejected. This is because there were twice as many variable nucleotides in the library we used compared to previous libraries with only five variable nucleotides, which means that a phred score of 30 was too strict. We observed similar results with a Phred score of 20 cutoff, indicating that this analysis is robust to the cut-off used.

The PAM analyses for NmeCas9 showed the expected NNNNGATT PAM (see Figure 3.3B-C). The fold-change plot (Figure 3.3B) shows the nucleotide frequency at each position of the 10N library. From this plot, the efficient NNNNGATT PAM was recapitulated, with the weaker PAMs of Y and W in the 7^{th} and 8^{th} position, respectively [29]. This plot also shows a bias of D in the 9^{th} and 10^{th} PAM position.

We calculated the average fold-change across all 10-mers matching selected motifs in the 5^{th} through the 10^{th} position (Figure 3.3C). For the NNNNGATTNN motif, an 'A' in the 9^{th} position increased cleavage efficiency compared to the other nucleotides. A 'T' in the 10^{th} position led to the greatest cleavage for all PAMs that started with GATT. The NmeCas9 motif was relatively specific; of the motifs that matched NNNNGAYWNN, only

NNNNGACTNN supported any measurable cleavage. Note that we only performed a single replicate of this PAM assay. If additional confidence that these subtle preferences reflect the real preferences of the nuclease, additional replicates would be warranted. Nevertheless, our results are consistent with previous work that observed similar PAMs when targeting DNA in mammalian cells [29]. We note that the subtle PAM preferences for NmeCas9 could be quantitatively assessed by cloning different PAMs flanking the NmeCas9 spacer in the reporter plasmid and repeating the reporter gene cleavage assay.

Though TXTL can rapidly characterize CRISPR-Cas systems, it does have some limitations. First, the temperature requirement for TXTL reactions is between 25-42˚C. While this range allowed for efficient cleavage using NmeCas9 and a number of other Cas nucleases we have worked with, this limit would restrict the analysis of other CRISPR-Cas systems that may require a higher temperature for efficient expression and/or cleavage. A TXTL environment also restricts the analysis of gRNA functionality in eukaryotic systems as it lacks the biological machinery native to those organisms (post-transcriptional/translational modifications, etc.). Finally, TXTL reactions lose activity after ~16 hours. This may limit the use our assay for Cas nucleases that are weakly expressed or are unstable in TXTL.

3.4 - CONCLUSIONS

Previous PAM characterization efforts have been hindered by the time required to perform existing assays. We have developed a method to characterize the PAM requirements of Cas nucleases that is much faster than previous methods. As myTXTL is commercially available for use, any lab or institution has access to use this powerful tool to characterize CRISPR-Cas systems. TXTL can also be used for other applications such as protein expression or, potentially, assessing gRNA activity.

3.4 - ACKNOWLEDGMENTS

We thank Erik Sontheimer for providing the plasmid encoding the NmeCas9.

3.5 - REFERENCES

[1] J.E. Garneau, M.È. Dupuis, M. Villion, D.A. Romero, R. Barrangou, P. Boyaval, C. Fremaux, P. Horvath, A.H. Magadán, S. Moineau, The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA, Nature. 468 (2010) 67–71.

[2] S.J. Brouns, M.M. Jore, M. Lundgren, E.R. Westra, R.J. Slijkhuis, A.P. Snijders, M.J. Dickman, K.S. Makarova, E.V. Koonin, J. van der Oost, Small CRISPR RNAs guide antiviral defense in prokaryotes, Science. 321 (2008) 960–964.

[3] H.P. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, CRISPR provides acquired resistance against viruses in prokaryotes, Science. 315 (2007) 1709–1712.

[4] L.A. Marraffini, E.J. Sontheimer, CRISPR interference limits horizontal gene transfer in *Staphylococci* by targeting DNA, Science. 322 (2008) 1843–1845.

[5] M.P. Terns, R.M. Terns, CRISPR-based adaptive immune systems, Curr Opin Microbiol. 14 (2011) 321–327.

[6] M.L. Luo, A.S. Mullis, R.T. Leenay, C.L. Beisel, Repurposing endogenous type I CRISPR-Cas systems for programmable gene repression, Nucleic Acids Res. 43 (2015) 674–681.

[7] P. Mali, L. Yang, K.M. Esvelt, J. Aach, M. Guell, J.E. DiCarlo, J.E. Norville, G.M. Church, RNA-guided human genome engineering via Cas9, Science. 339 (2013) 823–826.

[8] L. Cong, F.A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P.D. Hsu, X. Wu, W. Jiang, L.A. Marraffini, F. Zhang, Multiplex genome engineering using CRISPR/Cas systems, Science. 339 (2013) 819–823.

[9] B. Chen, L.A. Gilbert, B.A. Cimini, J. Schnitzbauer, W. Zhang, G.W. Li, J. Park, E.H. Blackburn, J.S. Weissman, L.S. Qi, B. Huang, Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system, Cell. 155 (2013) 1479–1491.

[10] R.T. Leenay, C.L. Beisel, Deciphering, communicating, and engineering the CRISPR PAM, J Mol Biol. 429 (2017) 177–191.

[11] J.M. Rouet P, Smih F, Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease, Mol Cell Biol. 14 (1994) 8096–8106.

[12] F.D. Urnov, E.J. Rebar, M.C. Holmes, H. Steve Zhang, P.D. Gregory, Genome editing with engineered zinc finger nucleases, Nat Rev Genet. 11 (2010) 636–646.

[13] J.K. Joung, J.D. Sander, TALENs: A widely applicable technology for targeted genome editing, Nat Rev Mol Cell Biol. 14 (2013) 49–55.

[14] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J.A. Doudna, E. Charpentier, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity, Science. 337 (2012) 816–821.

[15] R. Barrangou, J.A. Doudna, Applications of CRISPR technologies in research and beyond, Nat Biotechnol. 34 (2016) 933–941.

[16] B. Zetsche, J.S. Gootenberg, O.O. Abudayyeh, A. Regev, E. V Koonin, F. Zhang, I.M. Slaymaker, K.S. Makarova, P. Essletzbichler, S.E. Volz, J. Joung, J. Van Der Oost, Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system, Cell. 163 (2015) 759–771.

[17] E. V. Koonin, K.S. Makarova, F. Zhang, Diversity, classification and evolution of CRISPR-Cas systems, Curr Opin Microbiol. 37 (2017) 67–78.

[18] I. Grissa, G. Vergnaud, C. Pourcel, The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats, BMC Bioinformatics. 8 (2007) 172.

[19] O.O. Abudayyeh, J.S. Gootenberg, S. Konermann, J. Joung, I.M. Slaymaker, D.B.T. Cox, S. Shmakov, K.S. Makarova, E. Semenova, L. Minakhin, K. Severinov, A. Regev, E.S. Lander, E. V. Koonin, F. Zhang, C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector, Science. 353 (2016) aaf5573.

[20] B. Zetsche, M. Heidenreich, P. Mohanraju, I. Fedorova, J. Kneppers, E.M. Degennaro, N. Winblad, S.R. Choudhury, O.O. Abudayyeh, J.S. Gootenberg, W.Y. Wu, D.A. Scott, K. Severinov, J. Van Der Oost, F. Zhang, Multiplex gene editing by CRISPR-Cpf1 using a single crRNA array, Nat Biotechnol. 35 (2017) 31–34.

[21] E.R. Westra, P.B.G. van Erp, T. Künne, S.P. Wong, R.H.J. Staals, C.L.C. Seegers, S. Bollen, M.M. Jore, E. Semenova, K. Severinov, W.M. de Vos, R.T. Dame, R. de Vries, S.J.J. Brouns, J. van der Oost, CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3, Mol Cell. 46 (2012) 595–605.

[22] B.P. Kleinstiver, S.Q. Tsai, M.S. Prew, N.T. Nguyen, M.M. Welch, J.M. Lopez, Z.R. McCaw, M.J. Aryee, J.K. Joung, Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells, Nat Biotechnol. 34 (2016) 869–874.

[23] S. Shmakov, O.O. Abudayyeh, K.S. Makarova, Y.I. Wolf, J.S. Gootenberg, E. Semenova, L. Minakhin, J. Joung, S. Konermann, K. Severinov, F. Zhang, E. V. Koonin, Discovery and functional characterization of diverse class 2 CRISPR-Cas systems, Mol Cell. 60 (2015) 385–397.

[24] A. East-Seletsky, M.R. O'Connell, S.C. Knight, D. Burstein, J.H.D. Cate, R. Tjian, J.A. Doudna, Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection, Nature. 538 (2016) 270–273.

[25] S.H. Sternberg, S. Redding, M. Jinek, E.C. Greene, J.A. Doudna, DNA interrogation by the CRISPR RNA-guided endonuclease Cas9, Nature. 507 (2014) 62–67.

[26] F.J.M. Mojica, C. Díez-Villaseñor, J. García-Martínez, C. Almendros, Short motif sequences determine the targets of the prokaryotic CRISPR defense system, Microbiology. 155 (2009) 733–740.

[27] S.A. Shah, S. Erdmann, F.J.M. Mojica, R.A. Garrett, Protospacer recognition motifs: Mixed identities and functional diversity, RNA Biol. 10 (2013) 891–899.

[28] X.H. Zhang, L.Y. Tee, X.G. Wang, Q.S. Huang, S.H. Yang, Off-target effects in CRISPR/Cas9-mediated genome engineering, Mol Ther Nucleic Acids. 4 (2015) e264.

[29] C.M. Lee, T.J. Cradick, G. Bao, The *Neisseria meningitidis* CRISPR-Cas9 system enables specific genome editing in mammalian cells, Mol Ther. 24 (2016) 645–654.

[30] J. Elmore, T. Deighan, J. Westpheling, R.M. Terns, M.P. Terns, DNA targeting by the type I-G and type I-A CRISPR-Cas systems of *Pyrococcus furiosus*, Nucleic Acids Res. 43 (2015) 10353–10363.

[31] K.M. Esvelt, P. Mali, J.L. Braff, M. Moosburner, S.J. Yaung, G.M. Church, Orthogonal Cas9 proteins for RNA-guided gene regulation and editing, Nat Methods. 10 (2013) 1116–1121.

[32] W. Jiang, D. Bikard, D. Cox, F. Zhang, L.A. Marraffini, RNA-guided editing of bacterial genomes using CRISPR-Cas systems, Nat Biotechnol. 31 (2013) 233–239.

[33] B.P. Kleinstiver, M.S. Prew, S.Q. Tsai, V. V. Topkar, N.T. Nguyen, Z. Zheng, A.P.W. Gonzales, Z. Li, R.T. Peterson, J.R.J. Yeh, M.J. Aryee, J.K. Joung, Engineered CRISPR-Cas9 nucleases with altered PAM specificities, Nature. 523 (2015) 481–485.

[34] R.T. Leenay, K.R. Maksimchuk, R.A. Slotkowski, R.N. Agrawal, A.A. Gomaa, A.E. Briner, R. Barrangou, C.L. Beisel, Identifying and visualizing functional PAM diversity across CRISPR-Cas systems, Mol Cell. 62 (2016) 137–147.

[35] T. Karvelis, G. Gasiunas, J. Young, G. Bigelyte, A. Silanskas, M. Cigan, V. Siksnys, Rapid characterization of CRISPR-Cas9 protospacer adjacent motif sequence elements, Genome Biol. 16 (2015) 253.

[36] V. Pattanayak, S. Lin, J.P. Guilinger, E. Ma, J.A. Doudna, D.R. Liu, High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity, Nat Biotechnol. 31 (2013) 839–843.

[37] H. Hirano, J.S. Gootenberg, T. Horii, O.O. Abudayyeh, M. Kimura, P.D. Hsu, T. Nakane, R. Ishitani, I. Hatada, F. Zhang, H. Nishimasu, O. Nureki, Structure and Engineering of *Francisella novicida* Cas9, Cell. 164 (2016) 1–12.

[38] F.A. Ran, L. Cong, W.X. Yan, D.A. Scott, J.S. Gootenberg, A.J. Kriz, B. Zetsche, O. Shalem, X. Wu, K.S. Makarova, E. V. Koonin, P.A. Sharp, F. Zhang, In vivo genome editing using *Staphylococcus aureus* Cas9, Nature. 520 (2015) 186–191.

[39] T. Karvelis, G. Gasiunas, V. Siksnys, Methods for decoding Cas9 protospacer adjacent motif (PAM) sequences: A brief overview, Methods. 121–122 (2017) 3–8.

[40] A. Bolotin, B. Quinquis, A. Sorokin, S. Dusko Ehrlich, Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin, Microbiology. 151 (2005) 2551–2561.

[41] F.J.M. Mojica, C. Díez-Villaseñor, J. García-Martínez, E. Soria, Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements, J Mol Evol. 60 (2005) 174–182.

[42] C. Pourcel, G. Salvignol, G. Vergnaud, CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies, Microbiology. 151 (2005) 653–663.

[43] G.W. Tyson, J.F. Banfield, Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses, Environ Microbiol. (2008) 200–207.

[44] S. van Houte, A. Buckling, E.R. Westra, Evolutionary ecology of prokaryotic immune mechanisms, Microbiol Mol Biol Rev. 80 (2016) 745–763.

[45] W.E. England, R.J. Whitaker, Evolutionary causes and consequences of diversified CRISPR immune profiles in natural populations., Biochem Soc Trans. 41 (2013) 1431–1436.

[46] L.M. Childs, W.E. England, M.J. Young, J.S. Weitz, R.J. Whitaker, CRISPR-induced distributed immunity in microbial populations, PLoS One. 9 (2014) e101710.

[47] S.A. Shmakov, V. Sitnik, K.S. Makarova, Y.I. Wolf, K. V Severinov, E. V Koonin, The CRISPR spacer space is dominated by sequences from species-specific mobilomes, MBio. 8 (2017) e01397-17.

[48] C. Xue, A.S. Seetharam, O. Musharova, K. Severinov, S.J.J. Brouns, A.J. Severin, D.G. Sashital, CRISPR interference and priming varies with individual spacer sequences, Nucleic Acids Res. 43 (2015) 10831–10847.

[49] J. Garamella, R. Marshall, M. Rustad, V. Noireaux, The all *E. coli* TX-TL toolbox 2.0: A platform for cell-free synthetic biology, ACS Synth Biol. 5 (2016) 344–355. doi:10.1021/acssynbio.5b00296.

[50] J. Shin, V. Noireaux, An *E. coli* cell-free expression toolbox: Application to synthetic gene circuits and artificial cells, ACS Synth Biol. 1 (2012) 29–41.

[51] R. Marshall, C. Maxwell, S.P. Collins, T. Jacobsen, M.L. Luo, M.B. Begemann, B.N. Gray, E. January, A. Singer, Y. He, C.L. Beisel, V. Noireaux, Rapid and scalable characterization of CRISPR technologies using an *E. coli* cell-free transcription-translation system, Mol Cell. 69 (2018) 146-157.

[52] R. Marshall, C.S. Maxwell, S.P. Collins, C.L. Beisel, V. Noireaux, Short DNA containing chi sites enhances DNA stability and gene expression in *E. coli* cell-free transcription–translation systems, Biotechnol Bioeng. 114 (2017) 2137–2141.

[53] I. Fonfara, A. Le Rhun, K. Chylinski, K.S. Makarova, A.L. Lécrivain, J. Bzdrenga, E. V. Koonin, E. Charpentier, Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems, Nucleic Acids Res. 42 (2014) 2577–2590.

[54] Z. Hou, Y. Zhang, N.E. Propson, S.E. Howden, L.-F. Chu, E.J. Sontheimer, J.A. Thomson, Efficient genome engineering in human pluripotent stem cells using Cas9 from *Neisseria meningitidis*, PNAS. 110 (2013) 15644–15649.

[55] Z.Z. Sun, C.A. Hayes, J. Shin, F. Caschera, R.M. Murray, V. Noireaux, Protocols for implementing an *Escherichia coli* based TX-TL cell-free expression system for synthetic biology, J Vis Exp. (2013) e50762.

CHAPTER 4: The *Acidaminococcus* sp. Cas12a nuclease recognizes GTTV and GCTV as non-canonical PAMs

Thomas Jacobsen, Chunyu Liao, and Chase L. Beisel

Original publication:

T. Jacobsen, C. Liao, C.L. Beisel, The *Acidaminococcus* sp. Cas12a nuclease recognizes GTTV and GCTV as non-canonical PAMs, FEMS Microbiol Lett. In press (2019).

ABSTRACT

The clustered regularly interspaced short palindromic repeat (CRISPR)-associated (Cas) nuclease *Acidaminococcus sp*. Cas12a (AsCas12a, also known as AsCpf1) has become a popular alternative to Cas9 for genome editing and other applications. AsCas12a has been associated with a TTTV protospacer-adjacent motif (PAM) as part of target recognition. Using a cell-free transcription-translation (TXTL)-based PAM screen, we discovered that AsCas12a can also recognize GTTV and, to a lesser degree, GCTV motifs. Validation experiments involving DNA cleavage in TXTL, plasmid clearance in *E. coli*, and indel formation in mammalian cells showed that AsCas12a was able to recognize these motifs, with the GTTV motif resulting in higher cleavage efficiency compared to the GCTV motif. We also observed that the -5 position influenced the activity of DNA cleavage *in vitro* and in *E. coli*, with a C at this position resulting in the lowest activity. Together, these results show that wild-type AsCas12a can recognize non-canonical GTTV and GCTV motifs and exemplify why the range of PAMs recognized by Cas nucleases are poorly captured with a consensus sequence.

4.1 - INTRODUCTION

Clustered regularly interspaced short palindromic repeat (CRISPR)-CRISPR-associated (Cas) systems have become widespread tools used for various biotechnological applications [1,2]. In nature, these systems serve as adaptive immune systems that protect prokaryotes from invasion of mobile genetic elements by targeting and cleaving foreign DNA or RNA [3–6]. Cleavage is directed by non-coding guide RNAs (gRNAs) which possess a guide sequence that is complementary to a target sequence (known as the protospacer). These gRNAs complex with its associated Cas nuclease and bind to the target sequence, leading to cleavage and/or degradation of these targets [5–7]. The ease of reprogramming Cas nucleases and gRNAs has led to advancements in various fields, such as gene-editing, gene regulation, and diagnostics [8,9].

To successfully utilize Cas nucleases, each nuclease must be thoroughly characterized. This includes characterizing its gRNA structure, cleavage pattern (i.e. blunt vs. staggered), ability to tolerate mismatches across the target sequence, and other factors required for gRNA processing and DNA/RNA cleavage [10–15]. Arguably one of the most important requirements is determining the protospacer-adjacent motif (PAM), a short sequence comprised of 3-8 nucleotides that are initially recognized by the nuclease prior to assessment of base pairing between the guide and the target [16,17]. The PAM requirement is key for target recognition and cleavage [10,18] and also helps to prevent the system from targeting its own CRISPR locus that lacks a PAM sequence. Over the past decade, the PAMs of various Cas nucleases have been deciphered. For example, one of the pioneering Cas nucleases, the Type II-A Cas9 from *Streptococcus pyogenes* (SpCas9), has been determined to recognize an NGG PAM directly adjacent to the 3’ end of the protospacer [10,17,18]. While NGG remains the canonical PAM for this nuclease, it has been shown that SpCas9 more weakly recognizes non-consensus PAMs, such as NAG and NGA [15,19,20]. As the PAMs of various Cas nucleases have been reported, a common theme has emerged in which different Cas nucleases are associated with less-recognized PAM sequences that deviate from the consensus PAM [19–22]. These non-canonical PAMs have the potential to broaden the target space of a

given nuclease but also increase the potential of off-target sites associated with these PAMs.

Like SpCas9, the Type V-A Cas12a from *Acidaminococcus sp*. (AsCas12a, also known as AsCpf1) has become a widely-used Cas nuclease. However, unlike SpCas9, AsCas12a is able to cleave target sequences containing a TTTV PAM adjacent to the 5' end of the protospacer [11]. This nuclease has been repurposed for various applications, such as genome-editing, gene regulation, and diagnostics [11,23–25]. While the TTTV PAM has been reported as the consensus motif for AsCas12a, this nuclease is also able to recognize weaker CTTV and TCTV PAMs [11,26]. To further broaden the PAM specificity of AsCas12a, others have performed structure-guided mutagenesis to create variants of AsCas12a capable of recognizing TCCV and TATV PAMs [27,28]. While these variants have increased the targeting range of AsCas12a, they were developed from a wild-type version thought to only recognize the TTTV and, to a lesser extent, CTTV and TCTV motifs.

In this work, we discovered that the wild-type AsCas12a can recognize GYTV motifs. This insight came from a cell-free transcription-translation (TXTL)-based PAM screen [29,30]. DNA cleavage in TXTL, plasmid clearance in *E. coli*, and indel formation in HEK293T cells confirmed that AsCas12a could recognize GYTV motifs, with targets containing GTTV PAMs being more efficiently recognized compared to GCTV PAMs. We also observed a bias at the -5 position of both motifs, suggesting that AsCas12a could recognize a wider stretch of DNA as part of the PAM. Together, these results indicate that the range of PAM sequences recognized by AsCas12a is broader than originally reported, thus increasing the targeting range of AsCas12a for its various biotechnological applications.

4.2 - MATERIALS AND METHODS

4.2.1 - Strains, plasmids, and oligonucleotides

All strains, plasmids, primers, and gBlocks used in this work can be found in the Supplementary Table 4.1.

4.2.2 - TXTL-based PAM screen and DNA cleavage assay

For the PAM screen and DNA cleavage assays, we used a commercially available cell-free TXTL system developed from an all-*E. coli* lysate (Arbor Biosciences, Cat: 507096) [31]. The materials and methods to perform these assays, prepare the next-generation sequencing (NGS) library, and conduct data analysis have been described in thorough detail elsewhere [29,30]. The AsCas12a expression plasmid used for these experiments has been reported previously [32]. For the DNA cleavage assay, we used a modified version of pCB848 from previous work which expresses the GFP reporter [30]. The PAM sequence of pCB848 was mutated using the Q5 Site-Directed Mutagenesis Kit (NEB, Cat: E0554S). The same target site was used for all PAMs tested. The gRNA was expressed from the gBlock (custom gene fragment) TJ524 to target the GFP reporter plasmid and the 5N-randomized PAM-library. The non-targeting control was gBlock CSM-GB019, which contained a randomized, non-targeting guide RNA. Each gRNA was expressed in its processed form (20-nt repeat and 24-nt guide) from the J23119 promoter and terminated using a rho-independent terminator. GFP fluorescence was measured using a Synergy H1 plate reader from Biotek with excitation and emission wavelengths of 488nm and 553nm, respectively. The reported production of GFP was calculated using a linear standard calibration curve developed from recombinant eGFP [29]. While this calibration curve will vary from factors such as the plate reader and reagents used, our GFP production was calculated by dividing the raw fluorescence values by 9212.6. The PAM library was previously made using methods described in detail elsewhere [30]. The reaction setup for the TXTL reactions can be found in Supplementary Table 4.2. The PAMs and target sequences used for the DNA cleavage assay can be found in Supplementary Table 4.3. The NGS data, including the

raw data and post-processing reads, were deposited in the NCBI gene expression omnibus (accession # GSE123443).

4.2.3 - Plasmid clearance assay in E. coli

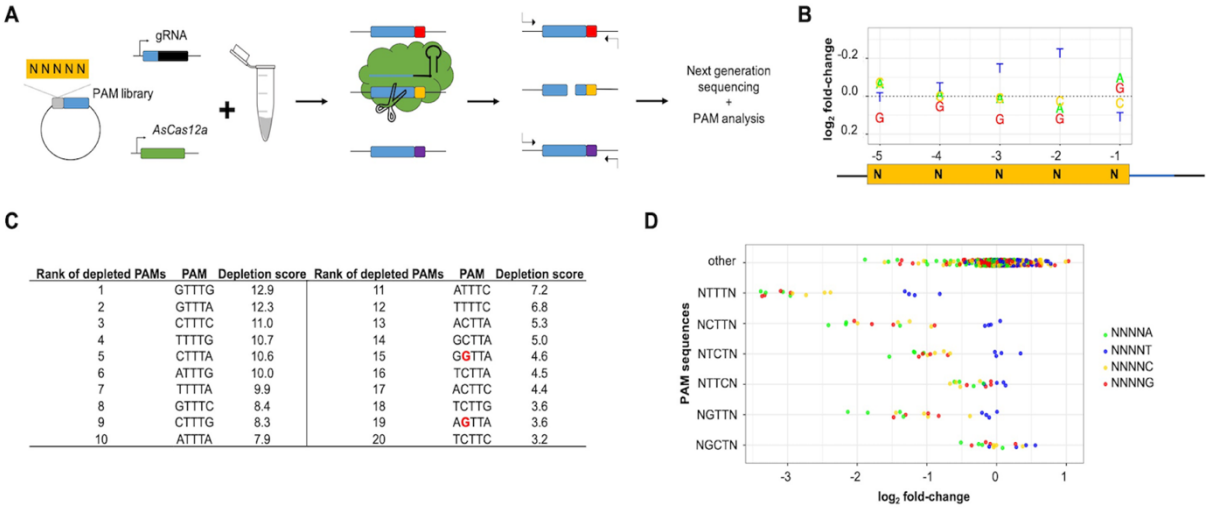
We used CBS-445 for expression of bacterial AsCas12a and CBS-444 for expressing the gRNA. The plasmid containing the gRNA target sequence flanked by NGYTC, TTTC, or GGCT PAMs were constructed using Q5 mutagenesis. We first transformed 50ng of CBS-444 into electrocompetent *E. coli* cells containing CBS-445 and a plasmid containing the gRNA target sequence. After a one-hour recovery at 37°C with shaking at 250 rpm in SOC, serial-diluted cells were plated on LB agar plates supplemented with ampicillin, kanamycin, and chloramphenicol. After a 16-hour incubation at 37°C, the colonies were counted for analysis.

4.2.4 - Indel formation in DNMT1.

The target sites and their PAMs for editing *DNMT1* can be found in Supplementary Table 4.3. The mammalian AsCas12a expression plasmid was obtained from Addgene (Cat: 69982). The gRNA expression plasmids, which encode a processed gRNA under the control of an hU6 promoter, were constructed as described elsewhere [33]. Briefly, the empty gRNA plasmid was digested with BbsI-HF (NEB, Cat: R3539S), and ligated with phosphorylated and annealed oligos, which contained the target sequence-of-interest. Transfection-grade DNA was prepared using the QIAGEN Plasmid Mini Kit (Qiagen, Cat: 12125). One day prior to the transient transfections, $2x10^5$ HEK293T cells were seeded in each well of a 12-well plate with 1mL of complete media (Dulbecco's Modified Eagle Medium (Invitrogen, Cat: 11965-092) supplemented with 10% fetal bovine serum (Invitrogen, Cat: A3840001) and 1% penicillin-streptomycin (Invitrogen, Cat: 15070063). For each gRNA tested, 160ng of the gRNA plasmid and 640ng of the AsCas12a plasmid were transfected using jetPRIME (Genesee Scientific, Cat: 55-132). Cells were then incubated for 20 hours at 37°C prior to replacing fresh growth media into each well. After media replacement, the transfected cells were incubated for another 52 hours at 37°C prior to genomic DNA isolation.

4.2.5 - Tracking of Indels by Decomposition (TIDE) analysis

Genomic DNA was isolated using the GeneJET Genomic DNA Purification Kit (ThermoFisher Scientific, Cat: K0721). An amplicon bridging all tested target sites were amplified by PCR from the genomic DNA using Q5 Hot Start High-Fidelity 2X Master Mix (NEB, Cat: M0494L) and primers TJ719/TJ722. After successful amplification, samples were prepared for Sanger sequencing using a DNA cleanup kit (Zymo Research, Cat: D4013). The primer closest to the predicted cleavage site for each target sequence was chosen for the Sanger sequencing reactions (either TJ719 or TJ722). The chromatograms for each sample were analyzed using TIDE [34]. Genomic DNA isolated from each PAM tested was analyzed against a non-PAM negative control. The target sequences and their PAMs can be found in Supplementary Table 4.3



Rank of depleted PAMs	PAM	Depletion score	Rank of depleted PAMs	PAM	Depletion score
1	GTTTG	12.9	11	ATTTC	7.2
2	GTTTA	12.3	12	TTTTC	6.8
3	CTTTC	11.0	13	ACTTA	5.3
4	TTTTG	10.7	14	GCTTA	5.0
5	CTTTA	10.6	15	GGTTA	4.6
6	ATTTG	10.0	16	TCTTA	4.5
7	TTTTA	9.9	17	ACTTC	4.4
8	GTTTC	8.4	18	TCTTG	3.6
9	CTTTG	8.3	19	AGTTA	3.6
10	ATTTA	7.9	20	TCTTC	3.2

Figure 4.1: A TXTL-based PAM screen with AsCas12a identifies a non-canonical GYTV motif. (A) Schematic of the TXTL-based PAM screen. The AsCas12a expression construct, a targeting or non-targeting gRNA expression construct, and the PAM-library plasmid were added into a TXTL reaction and incubated at 29°C for 16 hours. Plasmids containing recognized PAM sequences are cleaved, while the non-PAMs remain in the reaction mix. After incubation, the remaining PAM sequences in the library were amplified by PCR and subjected to next-generation sequencing. PAM sequences were then identified based on the extent of their depletion in the reaction with the targeting gRNA versus that with the non-targeting gRNA. (B) Plots showing the fold-change of each nucleotide at the different positions in the PAM library following cleavage by AsCas12a. The vertical axis has been inverted to emphasize depleted nucleotides. (C) List of the top 20 most depleted PAM sequences from the screen in rank order. The bolded, red text indicates a G in the PAM sequence. (D) The fold-change of selected 5-mers from the PAM screen.

