

# Single molecule analysis of CRISPR enzymes

by

Digvijay Singh

A dissertation submitted to Johns Hopkins University in conformity with  
the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

January, 2018

Doctoral dissertation Committee:

Dr. Geraldine Seydoux (Chair)

Dr. Taekjip Ha (Advisor)

Dr. Scott Bailey

Dr. Jie Xiao

Dr. Bin Wu

# ABSTRACT

CRISPR system provides adaptive immunity against foreign DNA in microbes. CRISPR nucleases in complex with a guide-RNA targets complementary DNA (protospacer) if they are next to protospacer adjacent motif (PAM), by the virtue of RNA-DNA base pairing, resulting in unwinding of DNA strands in protospacer. CRISPR has been re-purposed for wide-ranging biological applications. But lack of mechanistic understanding and specificity of both binding and cleavage continue to be challenges for CRISPR based applications. I employed single molecule fluorescence and biochemical assays to investigate the molecular mechanism of widely used CRISPR-enzymes (Cas9, its engineered derivatives (EngCas9) and Cpf1 orthologs). Following observations were made:

- CRISPR-enzymes bind DNA in 2 modes. In mode I, enzymes samples DNA non-specifically for transient PAM-detection. Upon successful PAM-recognition, binding shifts to mode II involving RNA-DNA duplex/DNA unwinding. Mode-II is longer lived, whose lifetime depends on number and location of on/off-target base-pairs (bp) (PAM-proximal off-targets bp; deleterious PAM-distal ones; tolerable).
- Cas9s require only 9-10 bp for stable binding and ~16 bp for cleavage. Cpf1s are significantly more specific, requiring ~17 bp for both.
- Release of cleaved-DNA products for genome-editing machinery is likely crucial. Cas9s do not release any, where Cpf1s releases one end.
- Cas9s cleavage happens from maximally unwound DNA state. PAM-distal off-target bp reduce extent and lifetime of unwinding thus delaying cleavage. EngCas9s mutations destabilize

unwound state and have a much lower intrinsic cleavage rate, which are the basis of their improved specificity.

- Significant effect of pH and reducing conditions on Cpf1 activity can explain higher performance variance (compared to Cas9) in different organisms. These can also be used as switches for Cpf1 activation/inactivation.

The work presented in this dissertation has already resulted in 3 first-author manuscripts and a review.

The work has been further extended into other fluorescent geometries and onto newer CRISPR enzymes. I am also applying the biophysical information from my PhD work to re-engineer CRISPR-Cas9 for precise control of its DNA binding and cleavage activity. These efforts have resulted in useful data and will likely lead to more manuscripts. Continued improvement in our understanding of CRISPR mechanism will assist in rational-design of new enzymes/strategies with higher specificity and efficiency.

*To my family*

# ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Prof. Taekjip Ha for giving me the opportunity to pursue my PhD training in his group. I joined graduate school in Fall 2012 and had my first rotation in his lab. I developed an instant liking to the group, its environment and the kind of questions that were being pursued; simple questions of profound fundamental importance. I wanted to join a group which could provide a great mix of training in biological protocol work, physical instrumentation and quantitative analysis, Ha Group was the perfect choice.

Prof. Ha enabled an extremely positive and patient environment in which I could blossom as a scientist. Just a week after joining the group, I met with an accident and broke a bone and tore ligaments. This accident immobilized me fully and partially for the next few months. The inability to attend the lab in those days meant that, in some critical instances I could not directly pick up the necessary single molecule experimental skills from extremely experienced lab members who were just about to graduate at the time. Hence my progress was a little slow in the beginning and I made many mistakes trying to pick up a lot of necessary skills by trial and error. But Prof. Ha was extremely supportive, informative and patient through this whole process. His enthusiasm for science, encouragement and profound insight were inspirational and always kept me going.