4.3 - RESULTS

4.3.1 - PAM screen of AsCas12a reveals non-canonical motifs

We were interested in the full range of PAM sequences recognized by AsCas12a. We began with a TXTL-based PAM screen that we recently developed [29,30]. As part of the assay, constructs separately encoding AsCas12a, a targeting gRNA or non-targeting gRNA, and a 5N-randomized PAM library flanking the target sequence were added to a TXTL reaction. Each reaction was incubated at 29°C for 16 hours, and the uncleaved library members were then amplified by PCR and subjected to next-generation sequencing. As part of the screen, the most strongly-recognized PAMs should exhibit the highest depletion in the reactions with the targeting versus non-targeting gRNAs (Figure 4.1A, see Supplementary Table 4.2 for the reaction setup). To determine the extent of cleavage of the PAM library, a GFP expression plasmid containing a TTTC PAM flanking the same target sequence was added to the reaction. GFP levels plateaued after four hours (Supplementary Figure 4.1), indicating that the PAM-library likely underwent extensive cleavage.

Figure 4.1B shows the fold-change of each nucleotide at each position in the PAM-library resulting from AsCas12a cleavage. From this plot, AsCas12a recognized the canonical TTTV PAM, with the T in the -4 position being more flexible compared to the -2 and -3 positions. While the data from Figure 4.1B confirmed the consensus motif, the top 20 depleted PAM scores included GGTTA and AGTTA (Figure 4.1C, Supplementary Table 4.4). Similarly, the PAM wheel of AsCas12a showed a high depletion of the consensus TTTV motif, though the GYTV can be observed when investigated deeper (Supplementary File 4.2). To probe the screening results more deeply, we generated dot plot showing the fold-change of 5-mers for the previously reported PAMs of (NCTTN, NTCTN, and NTTCN, where the latter is poorly recognized) [11,26] as well variants of the first two PAMs with G at the -4 position (NGTTN and NGCTN) (Figure 4.1D). While the canonical NTTTV PAM was the highest-depleted motif, the depletion of the NGTTV PAM was comparable to that of the NCTTV and NTCTV PAMs. Furthermore, the depletion of the NGCTV motif was less pronounced but comparable to

that of the NTTCV motif. Together, these data suggested that AsCas12a could recognize a non-canonical GYTV motif, with the GTTV motif preferred over the GCTV motif.

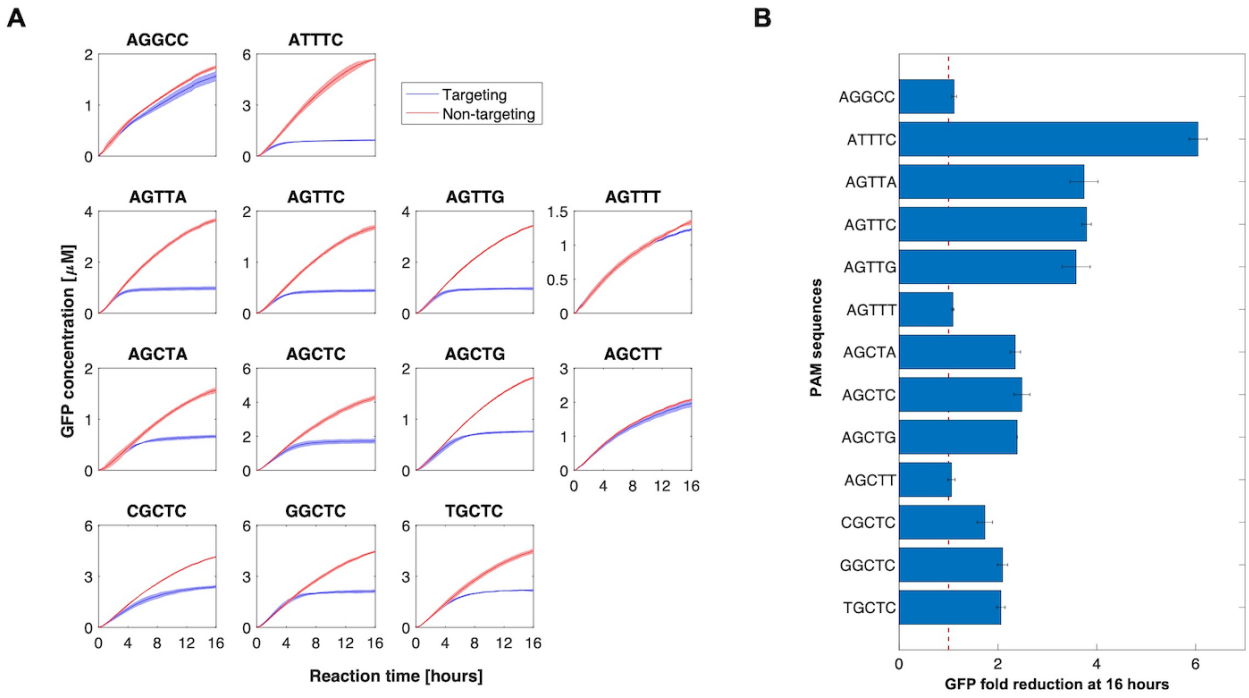


Figure 4.2: AsCas12a can recognize the GYTV motif as part of DNA cleavage in TXTL. (A) Time series of GFP expression in TXTL. The AsCas12a nuclease was expressed with a non-targeting gRNA (red line) or a targeting gRNA (blue line) designed to target a site upstream of the constitutive promoter controlling GFP. Cleavage leads to rapid degradation of the plasmid and loss of GFP expression. The PAM sequence is indicated above each time course. Each reaction was incubated at 29°C for 16 hours. The error bars represent the standard deviation from three separate TXTL reactions. (B) Fold-reduction in GFP for each PAM sequence. The fold-reductions were calculated using the GFP fluorescence data from the 16-hour time point from the reactions with the targeting gRNA and the non-targeting gRNA. The dotted red line indicates a fold-reduction of one (i.e. no reduction). The error bars represent the standard deviation from three separate TXTL reactions.

4.3.2 - AsCas12a can recognize the GYTV motif in vitro

To further investigate the output of the PAM screen, we tested the cleavage activity of AsCas12a using the GYTV motif in an *in vitro* DNA cleavage assay (Figure 4.2). We conducted the assay similarly to the TXTL-based PAM screen (Figure 4.1A) except that the GFP reporter plasmid was used in place of the PAM-library plasmid. The same target sequence as that in the PAM library was also used. We first tested the cleavage efficiency of AsCas12a using all potential sequences within an AGYTN motif as well as

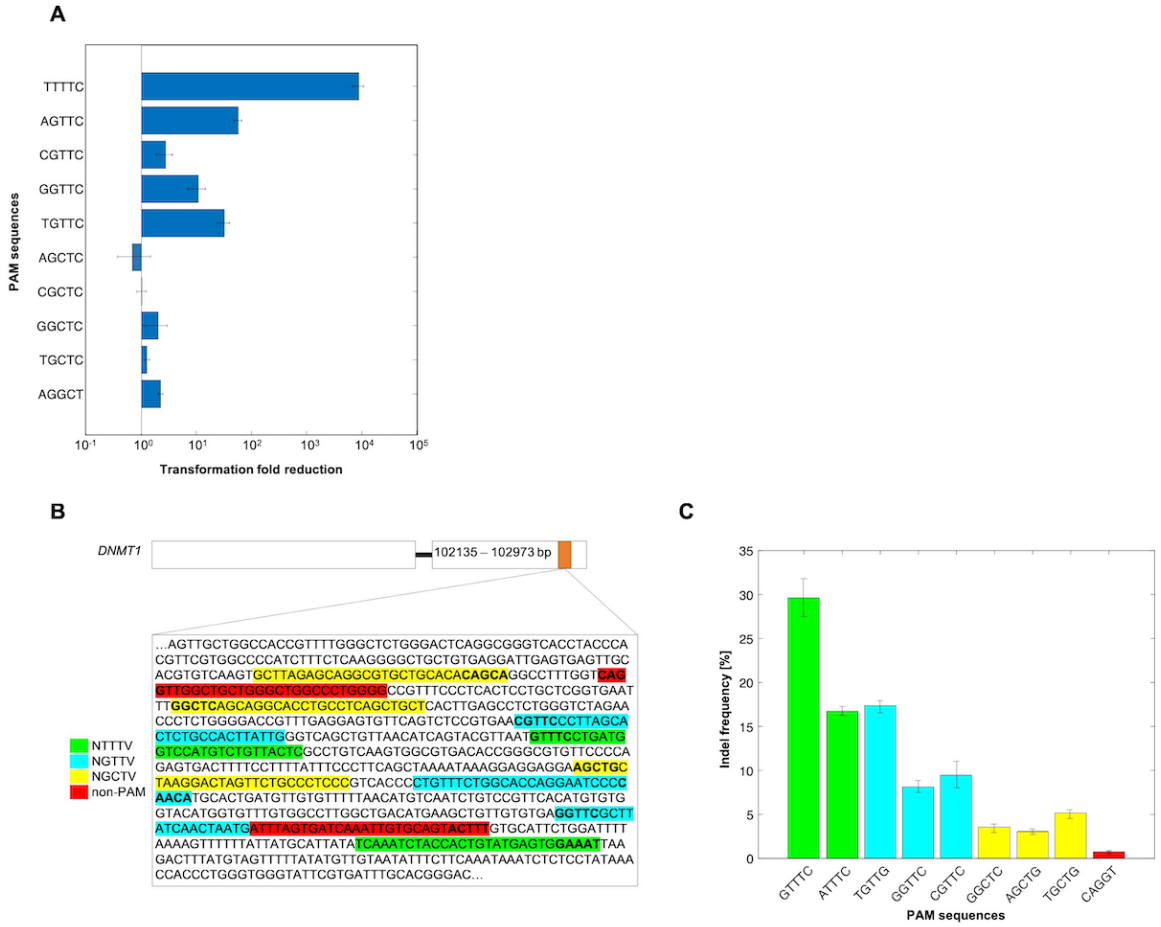


Figure 4.3: AsCas12a can recognize the GYTV motif *in vivo* as part of plasmid clearance in *E. coli* and DNA editing in mammalian cells. (A) The average fold reduction of the number of transformants in an *E coli* plasmid clearance assay. Cells that contain the AsCas12a plasmid and the target plasmid containing the protospacer flanked by various PAM sequences were transformed with a plasmid expressing the crRNA. The dotted red line indicates a fold-reduction of one (i.e. no reduction). The error bars represent the standard deviation from four independent experiments starting from separate colonies. (B) Section of *DNMT1* containing the sequences targeted in the indel-formation experiments. The red, green, cyan, and yellow highlighted text mark the targets and flanking PAMs, including non-PAMs (CAGGT, AAAGT) and the TTTV, GTTV, and GCTV motifs, respectively. The 5-mer PAM for each target is bolded. Note that some of the gRNAs were derived from the bottom strand, where each PAM is the reverse complement of the bolded sequence. (C) Indel frequencies for eight *DNMT1* target sites in HEK293T cells quantified using TIDE (Brinkman *et al.* 2014). As part of the analysis, sequencing data from each target sequence was compared to that of the target sequence flanked by the AAAGT non-PAM in the same, amplified region. Using a two-tailed T-test, the indel frequencies of all tested GCTV PAMs were statistically different compared to that of the non-PAM. The P-values for GGCTC, AGCTG, and TGCTG compared to the non-PAM were 0.0012, 0.0004, and 0.0002, respectively (n = 3). The error bars represent the standard deviation from three independent transfections.

two controls: a canonical TTTC PAM and unrecognized GGCC non-PAM. Performing the DNA cleavage assays, we observed that AGYTV sequences but not AGYTT sequences led to DNA cleavage, where the canonical TTTC PAM resulted in more rapid cleavage than the GTTV motif, while the GTTV motif led to more rapid cleavage than

the GCTV motif. These results further confirm the ability of AsCas12a to recognize the GYTV motif, with a preference for GTTV over GCTV.

Further interrogating the results from the PAM screen revealed that only the AGCTV motifs were present from the top 100 depleted PAMs. We therefore asked if the -5 position of the NGCTV motif contributes to DNA cleavage activity. Testing each possible nucleotide in the DNA cleavage assay resulted in a modest but noticeable bias at the -5 position within the NGCTC motif, with A being the most preferred, G and T being similarly preferred, and C being the least preferred (Figure 4.2). Therefore, the -5 position of the NGCTC motif can influence the *in vitro* DNA targeting activity of AsCas12a.

4.3.3 - The -5 PAM position influences target recognition in E coli

We next asked if the GTTV and GCTV motifs allow DNA cleavage by AsCas12a in cells, and how the -5 position influences cleavage activity. To answer these questions, we performed a plasmid clearance assay in *E. coli* by transforming a targeting or non-targeting gRNA plasmid into cells harboring the AsCas12a plasmid and a separate plasmid containing the gRNA target sequence flanked by various sequences (see Supplementary Table 4.3 for target sequence and PAMs). To assess if the bias at the -5 PAM position extends to multiple target sequences, we targeted a sequence different from the one used in the DNA cleavage assay in TXTL. After counting resistant colonies, the transformation fold-reduction was calculated in relation to that with the non-targeting gRNA control. We found that the NGTTC motif led to plasmid clearance that was at least two orders-of-magnitude lower than that for the canonical TTTC PAM (Figure 4.3A). We did notice that the colonies associated with many of the NGTTC sequences were smaller than those for the non-targeting control (Supplementary Figure 4.2), suggesting that cleavage had only partially cleared the plasmid. Sequences associated with the GCTC motif yielded negligible clearance (Figure 4.3A and Supplementary Figure 4.2). When comparing nucleotides at the -5 position of the NGTTC motif, we observed a bias toward an A and against a C, in agreement with the

in vitro DNA cleavage results (Figures 4.2 and 4.3A). These results indicate that the GTTC motif is a viable PAM for AsCas12a, while the GCTC motif may not be viable at

least for plasmid clearance in *E. coli*. Also, as different target sequences were used for the TXTL experiments and plasmid clearance assays, nucleotide bias at the -5 PAM position of the NGYTV motif can be observed across multiple target sequences.

4.3.4 - Indel formation can be achieved with AsCas12a using GYTV PAMs

We next sought to test AsCas12a's ability to recognize the GYTV motif in mammalian cells. We specifically chose to target sites within the *DNMT1* gene in HEK293T cells, as this gene had been used previously to assess indel formation with AsCas12a [11,25,35–37]. We transiently transfected a plasmid constitutively expressing AsCas12a along with a plasmid expressing a gRNA targeting *DNMT1* at various locations. After 72 hours post-transfection, we assessed the frequency of indel formation using TIDE analysis [34]. As part of the experiments, we targeted two different sites flanked by the TTTC motif, three different sites flanked by the GTTV and GCTV motifs, and one site flanked by an AGGT non-PAM (Figure 4.3A). An additional target site with an AAGT non-PAM was used as the reference for TIDE.

Targets containing the TTTV motif resulted in the highest indel frequency (Figure 4.3B, individual averages of two sites = 16.6%, 29.6%), while targets with the GTTV motif resulted in an efficiency either less than or comparable to the TTTV motif (individual averages of three sites = 17.3%, 9.2%, 8.1%). While using the selected GCTV PAMs resulted in a low indel frequency (individual averages of three sites = 5.1%, 3.5%, 3.1%), the GCTV PAMs resulted in indel frequencies that were statistically different compared to the AGGT non-PAM (0.63%). We thus conclude that AsCas12a can successfully edit targets flanked by GYTV motif in mammalian cells, with higher editing efficiencies for targets flanked by the GTTV motif than the GCTV motif.

4.4 - DISCUSSION

In this work, we found that AsCas12a recognized GTTV and GCTV as non-canonical PAMs for DNA targeting. DNA targeting with the GTTV motif was more efficient than that of the GCTV motif, as expected given AsCas12a's preference for CTTV over TCTV [11,26]. While the *in vitro* DNA cleavage assay and plasmid clearance indicated that TTTV was preferred over GTTV (Figures 4.2 and 4.3A), the indel-formation assays yielded at least one instance in which targets flanked by TTTV and GTTV motif exhibited similar editing efficiencies (Figure 4.3B). This occurrence is most likely due to the influence of the target site sequence, as reported in prior work [21]. Therefore, GTTV in particular may be a viable PAM sequence for genome editing with AsCas12a in mammalian cells.

In the DNA cleavage and plasmid clearance assays, we showed that the -5 position of the PAM could affect the cleavage efficiency (Figures 4.2 and 4.3A). This insight specifically came from interrogating the NGCTC motif in the *in vitro* DNA cleavage assay and the NGTTC motif in the plasmid clearance assay in *E. coli*, which both favored an A and disfavored a C at this position. Similar biases that extended beyond the standard PAM was observed in previous work for the Cas12a from *F. novicida* (FnCas12a) in its -4 position (consensus PAM of NTTV), as well as Cascade from the Type I-E CRISPR-Cas system (consensus PAM of AWG) [22]. This effect could be dependent on the target sequence as well as the sequences extending beyond the PAM.

Shortly following the first reports of genome editing with SpCas9, progress has been made to increase the targeting range of Cas nucleases and to reduce off-target effects. The former has been achieved in part by the discovery of Cas nucleases that recognize distinct PAMs as well as the engineering of Cas nucleases to increase PAM flexibility of wild-type Cas nucleases [27,28,38–40]. While increasing the targeting range has allowed for the targeting of almost every sequence of interest, this inevitably expands the number of potential off-target sites. Work has been done to increase PAM flexibility while reducing off-target effects, such as engineering high-fidelity Cas nucleases, tightly

controlling nuclease levels, and carefully selecting target sequences [39,41–43]. While one would assume that the targeting range and off-targeting would go hand-in-hand, one engineered variant of SpCas9 called xCas9 managed to achieve a wider targeting range and lower off-target effects compared to the parental SpCas9 [39]. Though the frequency of off-target effects can be decreased using available web-based tools [42], non-canonical PAMs that can be recognized by a particular Cas nuclease must be identified to better predict potential off-target sites.

Consensus sequences have been one of the most common methods of communicating PAMs [44,45]. This approach involves the reporting of a single sequence that captures the best recognized set of PAM sequences. For example, the consensus sequence of two commonly used Cas nucleases, SpCas9 and AsCas12a, have been reported as NGG and TTTV, respectively [10,11,17,18]. While the consensus sequence allows for a simple method of describing the PAM, it fails to reveal other PAM sequences Cas nucleases are able to recognize. Examples include SpCas9 recognizing weaker NAG and NGA PAMs and AsCas12a recognizing CTTV and TCTV PAMs [11,15,19,20]. While this work and others have reported PAMs using methods that include the sequence logo, consensus sequence, PAM wheel, and PAM table [16,22,29,44,45], there currently is no standard for conveying PAM sequences. Though reporting the consensus sequence has become the norm for communicating the PAM, more thorough methods or a set of methods are needed to fully describe the targeting range of Cas nucleases.

4.5 - ACKNOWLEDGEMENTS

We thank Benjamin Gray for critical discussions about Cas12a nucleases. The AsCas12a expression plasmid (pY010 (pcDNA3.1-hAsCpf1)) was a gift from Feng Zhang (Addgene, Cat: 69982). The gRNA expression plasmid (pU6-As-crRNA) was a gift from Jin-Soo Kim (Addgene, Cat: 78956). This work was funded by the National Institutes of Health (1R35GM119561) and the National Science Foundation (MCB-1413044).

4.6 - REFERENCES

[1] M. Adli, The CRISPR tool kit for genome editing and beyond, Nat. Commun. 9 (2018) 1911.

[2] Y. Li, S. Li, J. Wang, G. Liu, CRISPR/Cas systems towards next-generation biosensing, Trends Biotechnol. In press (2019).

[3] R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. Romero, P. Horvath, CRISPR provides acquired resistance against viruses in prokaryotes, Science. 315 (2007) 1709–1712.

[4] S.J. Brouns, M.M. Jore, M. Lundgren, E.R. Westra, R.J. Slijkhuis, A.P. Snijders, M.J. Dickman, K.S. Makarova, E.V. Koonin, J. van der Oost, Small CRISPR RNAs guide antiviral defense in prokaryotes, Science. 321 (2008) 960–964.

[5] C.R. Hale, P. Zhao, S. Olson, M.O. Duff, B.R. Graveley, L. Wells, R.M. Terns, M.P. Terns, RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex, Cell. 139 (2009) 945–956.

[6] J.E. Garneau, M.È. Dupuis, M. Villion, D.A. Romero, R. Barrangou, P. Boyaval, C. Fremaux, P. Horvath, A.H. Magadán, S. Moineau, The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA, Nature. 468 (2010) 67–71.

[7] E.R. Westra, P.B.G. van Erp, T. Künne, S.P. Wong, R.H.J. Staals, C.L.C. Seegers, S. Bollen, M.M. Jore, E. Semenova, K. Severinov, W.M. de Vos, R.T. Dame, R. de Vries, S.J.J. Brouns, J. van der Oost, CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3, Mol. Cell. 46 (2012) 595–605.

[8] D.S. Chertow, Next-generation diagnostics with CRISPR, Science. 360 (2018) 381–382.

[9] R. Barrangou, J.A. Doudna, Applications of CRISPR technologies in research and beyond, Nat. Biotechnol. 34 (2016) 933–941.

[10] M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J.A. Doudna, E. Charpentier, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity, Science. 337 (2012) 816–821.

[11] B. Zetsche, J.S. Gootenberg, O.O. Abudayyeh, A. Regev, E. V Koonin, F. Zhang,

I.M. Slaymaker, K.S. Makarova, P. Essletzbichler, S.E. Volz, J. Joung, J. Van Der Oost, Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system, Cell. 163 (2015) 759–771.

[12] E. Deltcheva, K. Chylinski, C.M. Sharma, K. Gonzales, Y. Chao, Z.A. Pirzada, M.R. Eckert, J. Vogel, E. Charpentier, CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III, Nature. 471 (2011) 602–607.

[13] E. Semenova, M.M. Jore, K.A. Datsenko, A. Semenova, E.R. Westra, Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence, Proc. Natl. Acad. Sci. 108 (2011) 10098–10103.

[14] B. Wiedenheft, E. van Duijn, J.B. Bultema, S.P. Waghmare, K. Zhou, A. Barendregt, W. Westphal, A.J.R. Heck, E.J. Boekema, M.J. Dickman, J.A. Doudna, RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions, Proc. Natl. Acad. Sci. 108 (2011) 10092–10097.

[15] W. Jiang, D. Bikard, D. Cox, F. Zhang, L.A. Marraffini, RNA-guided editing of bacterial genomes using CRISPR-Cas systems, Nat. Biotechnol. 31 (2013) 233–239.

[16] R.T. Leenay, C.L. Beisel, Deciphering, communicating, and engineering the CRISPR PAM, J. Mol. Biol. 429 (2017) 177–191.

[17] F.J.M. Mojica, C. Díez-Villaseñor, J. García-Martínez, C. Almendros, Short motif sequences determine the targets of the prokaryotic CRISPR defence system, Microbiology. 155 (2009) 733–740.

[18] W. Jiang, D. Bikard, D. Cox, F. Zhang, L.A. Marraffini, RNA-guided editing of bacterial genomes using CRISPR-Cas systems, Nat. Biotechnol. 31 (2013) 233–239.

[19] P.D. Hsu, D.A. Scott, J.A. Weinstein, F.A. Ran, S. Konermann, V. Agarwala, Y. Li, E.J. Fine, X. Wu, O. Shalem, T.J. Cradick, L.A. Marraffini, G. Bao, F. Zhang, DNA targeting specificity of RNA-guided Cas9 nucleases, Nat. Biotechnol. 31 (2013) 827–832.

[20] Y. Zhang, X. Ge, F. Yang, L. Zhang, J. Zheng, X. Tan, Z.B. Jin, J. Qu, F. Gu,

Comparison of non-canonical PAMs for CRISPR/Cas9-mediated DNA cleavage in human cells, Sci. Rep. 4 (2014) 1–5.

[21] K.M. Esvelt, P. Mali, J.L. Braff, M. Moosburner, S.J. Yaung, G.M. Church, Orthogonal Cas9 proteins for RNA-guided gene regulation and editing, Nat. Methods. 10 (2013) 1116–1121.

[22] R.T. Leenay, K.R. Maksimchuk, R.A. Slotkowski, R.N. Agrawal, A.A. Gomaa, A.E. Briner, R. Barrangou, C.L. Beisel, Identifying and visualizing functional PAM diversity across CRISPR-Cas systems, Mol. Cell. 62 (2016) 137–147.

[23] J.S. Chen, E. Ma, L.B. Harrington, M. Da Costa, X. Tian, J.M. Palefsky, J.A. Doudna, CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity, Science. 6245 (2018) eaar6245.

[24] Y.E. Tak, B.P. Kleinstiver, J.K. Nuñez, J.Y. Hsu, J.E. Horng, J. Gong, J.S. Weissman, J.K. Joung, Inducible and multiplex gene regulation using CRISPR–Cpf1-based transcription factors, Nat. Methods. 14 (2017) 1163–1166.

[25] B. Zetsche, M. Heidenreich, P. Mohanraju, I. Fedorova, J. Kneppers, E.M. DeGennaro, N. Winblad, S.R. Choudhury, O.O. Abudayyeh, J.S. Gootenberg, W.Y. Wu, D.A. Scott, K. Severinov, J. van der Oost, F. Zhang, Multiplex gene editing by CRISPR–Cpf1 using a single crRNA array, Nat. Biotechnol. 35 (2017) 31–34.

[26] H.K. Kim, M. Song, J. Lee, A.V. Menon, S. Jung, Y.M. Kang, J.W. Choi, E. Woo, H.C. Koh, J.W. Nam, H. Kim, In vivo high-throughput profiling of CRISPR-Cpf1 activity, Nat. Methods. 14 (2017) 153–159.

[27] L. Gao, D.B.T. Cox, W.X. Yan, J.C. Manteiga, M.W. Schneider, T. Yamano, H. Nishimasu, O. Nureki, N. Crosetto, F. Zhang, Engineered Cpf1 variants with altered PAM specificities, Nat. Biotechnol. 35 (2017) 789–792.

[28] B.P. Kleinstiver, A.A. Sousa, R.T. Walton, Y.E. Tak, J.Y. Hsu, K. Clement, M.M. Welch, J.E. Horng, J. Malagon-Lopez, I. Scarfò, M. V Maus, L. Pinello, M.J. Aryee, J.K. Joung, Engineered CRISPR–Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing, Nat. Biotechnol. (2019).

[29] R. Marshall, C.S. Maxwell, S.P. Collins, T. Jacobsen, M.L. Luo, M.B. Begemann,

B.N. Gray, E. January, A. Singer, Y. He, C.L. Beisel, V. Noireaux, Rapid and scalable characterization of CRISPR technologies using an *E. coli* cell-free transcription-translation system, Mol. Cell. 69 (2018) 146–157.

[30] C.S. Maxwell, T. Jacobsen, R. Marshall, V. Noireaux, C.L. Beisel, A detailed cell-free transcription-translation-based assay to decipher CRISPR protospacer-adjacent motifs, Methods. 143 (2018) 48–57.

[31] J. Garamella, R. Marshall, M. Rustad, V. Noireaux, The all *E. coli* TX-TL toolbox 2.0: A platform for cell-free synthetic biology, ACS Synth. Biol. 5 (2016) 344–355.

[32] C. Liao, F. Ttofali, R.A. Slotkowski, S.R. Denny, T.D. Cecil, R.T. Leenay, A.J. Keung, C.L. Beisel, One-step assembly of large CRISPR arrays enables multi-functional targeting and reveals constraints on array design, Manuscript submitted for publication. (2018).

[33] D. Kim, J. Kim, J.K. Hur, K.W. Been, S.H. Yoon, J.S. Kim, Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells, Nat. Biotechnol. 34 (2016) 863–868.

[34] E.K. Brinkman, T. Chen, M. Amendola, B. Van Steensel, Easy quantitative assessment of genome editing by sequence trace decomposition, Nucleic Acids Res. 42 (2014) 1–8.

[35] T. Yamano, B. Zetsche, R. Ishitani, F. Zhang, H. Nishimasu, O. Nureki, Structural basis for the canonical and non-canonical PAM recognition by CRISPR-Cpf1, Mol. Cell. 67 (2017) 633–645.

[36] M. Tu, L. Lin, Y. Cheng, X. He, H. Sun, H. Xie, J. Fu, C. Liu, J. Li, D. Chen, H. Xi, D. Xue, Q. Liu, J. Zhao, C. Gao, Z. Song, J. Qu, F. Gu, A ‘ new lease of life ’: FnCpf1 possesses DNA cleavage activity for genome editing in human cells, Nucleic Acids Res. 45 (2017) 11295–11304.

[37] B. Li, C. Zeng, W. Li, X. Zhang, X. Luo, W. Zhao, C. Zhang, Synthetic oligonucleotides Inhibit CRISPR-Cpf1- mediated genome editing, Cell Rep. 25 (2018) 3262–3272.

[38] B.P. Kleinstiver, M.S. Prew, S.Q. Tsai, V. V. Topkar, N.T. Nguyen, Z. Zheng, A.P.W. Gonzales, Z. Li, R.T. Peterson, J.R.J. Yeh, M.J. Aryee, J.K. Joung, Engineered CRISPR-Cas9 nucleases with altered PAM specificities, Nature. 523

(2015) 481–485.

[39] J.H. Hu, S.M. Miller, M.H. Geurts, W. Tang, L. Chen, N. Sun, C.M. Zeina, X. Gao, H.A. Rees, Z. Lin, D.R. Liu, Evolved Cas9 variants with broad PAM compatibility and high DNA specificity, Nature. 556 (2018) 57–63.

[40] H. Nishimasu, X. Shi, S. Ishiguro, L. Gao, S. Hirano, S. Okazaki, T. Noda, O.O. Abudayyeh, J.S. Gootenberg, H. Mori, S. Oura, B. Holmes, M. Tanaka, M. Seki, H. Hirano, H. Aburatani, R. Ishitani, M. Ikawa, N. Yachie, F. Zhang, O. Nureki, Engineered CRISPR-Cas9 nuclease with expanded targeting space, Science. 361 (2018) 1259–1262.

[41] B.P. Kleinstiver, V. Pattanayak, M.S. Prew, S.Q. Tsai, N.T. Nguyen, Z. Zheng, J.K. Joung, High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects, Nature. 529 (2016) 490–495.

[42] C.M. Lee, T.J. Cradick, E.J. Fine, G. Bao, Nuclease target site selection for maximizing on-target activity and minimizing off-target effects in genome editing, Mol. Ther. 24 (2016) 475–487.

[43] C. Shen, M. Hsu, C. Chang, M. Lin, J. Hwu, Y. Tu, Y. Hu, Synthetic switch to minimize CRISPR off-target effects by self-restricting Cas9 transcription and translation, Nucleic Acids Res. 47 (2018) e13.

[44] P. Horvath, D.A. Romero, A.C. Coûté-Monvoisin, M. Richards, H. Deveau, S. Moineau, P. Boyaval, C. Fremaux, R. Barrangou, Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*, J. Bacteriol. 190 (2008) 1401–1412.

[45] H. Deveau, R. Barrangou, J.E. Garneau, J. Labonté, C. Fremaux, P. Boyaval, D.A. Romero, P. Horvath, S. Moineau, Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*, J. Bacteriol. 190 (2008) 1390–1400.

CHAPTER 5: Characterization of Cas12a nucleases reveals diverse PAM profiles between closely-related orthologs

Thomas Jacobsen, Matthew B. Begemann, Benjamin N. Gray, Gloria Yi, and Chase L. Beisel

Original publication:

T. Jacobsen, M.B. Begemann, B.N. Gray, G. Yi, C.L. Beisel, Characterization of Cas12a nucleases reveals diverse PAM profiles between closely-related orthologs. Manuscript in preparation. (2019).

ABSTRACT

CRISPR-Cas systems comprise diverse adaptive immune systems in prokaryotes whose RNA-directed nucleases have been co-opted for various technologies. Recent efforts have focused on expanding the number of known CRISPR-Cas subtypes to identify nucleases with novel properties. However, the functional diversity of nucleases within each subtype remains poorly explored. Here, we used cell-free transcription-translation systems and human cells to characterize six Cas12a single-effector nucleases from the V-A subtype, including nucleases with high sequence homology. While these nucleases readily utilized each other's guide RNAs, they exhibited distinct PAM profiles and cleavage activities that did not track with their phylogenetic relationship. In particular, two Cas12a nucleases encoded by *Prevotella ihumii* (PiCas12a) and *Prevotella disiens* (PdCas12a) shared over 95% sequence homology yet recognized distinct PAM profiles, with PiCas12a, but not PdCas12a, accommodating a G at the -4 or -2 position and a T at the -1 position. Mutational analyses transitioning PiCas12a to PdCas12a deviated the PAM profile from either nuclease, allowing more flexible editing in human cells. Cas12a nucleases therefore can exhibit widely varying properties between otherwise related orthologs, suggesting selective pressure to diversify PAM recognition and supporting the expansion of the CRISPR toolbox through ortholog mining and PAM engineering.

5.1 - INTRODUCTION

Clustered regularly interspaced short palindromic repeats (CRISPR) and their CRISPR-associated (Cas) proteins comprise adaptive immune systems that protect bacteria and archaea from invading plasmids and bacteriophages [1–3]. These systems rely on RNA-directed effector nucleases that are directed by CRISPR-encoded guide RNAs (gRNAs) to bind and cleave complementary nucleic acids often flanked by a short protospacer-adjacent motif (PAM) [4,5]. The programmable nature of these nucleases lent to their direct use for genome-editing, gene regulation, and various other applications (reviewed in [6]). These applications have been spurred in part by the ongoing discovery of Cas nucleases with distinct properties such as DNA or RNA targeting, varying recognized PAM profiles, temperature ranges for optimal activity, and reduced propensity for off-targeting [7–14].

The available set of Cas nucleases are part of a remarkably diverse assortment of CRISPR-Cas systems encompassing various proteins, mechanisms, and functions. This diversity is hypothesized to have emerged from the ongoing arms race between bacteria and invasive genetic elements such as phages [15,16]. Attempts to capture this diversity are now reflected in a hierarchical classification scheme that groups systems into two classes, six types, and over 30 subtypes [7,8]. Ongoing bioinformatics and biochemical characterizations have mainly focused on expanding the list of subtypes, with recent reports expanding Type V systems to nine subtypes (V-A – V-I) and Type VI systems to five subtypes (VI-A, VI-B1/2, VI-C, VI-D) [7,17–19] . However, emerging evidence suggests that incredible diversity lies within each subtype. For instance, characterization of ranging single-effector Cas9 nucleases within the Type II-A subtype have shown that these nucleases not only share limited sequence homology but also can recognize distinct PAM profiles, exhibit ranging propensities to accept mismatches between the guide and target, and do not recognize each other's processed crRNA:tracrRNA duplexes serving as the gRNAs [20–23]. While these distinctions are normally observed for phylogenetically distinct nucleases, little is known about functional differences separating otherwise closely-related nucleases.

A unique opportunity to explore the functional diversity between related Cas nucleases rests within the V-A subtype of CRISPR-Cas systems [24]. This subtype contains Cas12a (also known as Cpf1) nucleases that exhibit unique properties compared to other known Cas nucleases. Specifically, these nucleases process gRNAs from a transcribed CRISPR array lacking accessory factors (e.g. tracrRNA), recognize T-rich PAMs located 5' of the displaced strand of target DNA, utilize a single RuvC domain to nick both strands of target DNA, and can non-specifically cleave single-stranded DNA upon target recognition [24–26]. In turn, these capabilities have led Cas12a to be harnessed for numerous applications in genome-editing, gene regulation, and nucleic acid sensing [24,27,28]. Ongoing characterization of Cas12a nucleases has also revealed variability between these V-A effectors, such as the ability to utilize each other's gRNAs, a propensity to recognize C or G at various PAM positions, as well as different temperature ranges in which these nucleases are active [11,14,24,29,30]. However, to what extent these properties impact these nucleases, particularly between related orthologs, remains largely unexplored.

Here, we characterized a set of six Cas12a nucleases, including nucleases exhibiting high homology with each other or with well-established nucleases. While these variants utilized each other's gRNAs, they exhibited ranging effective cleavage activities and recognized distinct PAM profiles. Further investigation of two Cas12a nucleases from *Prevotella ihumii* (PiCas12a) and *Prevotella disiens* (PdCas12a) revealed different PAM profiles despite sharing 95.7% homology. Furthermore, mutating PiCas12a toward PdCas12a revealed PAM profiles distinct from those associated with either parental nuclease. These findings demonstrate that otherwise closely-related CRISPR nucleases and the intervening mutants can exhibit divergent properties, with implications for the evolution of CRISPR-Cas systems and expanding the set of nucleases available for CRISPR technologies.

5.2 - MATERIALS AND METHODS

5.2.1 - Strains, plasmids, and oligonucleotides

All strains, plasmids, oligos, and gBlocks used in this work can be found in Table S1. All PCR amplifications were performed using the Q5 Hot Start High-Fidelity 2X Master Mix (NEB, Cat: M0494S).

The pET28b+ plasmids expressing the Cas12a nucleases were synthesized by GenScript and codon-optimized for plant systems. The PdCas12a bacterial expression plasmid was constructed by PCR-amplifying the PdCas12a from a previous construct (Addgene, Cat: 69990) and inserting it into the BamHI and NheI sites of the plasmid CB1095. All gRNAs used in this work were synthetized from IDT as custom gBlocks, which contain the constitutive J23119 promoter and a rho-independent terminator. The Q5 mutagenesis kit (NEB, Cat: E0554S) was used to generate the PiCas12a variants and the deGFP reporter plasmids following the manufacturer's instructions.

To construct the mammalian-expressed HkCas12a and PiCas12a plasmids, the nucleases and an N-terminal nuclear localization tag were PCR-amplified from their respective pET28b+ plasmids and inserted into the plasmid CB1067 using NEBuilder HiFi DNA Assembly Master Mix (NEB, Cat: E2621L). The empty crRNA expression plasmids were constructed using a previously built plasmid expressing an empty *Acidaminococcus* sp. BV3L6 Cas12a gRNA (Addgene, Cat: 78956). We used the Q5 mutagenesis kit to convert the direct repeat of AsCas12a to that of Hk and PiCas12a. The plasmids containing the empty Hk and PiCas12a gRNAs were then digested with BsmBI (NEB, Cat: R0580S) and ligated with 5' phosphorylated and annealed oligos containing a target sequence of interest.

5.2.2 - DNA cleavage assay using a cell-free transcription-translation (TXTL) system

The materials and methods to perform this assay was described in detail elsewhere [32,33]. Briefly, we used the commercially available cell-free TXTL system developed from an all-*E. coli* lysate (Arbor Biosciences, Cat: 507096) [34] to rapidly express Cas12a from a plasmid and targeting or non-targeting gRNAs from custom gBlocks. For

this assay, we used GFP reporter plasmids containing a target sequence flanked by potential PAM sequences. GFP fluorescence was measured with a Synergy H1 plate reader (BioTek) using excitation and emission wavelengths of 488 nM and 553 nM, respectively. The reactions were incubated for 16 hours at 29°C and the resulting fluorescence data were analyzed using end-point and time-course analyses. The reported production of GFP was calculated using a linear standard calibration curve developed from recombinant GFP as we have performed previously [29,33]. For the plate reader used for our experiments, the raw fluorescence values were divided by the conversion factor 9212.6 L/μmol.

5.2.3 - TXTL-based PAM screen

For the TXTL-based PAM screen, we constructed a plasmid containing a 5N-randomized PAM library flanked by a sequence targeted by the targeting gRNA using methods described previously [32,33]. Briefly, the PAM library was generated by PCR-amplifying CB847 and gBlock TJ460 with primer pairs CSMpr1308/1309 and CSMpr1310/1311, respectively, followed by assembling the two amplicons with NEBuilder HiFi DNA Assembly Master Mix. Following generation of the PAM library, we performed a high-throughput PAM determination screen [32,33] using the same protocol as the DNA cleavage assay in TXTL, though the GFP reporter plasmid was replaced with the PAM library plasmid. After incubating the samples for 16 hours at 29°C, the uncleaved 5N-randomized PAM library from the targeting and non-targeting reactions for each Cas12a were PCR-amplified and prepared for next-generation sequencing (NGS). The NGS data, including the raw data and post-processing reads, were deposited in the NCBI gene expression omnibus (accession #GSE130377). The code used for the processing and analyses of the NGS data can be found in the following public repository: https://bitbucket.org/csmaxwell/crispr-txtl-pam-counting-script/. The methods used to generate the PAM wheels are described in detail elsewhere [35].

5.2.4 - Indel formation in HEK293T cells

The target sites and their PAM sequences in the *DNMT1* gene in HEK293T cells can be found in Supplementary Table 2. The indel formation assays were conducted as described previously [29]. Briefly, $2x10^5$ HEK293T cells were seeded in 12-well plates

24 hours prior to performing transient transfections. For each reaction, we transiently transfected 160 ng of the crRNA and 640 ng of the Cas12a plasmid using jetPRIME (Genesee Scientific, Cat: 55-132). After a 20-hour incubation at 37°C, the growth media was replaced with fresh media was in each well. The cells were then incubated for an additional 52 hours at 37°C prior to the isolation of genomic DNA.

5.2.5 - Tracking of Indels by Decomposition (TIDE) analysis

Genomic DNA from transfected HEK293T cells were isolated using the GeneJET Genomic DNA Purification Kit (ThermoFisher Scientific, Cat: K0721) following the manufacturer's instructions. The genomic DNA was PCR-amplified using primer pairs TJ719/720 or TJ699/722. After validating amplification on a 1% agarose gel, the amplicon was prepared for Sanger sequencing with the primer closest to the expected cleavage site. The chromatograms obtained from each sequencing reaction were analyzed using TIDE analysis [36]. Each gRNA tested was analyzed against a non-PAM (AAAT) negative control targeting a site in the *DNMT1* gene (Supplementary Table 2).

5.2.6 - Statistical analyses

For a subset of the data, we performed statistical analyses to determine statistical significance between multiple datasets. The end-point fluorescence measurements were inputted in MATLAB's ttest2 function, which used a two-tailed t-test and a 95% confidence interval to discern statistical differences between the two samples tested.

5.3 - RESULTS

5.3.1 - A phylogenetically diverse set of Cas12a nucleases exhibit ranging effective activities in TXTL

To begin interrogating a cross-section of Cas12a nucleases found in nature, we identified a representative set from different organisms exhibiting varying extents of homology between each other and with well-established nucleases (Figure 5.1A). In particular, we were interested in identifying a set of nucleases that exhibited varying homology to the established and well-characterized Cas12a nucleases from *Francisella novicida* U112 (FnCas12a), *Lachnospiraceae bacterium* ND2006 (LbCas12a), and *Acidaminococcus* sp. BV3L6 (AsCas12a). The resulting set of nucleases are native to

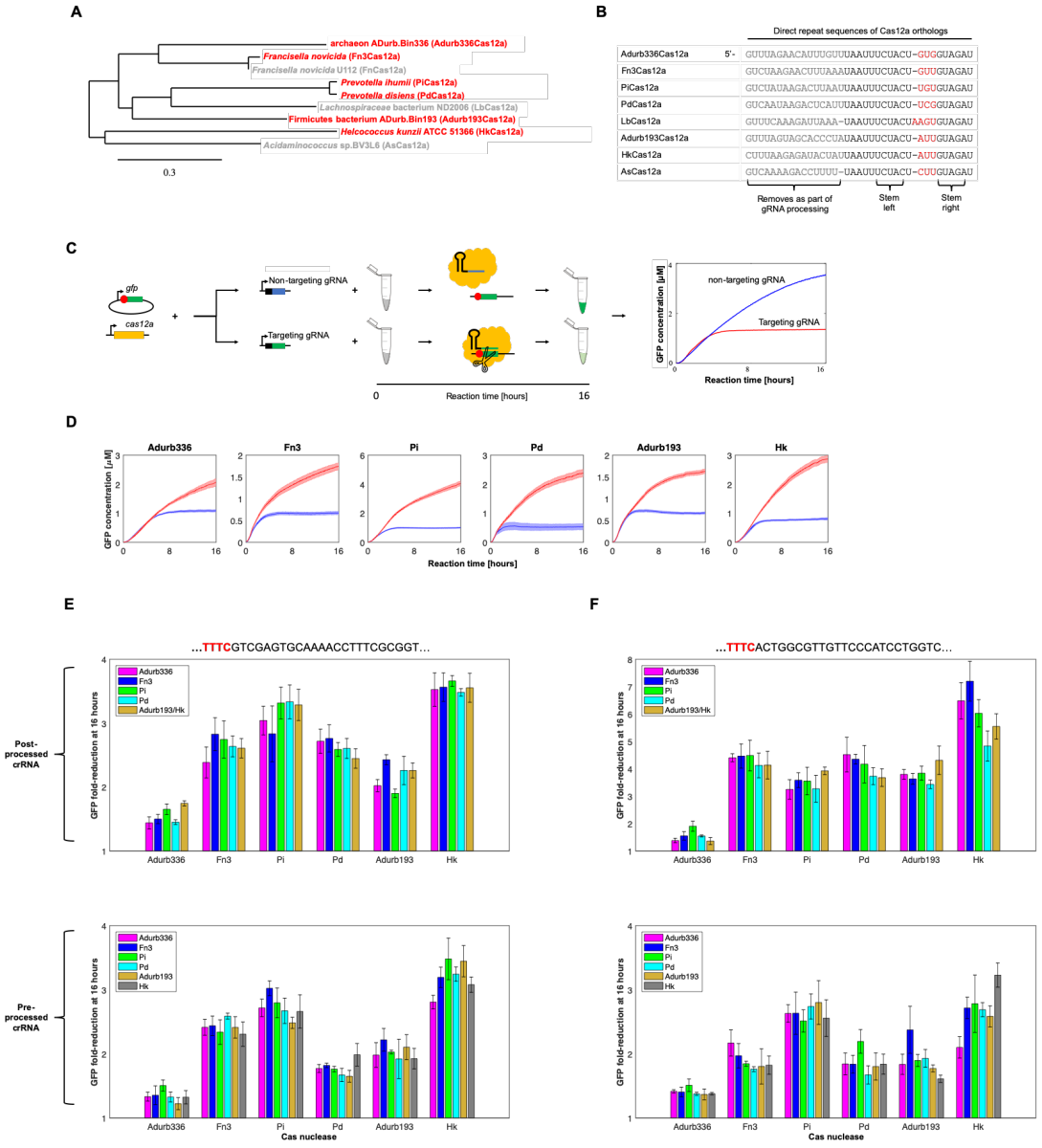


Figure 5.1: Phylogenetically diverse Cas12a nucleases exhibit varying cleavage activities and can process and utilize each other's gRNAs. (A) Phylogenetic tree of the Cas12a nucleases analyzed in this work (red) and other commonly used Cas12a nucleases (gray). (B) Direct repeat sequences of the nucleases represented in the phylogenetic tree. Sequences removed post crRNA maturation are indicated in gray text. The black text represents conserved bases of the processed crRNA across all shown Cas12a orthologs. Red text indicates non-conserved bases. Note that Fn3Cas12a shares the same crRNA as FnCas12a. (C) Representative figure showing the DNA cleavage assay in TXTL. A GFP reporter plasmid, Cas12a plasmid, and a targeting or non-targeting gRNA are added to a TXTL reaction. At the beginning of the reactions, each reaction expresses GFP, Cas12a, and the gRNA. The reaction containing the non-targeting gRNA will continue to express GFP as the gRNA will be unable to direct the Cas12a to cleave the GFP reporter plasmid, while the reaction containing the targeting gRNA will halt GFP production cleavage of the remaining reporter plasmid. (D) Time-series of GFP expression in TXTL for 16 hours at 29°C. Each Cas12a nuclease was expressed along with a non-targeting (red) or targeting crRNA (blue), with the latter designed to target a site upstream of a constitutive promoter expressing GFP. Cleavage of the reporter plasmid leads to rapid degradation and loss of GFP expression. The PAMs associated with each GFP reporter are labeled on each row, while the Cas12a nuclease species are labeled on the top. The error bars represent the standard deviation from three separate TXTL reactions. (E/F) Fold-reduction of GFP expression for each Cas12a nuclease targeting two different protospacers flanked by TTTC PAMs (bolded in red). For both target sequences, each Cas12a target and cleave the DNA using each other's full-length (bottom row) and mature (top row) crRNAs. Note that the Adurb193 and HkCas12a share the same processed crRNA sequence. The error bars represent the standard deviation from three separate TXTL reactions.

Firmicutes bacterium ADurb.Bin193 (Adurb193Cas12a), archaeon ADurb.Bin336 (Adurb336Cas12a), *Francisella novicida* (Fn3Cas12a), *Helcococcus kunzii* ATCC 51366 (HkCas12a), *Prevotella ihumii* (PiCas12a), and *Prevotella disiens* (PdCas12a) and exhibit varying similarities (Figure 5.1A) and associated direct repeat sequences (Figure 5.1B). Of these nucleases, PdCas12a was characterized *in vitro* as part of the original report of Cas12a [24], while HkCas12a was shown very recently to recognize a few PAM sequences *in vitro* adhering to a 5' YYN (Y = C/T) motif and enact genome-editing in human cells [37]. To our knowledge, the other four nucleases remain uncharacterized. Across this set, FnCas12a and Fn3Cas12a share 91.4% protein sequence homology, while PiCas12a and PdCas12a share 95.7% protein sequence homology. Though Adurb336Cas12a and Adurb193Cas12a are distinct from the other nucleases, they are most closely-related to FnCas12a (40.8% protein sequence homology) and LbCas12a (33.9% protein sequence homology). Also, while HkCas12a is distinct from the other orthologs, it is most closely-related to AsCas12a (33.8% protein sequence homology).

We first assessed the ability of these six nucleases to cleave target DNA using an all-*E. coli* cell-free TXTL assay we previously applied for characterizing CRISPR-Cas systems [32,33]. As part of the assay, DNA constructs each encoding the nuclease, a targeting or non-targeting gRNA, and a targeted GFP reporter were added to the TXTL mix, and GFP fluorescence was monitored over time using a plate reader (Figure 5.1C). Loss of GFP production reflects expression of the active nuclease:gRNA complex, cleavage of the GFP reporter plasmid, and subsequent plateauing of GFP expression. In this case, each Cas12a was expressed with a gRNA with processed repeat, 24-nt guide, and downstream rho-independent terminator. The target sequence was flanked by a 5' TTTC PAM recognized by all known Cas12a nucleases [24]. The resulting fluorescence time-courses from the assays confirmed that all nucleases could actively cleave the target DNA (Figure 5.1D). We do note that the cleavage efficiency—a reflection of the time to express the nuclease, form the ribonucleoprotein complex, and finally bind and cleave the target DNA—varied, with the fold-reduction in GFP levels (comparing

targeting and non-targeting gRNAs) after a 16-h reaction at 29°C ranging between 1.9-fold (for Adurb336Cas12a) and 4.5-fold (for PiCas12a).

5.3.2 - The Cas12a nucleases can process and utilize each other's gRNAs

Due to the phylogenetic diversity exhibited between these Cas12a nucleases, we asked whether each nuclease could utilize the other's gRNAs. Similar to prior work on characterizing Cas12a nucleases [24,37], the gRNAs associated with the Cas12a nucleases reported here share the same left and right stem sequences, with variability in the sequences removed post gRNA maturation (Figure 5.1B). The length and sequence of the connecting loop between the two stems also varied, though the total length of the full repeat was well conserved (35 – 36 nts) across these nucleases.

Cas12a binds a canonical hairpin formed near the 3' end of the repeat and cleaves immediately upstream of this hairpin through an endoribonuclease domain within the nuclease (Figure 5.1C) [25], where gRNAs have been expressed with the full-length repeat (~36 nts) or a processed repeat (~19 nts) [24,25]. We therefore used two sets of gRNAs as part of the TXTL-based DNA cleavage assay: one set containing the processed versions of the repeat, and another set containing the unprocessed gRNAs. We note that the processed repeat is identical for two nucleases (Adurb193Cas12a, HkCas12a). To account for a potential impact of the target sequence, we targeted two different sites flanked by a 5' TTTC PAM that were introduced into the GFP reporter plasmid (Figure 5.1E). We then performed the cleavage assay using each nuclease paired with each gRNA and target site. Based on the resulting fluorescence time-courses, all Cas12a nucleases tested were able to process as well as utilize each other's gRNAs (Figures 5.1E and Supplementary Figure 5.1). This finding matched that of previous work showing that AsCas12a and LbCas12a were able to utilize gRNAs associated with phylogenetically diverse Cas12a nucleases when the sequence length of the loop region was conserved [30]. We conclude that all of the tested Cas12a nucleases could utilize each other's gRNA with a full-length or processed version of the repeat, indicating one commonly shared property across the nucleases.

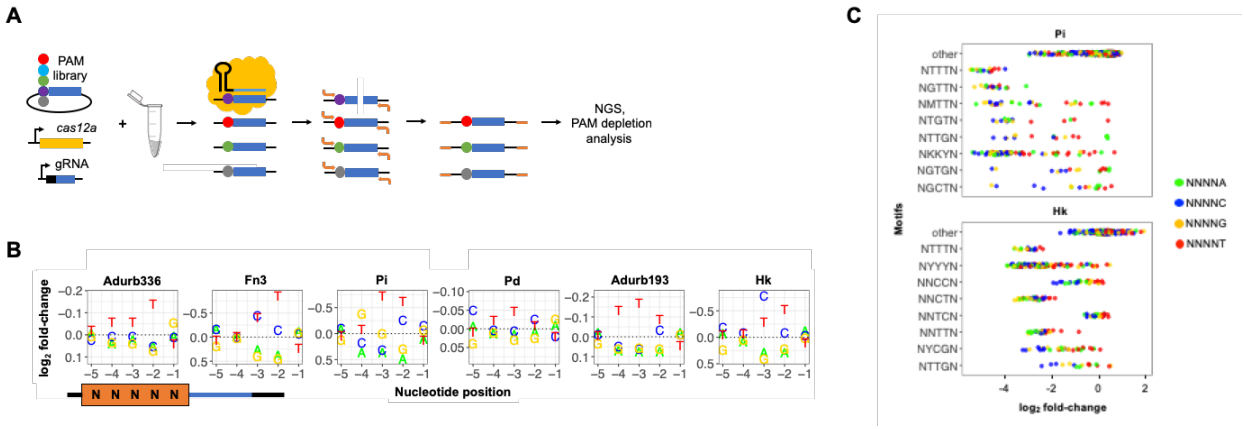


Figure 5.2: The characterized Cas12a nucleases are associated with various cleavage efficiencies and PAM sequences. (A) Representative figure showing the TXTL-based PAM screen. A 5N PAM library, Cas12a, and a targeting or non-targeting crRNA are added to a TXTL reaction. After a 16-hour incubation at 29°C, the library members containing a recognizable PAMs are cleaved, while the non-PAMs are left remaining in the reaction. The non-PAM members are PCR amplified and prepared for NGS. (B) Plots representing the fold-change in the nucleotide frequency at each PAM position in the 5N PAM library for each Cas12a tested. Note the inverted y-axis. (C) Fold-change of various depleted motifs seen in the PAM screen for PiCas12a and HkCas12a.

5.3.3 - PAM determination reveals distinct recognition profiles

We next were interested in deciphering the PAM profiles recognized by these Cas12a nucleases and how the profiles compared. To this end, we performed a TXTL-based PAM screen that we developed previously [32,33] (Figure 5.2A). As part of the assay, separate TXTL reactions were performed containing DNA encoding a Cas12a, a targeting or non-targeting gRNA, and a target sequence flanked by a 5N library of potential PAM sequences. The TXTL reactions containing each Cas12a resulted in the cleavage of library members containing a recognizable PAM. After a 16-hour incubation at 29°C, the uncleaved library members were PCR-amplified and subjected to NGS. The NGS data were analyzed for PAM depletion, with PAM recognition increasing with the extent of depletion.

The PAM profiles deciphered for each Cas12a (Figures 5.2B and 5.2C) showed that all six nucleases recognized the canonical TTTV motif commonly associated with Cas12a nucleases [24]. However, there were notable deviations from this PAM often unique to each nuclease. For instance, the PAM profiles for Adurb336Cas12a, PdCas12a, and Adurb193Cas12a all closely resembled those for LbCas12a and AsCas12a, with the consensus TTTV and the ability to accommodate a C at the -2 position. Separately, the PAM profile for Fn3Cas12a closely resembled that of its close ortholog FnCas12a, with

the consensus YTV (Y = C/T) and no strong bias at the -4 position [24]. Finally, the PAM profiles for HkCas12a and PiCas12a recognized highly flexible yet distinct PAM profiles. Specifically, HkCas12a recognized motifs with a mixture of C or T, with a clear bias toward C at the -3 position and T at the -2 position. There was also evidence that HkCas12a could weakly recognize a G at the -2 position when the -4 and -3 positions were C/T and T, respectively. Separately, PiCas12a recognized a multitude of motifs, including NTTN and KKYV (K = G/T) at an efficiency comparable to TTTV. At lower efficiencies, PiCas12a was also able to recognize TGTV, TTGB (B = C/G/T), and GYK motifs. Interestingly, both HkCas12a and PiCas12a exhibited less of a bias against a T at the -1 position, unlike the other Cas12a nucleases as well as the well-characterized Cas12a nucleases [24]. Most importantly, the different PAM profiles rarely related to the phylogeny of the Cas12a nucleases, as underscored by the contrasting PAM profiles of PiCas12a and PdCas12a despite sharing 95.7% protein sequence homology. This insight is further supported by accounting for a larger set of Cas12a nucleases we previously subjected to the TXTL-based PAM assay [33], revealing little correlation between PAM profiles and phylogenetic relationships (Supplementary Figure 5.2). Taken together, Cas12a nucleases can be associated with varying cleavage efficiencies and PAM specificities which can deviate between otherwise phylogenetically-related nucleases.