Under his mentorship, I learnt how to think big picture. And writing papers with him taught me how to organize my results into simple and concise stories. At the beginning of graduate school, my presentation skills involved me talking and presenting a lot of materials. My slides tended to be quite busy and thus my presentations were not sharp enough. Prof. Ha taught me how to improve the signal to noise ratio in my presentations. In my 2<sup>nd</sup> year, he used a phrase '*present information on a need to know basis and not on a good to know basis*'. Since then, this simple and incredibly useful phrase has been a guiding principle for all my presentations. I have also learnt a lot by watching his presentations in conferences and

seminars. There are still many things I need to improve to be as good a speaker as he is. Presentation skills are an extremely important part of a scientific career and I am grateful to Prof. Ha for his guidance in this regard. Life of scientists outside the lab, interaction with your peers in the department and in conferences are some of other important things I picked up by watching him in conferences and seminars. Not to mention, he was extremely supportive of my attendance in these conferences.

You would find many wonderful scientific attributes of Prof. Ha, both in writing and in words. One of many is his *minimalism*. He communicated with me in limited, short and precise sentences. This brought precision and accuracy in our conversations, helped me pick up the big picture that he was thinking, and also brought a lot of clarity in my own thoughts regarding my experiments. The density of wisdom in conversations with him has been extremely high. 10 minutes with him was greater than hours with others.

In Hopkins, we shared the building space with lab of Prof. Sua Myong. And I have learned a great deal in interactions with her and her lab members. Especially, I learnt a lot of useful Molecular Biology skills from Dr. Ramreddy Tipanna who is a post-doc in Myong lab. Prof. Myong and Prof. Ha are the wife-husband couple and they have hosted us multiple times in their house for holiday and graduation parties and those are some of my best memories of my PhD life. Lot of good food and interesting discussions!

I was able to form great collaborations both in and outside the Ha Group. Valuable inputs, discussions and healthy criticisms from these collaborations have greatly enhanced my knowledge and research output. The most notable of these collaborations are: Prof. Scott Bailey and his student John Mallon, Prof. Jennifer A. Doudna, her student Dr. Samuel H. Sternberg and Ha Group's former post-doc Dr. Jingyi Fei, and Prof. Venigalla B. Rao and his student Dr. Li Dai (Joyce). I would also like to thank my Doctoral thesis Committee for their extremely valuable advice and discussions.

I am extremely impressed by the enthusiasm of new graduate students that have joined the Ha Group. Yanbo Wang will be taking over most of my CRISPR work and he has already hit the ground running. Taylor Cottle and Ikenna Okafor are doing excellent CRISPR work in their rotations. It has been a great experience for me to mentor, teach CRISPR and related biophysical and biochemical skills to these young blood. It is a great time to employ cutting edge single molecule approaches to improve our understanding of the CRISPR toolbox. And I am quite positive that we will get to see excellent CRISPR work from the new CRISPR people of Ha Group.

Prior to my graduate school, I found exciting opportunities to work in many labs. Each one of them kept piquing my interest in science. Prof. Martin Gruebele, Prof. Robert B. Best, Prof. Charles Schroeder (rotation during grad school), Prof. Swagata Dasgupta and others were exactly the kind of scientists and ‘adults’ I wanted to be when I grew up. Having such role models and supportive mentors was extremely vital for a good start to my scientific career. I am very grateful for their contributions and they will be my heroes forever.

Sports and athletics were a vital part for my well-being, they helped me de-stress and re-energize from the daily rigors of academic research. I am extremely thankful to administration of Rec centers, Tennis courts and swimming pool. Some of my best thoughts came in at these places. I also made many friends here and enjoyed nerdy discussions with them about technical aspects of these sports. I also got a chance to play some rallies with Prof. Ha, who had technically sound and consistent shots. During my PhD, I made incredible friends who helped me see life outside the lab and absorb frequent setbacks of graduate school. Some of those amazing friends are: Saumya; who has been an incredible friend and one of the kindest humans I know. She has a great sense of clarity, simplicity and calm. She and I have a dog whose name is Leo. Leo does not live with either of us at the moment, but stories of his daily antics are a great reminder of how to live one’s life to the fullest. I hope to live with Leo soon. Leo has a [Facebook page](#) that I encourage you to check out. Pratik; *brother from another mother*. Very calm, composed and a super chill

guy to hang out with. He has gone through a huge tragedy during his PhD and his resolve to stay focused through all this is a source of great inspiration for me. Others, in no particular order are: Desi coterie of Illinois (Bhoomika, Aditi, Amruta and others), Ha-Myong lab company (Seongjin, Yanbo, Anustup, Salman, Olivia, Amir, my batch mates Boyang and Chang-Ting, and others), *angrezi warriors* of E. Clark Street and CS guys, and Tennis partners (Anthony, Anchit, Rakesh, Athindran, Kevin, Abid, Teodor and others).