5.3.4 - Variable bias against T at the -1 PAM position confirmed by TXTL

We next aimed to confirm the variable PAM recognition across the set of Cas12a nucleases. One of the more notable insights was the lack of bias against T at the -1 position of the PAM for HkCas12a and PiCas12a. We directly interrogated the impact of this PAM position for all six nucleases using the TXTL-based DNA cleavage assay (Figure 5.3). As part of the assay, we used a 5' TTTN PAM flanking the target sequence in the GFP reporter plasmid. We observed that all tested nucleases were able to recognize the TTTV motif, albeit at various cleavage efficiencies, with marginal differences based on the identity of the V. By contrast, four of the nucleases (Adurb336Cas12a, Fn3Cas12a, PdCas12a, Adurb193Cas12a) exhibited significantly reduced cleavage with T versus V at the -1 PAM position ($p = <0.0001, 0.0002, <0.0001$

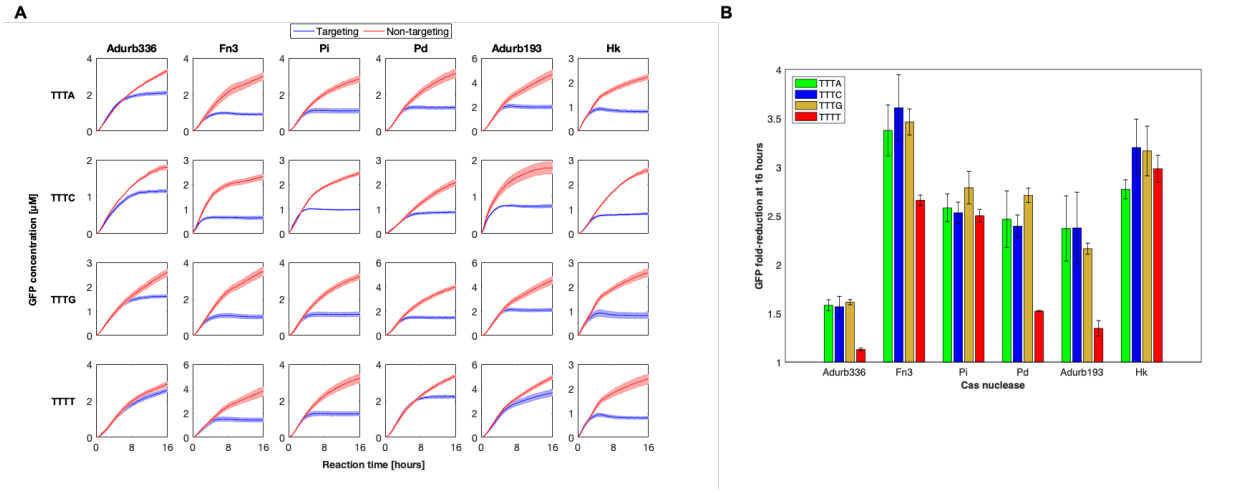


Figure 5.3: The Cas12a nucleases are associated with variable cleavage efficiencies and PAM sequences. (A) Time-series of GFP expression in TXTL for 16 hours at 29°C. Each Cas12a nuclease was expressed along with a non-targeting (red) or targeting crRNA (blue), with the latter designed to target a site upstream of a constitutive promoter expressing GFP. Cleavage of the reporter plasmid leads to rapid degradation and loss of GFP expression. The PAMs associated with each GFP reporter are labeled on each row, while the Cas12a nuclease species are labeled on the top. The error bars represent the standard deviation from three separate TXTL reactions. (B) Fold-reduction of GFP expression for each Cas12a and motif indicated. The fold-reduction was calculated using the GFP fluorescence data from the 16-hour time point from the reactions containing the targeting non-targeting crRNA. The error bars represent the standard deviation from three separate TXTL reactions.

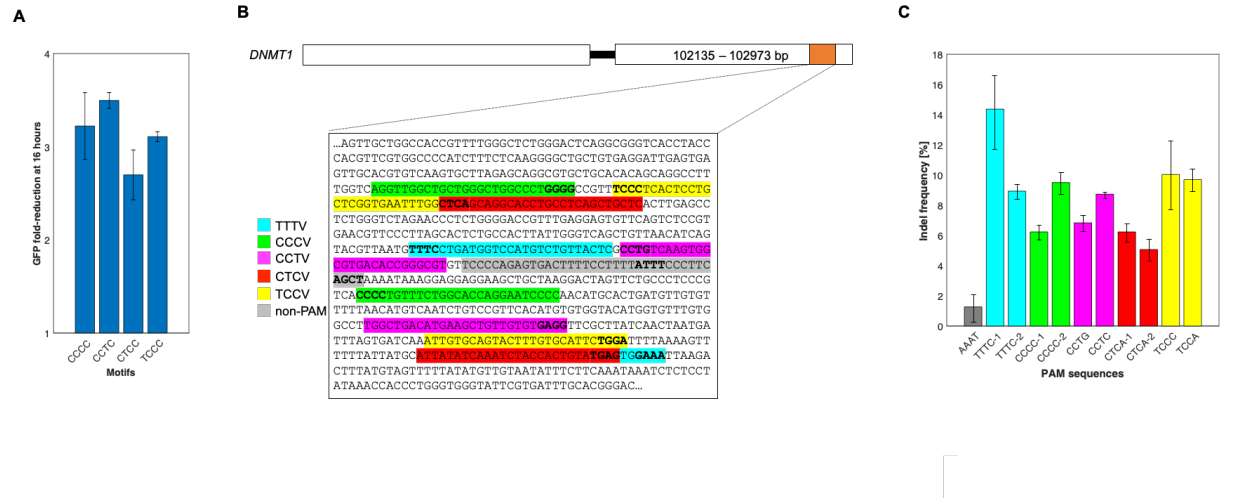


Figure 5.4: HkCas12a recognize C-rich PAM sequences in TXTL and mammalian cells. (A) Fold-reduction of GFP expression for HkCas12a targeting sequences containing various C-rich PAMs. The fold-reduction was calculated using the GFP fluorescence data from the 16-hour time point from the reactions containing the targeting non-targeting crRNA. The error bars represent the standard deviation from three separate TXTL reactions. (B) A section of the *DNMT1* gene containing the targeted sequences (highlighted) and their associated PAMs (bold) used for the indel formation assays. The sequences highlighted in cyan and gray represent the targets associated with the TTTC and unrecognized (AGCT, AAAT) motifs, respectively. The other colors represent targets associated with non-canonical motifs that were tested in the DNA cleavage assay (Figure 3B/D). Note that some of the target sequences and their associated PAMs are located in the bottom strand. (C) Indel frequencies for each *DNMT1* target site tested in HEK293T cells quantified using TIDE (PMID: 25300484). The indel frequency data for each target sequence were compared to that of a target sequence flanked by a non-PAM (AGCT). The error bars represent the standard deviation from three independent transfections.

, <0.0001, respectively comparing TTTT versus TTTV). The two exceptions were PiCas12a and HkCas12a, which exhibited cleavage that was statistically indistinguishable between T versus V at this PAM position (p = 0.1519, 0.6522, respectively comparing TTTT versus TTTV). These findings confirm that PiCas12 and HkCas12a do not possess an obvious bias against T at the -1 position, adding to the list of Cas12a nucleases with this distinct capability [37]. Given that PiCas12a and HkCas12a are phylogenetically distinct and separated by other Cas12a nucleases that were biased against T at the -1 PAM position, this finding provides further support that PAMs do not necessarily track with phylogenetic relationships.

5.3.5 - HkCas12a recognizes C-rich PAMs in TXTL and in human cells

From the PAM screen, HkCas12a exhibited an ability to efficiently recognize C-rich motifs (Figure 5.2), a capability not associated with other characterized Cas12a nucleases. To interrogate this unique recognition further, we applied the TXTL-based DNA cleavage assay to assess HkCas12a's ability to recognize CCCC, TCCC, CTCC, and CCTC as PAMs (Figure 5.4A). HkCas12a exhibited cleavage activity for all four sequences. We did notice that CTCC yielded the lowest cleavage efficiency, which matches the strong preference for C over T in the -3 position as indicated in the PAM wheel (Figure 5.2C). However, we note that the GFP fold-reduction was significantly different only when comparing CTCC to CCTC (p = 0.0079) and TCCC (p = 0.0021) but not to CCCC (p = 0.069) (Figure 5.4A).

We next asked if similar sequences could be recognized by HkCas12a in mammalian cells. Our results from the PAM screen and TXTL-based DNA cleavage assay suggested that this nuclease did not explicitly require T in the PAM, and a T at the -3 position could reduce nuclease activity. To assess this directly, we designed gRNAs targeting different sites in the *DNMT1* gene in HEK293T cells (Figure 5.4B and Supplementary Table 5.2) following our prior success generating indels in this gene and cell line with AsCas12a [29]. Sites were selected to capture a range of PAM sequences, including TTTC, CCCC, CCTG, CCTC, CTCA, TCCC, and TCCA. We transiently transfected a plasmid expressing HkCas12a and a plasmid expressing processed

gRNAs. After 72 hours post-transfection, we analyzed the frequency of indels formed by TIDE analysis [36] using the non-PAM AGCT as the reference (Figure 5.4C). HkCas12a generated indels for all targets that were all significantly different than a control with the non-PAM AAAT ($p < 0.05$). The lowest indel frequency was associated with a site flanked by CTCA, in line with the reduced preference for T at the -3 PAM position.

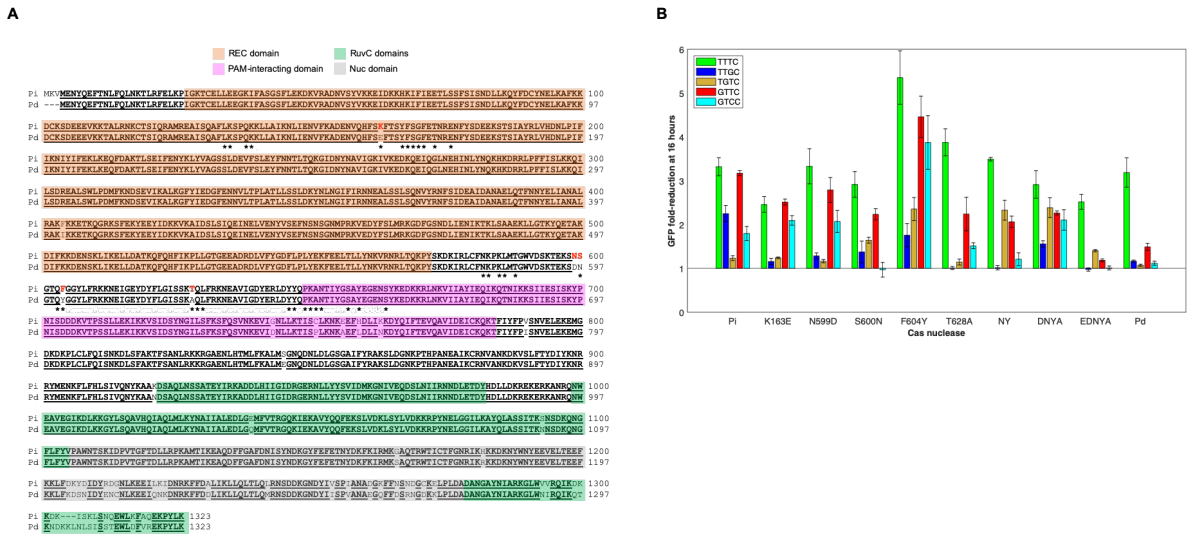


Figure 5.5: Altering residues in the PAM domain between Pd and PiCas12a modifies PAM recognition of PiCas12a. (A) Sequence alignment of Pd and PiCas12a. These nucleases share 95.3% sequence homology, with most of the variations stemming from the 3' end of the sequence. The black/bold/underlined sequences indicate matching sequences, while regular black text indicates unmatched residues. The red/bolded sequence indicate the residues investigated in this work. The asterisks indicate residues shown to alter PAM specificity when aligned with AsCas12a (PMID: 28581492, 30742127) (see Supplementary Document 5.1 for alignment with AsCas12a). The sequences were aligned using MUSCLE with default settings. (B) Fold-reduction of GFP expression for each PiCas12a variant and motif indicated. The fold-reduction was calculated using the GFP fluorescence data from the 16-hour time point from the reactions containing the targeting non-targeting crRNA (Supplementary Figure 5.3A). The error bars represent the standard deviation from three separate TXTL reactions.

However, another site flanked by CTCA yielded indel frequency statistically indistinguishable from those flanked by CCCC ($p = 1$) and CCTG ($p = 0.2563$), indicating that the target sequence is a major contributor to the editing efficiency as we and others observed previously [29,38]. These findings demonstrate the flexibility of HkCas12a for genome targeting in human cells and align with recent findings [37].

5.3.6 - Mutating PiCas12a toward PdCas12a reveals distinct PAM profiles in TXTL

We next turned our attention to PiCas12a, which shared 95.7% protein sequence homology to PdCas12a (Figure 5.5A) yet exhibited a distinct PAM profile that could

accommodate G at the -4 and -2 PAM positions and a G at the -3 position to a lesser extent (Figures 5.2B and 5.2C). The capacity to accommodate G at different positions particularly differentiates PiCas12a from other previously characterized Cas12a nucleases [24,37]. We first tested PiCas12a's ability to recognize GTTC, TGTC, TTGC, GTCC, GGTC, and GGCC motifs as part of the TXTL-based DNA cleavage assay (Figure 5.5B and Supplementary Figure 5.3). While PiCas12a was not able to recognize GGTC, GGCC, or GTGC (Supplementary Figure 5.3), the nuclease was able to recognize PAMs containing a single G in either the -4 or -2 positions (Figure 5.5B).

The disparate PAM profiles distinguishing PiCas12a and PdCas12a raised the question: which of the 57 mutations are responsible for these differences? Previous work showed that the PAM specificity of AsCas12a could be altered by mutating specific residues located within or between the recognition and PAM-interacting domains [39–41]. We therefore aligned the protein sequences of AsCas12a, PiCas12a, and PdCas12a to identify residues that altered the PAM preferences of AsCas12a and varied between PiCas12a and PdCas12a (Supplementary Document 5.1). This led us to identifying five residues in PiCas12a: K163, N599, S600, F604, and T628 (Figure 5.5A and Supplementary Document 5.1). Thus, we converted each or a combination of these residues separately in PiCas12a to match that of PdCas12a to create the PiCas12a variants with K163E, N599D, S600N, F604Y, T628A, NY (S600N/F604Y), DNYA (containing the four mutations N599D/S600N/F604Y/T628A), and EDNYA (containing all five mutations).

We hypothesized that these mutations would alter the PAM specificity of PiCas12a such that it could no longer recognize the non-canonical PAM sequences. We thus performed the TXTL-based DNA cleavage assay with these PiCas12a variants as well as PdCas12a along with the same single-G PAM sequences tested with PiCas12a (GTTC, TGTC, TTGC, GTCC) and the canonical TTTC (Figures 5.5B and Supplementary Figure 5.4). As expected, PdCas12a only strongly recognized TTTC and modestly recognized GTTC, similar to what we observed with AsCas12a [29]. We also found that introducing all five mutations (EDNYA) similarly yielded only strong recognition of TTTC.

The other variants, by contrast, each recognized at least one of the non-canonical sequences recognized by PiCas12a. In many cases, each mutation reduced or eliminated recognition of a non-canonical PAM, such as for recognition of TTGC by the K163E, N599D, S600N, T628A, and NY variants, or recognition of GTCC by the S600N and NY variants. However, the mutations also altered PAM recognition in ways that did not follow the expected progression from PiCas12a to PdCas12a. One example was that some mutations enhanced recognition of motifs poorly recognized by either PiCas12a or PdCas12a. Specifically, the S600N, F604Y, NY, and DNYA variants exhibited enhanced recognition of TGTC, while the K163E, F604Y, and DNYA variants exhibited enhanced recognition of GTCC. Another example was the non-additive effect of the mutations. In particular, the DNYA variant exhibited the most flexible PAM recognition, while the addition of the K163E mutation that exhibited enhanced recognition of GTCC yielded the EDNYA variant that only recognized the TTTC sequence. These findings reveal the complex landscape for PAM recognition that can occur even between closely-related orthologs of Cas12a.

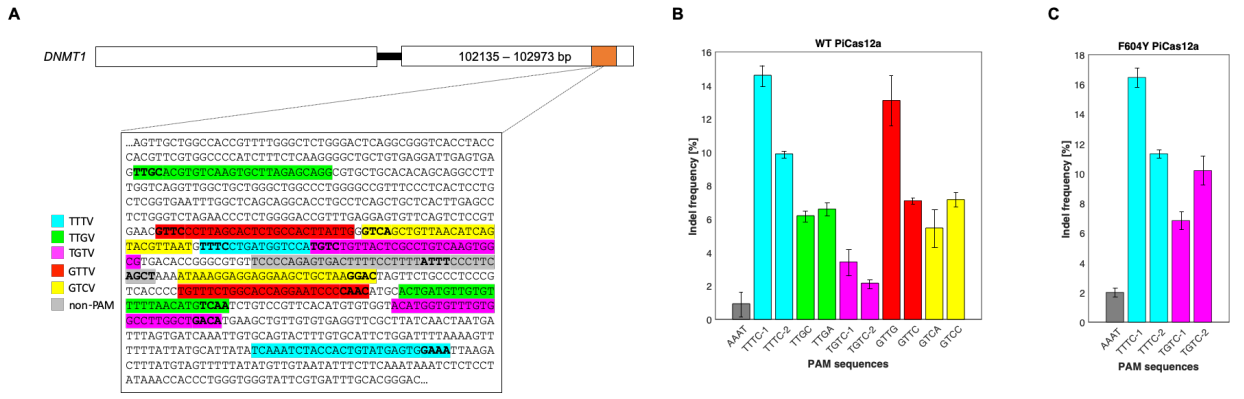


Figure 5.6: PiCas12a and one of its variants can recognize various non-canonical motifs in mammalian cells. (A) A section of the *DNMT1* gene containing the targeted sequences (highlighted) and their associated PAMs (bold) used for the indel formation assays. The sequences highlighted in cyan and gray represent the targets associated with the TTTC and unrecognized (AGCT, AAAT) motifs, respectively. The other colors represent targets associated with non-canonical motifs that were tested in the DNA cleavage assay (Figure 3B/D). Note that some of the target sequences and their associated PAMs are located in the bottom strand. (B) Indel frequencies for each *DNMT1* target site tested in HEK293T cells quantified using TIDE (PMID: 25300484). The indel frequency data for each target sequence were compared to that of a target sequence flanked by a non-PAM (AGCT). The error bars represent the standard deviation from three independent transfections.

5.3.7 - PiCas12a and the F604Y variant recognize distinct non-canonical PAMs in human cells

Finally, we sought to test PiCas12a and the F604Y variant in human cells due to their unique PAM sequences. We chose the F604Y PiCas12a variant to target sequences with the same TGTN motifs targeted with the wild-type nuclease, as it was one of the variants that effectively recognized the TGTC motif in the TXTL-based DNA cleavage assay (Figures 5.5B and Supplementary Figure 5.4A). We performed this assay following the experimental set up used for the indel formation assay with HkCas12a (Figure 5.4). For these experiments, we tested targets flanked by PAM sequences that were investigated as part of the TXTL-based DNA cleavage assay (Figures 5.5B, 5.6A, and Supplementary Table 5.2).

Both nucleases yielded indel formation across all tested PAMs, albeit with noticeable differences (Figure 5.6B). PiCas12a recognized PAMs associated with the TTGN, GTTN, and GTCN motifs. While wild-type PiCas12a weakly formed indels targeting sequences associated with the TGTN motif, the F604Y variant was able to form indels targeting the same sequence. Interestingly, some of the non-canonical PAMs tested for these nucleases were able to form indels more efficiently than the TTTC PAMs, though this phenomenon is likely due to the chosen target sequences as seen in prior work [38]. These data showed that PiCas12a and the F604Y variant can target unique non-canonical PAMs in mammalian cells, with the F604Y variant exhibiting a more flexible PAM profile.

5.4 - DISCUSSION

Here, we characterized six phylogenetically diverse Cas12a nucleases, including the previously reported Hk and PdCas12a [24,37], as well as previously unreported Adurb193, Adurb336, Fn3, and PiCas12a. While Fn and Fn3Cas12a were associated with high sequence homology and shared common PAM profiles, the PAM profiles of the other nucleases characterized in this work did not track with phylogeny (Figures 5.1 and 5.2). In particular, Hk and PiCas12a were associated with distinct PAM profiles that varied greatly compared to other closely-related orthologs, such as their ability to

tolerate a T in the -1 PAM position (Figure 5.3). HkCas12a recognized a C-rich PAM, with a strong bias for a C in the -3 PAM position and little bias in the -4 position (Figures 5.2B and 5.2C), which is in agreement with prior work [37]. PiCas12a recognized PAMs containing a G in the -4 and -2 PAM positions, with a strong bias for a G in the former position. While PiCas12a shared 95.7% sequence homology with the previously characterized PdCas12a, the PAM profiles between these closely-related orthologs varied significantly (Figures 5.2B and 5.5B). The distinct PAM profiles of Pi and PdCas12a highlight the remarkable diversity that can exist within CRISPR subtypes and suggest that other highly homologous nucleases could also recognize different PAM sequences.

To gain insight into the PAM profile of PiCas12a, we attempted to abolish the non-canonical PAM recognition of PiCas12a by mutating five residues, or a combination of the residues, that were previously shown to alter PAM recognition or were proximal to these residues [39,40] to match that of PdCas12a (Figure 5.5A, Supplementary Document 5.1). The PiCas12a variants investigated in this work includes the K163E, N599D, S600N, F604Y, and T628A variants, as well as the combinatorial NY (S600N/F604Y), DNYA (N599D, S600N, F604Y, T628A) and EDNYA variants. While many of these variants were associated with reduced non-canonical PAM recognition (e.g. reduced recognition of GTCC from S600N/NY variants), the effects from mutating these residues were not predictable (Figure 5.5B). In fact, some of these mutations led to broadening the PAM profile of wild-type PiCas12a (i.e. TGTC recognition of S600N, F604Y, NY, DNYA variants). These mutational analyses suggest that while mutations can alter the PAM profile of CRISPR nucleases, there exists a complex relationship between the mutated residues and their effects on PAM recognition. While we investigated five specific residues based on prior work altering PAM recognition in AsCas12a [39,40], other residues may influence PiCas12a's unique PAM profile. More exhaustive mutational analyses along with detailed biochemical and structural studies will provide insights into PAM engineering of Cas12a nucleases, which in turn may lead to CRISPR nucleases associated with diverse PAM profiles.

In this work, we characterized another Cas12a originating from *Prevotella* (PiCas12a) that shared high sequence homology with PdCas12a (Figure 5.5A). The varying PAM profiles between these aforementioned nucleases are remarkable when considering their 95.7% sequence similarity. While PdCas12a is associated with the canonical T-rich PAM, PiCas12a was able to recognize a diverse set of PAM sequences (Figure 5.2B and 5.2C). These characteristics (i.e. high sequence homology/diverse PAM profiles) of Pi and PdCas12a suggest that CRISPR nucleases may undergo selective pressure for PAM diversification that is distinct from other modes of evolution evident in CRISPR systems, such as increasing spacer diversity and horizontal gene transfer [42]. This claim is supported by prior work that revealed a phage's ability to interfere with CRISPR immunity through mutating the PAM sequence [16]. PAM diversification could also be driven by anti-CRISPR proteins (Acrs), which are proteins that inhibit the activity of Cas effectors [15]. Recently, Acrs were discovered for Type V CRISPR systems [43,44] and have been shown to inhibit Cas12a through the Acr directly binding to Cas12a [45]. The only major constraint on PAM diversification appears to be avoiding self-recognition as the PAM portion of the direct repeat sequence between Pi and PdCas12a was fully conserved (Figure 5.1B). The discovery of additional CRISPR nucleases associated with high homology and distinct PAM profiles will shed light on other systems that may have evolved through this mode.

The Cas12a nucleases characterized in this work reveal the diversity that exists within CRISPR subtypes, and also expands the targeting range of Cas12a nucleases for various biotechnological applications. The 95.7% sequence similarity exhibited between Pd and PiCas12a, along with their differing PAM profiles highlight the remarkable diversity that can exist between highly homologous CRISPR nucleases. Further work to understand the biochemical structure of Pd/PiCas12a, as well as a comprehensive comparison between the genomes of the *Prevotella disiens* and *ihumii* will provide insight into evolutionary relationship between these nucleases and lead to a more diversified CRISPR toolkit.

5.5 - ACKNOWLEDGEMENTS

The PdCas12a expression plasmid pY018 (pcDNA3.1-hPdCpf1) was a gift from Feng Zhang. The empty gRNA expression plasmid (pU6-As-crRNA) was a gift from Jin-Soo Kim.

5.6 - REFERENCES

[1] R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. Romero, P. Horvath, CRISPR provides acquired resistance against viruses in prokaryotes, Science. 315 (2007) 1709–1712.

[2] R. Sorek, C.M. Lawrence, B. Wiedenheft, CRISPR-mediated adaptive immune systems in bacteria and archaea, Annu. Rev. Biochem. 82 (2013) 237–266.

[3] R. Barrangou, L.A. Marraffini, CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity, Mol. Cell. 54 (2014) 234–244.

[4] R.T. Leenay, C.L. Beisel, Deciphering, communicating, and engineering the CRISPR PAM, J. Mol. Biol. 429 (2017) 177–191.

[5] F.J. Mojica, C. Diez-Villasenor, J. Garcia-Martinez, C. Almendros, Short motif sequences determine the targets of the prokaryotic CRISPR defence system, Microbiology. 155 (2009) 733–740.

[6] M. Adli, The CRISPR tool kit for genome editing and beyond, Nat. Commun. 9 (2018) 1911.

[7] E. V. Koonin, K.S. Makarova, F. Zhang, Diversity, classification and evolution of CRISPR-Cas systems, Curr. Opin. Microbiol. 37 (2017) 67–78.

[8] K.S. Makarova, Y.I. Wolf, E. V. Koonin, Classification and nomenclature of CRISPR-Cas systems: Where from here?, Cris. J. 1 (2018) 325–336.

[9] I. Mougiakos, P. Mohanraju, E.F. Bosma, V. Vrouwe, M. Finger Bou, M.I.S. Naduthodi, A. Gussak, R.B.L. Brinkman, R. van Kranenburg, J. van der Oost, Characterizing a thermostable Cas9 for bacterial genome editing and silencing, Nat. Commun. 8 (2017) 1647.

[10] L.B. Harrington, D. Paez-Espino, B.T. Staahl, J.S. Chen, E. Ma, N.C. Kyrpides, J.A. Doudna, A thermostable Cas9 with increased lifetime in human plasma, Nat. Commun. 8 (2017) 1424.

[11] M.A. Moreno-Mateos, J.P. Fernandez, R. Rouet, C.E. Vejnar, M.A. Lane, E. Mis, M.K. Khokha, J.A. Doudna, A.J. Giraldez, CRISPR-Cpf1 mediates efficient homology-directed repair and temperature-controlled genome editing, Nat. Commun. 8 (2017) 1–9.

[12] B.P. Kleinstiver, S.Q. Tsai, M.S. Prew, N.T. Nguyen, M.M. Welch, J.M. Lopez,

Z.R. McCaw, M.J. Aryee, J.K. Joung, Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells, Nat. Biotechnol. 34 (2016) 869–874.

[13] D. Kim, J. Kim, J.K. Hur, K.W. Been, S.H. Yoon, J.S. Kim, Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells, Nat. Biotechnol. 34 (2016) 863–868.

[14] A.A. Malzahn, X. Tang, K. Lee, Q. Ren, S. Sretenovic, Y. Zhang, H. Chen, M. Kang, Y. Bao, X. Zheng, K. Deng, T. Zhang, V. Salcedo, K. Wang, Y. Zhang, Y. Qi, Application of CRISPR-Cas12a temperature sensitivity for improved genome editing in rice, maize, and Arabidopsis, BMC Biol. 17 (2019) 9.

[15] J. Bondy-Denomy, A. Pawluk, K.L. Maxwell, A.R. Davidson, Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system, Nature. 493 (2013) 429–432.

[16] H. Deveau, R. Barrangou, J.E. Garneau, J. Labonté, C. Fremaux, P. Boyaval, D.A. Romero, P. Horvath, S. Moineau, Phage response to CRISPR-encoded resistance in Streptococcus thermophilus, J. Bacteriol. 190 (2008) 1390–1400.

[17] S. Konermann, P. Lotfy, N.J. Brideau, J. Oki, M.N. Shokhirev, P.D. Hsu, Transcriptome engineering with RNA-targeting Type VI-D CRISPR effectors, Cell. 173 (2018) 665–676.

[18] W.X. Yan, S. Chong, H. Zhang, K.S. Makarova, E. V Koonin, D.R. Cheng, D.A. Scott, Cas13d is a compact RNA-targeting Type VI CRISPR effector positively modulated by a WYL-domain-containing accessory protein, Mol. Cell. 70 (2018) 327–339.

[19] W.X. Yan, P. Hunnewell, L.E. Alfonse, J.M. Carte, E. Keston-Smith, S. Sothiselvam, A.J. Garrity, S. Chong, K.S. Makarova, E. V Koonin, D.R. Cheng, D.A. Scott, Functionally diverse type V CRISPR-Cas systems, Science. 363 (2019) 88–91.

[20] P. Chatterjee, N. Jakimo, J.M. Jacobson, Minimal PAM specificity of a highly similar SpCas9 ortholog, Sci. Adv. 4 (2018) eaau0766.

[21] I. Fonfara, A. Le Rhun, K. Chylinski, K.S. Makarova, A.L. Lécrivain, J. Bzdrenga, E. V. Koonin, E. Charpentier, Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas

systems, Nucleic Acids Res. 42 (2014) 2577–2590.

[22] A.E. Briner, P.D. Donohoue, A.A. Gomaa, K. Selle, E.M. Slorach, C.H. Nye, R.E. Haurwitz, C.L. Beisel, A.P. May, R. Barrangou, Guide RNA functional modules direct Cas9 activity and orthogonality, Mol. Cell. 56 (2014) 333–339.

[23] A.C. Komor, A.H. Badran, D.R. Liu, CRISPR-based technologies for the manipulation of eukaryotic genomes, Cell. 168 (2017) 20–36.

[24] B. Zetsche, J.S. Gootenberg, O.O. Abudayyeh, A. Regev, E. V Koonin, F. Zhang, I.M. Slaymaker, K.S. Makarova, P. Essletzbichler, S.E. Volz, J. Joung, J. Van Der Oost, Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system, Cell. 163 (2015) 759–771.

[25] I. Fonfara, H. Richter, M. Bratovič, A. Le Rhun, E. Charpentier, The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA, Nature. 532 (2016) 517–521.

[26] J.S. Chen, E. Ma, L.B. Harrington, M. Da Costa, X. Tian, J.M. Palefsky, J.A. Doudna, CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity, Science. 6245 (2018) eaar6245.

[27] Y.E. Tak, B.P. Kleinstiver, J.K. Nuñez, J.Y. Hsu, J.E. Horng, J. Gong, J.S. Weissman, J.K. Joung, Inducible and multiplex gene regulation using CRISPR–Cpf1-based transcription factors, Nat. Methods. 14 (2017) 1163–1166.

[28] J.S. Gootenberg, O.O. Abudayyeh, M.J. Kellner, J. Joung, J.J. Collins, F. Zhang, Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6, Science. 360 (2018) 439–444.

[29] T. Jacobsen, C. Liao, C.L. Beisel, The Acidaminococcus sp. Cas12a nuclease recognizes GTTV and GCTV as non-canonical PAMs, FEMS Microbiol. Lett. (2019).

[30] B. Li, W. Zhao, X. Luo, X. Zhang, C. Li, C. Zeng, Y. Dong, Engineering CRISPR–Cpf1 crRNAs and mRNAs to maximize genome editing efficiency, Nat. Biomed. Eng. 1 (2017) 66.

[31] A. Dereeper, F. Chevenet, G. Blanc, J.-F. Dufayard, J.-M. Claverie, M. Lescot, S. Audic, S. Buffet, S. Guindon, V. Guignon, V. Lefort, O. Gascuel, Phylogeny.fr: robust phylogenetic analysis for the non-specialist, Nucleic Acids Res. 36 (2008)

W465–W469.

[32] C.S. Maxwell, T. Jacobsen, R. Marshall, V. Noireaux, C.L. Beisel, A detailed cell-free transcription-translation-based assay to decipher CRISPR protospacer-adjacent motifs, Methods. 143 (2018) 48–57.

[33] R. Marshall, C.S. Maxwell, S.P. Collins, T. Jacobsen, M.L. Luo, M.B. Begemann, B.N. Gray, E. January, A. Singer, Y. He, C.L. Beisel, V. Noireaux, Rapid and scalable characterization of CRISPR technologies using an E. coli cell-free transcription-translation system, Mol. Cell. 69 (2018) 146–157.

[34] J. Garamella, R. Marshall, M. Rustad, V. Noireaux, The all E. coli TX-TL toolbox 2.0: A platform for cell-free synthetic biology, ACS Synth. Biol. 5 (2016) 344–355.

[35] R.T. Leenay, K.R. Maksimchuk, R.A. Slotkowski, R.N. Agrawal, A.A. Gomaa, A.E. Briner, R. Barrangou, C.L. Beisel, Identifying and visualizing functional PAM diversity across CRISPR-Cas systems, Mol. Cell. 62 (2016) 137–147.

[36] E.K. Brinkman, T. Chen, M. Amendola, B. Van Steensel, Easy quantitative assessment of genome editing by sequence trace decomposition, Nucleic Acids Res. 42 (2014) 1–8.

[37] F. Teng, J. Li, T. Cui, K. Xu, L. Guo, Q. Gao, G. Feng, C. Chen, D. Han, Q. Zhou, W. Li, Enhanced mammalian genome editing by new Cas12a orthologs with optimized crRNA scaffolds, Genome Biol. 20 (2019) 15.

[38] K.M. Esvelt, P. Mali, J.L. Braff, M. Moosburner, S.J. Yaung, G.M. Church, Orthogonal Cas9 proteins for RNA-guided gene regulation and editing, Nat. Methods. 10 (2013) 1116–1121.

[39] L. Gao, D.B.T. Cox, W.X. Yan, J.C. Manteiga, M.W. Schneider, T. Yamano, H. Nishimasu, O. Nureki, N. Crosetto, F. Zhang, Engineered Cpf1 variants with altered PAM specificities, Nat. Biotechnol. 35 (2017) 789–792.

[40] B.P. Kleinstiver, A.A. Sousa, R.T. Walton, Y.E. Tak, J.Y. Hsu, K. Clement, M.M. Welch, J.E. Horng, J. Malagon-Lopez, I. Scarfò, M. V Maus, L. Pinello, M.J. Aryee, J.K. Joung, Engineered CRISPR–Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing, Nat. Biotechnol. (2019).

[41] T. Yamano, B. Zetsche, R. Ishitani, F. Zhang, H. Nishimasu, O. Nureki, Structural

basis for the canonical and non-canonical PAM recognition by CRISPR-Cpf1, Mol. Cell. 67 (2017) 633–645.

[42] E.R. Westra, A.J. Dowling, J.M. Broniewski, S. van Houte, Evolution and Ecology of CRISPR, Annu. Rev. Ecol. Evol. Syst. 47 (2016) 307–331.

[43] K.E. Watters, C. Fellmann, H.B. Bai, S.M. Ren, J.A. Doudna, Systematic discovery of natural CRISPR-Cas12a inhibitors, Science. 362 (2018) 236–239.

[44] N.D. Marino, J.Y. Zhang, A.L. Borges, A.A. Sousa, L.M. Leon, B.J. Rauch, R.T. Walton, J.D. Berry, J.K. Joung, B.P. Kleinstiver, J. Bondy-Denomy, Discovery of widespread type I and type V CRISPR-Cas inhibitors, Science. 362 (2018) 240–242.

[45] G.J. Knott, B.W. Thornton, M.J. Lobba, J.-J. Liu, B. Al-Shayeb, K.E. Watters, J.A. Doudna, Broad-spectrum enzymatic inhibition of CRISPR-Cas12a, Nat. Struct. Mol. Biol. (2019).

CHAPTER 6: Conclusions and future work

6.1 - CONCLUSIONS

This thesis focused on the development of tools and techniques to improve upon the current synthetic biology toolkit. **Chapter 1** provided an overview on the state of gene regulatory tools in *Drosophila*. To increase the availability of gene regulatory tools in eukaryotic systems, **Chapter 2** focused on the development of these tools based on self-cleaving ribozymes. In **Chapter 3**, a rapid technique to characterize PAMs of Cas nucleases was developed. Using the technique developed in **Chapter 3**, **Chapters 4 and 5** focused on the discovery of non-canonical PAM sequences of a previously characterized Cas nuclease (**Chapter 4**) and the characterization of a set of previously unreported Cas nucleases, with some of these nucleases being associated with unique, non-canonical PAM sequences (**Chapter 5**). While this thesis helped to expand the tools available for various biological applications, more work can be done to further expand the toolbox.

6.2 - PREDICTABLE TUNING OF GENE EXPRESSION

The work in **Chapter 2** focused on the engineering of self-cleaving ribozymes to regulate gene expression in eukaryotic systems. While this goal was achieved by creating a set of self-cleaving ribozymes/upstream competing sequences that resulted in consistent reduction of gene expression in HEK293T cells and *Drosophila* embryos, this tool could be further improved by using RNA prediction algorithms that can accurately predict RNA secondary structures. Successfully predicting RNA secondary structures from the ribozyme/upstream competing sequences would allow for predictable control of gene expression [1]. In **Chapter 2**, Mfold and Sfold were used for the predictions, but these tools did not correlate well with the experimental data. This variability could be due to effects from the flanking sequences surrounding the ribozyme and/or upstream competing sequences, mRNA stability, or from the action of endogenous biological machinery. One way to test the former hypothesis would be to insert longer insulating sequences to flank both the 5’ and 3’ ends of the ribozyme/upstream competing sequences. In theory, the addition of insulating sequences should either normalize the effects from the flanking sequences across all upstream competing sequences used or prevent any interaction between the ribozyme/upstream competing sequence with their

flanking sequences. The ability to precisely predict fold reduction from secondary structure predictions from the ribozyme/upstream competing sequences would allow for the rapid design of novel ribozyme/upstream competing sequence constructs without the need to first test various competing sequences.

6.3 - FURTHER INCREASING THE PAM FLEXIBILITY OF PICAS12A

In **Chapter 5**, PiCas12a was shown to recognize the canonical T-rich motifs (TTTN), as well as non-canonical motifs containing a G in two different positions of the PAM (i.e. **G**TTC and TT**G**C). By mutating residues shown to alter PAM recognition [2,3], the PAM flexibility of this nuclease was inadvertently increased to include PiCas12a variants with the ability to recognize the TGTC motif. While the mutated variants tested in the study increased the targeting range of this nuclease, further increasing PAM flexibility of PiCas12 would result in a nuclease with the ability to target a wide range of target sequences in a genome of interest. Previous studies have developed variants of commonly used Cas12a nucleases associated with increased PAM specificity, though these variants only resulted in a single base-pair difference when compared to the canonical TTTV motif (TT**Y**N, **V**TTV, T**R**TV) [2,3]. A structure-guided mutagenesis screen of PiCas12a could potentially result in a variant with the ability to target both T/A-rich target sequences, as well as G/C rich sequences. If successful, this nuclease has the potential to become a more prevalent tool for CRISPR technologies compared to the widely used Cas9.

6.4 - LIVE RNA IMAGING IN *DROSOPHILA*

The dynamics of mRNA plays a critical role in the spatial and temporal expression of various proteins (reviewed in [4,5]). Previous work has shown that mRNA localization is important for various aspects of multicellular development, such as cell fate determination, proper tissue functionality, and cell movement [6–8]. Advancements in imaging technologies have allowed for the visualization and quantification of mRNA dynamics in various model systems, such as *Drosophila*. While the importance of studying mRNA dynamics has been established, the toolbox of mRNA imaging is currently limited.

Currently, two imaging techniques have been mainly used to study mRNA dynamics in *Drosophila*: fluorescent *in situ* hybridization (FISH) and the MS2/PP7-MS2 coat protein (MCP)/PP7 coat protein (PCP) systems. FISH is a technique developed in *Drosophila* that uses fluorescent probes to bind to specific mRNA sequences of interest [9]. While this technique allows for spatial determination of mRNA transcripts, it is limited as a technique to image snapshots in time. More recently, the MS2/PP7-MCP/PCP was engineered to live-image nascent transcription in *Drosophila* [10,11]. This technique involves tagging the untranslated regions of specific genes with multiple MS2 or PP7 sequences and tagging the proteins that bind to the MS2 or PP7 sequences (i.e. MCP or PCP) with a fluorescent reporter. When transcription of the gene of interest initiates, multiple reporter gene-tagged MCP or PCP bind to their respective sequences and fluoresces at the site of transcription. While this tool has been widely used to study nascent transcription in *Drosophila*, a key drawback has been the failure to track mRNA transcripts outside of nuclei [11].

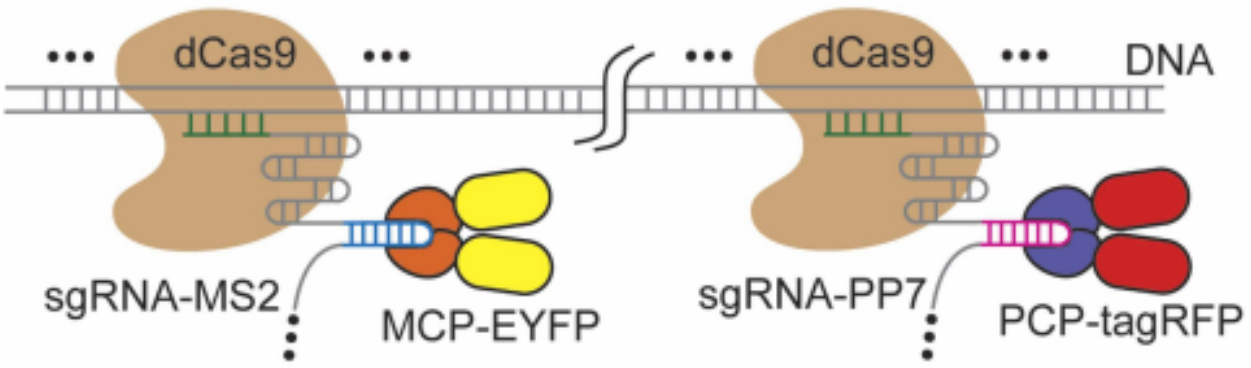


Figure 6.1: Tagging aptamers to sgRNA for live DNA imaging. Multiple MS2 (or PP7) repeat loops can be tagged onto the sgRNA of catalytically-inactive Cas9 (dCas9), while maintaining functionality (i.e. DNA targeting/binding). When a fluorescent reporter-tagged MCP (or PCP) binds to the MS2 (or PP7) loops, imaging of the targeted DNA of interest is possible. Figure adapted from [15].

In 2016, a novel Cas nuclease, named Cas13a, was discovered and characterized as an RNA-targeting CRISPR effector [12,13]. Like other Cas nucleases, specific mutations in its catalytic domain resulted in the creation of a catalytically-deactivated version of Cas13a (dCas13a) that acts as an RNA-binding protein [12,13]. Previous work has shown dCas13a's ability to image RNA in bacteria [14]. In order to harness this nuclease for RNA imaging in *Drosophila*, combining dCas13a with the MS2-MCP (or PP7-PCP) system can

be used as a tool to image both nascent and mature mRNA transcripts. Previously, another group performed similar work, but used dCas9 to image DNA (Figure 6.1) [15,16]. Briefly, by introducing 6 or 12 MS2 (or PP7) loops to the crRNA of Cas13a and tagging MCP (or PCP) with GFP, the combination of these two systems has the potential to allow for live mRNA imaging that is less prone to photobleaching [15,16]. Also, this tool should be able to track mRNA localization outside of the nucleus due to the strong binding of the Cas13a to the mRNA of interest, as well as the ability of the 6 or 12 MS2 (or PP7) loops to bind with MCP (or PCP)-GFP. If successful, this tool would allow for the tracking of mRNA transcripts in *Drosophila*, and potentially in other multicellular systems.

6.5 - REFERENCES

[1] J.M. Carothers, J. a Goler, D. Juminaga, J.D. Keasling, Model-driven engineering of RNA devices to quantitatively program gene expression, Science. 334 (2011) 1716–1719.

[2] B.P. Kleinstiver, A.A. Sousa, R.T. Walton, Y.E. Tak, J.Y. Hsu, K. Clement, M.M. Welch, J.E. Horng, J. Malagon-Lopez, I. Scarfò, M. V Maus, L. Pinello, M.J. Aryee, J.K. Joung, Engineered CRISPR–Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing, Nat. Biotechnol. (2019).

[3] L. Gao, D.B.T. Cox, W.X. Yan, J.C. Manteiga, M.W. Schneider, T. Yamano, H. Nishimasu, O. Nureki, N. Crosetto, F. Zhang, Engineered Cpf1 variants with altered PAM specificities, Nat. Biotechnol. 35 (2017) 789–792.

[4] K.C. Martin, A. Ephrussi, mRNA Localization: Gene Expression in the Spatial Dimension, Cell. 136 (2009) 719–730.

[5] C. Eliscovich, A.R. Buxbaum, Z.B. Katz, R.H. Singer, mRNA on the move: The road to its biological destiny, J. Biol. Chem. 288 (2013) 20361–20368.

[6] A. Jedrusik, D.E. Parfitt, G. Guo, M. Skamagki, J.B. Grabarek, M.H. Johnson, P. Robson, M. Zernicka-Goetz, Role of Cdx2 and cell polarity in cell allocation and specification of trophectoderm and inner cell mass in the mouse embryo, Genes Dev. 22 (2008) 2692–2706.

[7] E.H. Kislauskis, X. Zhu, R.H. Singer, Beta-Actin messenger RNA localization and protein synthesis augment cell motility, J. Cell Biol. 136 (1997) 1263–1270.

[8] M. Tolino, M. Kohrmann, M.A. Kiebler, RNA-binding proteins involved in RNA localization and their implications in neuronal diseases, Eur. J. Neurosci. 35 (2012) 1818–1836.

[9] E. Lécuyer, N. Parthasarathy, H.M. Krause, Fluorescent in situ hybridization protocols in Drosophila embryos and tissues., Methods Mol Biol. 420 (2008) 289–302.

[10] T.T. Weil, K.M. Forrest, E.R. Gavis, Localization of bicoid mRNA in late oocytes Is maintained by continual active transport, Dev. Cell. 11 (2006) 251–262.

[11] K.M. Forrest, E.R. Gavis, Live Imaging of endogenous RNA reveals a diffusion and

entrapment mechanism for nanos mRNA localization in Drosophila, Curr. Biol. 13 (2003) 1159–1168.

[12] O.O. Abudayyeh, J.S. Gootenberg, S. Konermann, J. Joung, I.M. Slaymaker, D.B.T. Cox, S. Shmakov, K.S. Makarova, E. Semenova, L. Minakhin, K. Severinov, A. Regev, E.S. Lander, E. V. Koonin, F. Zhang, C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector, Science. 353 (2016) aaf5573.

[13] A. East-Seletsky, M.R. O'Connell, S.C. Knight, D. Burstein, J.H.D. Cate, R. Tjian, J.A. Doudna, Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection, Nature. 538 (2016) 270–273.

[14] O.O. Abudayyeh, J.S. Gootenberg, P. Essletzbichler, S. Han, J. Joung, J.J. Belanto, V. Verdine, D.B.T. Cox, M.J. Kellner, A. Regev, E.S. Lander, D.F. Voytas, A.Y. Ting, F. Zhang, RNA targeting with CRISPR-Cas13, Nature. 550 (2017) 280–284.

[15] S. Shao, W. Zhang, H. Hu, B. Xue, J. Qin, C. Sun, Y. Sun, W. Wei, Y. Sun, Long-term dual-color tracking of genomic loci by modified sgRNAs of the CRISPR/Cas9 system, Nucleic Acids Res. 44 (2016) e86.

[16] S. Wang, J.H. Su, F. Zhang, X. Zhuang, An RNA-aptamer-based two-color CRISPR labeling system, Sci. Rep. 6 (2016) 1–7.

APPENDICES

Appendix A - Chapter 2 Supplementary Material

Supplementary Table 2.1 *- DNA constructs and fly lines used in Chapter 2.*

Plasmids			
CB number	*Other name*	*Short description*	*Sequence map*
pCB1132	pTJ10	pcDNA3.1+	https://benchling.com/s/seq-gFWrR93gaiEh2bTpWUDE
pCB1133	pTJ36	pcDNA3.1+GFP	https://benchling.com/s/seq-iJoRjxZ9pd8L4nF0saMr
pCB1134	pTJ102	pUAST-I0-5	https://benchling.com/s/seq-KqT47IEJSReB4Jj73hgO
pCB1135	pTJ103	pUAST-A0-5	https://benchling.com/s/seq-aurdVi8EZAUun9mNeDcv
pCB1136	pTJ6	pcDNA3.1+GFP-I0-5	https://benchling.com/s/seq-04iZ7J1oxUioS1GpHgBk
pCB1137	pTJ7	pcDNA3.1+GFP-A0-5	https://benchling.com/s/seq-dkatsMbasyeJUWLDJH0Y
pCB1138	pTJ5	pcDNA3.1+GFP-I1-5	https://benchling.com/s/seq-HZBjyRWYAcUU7FjpMALw
pCB1139	pTJ14	pcDNA3.1+GFP-A1-5	https://benchling.com/s/seq-ItGGi4uySw3b7Z2vVsH2
pCB1140	pTJ16	pcDNA3.1+GFP-I2-5	https://benchling.com/s/seq-B5Z9mXCZE7TYNurIKaZe
pCB1141	pTJ15	pcDNA3.1+GFP-A2-5	https://benchling.com/s/seq-g3saSnQOyxKnvBQgzKPm
pCB1142	pTJ18	pcDNA3.1+GFP-I3-5	https://benchling.com/s/seq-FWxc8Q1CPUIQ5FgJvwyP
pCB1143	pTJ17	pcDNA3.1+GFP-A3-5	https://benchling.com/s/seq-Kk03ANx15mrxAVaH1wOK
pCB1144	pTJ21	pcDNA3.1+GFP-I4-5	https://benchling.com/s/seq-cddbG258tNuYwHRm6rlm
pCB1145	pTJ19	pcDNA3.1+GFP-A4-5	https://benchling.com/s/seq-WBsMv3nKwDlJOcZid57M
pCB1146	pTJ22	pcDNA3.1+GFP-I5-5	https://benchling.com/s/seq-o44r3qbuGe4PzREM8mEh
pCB1147	pTJ47	pcDNA3.1+GFP-A5-5	https://benchling.com/s/seq-RkVdZKIueCyqiUwIK5mP
pCB1148	pTJ24	pcDNA3.1+GFP-I6-5	https://benchling.com/s/seq-NJE9O2DlIVe5WpCFL3qi
pCB1149	pTJ23	pcDNA3.1+GFP-A6-5	https://benchling.com/s/seq-7er2HuMHfGloyoeR4N4I
pCB1150	pTJ28	pcDNA3.1+GFP-I7-5	https://benchling.com/s/seq-GODBBoWdFXjpbVaPjYuU
pCB1151	pTJ27	pcDNA3.1+GFP-A7-5	https://benchling.com/s/seq-GODBBoWdFXjpbVaPjYuU
pCB1152	pTJ26	pcDNA3.1+GFP-I8-5	https://benchling.com/s/seq-vR02pi8FAYkIDx3MlPZC
pCB1153	pTJ25	pcDNA3.1+GFP-A8-5	https://benchling.com/s/seq-IRUuMXX58xVdm79drI6V
pCB1154	pTJ32	pcDNA3.1+GFP-I9-5	https://benchling.com/s/seq-BHslbOqpEJybmyf0VO4t
pCB1155	pTJ31	pcDNA3.1+GFP-A9-5	https://benchling.com/s/seq-2exwl9y2tF5cOQYb0BFC

CB number	Other name	Short description	Sequence map
pCB1156	pTJ30	pcDNA3.1+GFP-I10-5	https://benchling.com/s/seq-CIWcPoRZ4RbWAMJs1PEv
pCB1157	pTJ29	pcDNA3.1+GFP-A10-5	https://benchling.com/s/seq-nQqqSN9qHmSsQ0nXVgw3
pCB1158	pTJ39	pcDNA3.1+GFP-I0-3	https://benchling.com/s/seq-9lZ9941GuYI9CSILPnau
pCB1159	pTJ38	pcDNA3.1+GFP-A0-3	https://benchling.com/s/seq-29YPCeulzItgVKJfRgjC
pCB1160	pTJ44	pcDNA3.1+GFP-I1-3	https://benchling.com/s/seq-r3LEn1bw9L7U0hXh6APx
pCB1161	pTJ43	pcDNA3.1+GFP-A1-3	https://benchling.com/s/seq-XaCN899dCmNH10ivigzt
pCB1162	pTJ49	pcDNA3.1+GFP-I2-3	https://benchling.com/s/seq-GTmxtZTOgMt7Q4kV9o6X
pCB1163	pTJ48	pcDNA3.1+GFP-A2-3	https://benchling.com/s/seq-L1K4cXbkuUI8E1eQXUiA
pCB1164	pTJ57	pcDNA3.1+GFP-I3-3	https://benchling.com/s/seq-LDi6WzVeDQoGa48c3Q7b
pCB1165	pTJ56	pcDNA3.1+GFP-A3-3	https://benchling.com/s/seq-AIjxSIIEUckblvZIbf1k
pCB1166	pTJ59	pcDNA3.1+GFP-I4-3	https://benchling.com/s/seq-vbfyvrSc4XfBNqAcZHih
pCB1167	pTJ58	pcDNA3.1+GFP-A4-3	https://benchling.com/s/seq-wZW6lU2q8u8dFP0WiypX
pCB1168	pTJ61	pcDNA3.1+GFP-I5-3	https://benchling.com/s/seq-ya203vFz4MSCenLID7gn
pCB1169	pTJ60	pcDNA3.1+GFP-A5-3	https://benchling.com/s/seq-A3wW07sd1dTnY7NEiQFD
pCB1170	pTJ66	pcDNA3.1+GFP-I6-3	https://benchling.com/s/seq-xL2NcyYNQmpH1GsddOeN
pCB1171	pTJ65	pcDNA3.1+GFP-A6-3	https://benchling.com/s/seq-L8ENGLH3nlhshfAI9pfR
pCB1172	pTJ63	pcDNA3.1+GFP-I7-3	https://benchling.com/s/seq-xjLF1s8rUIlYajWo11li
pCB1173	pTJ62	pcDNA3.1+GFP-A7-3	https://benchling.com/s/seq-7TjWifaGeSkjQsKlRXb3
pCB1174	pTJ68	pcDNA3.1+GFP-I8-3	https://benchling.com/s/seq-5EwcQ2GsCfMXP4qdO0NS
pCB1175	pTJ67	pcDNA3.1+GFP-A8-3	https://benchling.com/s/seq-cLI3avsAHKQvSqkOp2rm
pCB1176	pTJ70	pcDNA3.1+GFP-I9-3	https://benchling.com/s/seq-X0cljWNuZ8URLmaC3MhR
pCB1177	pTJ69	pcDNA3.1+GFP-A9-3	https://benchling.com/s/seq-4ib3BddngSdmSfVr31pC
pCB1178	pTJ72	pcDNA3.1+GFP-I10-3	https://benchling.com/s/seq-8tfTQpXNZJ86ms2M05GD
pCB1179	pTJ71	pcDNA3.1+GFP-A10-3	https://benchling.com/s/seq-LcRSiCc9iIJkezKP4sYy
pCB1180	pTJ353	pUAST-attB	https://benchling.com/s/seq-KVGwmQKoYyk7RqlBKFDV
pCB1181	pTJ259	pUAST-hbpe-lacZ	https://benchling.com/s/seq-hV4IbQLQ1cfS2fxByXox
pCB1182	pTJ261	pUAST-hbpe-I0-5	https://benchling.com/s/seq-SkFGfa4HmaHSA6FHUaHR

CB number	*Other name*	*Short description*	*Sequence map*
pCB1183	pTJ262	pUAST-hbpe-A0-5	https://benchling.com/s/seq-2ytUMR0U1dtnOxFLvYlh
pCB1184	pTJ268	pUAST-hbpe-I2-5	https://benchling.com/s/seq-4IJOcdiJs4pK4sRJxfxs
pCB1185	pTJ266	pUAST-hbpe-A2-5	https://benchling.com/s/seq-WP5AE9kkuDln7bgSEexF
pCB1186	pTJ269	pUAST-hbpe-I5-5	https://benchling.com/s/seq-142rU3rQIA56BkH1Wkoo
pCB1187	pTJ267	pUAST-hbpe-A5-5	https://benchling.com/s/seq-0X141lKCHJxqvltlwDr7
pCB1188	pTJ263	pUAST-hbpe-I0-3	https://benchling.com/s/seq-PjFZsxllJO1lsJWApdRO
pCB1189	pTJ264	pUAST-hbpe-A0-3	https://benchling.com/s/seq-Bc8AIrzQ8F9PaKchxUVJ
pCB1190	pTJ271	pUAST-hbpe-I5-3	https://benchling.com/s/seq-J44k2alf80h88bVNmNs9
pCB1191	pTJ270	pUAST-hbpe-A5-3	https://benchling.com/s/seq-3hyYldVKIEZBc099pL9W

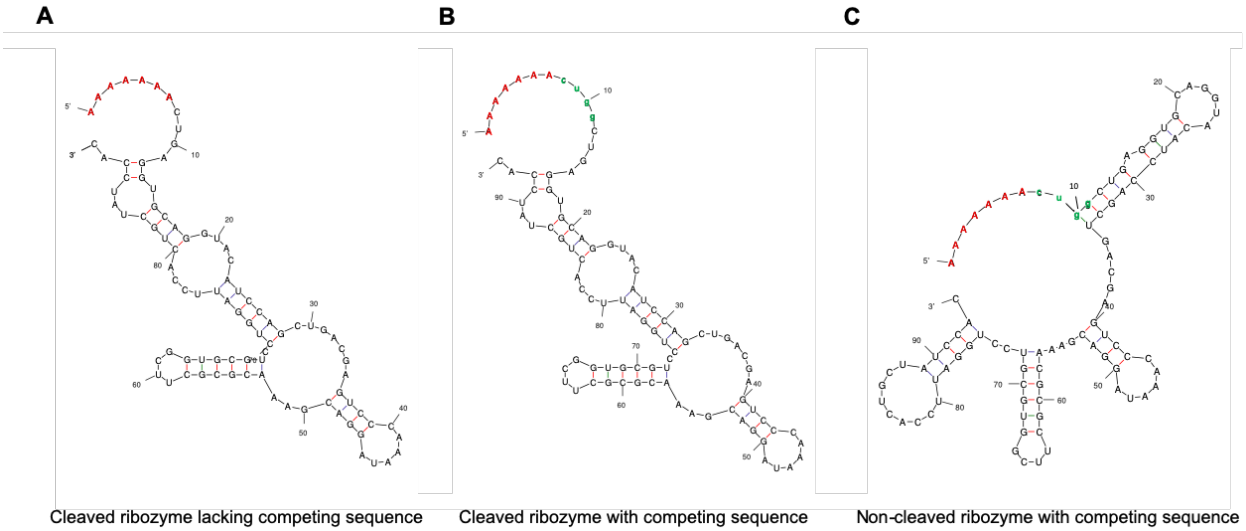
Oligos	
Name	*Sequence*
TJ4	ggctagcgtttaaacttaaagcttACGGTGTGCGTCGTAACA
TJ13	ggtttaaacgggcccttctagaTTACTTGTACAGCTCGTCCATG
TJ53	cgagctgtacaagtaatctagaACGGTGTGCGTCGTAACA
TJ54	ggtttaaacgggccctctagaGTGGATAGCAGTGGAATCC
TJ441	agtcccgtggacctagaagggaattcaaaaacctaggCTGAGGTGCAGGTACATC
TJ442	taaaacgacgggatccaaggctgcagtttttggcgcgccGTGGATAGCAGTGGAATC
TJ451	tggtgtcaaaaataaggtaccgaattcaaaaacctaggCTGAGGTGCAGGTACATC
TJ452	caaagatcctctagaggtaccGTGGATAGCAGTGGAATC
TJ19	AATTCcgactgcgcagtcgAAAAAAActcgaC
TJ20	TCGAGtcgagTTTTTTTcgactgcgcagtcgG
TJ21	AATTCcgactgcgcagtcgAAAAAAActggatC
TJ22	TCGAGatccagTTTTTTTcgactgcgcagtcgG
TJ23	AATTCcgactgcgcagtcgAAAAAAAcgggttC
TJ24	TCGAGaacccgTTTTTTTcgactgcgcagtcgG
TJ25	AATTCcgactgcgcagtcgAAAAAAAgctggaagC
TJ26	TCGAGcttccagcTTTTTTTcgactgcgcagtcgG
TJ27	AATTCcgactgcgcagtcgAAAAAAActggC
TJ28	TCGAGccagTTTTTTTcgactgcgcagtcgG
TJ29	AATTCcgactgcgcagtcgAAAAAAAggatC
TJ30	TCGAGatccTTTTTTTcgactgcgcagtcgG
TJ31	AATTCcgactgcgcagtcgAAAAAAAgctagaagC
TJ32	TCGAGcttctagcTTTTTTTcgactgcgcagtcgG
TJ33	AATTCcgactgcgcagtcgAAAAAAAgctggaC
TJ34	TCGAGtccagcTTTTTTTcgactgcgcagtcgG
TJ35	AATTCcgactgcgcagtcgAAAAAAAgtggtC
TJ36	TCGAGaccacTTTTTTTcgactgcgcagtcgG
TJ37	AATTCcgactgcgcagtcgAAAAAAAttggC