Last, but most important people to thank are the most important people in my life; my parents. I have found no greater than inspiration the hard work put up by parents in making sure that I had a great childhood and received education that led me to where I am today. I would not be anything without them. I could spend many thesis writing about their contributions and positive influence on me. They have always taught me to work hard, dream big, and stay positive and honest. A big motivation behind my many ambitions is to bring joy and pride to my parents, who will always be my biggest supporters. I have been away from my home for >12 years now (high school, college, grad school) and one of my immediate goals is to setup a life where I could spend far-far greater time with my parents. And the one and the only one thing which I have away sought from the Lord Shiva (the God of Gods) is good health for my parents. And my successes in PhD has given them happiness and happiness helps with good health.

The space here is too small to list all my incredible friends, mentors and their influence on me. Omissions in the list are purely for space reasons. I hope to continue to learn many things from you.



# Table of Contents

ABSTRACT.....	ii
ACKNOWLEDGEMENTS.....	v
Chapter 1. Introduction to CRISPR and single molecule technologies.....	1
1.1 ABSTRACT.....	2
1.2 INTRODUCTION to CRISPR.....	2
1.3 FLUORESCENCE and FORSTER RESONANCE ENERGY TRANSFER.....	5
1.4 INTRODUCTION to SINGLE-MOLECULE TECHNOLOGIES.....	8
1.5 SINGLE MOLECULE FRET (smFRET) and ITS IMPLEMENTATION.....	11
Chapter 2. Real-time observation of DNA recognition and rejection by the RNA-guided nuclease Cas9	14
2.1 ABSTRACT.....	15
2.2 INTRODUCTION.....	15
2.3 RESULTS.....	16
2.3.1 Real-time single molecule assay for DNA interrogation by Cas9-RNA.....	16
2.3.2 Effect of DNA target mismatches on Cas9-RNA binding.....	17
2.3.3 Different DNA binding modes of Cas9-RNA.....	25
2.3.4 Effect of DNA target mismatches on Cas9-RNA binding and dissociation kinetics.....	26
2.4 DISCUSSION.....	41
2.5 MATERIALS AND METHODS.....	42
2.5.1 Preparation of DNA targets.....	42

2.5.2	Expression and purification of Cas9 and dCas9 .....	43
2.5.3	Preparation of guide-RNA and Cas9-RNA.....	44
2.5.4	Single-molecule detection and data analysis.....	44
2.5.5	FRET histograms and Cas9-RNA bound DNA fraction.....	45
2.5.6	Lifetime analysis of bound and unbound states via thresholding .....	45
2.5.7	Counts of DNA target sequences in human genome.....	47
2.6	AUTHOR CONTRIBUTIONS.....	47
Chapter 3.	Mechanisms of improved specificity of engineered Cas9s revealed by single molecule analysis	50
3.1	ABSTRACT.....	51
3.2	INTRODUCTION .....	51
3.3	RESULTS .....	53
3.3.1	Real-time single molecule FRET assay for DNA interrogation .....	53
3.3.2	Cas9-RNA induced DNA unwinding.....	67
3.3.3	DNA cleavage vs. mismatches.....	69
3.3.4	Mechanism of mismatch sensitivity of EngCas9s .....	70
3.3.5	PAM-proximal DNA unwinding .....	70
3.4	DISCUSSION .....	93
3.5	MATERIALS AND METHODS.....	96
3.5.1	Preparation of DNA targets.....	96
3.5.2	Expression and purification of Cas9 .....	97
3.5.3	Preparation of guide-RNA and Cas9-RNA.....	98