Name	*Sequence*
TJ38	TCGAGccaaTTTTTTTcgactgcgcagtcgG
TJ443	AATTCcgactgcgcagtcgAAAAAAActggatC
TJ444	CTAGGatccagTTTTTTTcgactgcgcagtcgG
TJ445	AATTCcgactgcgcagtcgAAAAAAActggC
TJ446	CTAGGccagTTTTTTTcgactgcgcagtcgG
TJ644	CAACCCGTGGTCGGCTTACG
TJ645	gcaacgaattaaccctcactaaagGGCGTTAGGGTCAATGCGGG

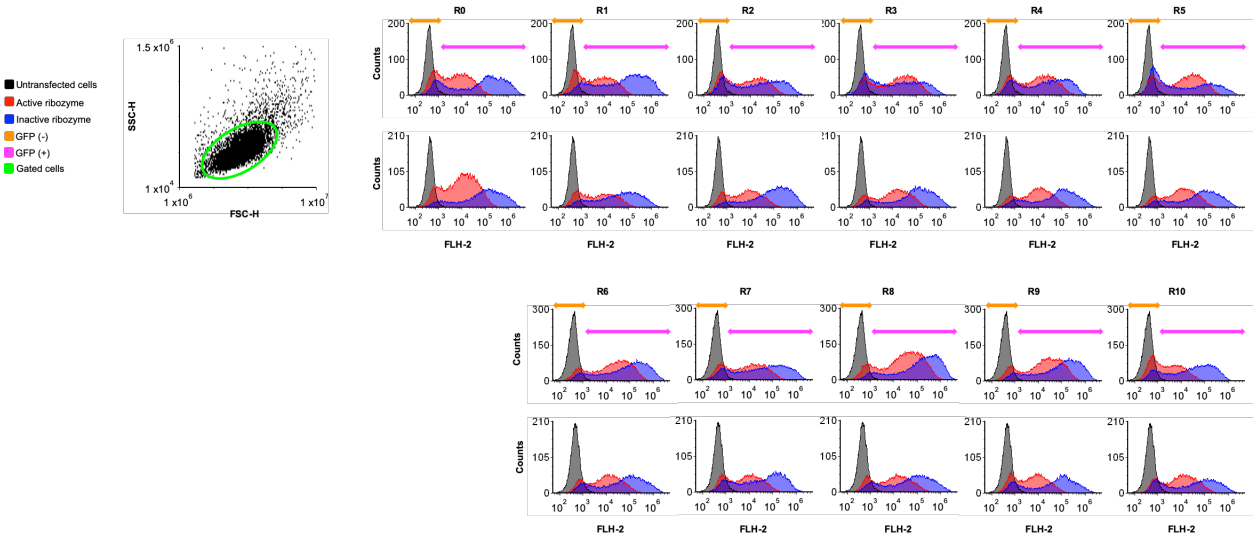
Indicates oligos for PCR amplification (uppercase indicates annealing region)
Indicates oligoes used to anneal/ligate (red text indicates upstream competing sequence)
Indicates oligos to make lacZ anti-sense probe

Supplementary Table 2.2 *- Transfections conditions used in work for Chapter 2.*

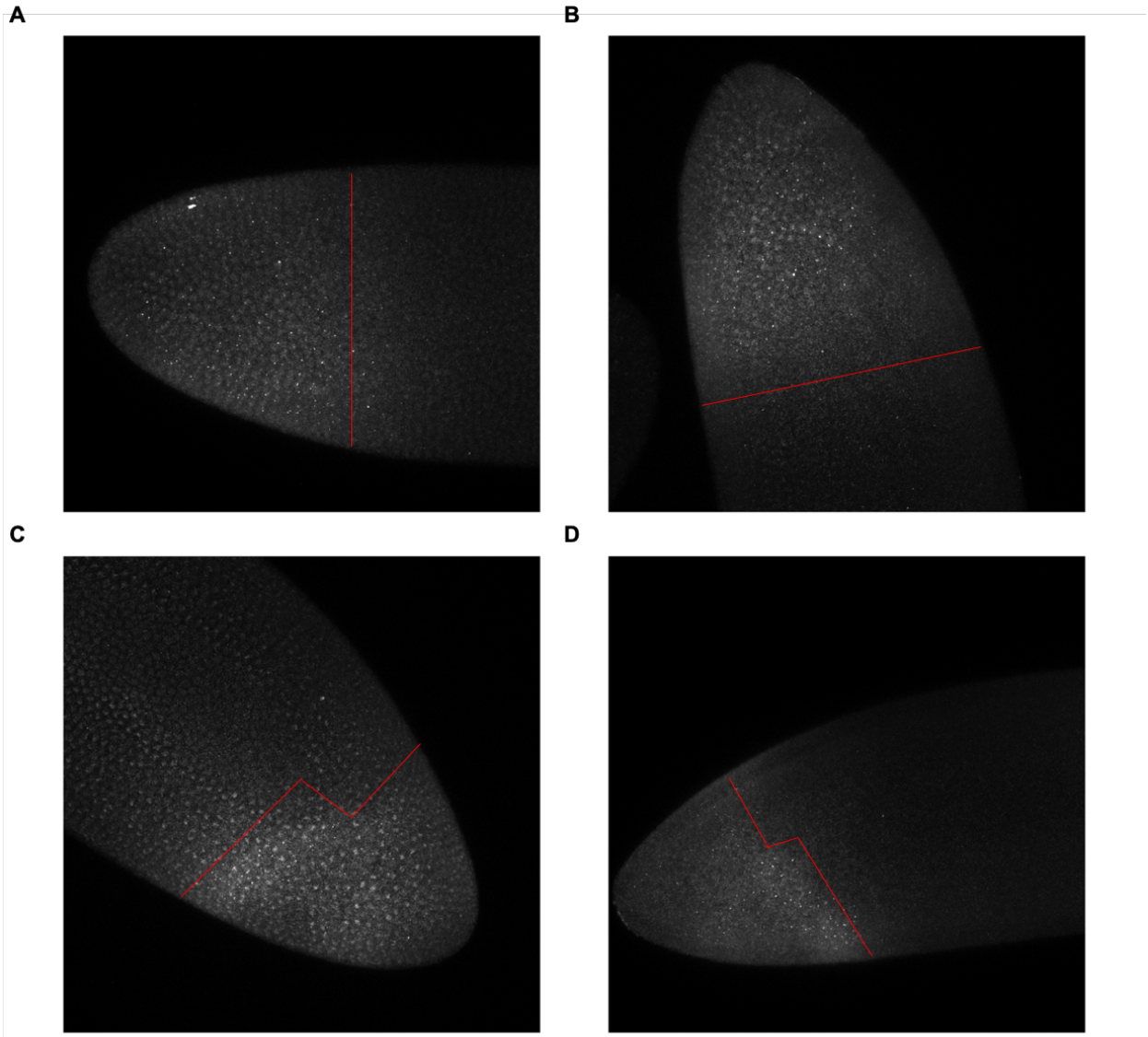
Transfection conditions		
Vessel	35mm plate	24-well plate
Working volume (µl)	2,000	500
Cells seeded	400,000	80,000
DNA (ng)	2,800	550
DNA volume (µl)	129	26
FugeneHD reagent (µl)	8	1.7
Total volume added per plate/well (µl)	125	25



Supplementary Figure 2.1 *- Representative figures depicting secondary structures of self-cleaving ribozymes. (A)Self-cleaving ribozyme that lacks a competing sequence in a cleavable conformation. (B) Self-cleaving ribozyme that contains a competing sequence in a cleavable conformation. (C) Self- cleaving ribozyme that contains a competing sequence in a non-cleavable conformation. The red text indicates the insulating sequence, green text indicates the competing sequence, and black text indicates the ribozyme.*



Supplementary Figure 2.2 - *Flow cytometry data of transiently transfected HEK293T cells. (A) Representative forward and side scatter plot of HEK293T cells transiently transfected with ribozyme constructs. The cell population was gated in green. (B) Histograms of transiently transfected HEK293T cells. Plotted are the number of cells at corresponding fluorescent values of untransfected cells (black), cells containing an active ribozyme/competing sequence (red), and cells containing an inactive ribozyme/competing sequence (blue) in the 5'UTR (top row) or 3'UTR (bottom row) of gfp.*



Supplementary Figure 2.3 *- Representative embryos labeled with lacZ gradient width associated with (A/B) symmetric and (C/D) asymmetric lacZ gradients. Red line indicates end of lacZ gradient. Multiple red lines indicate the width of the lacZ gradient at a particular anterior-posterior axis length.*

1. Download, install, and open Fiji program.

2. Import the image into Fiji by dragging the image file to the opened window.

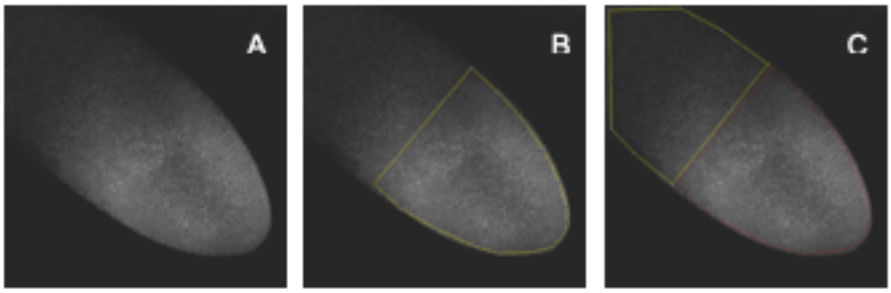
3. Select Image > Stacks > Z Project…

4. The default "Start slice" and "Stop slice" is fine. It will default to the minimum and maximum number of slices, respectively. Select "Projection type" to "Max Intensity" and click "Okay". See Figure A for representative image.

5. Use the "Polygon selections" to draw an area of *hunchback* (*hb*) signal. See Figure B below for representative embryo bounded by *hb* domain (yellow).

6. Select Analyze > Measure and record the "Mean" value to obtain the average intensity of the *hb* domain.

7. Select Edit > Draw to keep the outline of the *hb* domain but allow for the addition of the background noise selection. See Figure C below for representative embryo bounded by background noise domain (yellow). The *hb* domain is drawn (red) to ensure signal is differentiated from background noise.



8. Repeat Steps 5 and 6 to obtain the average intensity of background noise.

9. Calculate the fold reduction of *lacZ* using the equation below.

$$\text{Fold reduction} = \frac{(hb - noise)_I}{(hb - noise)_A}$$

hb = average intensity of hunchback domain
noise = average intensity of background noise
I = measurements from embryos containing an inactive ribozyme
A = measurements from embryos containing an active ribozyme

Supplementary Document 2.1 *- In-depth protocol for measuring fluorescence intensity of Drosophila embryos.*

Appendix B - Chapter 3 Supplementary Material

Supplementary Table 3.1 *- DNA constructs used in Chapter 3.*

Plasmids			
CB name	*Other name*	*Short description*	*Sequence map*
CB849	pTJ258	10N library	https://benchling.com/s/seq-bW1V6MQQ0iEhruRxZUQk
CB847	CSM160	ColE1-Chlor-mRFP-dropout	https://benchling.com/s/seq-lL0L3l0okiuiMciHksh3
CB845	SPC028	DS-Nmcas	https://benchling.com/s/seq-kDY4PqPxVTtLqsZy0K8X
CB848	pTJ247	P70a-deGFP-NmeCas9-spacer	https://benchling.com/s/seq-icn6vwD0IKsZbGCiBYcJ
CB556	CSM254	p70a-deGFP	https://benchling.com/s/seq-xvSlkNXccbKBARdCT3sZ
CB850	CSM184	p70a-T7RNAP	https://benchling.com/s/seq-dR6GKUnISaRRHmXwWsup

Oligos	
Name	*Sequence*
Chi6.FWD	TCACTTCACTGCTGGTGGCCACTGCTGGTGGCCACTGCTGGTGGCCACTGCTGGTGGCCACTGCTGGTGGCCA
Chi6.REV	TGGCCACCAGCAGTGGCCACCAGCAGTGGCCACCAGCAGTGGCCACCAGCAGTGGCCACCAGCAGTGAAGTGA
CSMpr1308	TTTTCTGTTCAGCTCACTCAAAGGCGGTAAACCAATAAAAAACGCCCGGCGGCAACCGAG
CSMpr1309	TGGTGAATATGTCAAACATCCACTCTGCGGACGGTTATCCACAGAATCAGGGGATAACGC
CSMpr1310	GCGTTATCCCCTGATTCTGTGGATAACCGTCCGCAGAGTGGATGTTTGACATATTCACCA
CSMpr1311	CTCGGTTGCCGCCGGGCGTTTTTTATTGGTTTACCGCCTTTGAGTGAGCTGAACAGAAAA
RL133	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCATTAAAAATTGAATTGACATTAACCT
RL134	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATTCACCACCCTGAATTGACT
RL135	AATGATACGGCGACCACCGAGATCTACACTAGATCGCTCGTCGGCAGCGTC
TJpr371	GCGAATTAATACGACTCACTATAGGGCTTAAGTATAAGGAGGAAAAAATATGGCCGCCTTCAAGCCCAACCCCATC
TJpr372	AAACCCCTCCGTTTAGAGAGGGGTTATGCTAGTTAGCGCACAGGAGGGCGCTTCTTCAG
TJpr373	CAAAACCTTTCGCGGTGAGCTAACACCGTGCGT
TJpr374	CACTCGACCGTAGATTCGGTGGTTACGTCCGTCA
TJpr416	ATACGAATATCTGGACCCGACC
TJPr417	CTTTTTACGGTTCCTGGCCTTT
TJpr422	CAAGCAGAAGACGGCATACGAGATGCTCAGGAGTCTCGTGGGCTCGG
TJpr423	CAAGCAGAAGACGGCATACGAGATAGGAGTCCGTCTCGTGGGCTCGG

Name	*Sequence*
T7 FWD	GCGAATTAATACGACTCACTATAGGGCTTAAGTATAAGGAGGAAAAAATATG[5' binding site]
T7 REV	AAACCCCTCCGTTTAGAGAGGGGTTATGCTAGTTA[3' binding site]

Red text indicates the index used for NGS.
Underlined sequence indicates priming region.
Green text indicates start and stop codons for linear DNA expression.
Blue text indicates the Shine-Delgarno sequence for linear DNA expression.

gBlocks	
Name	*Sequence Map*
CSM-GB019	https://benchling.com/s/seq-otMMHB6DWWex9FTnjGJZ
CSM-GB089	https://benchling.com/s/seq-jnRn9HoRKajEQiNxTxgO
CSM-GB191	https://benchling.com/s/seq-9MUyfmFgh1EECoofI4wn
Cas-GB	https://benchling.com/s/seq-QxfdyNCwobR9zkOrPDic

Appendix C - Chapter 4 Supplementary Material

Supplementary Table 4.1 *- DNA constructs used in Chapter 4.*

Plasmids			
CB name	*Other name*	*Short description*	*Sequence map*
pCB1079	pTJ334	5N PAM library	https://benchling.com/s/seq-p1iHLqF9UsxrUYon2z00
pCB951	RS72	bacterial AsCas12a	https://benchling.com/s/seq-7B36UCyOWb4kvASNjhyt
pCB1058	pTJ419	p70a-deGFP w/ GCTA PAM	https://benchling.com/s/seq-gHm9yPTPEDxr7E0YuFpg
pCB1059	pTJ276	p70a-deGFP w/ GCTC PAM	https://benchling.com/s/seq-JPI401l6luHalN5AEPT5
pCB1060	pTJ418	p70a-deGFP w/ GCTG PAM	https://benchling.com/s/seq-6prfWDroFPADKbRCYWI7
pCB1061	pTJ369	p70a-deGFP w/ GTTA PAM	https://benchling.com/s/seq-L12fVmeUmYK2RS0SfOqF
pCB1062	pTJ417	p70a-deGFP w/ GTTC PAM	https://benchling.com/s/seq-P1dVsFRtZGwXudUnuSo4
pCB1063	pTJ426	p70a-deGFP w/ GTTG PAM	https://benchling.com/s/seq-rCmjtwe6rPSdxDWhx4yb
pCB1064	pTJ427	p70a-deGFP w/ CGCTC PAM	https://benchling.com/s/seq-eARQtd5LnzNUckDipV55
pCB1065	pTJ428	p70a-deGFP w/ GGCTC PAM	https://benchling.com/s/seq-4qH1F7nQPSB6tAiJxz4o
pCB1066	pTJ429	p70a-deGFP w/ TGCTC PAM	https://benchling.com/s/seq-7ItvpkmYiZU36MWhaf91
pCB1067	pTJ93	pcDNA3.1-hAsCpf1	https://www.addgene.org/browse/sequence/202532/
pCB1068	pTJ433	pU6-As-crRNA	https://www.addgene.org/browse/sequence/150228/
pCB1069	pTJ438	pU6-As-crRNA-TTTC-1	https://benchling.com/s/seq-IhISvm5TfaqYytl15j2a
pCB1070	pTJ439	pU6-As-crRNA-TTTC-2	https://benchling.com/s/seq-xcX6aDgfWSCNHhuxqB3K
pCB1071	pTJ445	pU6-As-crRNA-NGCTV-1	https://benchling.com/s/seq-TFV0s9yHdJoToVq3aKuN
pCB1072	pTJ446	pU6-As-crRNA-NGCTV-2	https://benchling.com/s/seq-mx1jNqIu5L6M0KIfOVKr
pCB1073	pTJ447	pU6-As-crRNA-NGCTV-3	https://benchling.com/s/seq-kijpvk5BYcvwXpgdACqy
pCB1074	pTJ448	pU6-As-crRNA-NGTTV-1	https://benchling.com/s/seq-AC0zMfDPzJ8tFxPWueUU
pCB1075	pTJ449	pU6-As-crRNA-NGTTV-2	https://benchling.com/s/seq-AC0zMfDPzJ8tFxPWueUU
pCB1076	pTJ450	pU6-As-crRNA-NGTTV-3	https://benchling.com/s/seq-65SrLlmvQqVesBDOdsiy
pCB1077	pTJ466	pU6-As-crRNA-NT-1	https://benchling.com/s/seq-w2BROqEcNTS3KVchDyrU
pCB1078	pTJ467	pU6-As-crRNA-NT-2	https://benchling.com/s/seq-uUbxKTHNpWgIbi60BjAV
pCB869	pCL213	pUA66_GFP1s	https://benchling.com/s/seq-yJZiMYAaOyJBMTQn9ZXz
CBS-443	pCL473	pNT	https://benchling.com/s/seq-CASemLIA4uou1FZKS0LJ
CBS-444	pCL474	pcrRNA	https://benchling.com/s/seq-ZjOiOZm0ll9v9vZqenMx
CBS-445	pCL475	pBAD_AsCpf1 for *E. coli*	https://benchling.com/s/seq-AUeJJANQJf2XtJs1qg3l
CBS-446	pCL476	pCL213_AGTTC	https://benchling.com/s/seq-qoJhyEXqfJ7zuu21XCX5
CBS-447	pCL477	pCL213_TGTTC	https://benchling.com/s/seq-qoJhyEXqfJ7zuu21XCX5
CBS-448	pCL478	pCL213_CGTTC	https://benchling.com/s/seq-tuUtpnoiBZn2Oo0vCxlY
CBS-449	pCL479	pCL213_GGTTC	https://benchling.com/s/seq-TnbYk0NBADdVIxrPlrQ2
CBS-450	pCL480	pCL213_AGCTC	https://benchling.com/s/seq-TnbYk0NBADdVIxrPlrQ2
CBS-451	pCL481	pCL213_TGCTC	https://benchling.com/s/seq-9pdZ054uqV7JKgYTy9cH
CBS-452	pCL482	pCL213_CGCTC	https://benchling.com/s/seq-MMposXPCBSiOFkMeh2a6
CBS-453	pCL483	pCL213_GGCTC	https://benchling.com/s/seq-M7jJdOos7IgekQO0yEGM
CBS-454	pCL484	pCL213_AGGCT	https://benchling.com/s/seq-MYM044XZMS9K2Riu4zhi

Oligos*	
Name	*Sequence*
TJ18	GTACAAAATACGTGACGTAGAAAG
TJ472	CGTCACGTTCACGCATCAG

Name	*Sequence*
N/A	ggacgtaaccnnnnnGTCGAGTGCAAAACCTTTC**
TJ473	CCGCAGAGTGGATGTTTGACAT
TJ662	agatCTGATGGTCCATGTCTGTTACTC
TJ663	aaaaGAGTAACAGACATGGACCATCAG
TJ664	agatCACTCATACAGTGGTAGATTTGA
TJ665	aaaaTCAAATCTACCACTGTATGAGTG
TJ689	agatAGCAGGCACCTGCCTCAGCTGCT
TJ690	aaaaAGCAGCTGAGGCAGGTGCCTGCT
TJ691	agatCTAAGGACTAGTTCTGCCCTCCC
TJ692	aaaaGGGAGGGCAGAACTAGTCCTTAG
TJ693	agatTGTGCAGCACGCCTGCTCTAAGC
TJ694	aaaaGCTTAGAGCAGGCGTGCTGCACA
TJ695	agatGGGATTCCTGGTGCCAGAAACAG
TJ696	aaaaCTGTTTCTGGCACCAGGAATCCC
TJ697	agatGCTTATCAACTAATGATTTAGTG
TJ698	aaaaCACTAAATCATTAGTTGATAAGC
TJ699	agatCCTTAGCACTCTGCCACTTATTG
TJ700	aaaaCAATAAGTGGCAGAGTGCTAAGG
TJ723	agatTGGCTGCTGGGCTGGCCCTGGGG
TJ724	aaaaCCCCAGGGCCAGCCCAGCAGCCA
TJ725	agatACTGCACAATTTGATCACTAAAT
TJ726	aaaaATTTAGTGATCAAATTGTGCAGT
TJ719	AGTTGCTGGCCACCGTTTTG
TJ720	CACAACATCAGTGCATGTTGGG
TJ721	GGTCAGCTGTTAACATCAGTACG
TJ722	GTCCCGTGCAAATCACGAATAC
CL765	agttcTGTCAGTGGAGAGGG
CL766	TTTGTGCCCATTAACATCAC
CL767	tgttcTGTCAGTGGAGAGGG
CL768	cgttcTGTCAGTGGAGAGGG
CL769	ggttcTGTCAGTGGAGAGGGT
CL770	agctcTGTCAGTGGAGAGGGT
CL771	tgctcTGTCAGTGGAGAGGGT
CL772	cgctcTGTCAGTGGAGAGGGT
CL773	ggctcTGTCAGTGGAGAGGGT
CL774	aggctTGTCAGTGGAGAGGGTGA

*Upper-case sequence represents annealed portion of primer

**The nnnnn sequence will vary on desired PAM sequence

gBlocks	
Name	*Sequence map*
TJ524	https://benchling.com/s/seq-nOh94pdoQMUYM5SoQzmk
CSM-GB019	https://benchling.com/s/seq-otMMHB6DWWex9FTnjGJZ

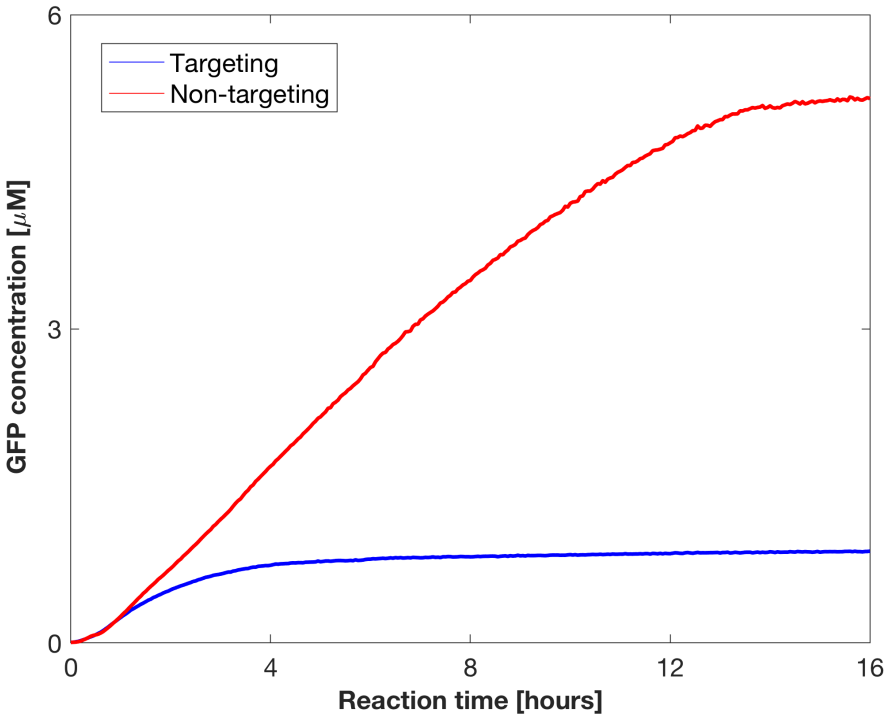
Supplementary Table 4.2 *- TXTL reaction setup for the DNA cleavage assay and PAM screen. The master mix for both the DNA cleavage assay and PAM screen were supplemented with components shared for all reaction conditions (e.g. AsCas12a plasmid), while unique components were added separately to each reaction condition (e.g. targeting vs. non-targeting gRNA).*

DNA cleavage assay			
Component	***Volume (µL)***	***Final conc.***	***Stock conc.***
myTXTL Sigma 70 Master Mix	75		
AsCas12a plasmid	1	2nM	200nM
Chi6 DNA	2	1µM	50µM
water	2		
Reaction condition	***DNA 1***	***DNA 2***	
1	p70a-deGFP w/ PAM 1	targeting gRNA	
2	p70a-deGFP w/ PAM 1	non-targeting gRNA	
3	p70a-deGFP w/ PAM 2	targeting gRNA	
4	p70a-deGFP w/ PAM 2	non-targeting gRNA	
5	p70a-deGFP w/ PAM 3	targeting gRNA	
6	p70a-deGFP w/ PAM 3	non-targeting gRNA	
7	p70a-deGFP w/ PAM 4	targeting gRNA	
8	p70a-deGFP w/ PAM 4	non-targeting gRNA	
PAM screen			
Component	***Volume (µL)***	***Final conc.***	***Stock conc.***
myTXTL Sigma 70 Master Mix	75		
AsCas12a plasmid	1	2nM	200nM
p70a-deGFP	1	0.5nM	50nM
5N PAM library	1	0.5nM	50nM
Chi6 DNA	2	1µM	50µM
Reaction condition	***DNA 1***		
1	targeting gRNA		
2	non-targeting gRNA		

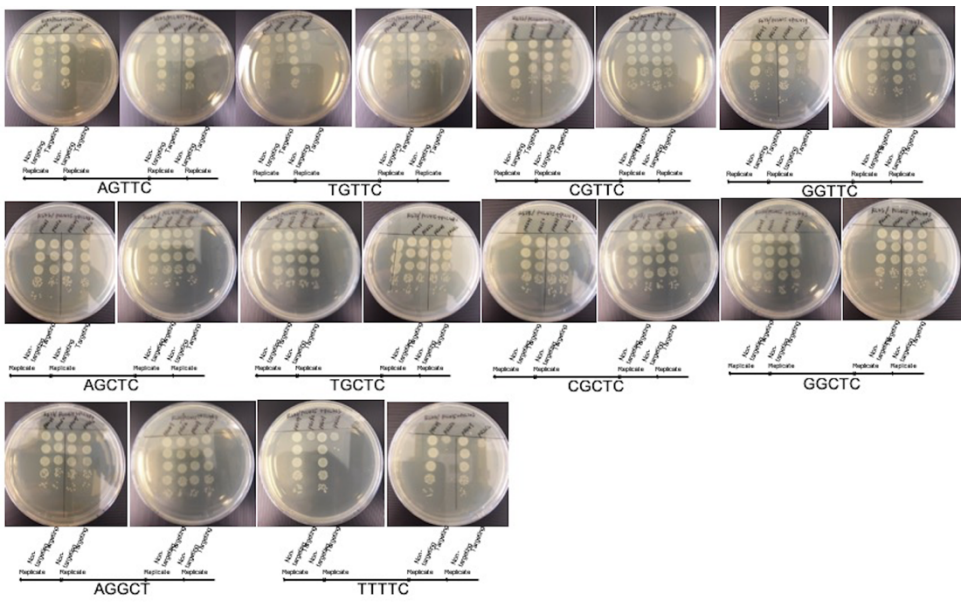
Supplementary Table 4.3 *- List of target sequences and their associated PAMs used for the DNA cleavage assay in vitro and indel formation in DMNT1 in HEK293T cells.*