3.5.4	Single-molecule fluorescence imaging and data analysis .....	99
3.5.5	FRET histograms, Cas9-RNA bound DNA fraction and unwound fraction.....	100
3.5.6	Kinetic analysis of DNA interrogation by Cas9-RNA.....	100
3.5.7	Kinetic analysis of Cas9-RNA induced DNA unwinding and rewinding.....	101
3.5.8	Gel-electrophoresis to investigate Cas9-RNA induced cleavage.....	102
3.6	AUTHOR CONTRIBUTIONS and ACKNOWLEDGEMENTS.....	106
Chapter 4. Real-time observation of DNA target interrogation and product release by the RNA-guided endonuclease CRISPR Cpf1 .....		
		107
4.1	ABSTRACT .....	108
4.2	INTRODUCTION .....	108
4.3	RESULTS .....	109
4.3.1	Real-time DNA interrogation by Cpf1-RNA .....	109
4.3.2	DNA cleavage by Cpf1 as a function of mismatches.....	130
4.3.3	Fate of cleaved DNA.....	131
4.4	DISCUSSION .....	144
4.5	MATERIALS AND METHODS.....	152
4.5.1	Preparation of DNA target and guide-RNA.....	152
4.5.2	Preparation of Cpf1-RNA.....	154
4.5.3	Single-molecule fluorescence imaging and data analysis.....	154
4.5.4	Expression and purification of Cpf1.....	155
4.5.5	FRET histograms, Cas9-RNA bound DNA fraction and unwound fraction.....	156
4.5.6	FRET efficiency histograms and Cpf1-RNA bound DNA fraction.....	157

4.5.7	Determination of binding kinetics.....	157
4.5.8	Estimation of dissociation constant ( $K_d$ ).....	158
4.5.9	Overall lifetime of release of cleavage products.....	158
4.5.10	Gel electrophoresis experiments.....	159
4.6	AUTHOR CONTRIBUTIONS.....	166
CONCLUSION AND OUTLOOK.....		167
Chapter 5.	Protocols.....	169
5.1	INTRODUCTION.....	170
5.2	EXPRESSION and PURIFICATION of CRISPR proteins.....	171
5.2.1	Materials.....	171
5.2.2	Basic introduction about the pET based vector for Cas9 protein production.....	173
5.2.3	Transformation of the cells with the vector containing your gene of interest.....	174
5.2.4	Preparing the media for large-scale culture.....	177
5.2.5	Preparing the inoculant for large-scale culture.....	178
5.2.6	Transferring the inoculant from 4mL culture to the large-scale culture.....	179
5.2.7	Setting up the OD <sub>600</sub> measurement.....	180
5.2.8	OD <sub>600</sub> check.....	181
5.2.9	Induction with (Isopropyl $\beta$ -D-1-thiogalactopyranoside) IPTG.....	181
5.2.10	Harvesting the cells.....	182
5.2.11	Lysing the cells.....	183
5.2.12	Ni-NTA Chromatography for protein purification.....	184

5.2.13	Elution of the protein of interest .....	187
5.2.14	Dialysis and obtaining Apo-Cas9 by utilizing TEV protease to remove Hist-Tag and MBP-Tag	188
5.3	IN VITRO TRANSCRIPTION AND PURIFICATION OF RNA.....	191
5.3.1	Introduction.....	191
5.3.2	Materials .....	191
5.3.3	Procedure .....	191
5.3.4	Preparation of partial dsDNA template for invitro transcription .....	193
5.3.5	Setting up the invitro transcription reaction.....	194
5.3.6	DNase treatment to degrade the partial dsDNA template.....	195
5.3.7	Purification of the RNA .....	195
5.3.8	Checking the integrity of the RNA .....	197
5.4	NHS ESTER LABELING OF NUCLEIC ACIDS WITH FLUORESCENT LABELS.....	198
5.4.1	Materials .....	198
5.4.2	Procedure .....	199
5.4.3	Purifying the labeled nucleic acids from the labeling reaction.....	202
5.5	CYSTEINE MALEIMIDE LABELING OF PROTEINS WITH FLUORESCENT LABELS	205
5.5.1	Materials .....	205
5.5.2	Procedure .....	206
5.5.3	Purifying the labeled protein from free R or dyes.....	208
5.6	PEGYLATION PROTOCOL .....	209
5.6.1	Introduction.....	209

5.6.2	Materials .....	209
5.6.3	Procedure .....	210
5.7	A VERSATILE PROTOCOL FOR SINGLE MOLECULE FRET EXPERIMENT USING CRISPR ENZYMES .....	218
5.7.1	Introduction.....	218
5.7.2	Materials .....	218
5.7.3	Procedure .....	220
5.7.4	Preparing the chambers for smFRET experiments .....	220
5.7.5	TIR Condition and TIR Spot.....	221
5.7.6	Immobilizing DNA target molecules on the surface.....	222
5.7.7	Preparation of Guide-RNA .....	224
5.7.8	Preparation of Cas9-RNA .....	226
5.7.9	Single Molecule Imaging/Data Acquisition.....	227
5.7.10	Double checking the consistency of the focus in the long-duration movies .....	230
5.7.11	Analysis of the single molecule movies.....	232
	BIBLIOGRAPHY .....	234
	CURRICULUM VITAE .....	248

# Table of Figures

Figure 1.1   Schematic of DNA targeting by CRISPR enzymes commonly used in genome engineering. ...	4
Figure 1.2 Molecular energy levels and transitions between them for fluorescence and FRET spectroscopy.....	8
Figure 1.3   Common strategies for single molecule fluorescence. ....	12
Figure 2.1   Cas9-RNA binding to a cognate sequence. ....	18
Figure 2.2   FRET probe labeling locations in the Cas9-RNA-DNA complex.....	20
Figure 2.3   Target cleavage activity is not impaired by fluorescent labeling of guide RNA or DNA target. ....	21
Figure 2.4   Comparison between Cas9 and dCas9, and high stability of Cas9-RNA-DNA for certain DNA targets.....	22
Figure 2.5   Cas9-RNA binding to DNA with proximal or distal mismatches. ....	24
Figure 2.6   Cas9-RNA bound state lifetimes for different DNA targets.....	27
Figure 2.7   Hidden Markov model analysis and transition density plots reveal transitions between three different FRET states. ....	29
Figure 2.8   Transition density plots for 5-20 <sub>mm</sub> DNA target at two different frame rates of image acquisition.....	31
Figure 2.9   Transition density plots for 9-20 <sub>mm</sub> , 5-20 <sub>mm</sub> and 1-2 <sub>mm</sub> DNA targets with different inputs for hidden Markov modeling. ....	32
Figure 2.10   Determination of dwell times in the unbound state. ....	33
Figure 2.11   Determination of the rates of FRET appearance and disappearance. ....	34
Figure 2.12   The binding rates. ....	35
Figure 2.13   Binding dynamics for ‘PAM-less’ cognate DNA target and non-cognate target with PAM. ....	36
Figure 2.14   Fitting survival probability of Cas9-RNA bound states vs. time.....	38

Figure 2.15   Kinetic model describing the transitions between various states of Cas9-RNA DNA targeting. ....	40
Figure 2.16   FRET histograms for roadblock constructs and ‘PAM-less’ constructs.....	40
Figure 2.17   The proposed model of bimodal Cas9-RNA binding along with the kinetics of Cas9-RNA DNA targeting as a function of mismatches. ....	42
Figure 3.1   smFRET assay to study DNA interrogation by engineered Cas9-RNA. ....	57
Figure 3.2   FRET probe locations for DNA interrogation by Cas9-RNA. ....	58
Figure 3.3   Determination of $K_d$ between Cas9-RNA and DNA. ....	59
Figure 3.4   $E$ histograms for different DNA targets obtained smFRET DNA interrogation experiments at 20 nM Cas9-RNA. ....	61
Figure 3.5   Ultrastable binding of EngCas9-RNA to DNA. ....	62
Figure 3.6   Binding and kinetic parameters of DNA interrogation (20 nM Cas9-RNA). ....	64
Figure 3.7   Transition density plots for smFRET DNA interrogation experiments. ....	66
Figure 3.8   FRET probe locations for smFRET DNA unwinding experiments. ....	72
Figure 3.9   Fluorescent labeling for smFRET DNA unwinding experiments does not affect cleavage. ...	74
Figure 3.10   Internal DNA unwinding/rewinding dynamics modulated by mismatches and Cas9 mutations.....	76
Figure 3.11   smFRET DNA unwinding experiments at different Cas9-RNA concentrations and different frame rates of image acquisition. ....	78
Figure 3.12   Transition density plots for smFRET DNA unwinding experiments. ....	79
Figure 3.13   smFRET DNA unwinding experiments using catalytically active Cas9-RNA. ....	80
Figure 3.14   Schematic of smFRET DNA unwinding experiments using surface-tethered Cas9-RNA. ...	82
Figure 3.15   DNA unwinding dynamics with surface immobilized Cas9-RNA. ....	83
Figure 3.16   DNA unwinding dynamics upon initial formation of Cas9-RNA-DNA complex.....	85
Figure 3.17   Appearance of unwound state over time (dCas9-HF1-RNA on DNA with $n_{PD}=1$ ).....	86
Figure 3.18   Cleavage vs mismatches and its relation with DNA unwinding.....	87