	PAM	Target sequence
PAMs and target sequence(s) used in DNA cleavage assay	AGGCC	GTCGAGTGCAAAACCTTTCGCGGT
	ATTTC	
	AGTTA	
	AGTTC	
	AGTTG	
	AGCTA	
	AGCTC	
	AGCTG	
	CGCTC	
	GGCTC	
	TGCTC	
PAMs and target sequence(s) used in plasmid clearance assay	TTTTC	TGTCAGTGGAGAGGGTGAAGGTGATG
	AGTTC	
	CGTTC	
	GGTTC	
	AGCTC	
	CGCTC	
	GGCTC	
	TGCTC	
	AGGCT	
PAMs and target sequence(s) used in *DNMT1* indel formation assay	GTTTC	CTGATGGTCCATGTCTGTTACTC
	ATTTC	CACTCATACAGTGGTAGATTTGA
	TGTTG	GGGATTCCTGGTGCCAGAAACAG
	GGTTC	GCTTATCAACTAATGATTTAGTG
	CGTTC	CCTTAGCACTCTGCCACTTATTG
	GGCTC	AGCAGGCACCTGCCTCAGCTGCT
	AGCTG	CTAAGGACTAGTTCTGCCCTCCC
	TGCTG	TGTGCAGCACGCCTGCTCTAAGC
	CAGGT	TGGCTGCTGGGCTGGCCCTGGGG

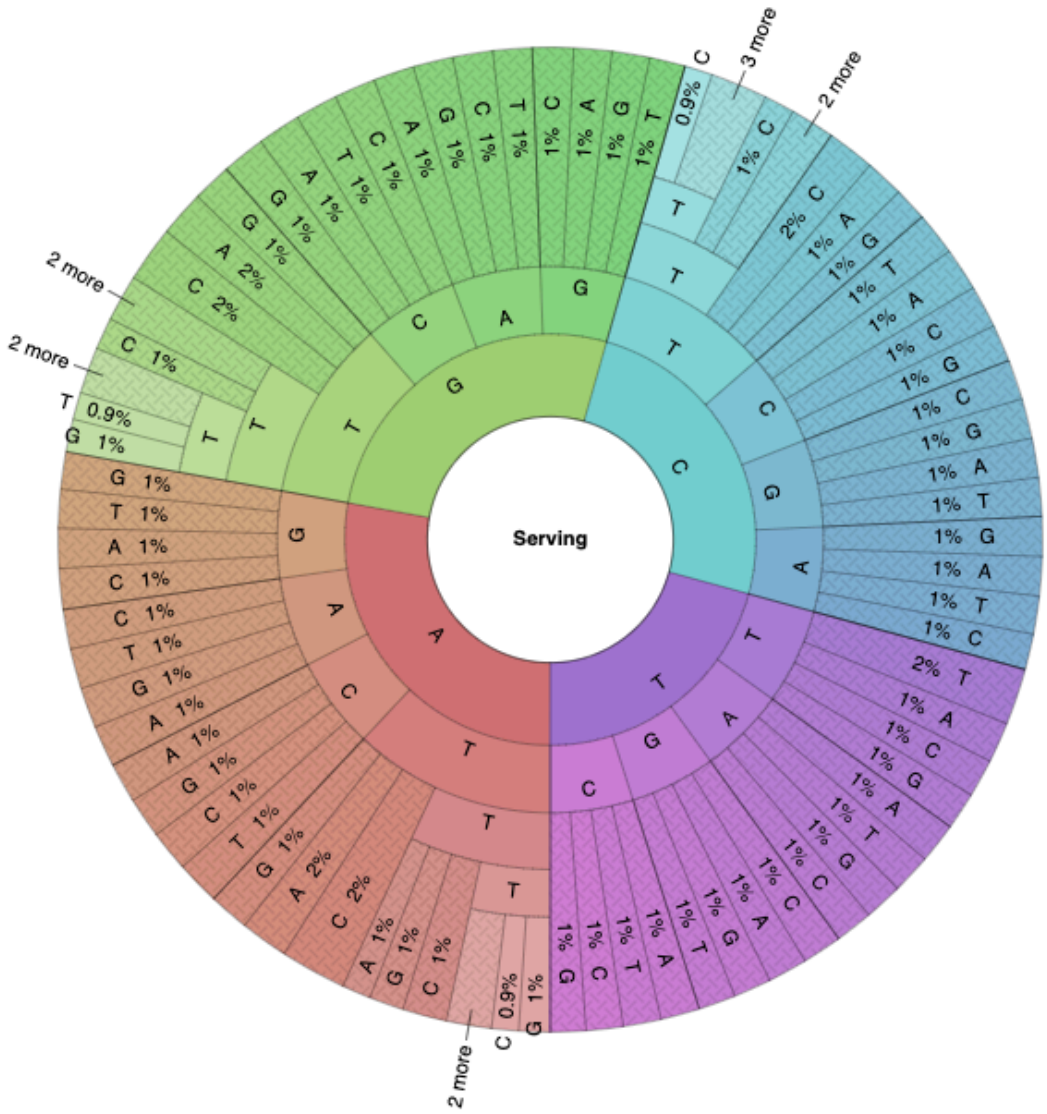
Supplementary Table 4.4 *- List of all possible 5N PAM sequences and their corresponding depletion scores. Due to the size of this table, it was not included here. Refer to Supplementary Table 4 of (PMID: 31004485) for the table.*



Supplementary Figure 4.1 *- Time series of GFP expression from the PAM screen (Related to Figure 1A). The GFP reporter plasmid was added to the TXTL PAM screen reaction to indirectly assess the extent of cleavage of the PAM library. The GFP plasmid contained the same protospacer used in the PAM library but was flanked by an ATTTC PAM sequence. GFP expression from the reaction containing the targeting gRNA (blue line) ceased after approximately four hours, while GFP was continuously expressed using the non-targeting gRNA (red line).*



Supplementary Figure 4.2 *- Images of the plates containing the colonies from the plasmid clearance assays in E. coli. Cells harboring the AsCas12a expression plasmid and a plasmid containing the target sequence flanked by the indicated sequence were transformed with a plasmid encoding the targeting or non-targeting gRNA. After recovering in SOC at 37°C for one hour, the recovered cells were plated drop-wise onto LB agar plates supplemented with ampicillin, kanamycin, and chloramphenicol. The colony numbers were calculated after the agar plates were incubated at 37°C for 16 hours. Note the smaller colonies for many of the experiments with the NGTTC sequences.*



Supplementary File 4.1 *- PAM wheel of AsCas12a developed from the TXTL-based PAM screen. Larger "pie slices" indicate greater depletion. A screenshot of Supplementary File 4.1 is above. Refer to Supplementary File 1 of (PMID: 31004485) for the file.*

Appendix D – Chapter 5 Supplementary Material

Supplementary Table 5.1 - *DNA constructs used in Chapter 5.*

Plasmids			
CB name	*Other name*	*Short description*	*Sequence map*
pCB1079	pTJ334	5N-randomized PAM library	https://benchling.com/s/seq-p1iHLqF9UsxrUYon2z00
pCB1023	pFT106	p70a-T7pol	https://benchling.com/s/seq-DqIfjGAc6W6G7rJ5ACZ5
pCB1081	pTJ195	p70a-deGFP w/ TTTC PAM	https://benchling.com/s/seq-HEjG45Fhppr5qewqvG42
pCB1082	pTJ413	p70a-deGFP w/ CCCC PAM	https://benchling.com/s/seq-3RdD1XxucgzPMN8yo8ip
pCB1083	pTJ420	p70a-deGFP w/ CCTC PAM	https://benchling.com/s/seq-jPWZI5HYWUt27JaRm0aF
pCB1084	pTJ272	p70a-deGFP w/ CTCC PAM	https://benchling.com/s/seq-QjlPPn8WYWIcoxEw6sac
pCB1085	pTJ421	p70a-deGFP w/ TCCC PAM	https://benchling.com/s/seq-CCxOMt8fdJSHd9BFa8XZ
pCB1062	pTJ417	p70a-deGFP w/ GTTC PAM	https://benchling.com/s/seq-P1dVsFRtZGwXudUnuSo4
pCB1086	pTJ352	p70a-deGFP w/ TGTC PAM	https://benchling.com/s/seq-XXL7vXMYYupuZMCHc9ii
pCB1087	pTJ274	p70a-deGFP w/ TTGC PAM	https://benchling.com/s/seq-XXL7vXMYYupuZMCHc9ii
pCB1088	pTJ275	p70a-deGFP w/ GTCC PAM	https://benchling.com/s/seq-IqwsO14HspmA0pN84SRk
pCB1089	pTJ283	p70a-deGFP w/ GGCC PAM	https://benchling.com/s/seq-N1AmbYRzWVlSuZ3ptQFN
pCB1090	pTJ285	p70a-deGFP w/ GGTC PAM	https://benchling.com/s/seq-ncFBrtQFxjSEW8yJboIz
pCB1091	pTJ344	pET28b+PdCas12a	https://benchling.com/s/seq-12QM6CPS5iVIjFzFqvHL
pCB1092	pTJ218	pET28b+Adurb193Cas12a	https://benchling.com/s/seq-UUogLf9zQg0veTv3RyPq
pCB1093	pTJ221	pET28b+Adurb336Cas12a	https://benchling.com/s/seq-rUIKKyEdT4dnLnOHEUWv
pCB1094	pTJ222	pET28b+Fn3Cas12a	https://benchling.com/s/seq-LPcwir8b4VSo8agAiXh3
pCB1095	pTJ389	pET28b+HkCas12a	https://benchling.com/s/seq-oJGmp0u7JkJV1G0Uroub
pCB1096	pTJ219	pET28b+PiCas12a	https://benchling.com/s/seq-RAQQTCcEdeQaZP0PIyTP
pCB1097	pTJ339	pET28b+PiCas12a N599D	https://benchling.com/s/seq-XsIMrrVgjDP9Lt91z1Uu
pCB1098	pTJ341	pET28b+PiCas12a S600N	https://benchling.com/s/seq-dUQ4zKF12MZkkCuUgZLB
pCB1099	pTJ338	pET28b+PiCas12a F604Y	https://benchling.com/s/seq-54m2v996qSZ139PleSkZ
pCB1100	pTJ340	pET28b+PiCas12a T628A	https://benchling.com/s/seq-yarD1GNaPOIg7hA129Ye
pCB1101	pTJ356	pET28b+PiCas12a NY	https://benchling.com/s/seq-hvYMytu1dYz16gD5kMDz

CB name	*Other name*	*Short description*	*Sequence map*
pCB1102	pTJ343	pET28b+PiCas12a DNYA	https://benchling.com/s/seq-mZD3ZlPSvJcON7VgxWWd
pCB1067	pTJ93	pcDNA3.1-hAsCpf1	https://www.addgene.org/browse/sequence/202532/
pCB1103	pTJ506	pcDNA3.1+HkCas12a+NLS	https://benchling.com/s/seq-4YZQfJiZwnUPwBySxEgn
pCB1104	pTJ507	pcDNA3.1+PiCas12a+NLS	https://benchling.com/s/seq-EKY6sdLJSIJcFDgSOwYd
pCB1105	pTJ508	pcDNA3.1+PiCas12a F604Y+NLS	https://benchling.com/s/seq-8dhkN9Z4yvVpgx2zg1qS
pCB1068	pTJ433	pU6-As-crRNA	https://www.addgene.org/browse/sequence/150228/
pCB1106	pTJ476	pU6-Hk-crRNA	https://benchling.com/s/seq-NbsFXRuEONbr2QIbsAan
pCB1107	pTJ494	pU6-Hk-crRNA-TTTV-1	https://benchling.com/s/seq-9dWpEXIC4TtQaWNWYKtu
pCB1108	pTJ487	pU6-Hk-crRNA-TTTV-2	https://benchling.com/s/seq-0ESs6jaBwq9q6B1CKe9R
pCB1109	pTJ497	pU6-Hk-crRNA-CCCV-1	https://benchling.com/s/seq-BuNJEBgi7wl0aCyMr9CV
pCB1110	pTJ504	pU6-Hk-crRNA-CCCV-2	https://benchling.com/s/seq-5ZlE2w5ryQwzMDwtWoXG
pCB1111	pTJ495	pU6-Hk-crRNA-CCTV-1	https://benchling.com/s/seq-62IOkJmvfyoq41iCblzh

Oligos*	
Name	*Sequence*
TJ18	GTACAAAATACGTGACGTAGAAAG
TJ473	CCGCAGAGTGGATGTTTGACAT
TJ472	CGTCACGTTCACGCATCAG
N/A	ggacgtaaccnnnnnGTCGAGTGCAAAACCTTTC
TJ735	AATTTCTACTtgTGTAGATTGAGACGGG
TJ736	AATTTCTACTaTTGTAGATTGAGACG
TJ737	ACGGTGTTTCGTCCTTTC
TJ662	agatCTGATGGTCCATGTCTGTTACTC
TJ663	aaaaGAGTAACAGACATGGACCATCAG
TJ664	agatCACTCATACAGTGGTAGATTTGA
TJ665	aaaaTCAAATCTACCACTGTATGAGTG
TJ695	agatGGGATTCCTGGTGCCAGAAACAG
TJ696	aaaaCTGTTTCTGGCACCAGGAATCCC
TJ699	agatCCTTAGCACTCTGCCACTTATTG
TJ700	aaaaCAATAAGTGGCAGAGTGCTAAGG
TJ761	agatTGTTACTCGCCTGTCAAGTGGCG
TJ762	aaaaCGCCACTTGACAGGCGAGTAACA
TJ763	agatAGCCAAGGCCACAAACACCATGT
TJ764	aaaaACATGGTGTTTGTGGCCTTGGCT
TJ765	agatACGTGTCAAGTGCTTAGAGCAGG
TJ766	aaaaCCTGCTCTAAGCACTTGACACGT

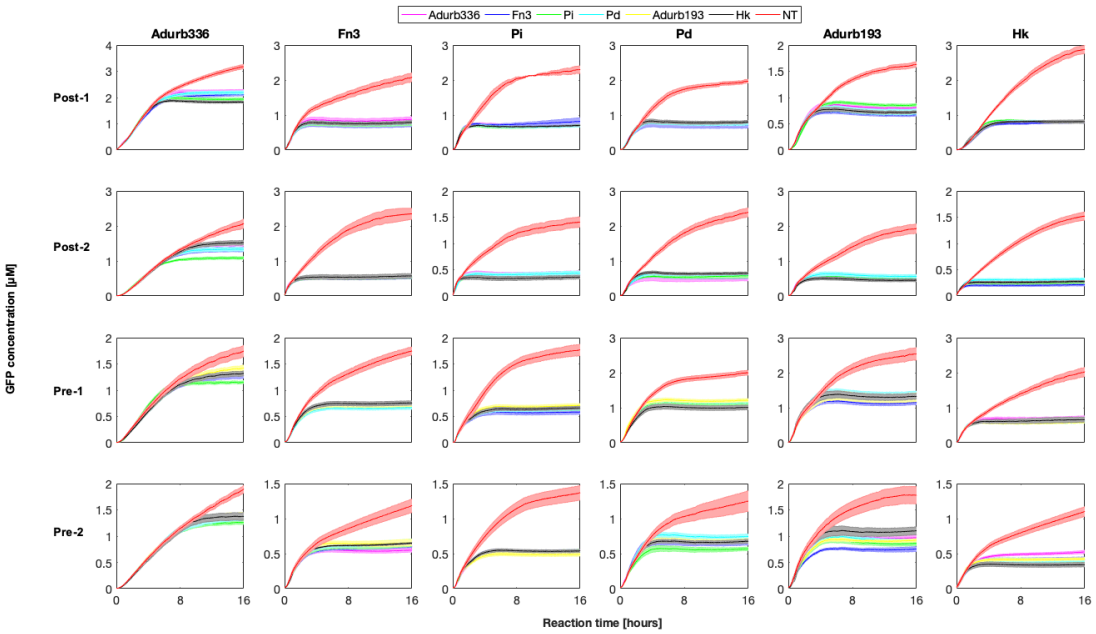
Name	*Sequence*
TJ767	agatCATGTTAAAAACACAACATCAGT
TJ768	aaaaACTGATGTTGTGTTTTTAACATG
TJ769	agatGCTGTTAACATCAGTACGTTAAT
TJ770	aaaaATTAACGTACTGATGTTAACAGC
TJ771	agatTTAGCAGCTTCCTCCTCCTTTAT
TJ772	aaaaATAAAGGAGGAGGAAGCTGCTAA
TJ745	agatTCACTCCTGCTCGGTGAATTTGG
TJ746	aaaaCCAAATTCACCGAGCAGGAGTGA
TJ747	agatGAATGCACAAAGTACTGCACAAT
TJ748	aaaaATTGTGCAGTACTTTGTGCATTC
TJ749	agatGCAGGCACCTGCCTCAGCTGCTC
TJ750	aaaaGAGCAGCTGAGGCAGGTGCCTGC
TJ751	agatTACAGTGGTAGATTTGATATAAT
TJ752	aaaaATTATATCAAATCTACCACTGTA
TJ753	agatTCAAGTGGCGTGACACCGGGCGT
TJ754	aaaaACGCCCGGTGTCACGCCACTTGA
TJ755	agatACACAACAGCTTCATGTCAGCCA
TJ756	aaaaTGGCTGACATGAAGCTGTTGTGT
TJ757	agatAGGGCCAGCCCAGCAGCCAACCT
TJ758	aaaaAGGTTGGCTGCTGGGCTGGCCCT
TJ759	agatTGTTTCTGGCACCAGGAATCCCC
TJ760	aaaaGGGGATTCCTGGTGCCAGAAACA
TJ773	agatAAAAGGAAAAGTCACTCTGGGGA
TJ774	aaaaTCCCCAGAGTGACTTTTCCTTTT
TJ777	agatGAAGGGAAATAAAAGGAAAAGTC
TJ778	aaaaGACTTTTCCTTTTATTTCCCTTC
TJ719	AGTTGCTGGCCACCGTTTTG
TJ720	CACAACATCAGTGCATGTTGGG
TJ722	GTCCCGTGCAAATCACGAATAC
TJ780	AAAAGGCCGGCGGCCACG
TJ781	GGTGGCGGTACCAAGCTTAAGTTTAAACG
TJ782	ttaagcttggtaccgccaccATGGCCCCAAAGAAGAAAC
TJ783	ttcgtggccgccggcctttTGAAATAATGAAATTAATCCAGTCC
TJ784	ttaagcttggtaccgccaccATGAAAGTGATGGAAAACTATC
TJ785	ttcgtggccgccggcctttTTTCAGGTACGGCTTTTC

*Note that uppercase portions of oligo sequence indicate a sequence annealed for spacer sequence cloning or annealed for PCR amplification.

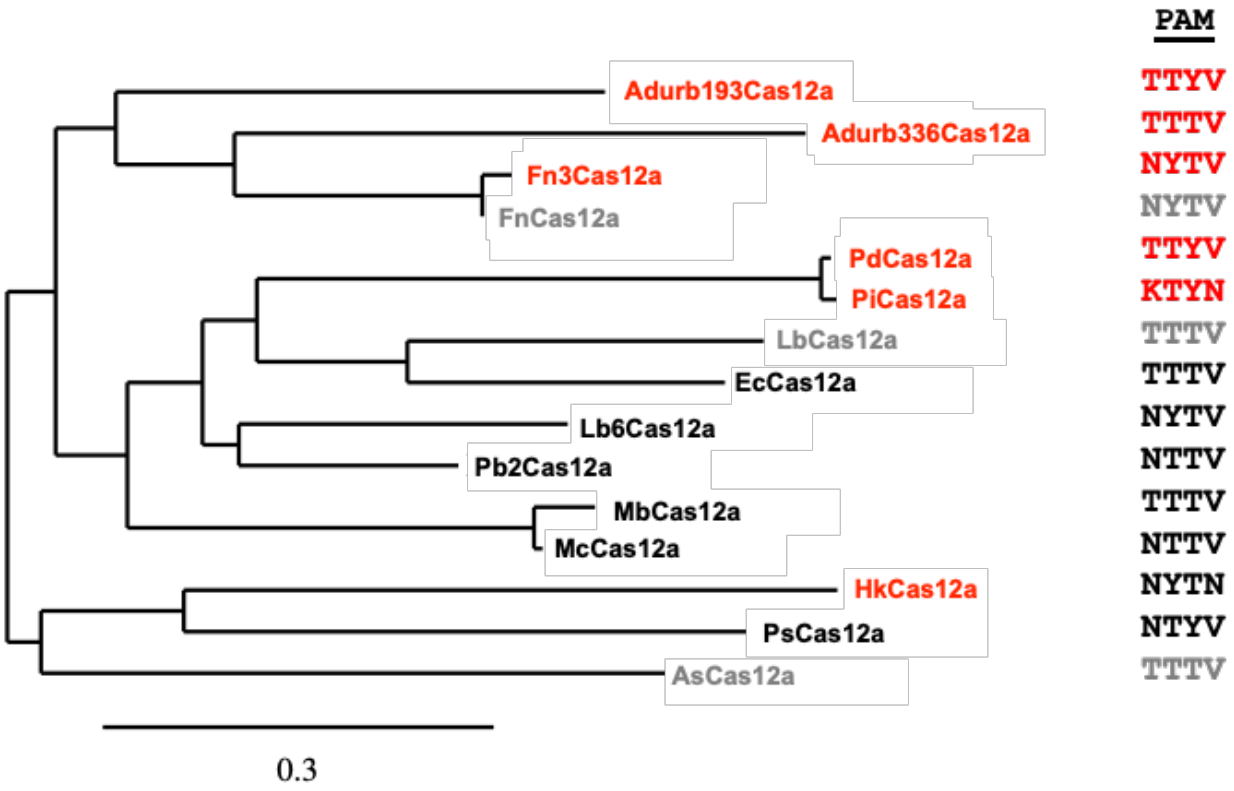
gBlocks	
Name	*Sequence map*
CSM-GB019	https://benchling.com/s/Rmz8lGii
TJ625	https://benchling.com/s/seq-pwungqalkQGC4ApnJvSy
TJ627	https://benchling.com/s/seq-OfioRA9bkEzrOpTccL69
TJ628	https://benchling.com/s/seq-d99L4tsndHNsoipyvhbs
TJ630	https://benchling.com/s/seq-e4awD5YAOSWeSXh4TdvG
TJ568	https://benchling.com/s/seq-hzwSQj0C8sM0I5DOMX4q
TJ629	https://benchling.com/s/seq-g2djI7T1yFVD1qJ3iDJ5
TJ626	https://benchling.com/s/seq-qTSEKpOjjrN3U0mRnMWV
CSM-GB101	https://benchling.com/s/seq-1Oc14ag0DayaU8AXDFA8
CSM-GB181	https://benchling.com/s/seq-NmOrcDFdsxAXvbr6rh6M
CSM-GB096	https://benchling.com/s/seq-mfowzBrQEzGgloAuNHap
CSM-GB187	https://benchling.com/s/seq-28ZjXMLPnjrKajndmZKn
TJ739	https://benchling.com/s/seq-Jb5i27Xdn96yXtHj7vHZ
TJ740	https://benchling.com/s/seq-19qoZSR8LwFQ1x2o566x
TJ741	https://benchling.com/s/seq-fpAyPT2OAPymrtKtrKBS
TJ742	https://benchling.com/s/seq-4PNW0E2UCJTjnxNBcOuw
TJ743	https://benchling.com/s/seq-GiI8szzDy3eow6V7U3NR
TJ744	https://benchling.com/s/seq-YIYq8cV4u27p3mGXIbtW
TJ738	https://benchling.com/s/seq-Q0KvLHyRkUAg8HEuTya2
CSM-GB100	https://benchling.com/s/seq-oChuXx1KbfWDNrvttAKr
CSM-GB182	https://benchling.com/s/seq-hZUY9pkctHdMS7oOPEmQ
CSM-GB104	https://benchling.com/s/seq-BnaA3z0zCFcnkGVsA6kf
CSM-GB188	https://benchling.com/s/seq-bmxJZDZKTc40Gs4exQvO

Supplementary Table 5.2 - *List of all target sequences and their associated PAMs used for the DNA cleavage assays in TXTL and indel formation assay of the DNMT1 gene in HEK293T cells.*

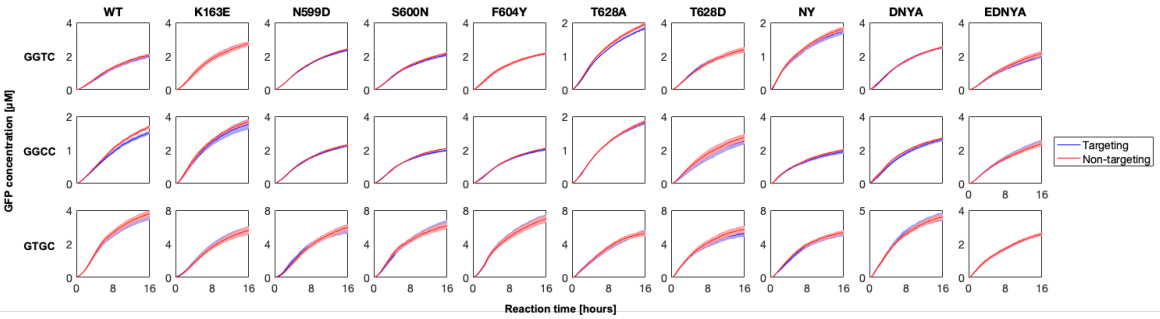
	PAM	Target sequence
PAMs and target sequence(s) used in DNA cleavage assay	TTTC	GTCGAGTGCAAAACCTTTCGCGGT
	CCCC	
	CCTC	
	CTCC	
	TCCC	
	TTGC	
	TGTC	
	GTCC	
	GTCC	
	GGTC	
	GGCC	
PAMs and target sequence(s) used in orthognality testing	TTTC	GTCGAGTGCAAAACCTTTCGCGGT
		ACTGGCGTTGTTCCCATCCTGGTC
PAMs and target sequence(s) used in DNMT1 indel formation assay	TTTC 1	CTGATGGTCCATGTCTGTTACTC
	TTTC 2	CACTCATACAGTGGTAGATTTGA
	CCCC 1	AGGGCCAGCCCAGCAGCCAACCT
	CCCC 2	TGTTTCTGGCACCAGGAATCCCC
	CCTG	TCAAGTGGCGTGACACCGGGCGT
	CCTC	ACACAACAGCTTCATGTCAGCCA
	CTCA 1	GCAGGCACCTGCCTCAGCTGCTC
	CTCA 2	TACAGTGGTAGATTTGATATAAT
	TCCC	TCACTCCTGCTCGGTGAATTTGG
	TCCA	GAATGCACAAAGTACTGCACAAT
	TTGC	ACGTGTCAAGTGCTTAGAGCAGG
	TTGA	CATGTTAAAAACACAACATCAGT
	TGTC 1	TGTTACTCGCCTGTCAAGTGGCG
	TGTC 2	AGCCAAGGCCACAAACACCATGT
	GTTG	GGGATTCCTGGTGCCAGAAACAG
	GTTC	CCTTAGCACTCTGCCACTTATTG
	GTCA	GCTGTTAACATCAGTACGTTAAT
	GTCC	TTAGCAGCTTCCTCCTCCTTTAT



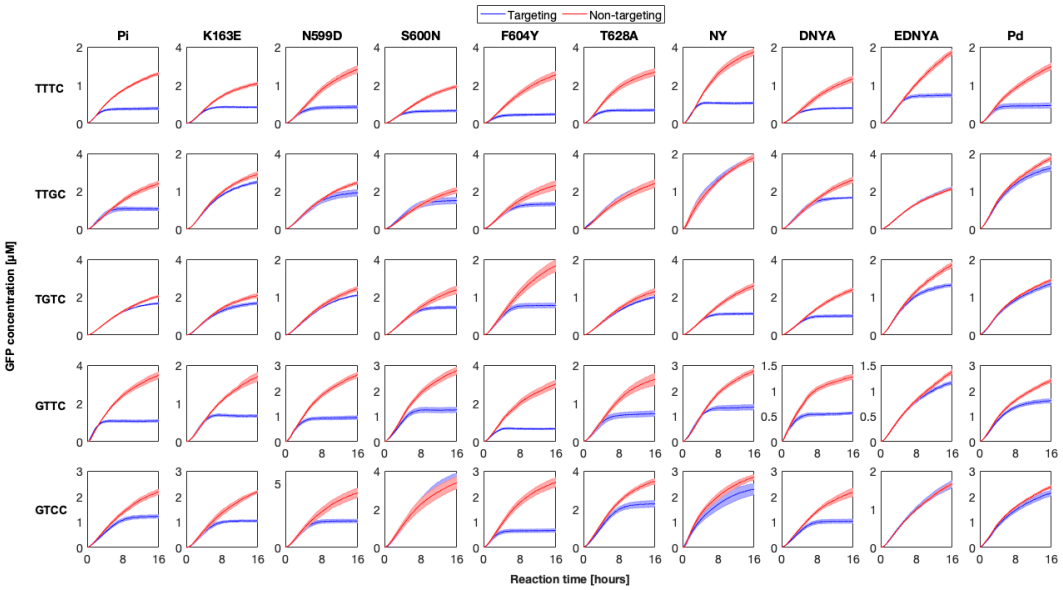
Supplementary Figure 5.1 - *Time-series of GFP expression in TXTL reactions containing a Cas12a, targeting or non-targeting (red) cRNA, and a GFP reporter containing a protospacer. For this experiment, we tested both full-length (Pre) and processed (Post) gRNA that targeted two protospacers containing a TTTC PAM (Pre-1, Pre-2, etc.). Note that the Adurb193 and HkCas12a share the same post-processed gRNA sequence. The error bars represent the standard deviation from three separate TXTL reactions.*