Figure 3.19   Cas9-RNA induced unwinding of various DNA and mechanisms of increased specificity by EngCas9s. ....	88
Figure 3.20   smFRET unwinding experiments using pre-unwound DNA. ....	90
Figure 3.21   Locations of EngCas9 mutations in dCas9-RNA-DNA complex (PDB ID: 4UN3). ....	91
Figure 3.22   Probe locations for smFRET assay to investigate Cas9-RNA induced DNA unwinding in PAM-proximal site. ....	93
Figure 3.23   DNA unwinding occurs in the absence of divalent cations. ....	96
Figure 4.1   Design of DNA targets and guide-RNA along with FRET probes labeling locations in the Cpf1-RNA-DNA complex for smFRET assay for DNA interrogation by Cpf1-RNA. ....	114
Figure 4.2   smFRET assay to study DNA interrogation by Cpf1-RNA. ....	115
Figure 4.3   FnCpf1-RNA activity is not impaired by fluorescent labeling of guide-RNA and DNA target. ....	116
Figure 4.4   Bound fraction and rates of FRET appearance and disappearance with increasing Cpf1-RNA concentration. ....	118
Figure 4.5   Cleavage activity of AsCpf1 at different pH conditions. ....	119
Figure 4.6   <i>E</i> histograms during DNA interrogation by Cpf1-RNA. ....	121
Figure 4.7   Dynamic interaction of Cpf1-RNA with DNA as a function of mismatches. ....	122
Figure 4.8   Representative smFRET time-trajectories from smFRET experiments to study DNA interrogation by AsCpf1-RNA. ....	124
Figure 4.9   Representative smFRET time-trajectories from smFRET experiments to study DNA interrogation by FnCpf1-RNA. ....	126
Figure 4.10   Representative smFRET time-trajectories from smFRET experiments to study DNA interrogation by LbCpf1-RNA. ....	127
Figure 4.11   Bimodal nature of DNA interrogation by Cpf1-RNA and its parameters (50 nM Cpf1-RNA). ....	129
Figure 4.12   FnCpf1 activity at different temperatures and divalent cation conditions. ....	132

Figure 4.13   AsCpf1-RNA activity and effect of divalent cation conditions.....	134
Figure 4.14   Cpf1-RNA activity at different pH conditions.....	136
Figure 4.15   DNA cleavage and product release.....	138
Figure 4.16   Time-lapse of cleavage byCpf1. ....	139
Figure 4.17   Catalytic activity of Cpf1 increases stability of Cpf1-RNA-DNA. ....	141
Figure 4.18   Design of DNA targets and guide-RNA for single-molecule cleavage product release assay to study Cpf1-RNA induced cleavage product release of DNA targets.....	143
Figure 4.19   Model of Cpf1-RNA DNA targeting, cleavage and product release. ....	147
Figure 4.20   FnCpf1 activity for DNA targets in pre-unwound configuration and DNA targets with partially or completely single stranded target strand analyzed by 4% native gel electrophoresis and SYBR Gold II staining of nucleic acids. ....	149
Figure 4.21   Effect of ‘nick’ and reducing conditions on the Cpf1 induced DNA cleavage and release of cleavage products.....	150
Figure 4.22   Effect of an extra guanine (G) base in guide-RNA for LbCpf1-RNA activity.....	151
Figure 4.23   FnCpf1 activity is not affected by the imaging buffer components.....	152
Figure 5.1   Schematic of DNA template for invitro transcription.....	192
Figure 5.2   Importance of C/CC at +1 and +2 sites for high efficiency invitro transcription .....	193
Figure 5.3   Schematic describing the reaction between NHS ester reagent (conjugated to fluorophore R) with the –NH <sub>2</sub> group in the nucleic acid or other target of interest. ....	199
Figure 5.4   –NH <sub>2</sub> (amine) group present in the modified thymine. ....	200
Figure 5.5   Hydrolysis of the ester.....	201
Figure 5.6   Schematic describing the reaction between Maleimide reagent (conjugated to fluorophore R) with the –SH group of cysteine in the protein of interest which is to be labeled. ....	205
Figure 5.7   Complete chemical schematic of the PEGylation protocol .....	211
Figure 5.8  Representative image of Cognate DNA target in Cas9-RNA imaging buffer.....	223
Figure 5.9   FRET probe labeling locations in the Cas9-RNA-DNA complex.....	225