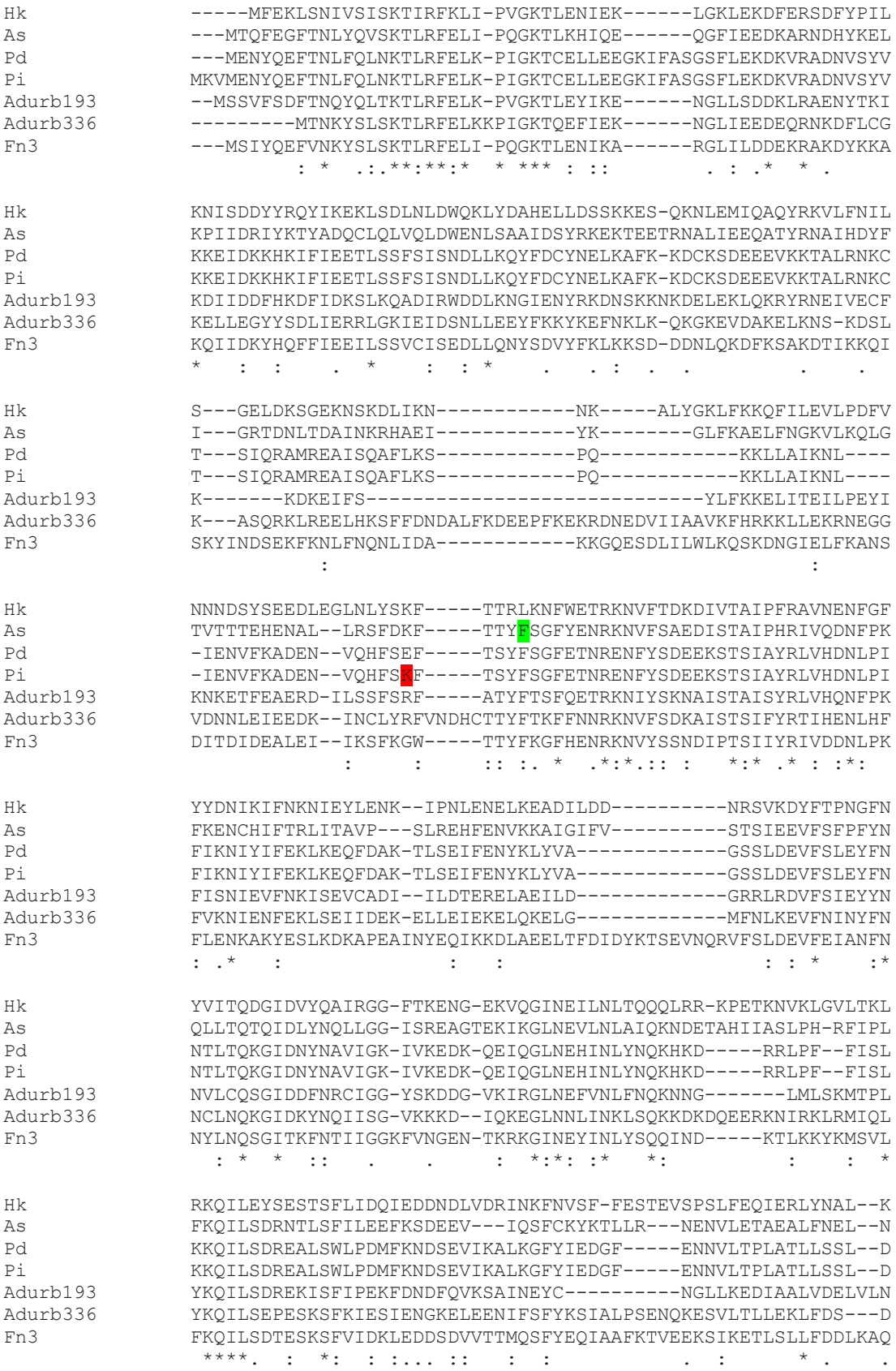
Supplementary Figure 5.2 - *PAM profiles do not trend with phylogeny. Phylogenetic tree and consensus PAMs of Cas12a-containing species. Red indicates nucleases investigated in this study, black indicates nucleases previously investigated in our prior work (PMID: 29304331), and gray indicates well-characterized Cas12a nucleases (PMID: 26422227).*



Supplementary Figure 5.3 - *PiCas12a variants cannot recognize GGYC or GTGC motifs in TXTL. Time series of GFP expression in TXTL reactions expressing a PiCas2a variant, a targeting (blue) or non-targeting (red) gRNA, and a GFP reporter containing the protospacer flanked by a GGTC (top row), GGCC (middle row), or GTGC (bottom row) motifs. Each column represents a different PiCas12a variant analyzed in this study. The error bars represent the standard deviation from three separate TXTL reactions.*



Supplementary Figure 5.4 - *Time-series of GFP expression in TXTL reactions containing a PiCas12a variant (labeled on top), a targeting (blue) or non-targeting (red) gRNA, and a GFP reporter containing the protospacer flanked by various non-canonical motifs (labeled on rows). The error bars represent the standard deviation from three separate TXTL reactions.*


Hk -----MFEKLSNIVSISKTIRFKLI-PVGKTLENIEK------LGKLEKDFERSDFYPIL
As ---MTQFEGFTNLYQVSKTLRFELI-PQGKTLKHIQE------QGFIEEDKARNDHYKEL
Pd ---MENYQEFTNLFQLNKTLRFELK-PIGKTCELLEEGKIFASGSFLEKDKVRADNVSYV
Pi MKVMENYQEFTNLFQLNKTLRFELK-PIGKTCELLEEGKIFASGSFLEKDKVRADNVSYV
Adurb193 --MSSVFSDFTNQYQLTKTLRFELK-PVGKTLEYIKE------NGLLSDDKLRAENYTKI
Adurb336 ---------MTNKYSLSKTLRFELKKPIGKTQEFIEK------NGLIEEDEQRNKDFLCG
Fn3 ---MSIYQEFVNKYSLSKTLRFELI-PQGKTLENIKA------RGLILDDEKRAKDYKKA
KNISDDYYRQYIKEKLSDLNLDWQKLYDAHELLDSSKKES-QKNLEMIQAQYRKVLFNIL
KPIIDRIYKTYADQCLQLVQLDWENLSAAIDSYRKEKTEETRNALIEEQATYRNAIHDYF
KKEIDKKHKIFIEETLSSFSISNDLLKQYFDCYNELKAFK-KDCKSDEEEVKKTALRNKC
KKEIDKKHKIFIEETLSSFSISNDLLKQYFDCYNELKAFK-KDCKSDEEEVKKTALRNKC
KDIIDDFHKDFIDKSLKQADIRWDDLKNGIENYRKDNSKKNKDELEKLQKRYRNEIVECF
KELLEGYYSDLIERRLGKIEIDSNLLEEYFKKYKEFNKLK-QKGKEVDAKELKNS-KDSL
KQIIDKYHQFFIEEILSSVCISEDLLQNYSDVYFKLKKSD-DDNLQKDFKSAKDTIKKQI
S---GELDKSGEKNSKDLIKN----------NK-----ALYGKLFKKQFILEVLPDFV
I---GRTDNLTDAINKRHAEI----------YK--------GLFKAELFNGKVLKQLG
T---SIQRAMREAISQAFLKS----------PQ----------KKLLAIKNL----
T---SIQRAMREAISQAFLKS----------PQ----------KKLLAIKNL----
K-------KDKEIFS----------YLFKKELITEILPEYI
K---ASQRKLREELHKSFFDNDALFKDEEPFKEKRDNEDVIIAAVKFHRKKLLEKRNEGG
SKYINDSEKFKNLFNQNLIDA----------KKGQESDLILWLKQSKDNGIELFKANS
NNNDSYSEEDLEGLNLYSKF-----TTRLKNFWETRKNVFTDKDIVTAIPFRAVNENFGF
TVTTTEHENAL--LRSFDKF-----TTYESGFYENRKNVFSAEDISTAIPHRIVQDNFPK
-IENVFKADEN--VQHFSEF-----TSYFSGFETNRENFYSDEEKSTSIAYRLVHDNLPI
-IENVFKADEN--VQHFSKF-----TSYFSGFETNRENFYSDEEKSTSIAYRLVHDNLPI
KNKETFEAERD-ILSSFSRF-----ATYFTSFQETRKNIYSKNAISTAISYRLVHQNFPK
VDNNLEIEEDK--INCLYRFVNDHCTTYFTKFFNNRKNVFSDKAISTSIFYRTIHENLHF
DITDIDEALEI--IKSFKGW-----TTYFKGFHENRKNVYSSNDIPTSIIYRIVDDNLPK
YYDNIKIFNKNIEYLENK--IPNLENELKEADILDD----------NRSVKDYFTPNGFN
FKENCHIFTRLITAVP---SLREHFENVKKAIGIFV----------STSIEEVFSFPFYN
FIKNIYIFEKLKEQFDAK-TLSEIFENYKLYVA----------GSSLDEVFSLEYFN
FIKNIYIFEKLKEQFDAK-TLSEIFENYKLYVA----------GSSLDEVFSLEYFN
FISNIEVFNKISEVCADI--ILDTERELAEILD----------GRRLRDVFSIEYYN
FVKNIENFEKLSEIIDEK-ELLEIEKELQKELG----------MFNLKEVFNINYFN
FLENKAKYESLKDKAPEAINYEQIKKDLAEELTFDIDYKTSEVNQRVFSLDEVFEIANFN
YVITQDGIDVYQAIRGG-FTKENG-EKVQGINEILNLTQQQLRR-KPETKNVKLGVLTKL
QLLTQTQIDLYNQLLGG-ISREAGTEKIKGLNEVLNLAIQKNDETAHIIASLPH-RFIPL
NTLTQKGIDNYNAVIGK-IVKEDK-QEIQGLNEHINLYNQKHKD-----RRLPF--FISL
NTLTQKGIDNYNAVIGK-IVKEDK-QEIQGLNEHINLYNQKHKD-----RRLPF--FISL
NVLCQSGIDDFNRCIGG-YSKDDG-VKIRGLNEFVNLFNQKNNG-------LMLSKMTPL
NCLNQKGIDKYNQIISG-VKKKD--IQKEGLNNLINKLSQKKDKDQEERKNIRKLRMIQL
NYLNQSGITKFNTIIGGKFVNGEN-TKRKGINEYINLYSQQIND-----KTLKKYKMSVL
RKQILEYSESTSFLIDQIEDDNDLVDRINKFNVSF-FESTEVSPSLFEQIERLYNAL--K
FKQILSDRNTLSFILEEFKSDEEV---IQSFCKYKTLLR---NENVLETAEALFNEL--N
KKQILSDREALSWLPDMFKNDSEVIKALKGFYIEDGF-----ENNVLTPLATLLSSL--D
KKQILSDREALSWLPDMFKNDSEVIKALKGFYIEDGF-----ENNVLTPLATLLSSL--D
YKQILSDREKISFIPEKFDNDFQVKSAINEYC----------NGLLKEDIAALVDELVLN
YKQILSEPESKSFKIESIENGKELEENIFSFYKSIALPSENQKESVLTLLEKLFDS---D
FKQILSDTESKSFVIDKLEDDSDVVTTMQSFYEQIAAFKTVEEKSIKETLSLLFDDLKAQ

```
Hk              --SIKKEEVYID-ARNTQKFSQMLFGQWDVIRRGY-------------------TVKITE
As              --SIDLTHIFIS-HKKLETISSALCDHWDTLRNAL-----------YERRISELTGKITK
Pd              --KYNLNGIFIRNNEALSSLSQNVYRNFS-IDEAIDANAELQTFNNYELIANALRAKIKK
Pi              --KYNLNGIFIRNNEALSSLSQNVYRNFS-IDEAIDANAELQTFNNYELIANALRAKFKK
Adurb193        YAQGNSNNIYIV-GKEFTNLSKMLTDSWSTINDSL-----------YLFAVSLFGDNDKK
Adurb336        --ELDFDGIYVK-QKSINKISNKIHSDWYLIESAL---------------KERYKSKIKS
Fn3             --KLDLSKIYFKNDKSLTDLSQQVFDDYSVIGTAV-----------LEYITQQVAPKNLD
                  . .   ::.        :*. :   :  :  .                     :  .

Hk              GSKEEKKKYKEYLE-LDETSKAKRYLNIREIEELV----------------NLVEGFEEV
As              SAKEK----------VQRSLKHED-INLQEIISAA-------------------------
Pd              ETKQGRKSFEKYEEYIDKKVKAIDSLSIQEINELVENYVSEFNSNSGNMPRKVEDYFSLM
Pi              ETKQGRKSFEKYEEYIDKKVKAIDSLSIQEINELVENYVSEFNSNSGNMPRKVEDYFSLM
Adurb193        STKEK----------INRWIKSAE-FSVKTMKDAL--------------------LMNGI
Adurb336        TGKKDKTDLQKEKE-TKKWFKETKHFSINEINSAIEFTKNDLITKEN---KDIWKYFKGV
Fn3             NPSKKEQDL------IAKKTEKAKYLSLETIKLALEEFNKHRDIDKQCRFEEILSNFAAI
                  .:               :    :.:  :

Hk              ---DV-FSVLLEKFKMNNIERSEFEAPIYGSPIK----------LEAIKEYLEKHLEEYH
As              ------GKELSEAFKQKTSEILSHAHAALDQPLPTTLK--KQEEKEILKSQLDSLLGLYH
Pd              RKGDFGSNDLIENIKTKLSAAEKLLGTKYQETAKDIFK--KDENSKLIKELLDATKQFQH
Pi              RKGDFGSNDLIENIKTKLSAAEKLLGTKYQETAKDIFK--KDENSKLIKELLDATKQFQH
Adurb193        ---DVKIEKLFDTIKQKTDVIIKEYEAIKPYLCNDEKFLGNETGIERVKSLLDAIMELMH
Adurb336        ---KTKEKNLFNEIQTSFGDLKKVKFGKERELLYDD----NEENVVKIKKALDYVQELFW
Fn3             -------PMIFDEIAQNKDNLAQISIKYQNQGKKDLLQASAEEDVKAIKDLLDQTNNLLH
                        : : :  .     .                         :*. *:      .

Hk              KWKLLLIGN-----DDLDTDETFYPLLNEVISDYY-IIPLYNLTRNYLTRKHSDKDKIKV
As              LLDWFAVD------ESNEVDPEFSARLTGIKLEMEPSLSFYNKARNYATKKPYSVEKFKL
Pd              FIKPLLGTG-----EEADRDLVFYGDFLPLYEKFEELTLLYNKVRNRLTQKPYSKDKIRL
Pi              FIKPLLGTG-----EEADRDLVFYGDFLPLYEKFEELTLLYNKVRNRLTQKPYSKDKIRL
Adurb193        MLKVFNVS------NELDRDMSFYSVYDTIYNNLTEVIQLYNKVRNYATQKPYSEDKYKL
Adurb336        LISPLMYEKPKEGVFDLDLNPDFYEKFVQIYEELQQITPLYNKTRNFISQKPHCESKFKL
Fn3             RLKIFHISQSEDKANILDKDEHFYLVFEECYFELANIVPLYNKIRNYITQKPYSDEKFKL
                  . :            : :  *         .     :**  **  :.*    .* .:

Hk              NF-DFPTLADGWSE------SKISDNRSIILRKGG-----YYYLGILIDNKLLINKKN--
As              NF-QMPTLASGW-D-----VNKEKNNGAILFVKNG-----LYYLGIMPKQK----GRYKA
Pd              CF-NKPKLMTGWVDSKTEKSDNGTQYGGYLFRKKNEIGEYDYFLGISSKAQLF--RKNEA
Pi              CF-NKPKLMTGWVDSKTEKSNSGTQFGGYLFRKKNEIGEYDYFLGISSKTQLF--RKNEA
Adurb193        NFENKGNFLNGWVDSKTESNDNGTQYGGYLFRKRNLLNQYDYYLGVSSDVKLFRECKDSN
Adurb336        NF-GDGYLLNGWAE-----KDGRCGYRAVIFRQGN-----KYYLGIIAKG-----KKNTK
Fn3             NF-ENSTLASGW-D-----KNKESANTAILFIKDD-----KYYLGIMDK------KHNKI
                 *     :  ** :      .      . :: : .      *:**:  .       .

Hk              --------KSKKIYE---ILIYNQIPEFSKSIPNYPFT-KKVKEHFKNNVS-DFQLIDGY
As              LSFE-PTEKTSEGFD---KMYYDYFPDAAKMIPKCSTQLKAVTAHFQTHTT-PILLSNNF
Pd              VIGD---YERLDYYQPKANTIYGSAYEGENSYKEDKKRLNKVIIAYIEQIK-QTNIKKSI
Pi              VIGD---YERLDYYQPKANTIYGSAYEGENSYKEDKKRLNKVIIAYIEQIK-QTNIKKSI
Adurb193        VFCD---YERLYYYQPKVSTIYGSAYKGEKTYAKDKSELITIIDDFVSCCDIDNDKLIAF
Adurb336        MFDKIPNYKSGDYYE---KMKYNQLQKPDQNLP------------------------RIF
Fn3             FSDKAIEENKGEGYK---KIVYKQIADASKDIQ------NLMIIDGKTVCKKGRKDRNGV
                        :    :.      *    .  :

Hk              VSPLI-----ITKEIYD----IKKEKKYKKDFYKDNNTNKNYLYTIYKWIEFCKQFLYKY
As              IEPLE-----ITKEIYDLNNPEKEPKKFQTAYAKKTGDQKGYREALCKWIDFTRDFLSKY
Pd              IESISKYPNISDDDKVTPSSLLEKIKKVSIDSYNGILSFKSFQSVNKEVIDNLLKTISPL
Pi              IESISKYPNISDDDKVTPSSLLEKIKKVSIDSYNGILSFKSFQSVNKEVIGNLLKTISCL
Adurb193        KEKVE----------LTPNGYINWLKKEYPSIFEKLICNQEFKNKNNEIIGNLKKALKDL
Adurb336        FAP-------SNEKYYNPSEEIKSIKN--------KGSFRTNIEDCHKMIEFYKKSISKK
Fn3             NRQLL---SLKRKHLPENIYRIKETKSYLKN------EARFSRKDLYDFIDYYKDRL---
                              :  *.            .       . *    . :
```

```
Hk              KGPNKESYKEMY-DFS-TLKDTSLYVNLNDFYADVNSCAYR--VLFNKIDENTIDNAVED
As              ------TKTTSI-DLS-SLRPSSQYKDLGEYYAELNPLLYH--ISFQRIAEKEIMDAVET
Pd              K-----NKAEFL-DLI-----NKDYQIFTEVQAVIDEICKQKTFIYFPISNVELEKEMGD
Pi              K-----NKEEFH-DLI-----KKDYQIFTEVQAVIDEICKQKTFIYFPVSNVELEKEMGD
Adurb193        I-----PKMPYVQQYI-----EKQYTIVNEIISDIEEISHNKILTYIKINRREMENALSR
Adurb336        ------SEWDIY-KFK--FKETNDYVSIDEFYKDFSNQAYK--IEFQKISVDDIHELVDS
Fn3             ---------DYY-DFEFELKPSNEYSDFNDFTNHIGSQGYK--LTFENISQDYINSLVNE
                             .       . *  . :    .     .  . :  :    : . :

Hk              G--KLLLFQIYNKDFS---------PESKGKKNLHTLYWLSMFSEENLRTRKLKLNGQAE
As              G--KLYLFQIYNKDFA---------KGHHGKPNLHTLYWTGLFSPENLAKTSIKLNGQAE
Pd              KDKPLCLFQISNKDLSFAKTFSANLRKKRGAENLHTMLFKALMEGNQ---DNLDL-GSGA
Pi              KDKPLCLFQISNKDLSFAKTFSANLRKKRGAENLHTMLFKALMSGNQ---DNLDL-GSGA
Adurb193        GTKRLYLFKITNKDLSFAESFEKGYRKEIKNKNLHTLYFEQLMSGGN---SVFDL-GTGS
Adurb336        G--KLYLFTIYNKDFSEKKFKK---DFSKRKPNLHTMYWKELFSDENLKNVVYQLNGGAE
Fn3             G--KLYLFQIYSKDFS---------AYSKGRPNLHTLYWKALFDERNLQDVVYKLNGEAE
                    * ** * .**::               ****: :  ::.  :      .* * .

Hk              IFYRKK-LEKKPIIHKEGSILLNKIDKEGN-TIPENIYHECYRYLNKKIGREDLSDEAIA
As              LFYRPKSRM-KRMAHRLGEKMLNKKLKDQKTPIPDTLYQELYDYVNHRLSH-DLSDEARA
Pd              IFYRAKSLDGNKPTHPANEAIKCRNVANKD-K------VSLF------------------
Pi              IFYRAKSLDGNKPTHPANEAIKCRNVANKD-K------VSLF------------------
Adurb193        IFYRPASIE-SPFVHKKGEILIDRKDKNGN-PIPMDIYDQLFNHFNY--GS-TLSEEAKK
Adurb336        IFFRKKSPSNKKIIHIKNKAVPNEKPINGR-K------ESTF------------------
Fn3             LFYRKQSIP-KKITHPAKETIANKNKDNPK-K------ESVF------------------
                :*:*      .   *   . :     :           . :

Hk              LFNKDVLKYKEARFDIIKDRRYSESQFFFHVPITFNWDIKTNK-N-VNQIVQGMIKD-GE
As              L--LPNVITKEVSHEIIKDRRFTSDKFFFHVPITLNYQAANSPSK-FNQRVNAYLKE-HP
Pd              ------------TYDIYKNRRYMENKFLFHLSIVQNYKAANDSAQ-LNSSATEYIRK-AD
Pi              ------------TYDIYKNRRYMENKFLFHLSIVQNYKAAKDSAQ-LNSSATEYIRK-AD
Adurb193        Y--KDIVVVKNAKHDIIKDKRFTQDKFLFHLSITINYKATGSG-Y-INAKVLNAISN-NP
Adurb336        ------------EYDLIKDKRYFEDKYLFHCPITLNFKSEKNNVESFNNYINEFIVKNND
Fn3             ------------EYDLIKDKRFTEDKFFFHCPITINFKSSGAN-K-FNDEINLLLKEKAN
                            .:: *:.*: ..:::** .*. *:.        .*      : .

Hk              IKHIIGIDRGERHLLYYSVIDLEGNIVEQGSLNTLEQNRFDNSTVKVDYQNKLRTREEDR
As              ETPIIGIDRGERNLIYIITVIDSTGKILEQRSLNTIQ---------QFDYQKKLDNREKER
Pd              DLHIIGIDRGERNLLYYSVIDMKGNIVEQDSLNIIR----NN-DLETDYHDLLDKREKER
Pi              DLHIIGIDRGERNLLYYSVIDMKGNIVEQDSLNIIR----NN-DLETDYHDLLDKREKER
Adurb193        NVNIIGIDRGERNLLYVSLINQRGDIITQKSFNIAG---------GFDYAAKLTQYEKNR
Adurb336        EINILGLDRGERNLLYYYLLNPNGEKIKSGSFNIFK----DEFGREFDYHKMLSEREDER
Fn3             DVHILSIDRGERHLAYYTLVDGKGNIIKQDNFNIIG----ND-RMKTNYHDKLAAIEKDR
                   *:.:*****:* *  :::  *. : . .:*             :*   *   *.:*

Hk              DRARKNWTNINKIKELKDGYLSHVVHKLSRLIIKYEAIVIMENLNQGFKRGRFKV-ERQV
As              VAARQAWSVVGTIKDLKQGYLSQVIHEIVDLMIHYQAVVVLENLNFGFKSKRTGIAEKAV
Pd              KANRQNWEAVEGIKDLKKGYLSQAVHQIAQLMLKYNAIIALEDLGQMFVTRGQKI-EKAV
Pi              KANRQNWEAVEGIKDLKKGYLSQAVHQIAQLMLKYNAIIALEDLGEMFVTRGQKI-EKAV
Adurb193        DIARKDWKQIGNIKELKEGYLSVVIHEIIQMMIKNNAIIVMEDLNFGFKRGRFKF-ERQV
Adurb336        DKDQKEWKNIKNIKDLKEGYLSRIIHEISKIVIDNNAIVVLEDLNPDFKRNRFKF-ENQV
Fn3             DSARKDWKKINNIKEMKEGYLSQVVHEIAKLVIEYNAIVVFEDLNFGFKRGRFKV-EKQV
                   .: *  :  **::*.****  :*::  :::  :*:: :*:*.  *      . *. *

Hk              YQKFELALMNKLSALSFKEKYDERKNLEPSGILNPIQACYPVDAYQEL--QGQNGIVFYL
As              YQQFEKMLIDKLNCLVLK----DYPAEKVGGVLNPYQLTDQFTSFAKM--GTQSGFLFYV
Pd              YQQFEKSLVDKLSYLVDK----KRPYNELGGILKAYQLASSITKNNS---DKQNGFLFYV
Pi              YQQFEKSLVDKLSYLVDK----KRPYNELGGILKAYQLASSITKSNS---DKQNGFLFYV
Adurb193        YQKFEKMLIDKLNYLADK----KKNPADEGGILRGYQLTEAFTSFQKL--GKQSGFIFYV
Adurb336        YQKFEEKLIKKLNYLVFK----NRKSTDSGGLFRAYQLTSKFVSFKNLREQKQSGILYYV
Fn3             YQKLEKMLIEKLNYLVFK----DNEFDKTGGVLRAYQLTAPFETFKKM--GKQTGIIYYV
                **::*  *:.**. *  *    .    . .*::.  *    .    .     *.*:::*:
```

```
Hk              PAAYTSVIDPVTGFTNLFRLKSINSSKYEE-FIKKFKNIYFDNEEEDFKFIFNYKD---F
As              PAPYTSKIDPLTGFVDPFVWKTIKNHESRKHFLEGFDFLHYDVKTGDFILHFKMNRNLSF
Pd              PAWNTSKIDPVTGFTDLLRPKAMTIKEAQD-FFGAFDNISYN-DKGYFEFETNYDK---F
Pi              PAWNTSKIDPVTGFTDLLRPKAMTIKEAQD-FFGAFDNISYN-DKGYFEFETNYDK---F
Adurb193        PAGYTSKIDPTTGFVSLINLKWTSIENAQK-IISAMDFIRYNSSEDYFEFGIDYDK---L
Adurb336        NANYTSKIDPKTGFVNLLYPNYKNIEKSKL-FFDKFDSIKYNKSENMFEFEFDYSN---F
Fn3             PAGFTSKICPVTGFVNQLYPKYESVSKSQE-FFSKFDKICYNLDKGYFEFSFDYKN---F
                 *  ** * * ***.. :  :  .  :    ::  :. : :: .   * :  . .    :

Hk              AKANLVILNNIKSKDWKISTR---------------GERISYNSK--KKEYFY-----VQ
As              QRG----LPGFMPA-WDIVFEKNETQFDAKGTPFIAGKRIVPVIE---NHRFTGRYRDLY
Pd              KIR----MKSAQTR-WTICTF---------------GNRIKRKKD---KNYWN--YEEVE
Pi              KIR----MKGAQTR-WTICTF---------------GNRIKHKKD---KNYWN--YEEVE
Adurb193        RTA----QTDFRKK-WTVCTF---------------GDRYTHTQN---KATSNHTTEKVN
Adurb336        TKN----LKLKKNN-WTVFTN---------------GERIIQVKT--KNNIYE--PKKIN
Fn3             G------DKAAKGK-WTIASF---------------GSRLINFRNSDKNHNWD--TREVY
                              * :                  *.*        :         :

Hk              PTEFLINKLKELNIDYENIDIIPLIDNLEEKAKRKILKALFDTFKYSVQLRNYDF--END
As              PANELIALLEEKGIVFRDGS--NILPKLLENDDSHAIDTMVALIRSVLQMRNSNAATGED
Pd              LTEEFKKLFKDSNIDYENC---NLKEEIQNKDNRKFFDDLIKLLQLTLQMRNSDDK-GND
Pi              LTEEFKKLFDKYDIDYRDG---NLKEEILKIDNRKFFDALIKLLQLTLQLRNSDDK-GND
Adurb193        LTAKLKAVLDKYSIDYDKGE--DIRDILCKSNSKELLETVLYVLKLAMQMRSTSPEDDVD
Adurb336        LTNEMKNLFESEGISFNDGK--NLKEEIINSNSKKLHTNLTNLLKYTLQLRNSNNKTGED
Fn3             PTKELEKLLKDYSIEYGHGE--CIKAAICGESDKKFFAKLTSVLNTILQMRNSKTGTELD
                 :  :   :.. .* :       :   :    .      :   :.  :*:*. .     *

Hk              YIISPTADDNGNYYNSNEIDIDKTNLPNNGDANGAFNIARKGLLLKDRIVNSNE--SKVD
As              YINSPVRDLNGVCFDSRFQN---PEWPMDADANGAYHIALKGQLLLNHLKESKD--LKLQ
Pd              YIISPVANAEGQFFDSRNGD---KKLPLDADANGAYNIARKGLWNIRQIKQTKN-DKKLN
Pi              YIVSPIANADGKFFNSNDGC---KELPLDADANGAYNIARKGLWVVRQIKDKKD---KIS
Adurb193        MIISPVKNDDGKFFVSGN-D---EKLPIDADANGAFHIALKGLMYIKRIKNGKI--ECFD
Adurb336        YILSCVKDDEKNFFNSNNAK---DSEPENADANGAYHIGLKGIMLIKKMKKQSDKNQKID
Fn3             YLISPVADVNGNFFDSRQAP---KNMPQDADANGAYHIGLKGLMLLDRIKNNQE-GKKLN
                 : *   : :   : *        . * :.*****::*. **     .: . .     ..

Hk              LKIKNEDWINFIIS---------
As              NGISNQDWLAYIQELRN------
Pd              LSISSTEWLDFVREKPYLK----
Pi              -KLSNQEWLKFAQEKPYLK----
Adurb193        KGMATYEWLKFMQNREYKS----
Adurb336        LRISNEEYFNEMCSKEQSKKKES
Fn3             LVIKNEEYFEFVQNRNN------
                  : . :::    .
```

Green highlight indicates mutated AsCas12a residues in previous work.
Red highlight indicates mutated PiCas12a residues.

Supplementary Document 5.1 - *Sequence alignments of various Cas12a nucleases investigated in this work, as well as AsCas12a. The green highlighted residues indicate mutated residues of AsCas12a shown to alter PAM specificity (PMID: 28581492, 30742127). Residues highlighted in red indicates the mutated PiCas12a residues tested in this work. The sequences were aligned using MUSCLE with default settings.*