Figure 5.10 | Representative image of Cognate DNA target (Cy3) incubated with 20nM Cas9-RNA complex (Cy5). The Spots on the right indicate FRETing molecules which indicate the interaction between DNA (with Cy3) and Cas9-RNA (with Cy5) bound to the DNA target..... 228

Figure 5.11 | A representative smFRET time-trajectory from a long duration movie ..... 231

Figure 5.12 | A representative image showing the various parameters for the movie analysis in the smCamera Software. .... 232

# **Chapter 1.**

## **Introduction to CRISPR and single molecule technologies**

\*Some contents of this chapter is under review at:

Singh, D., Ha, T. Understanding the molecular mechanisms of CRISPR toolbox using single molecule approaches. ACS Chemical Biology.

## 1.1 ABSTRACT

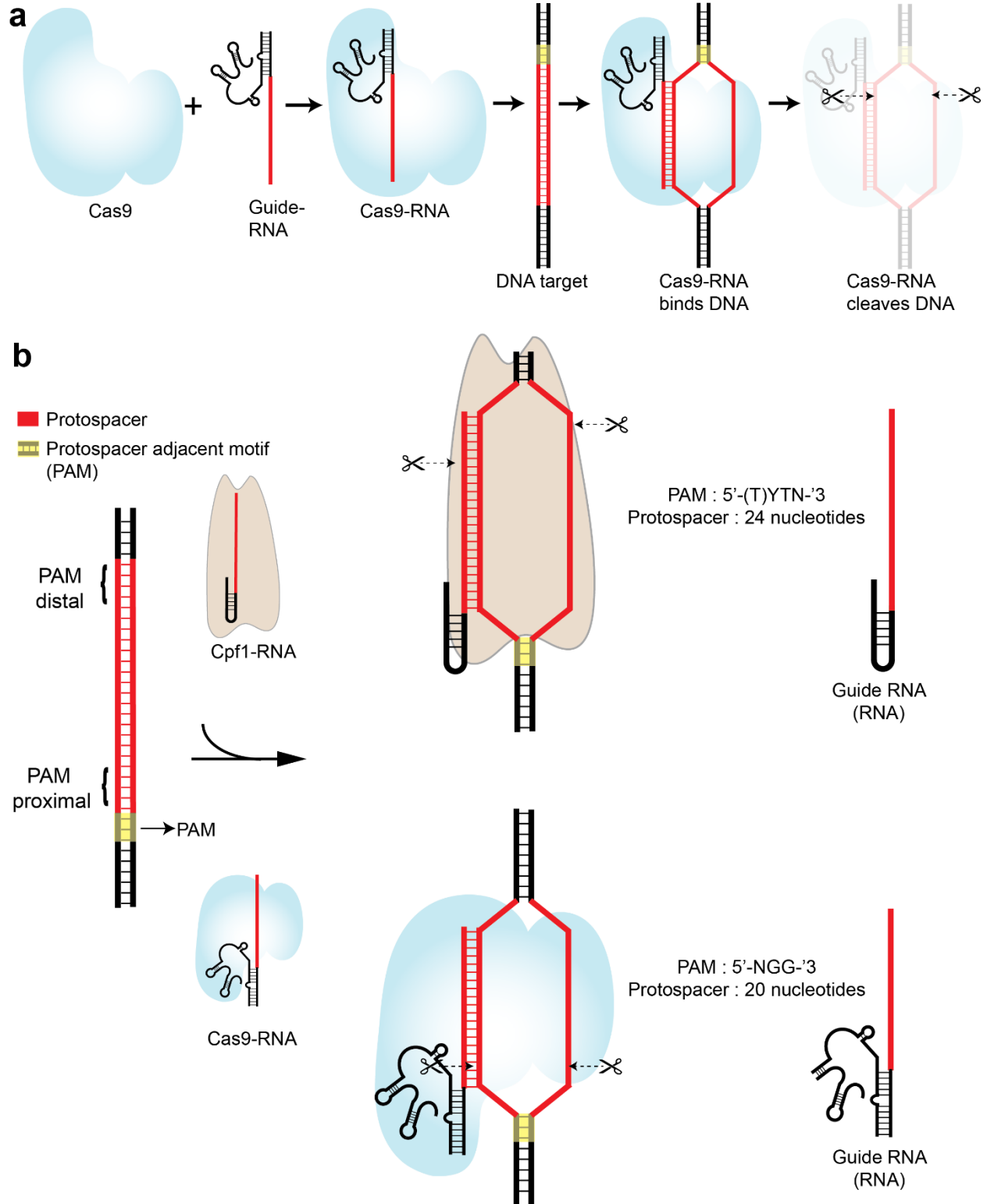
The adaptive immunity against foreign genetic elements conferred by CRISPR systems in microbial species has been repurposed as a revolutionary technology for wide-ranging biological applications, chiefly genome engineering. Biochemical, structural, genetic and genomics studies have revealed important insights into their function and mechanisms, but most ensemble studies cannot observe structural changes of these molecules during their function and are often blind to key reaction intermediates. Here I discuss how single molecule technologies can help us understand the molecular mechanism of CRISPR tool box.

## 1.2 INTRODUCTION to CRISPR

In certain bacteria and archaea, CRISPR (clustered regularly interspaced short palindromic repeats)–Cas systems form an adaptive defense against attacks by foreign genetic elements such as viruses<sup>1</sup>. The system functions by storing memory of attacks by acquiring sequences of the foreign genetic elements into host genome in regions known as CRISPR loci<sup>2</sup>. Short RNA transcripts (Guide-RNA) from CRISPR loci forms a complex with the CRISPR proteins (CRISPR-RNA)<sup>3</sup>. During future invasions, the CRISPR-RNA is directed by the guide-RNA to target foreign genetic elements for its nucleolytic impairment, chiefly by the virtue of base-pairing between the guide-RNA and DNA sequences (DNA target) in foreign genetic elements<sup>4,5</sup>. Region in DNA target spanning the canonical base-pairing between the guide-RNA and DNA target is called the protospacer. The single most important requirement of such DNA targeting is that the protospacer be followed by a special motif called PAM (protospacer adjacent motif)<sup>5,6</sup>. Sequences in protospacer closest to PAM are referred to as being PAM-proximal whereas those farthest away from PAM are PAM-distal. Throughout this dissertation, the base pairs in the protospacer that are complementary to the guide-RNA of CRISPR-RNA are referred to as matches whereas the others are called mismatches.  $n_{PD}$  (the number of PAM-distal mismatches) and  $n_{PP}$  (the number of PAM-proximal

mismatches) are used to denote the location and extent of mismatches.

Abilities to program a nuclease to induce a cut at a desired genomic site and to program a nucleolytically dead CRISPR fused with a marker or an effector to bind any genomic site has revolutionized biology<sup>7,8</sup>. Minimalistic versions of CRISPR systems are employed for such applications because they provide the modularity and ease of programming. Class II of CRISPR is one example as it uses a single CRISPR endonuclease in complex with guide-RNA to target DNA<sup>9</sup>. The most commonly used type is SpCas9 (Cas9 from *Streptococcus Pyogenes*), followed by SaCas9 from *Staphylococcus Aureus*<sup>10</sup>, and Cpf1 variants; AsCpf1 from *Acidaminococcus sp* and LbCpf1 from *Lachnospiraceae bacterium*<sup>11</sup>. In many other CRISPR systems, multiple CRISPR proteins form a Cascade complex with the guide-RNA to target DNA and then recruit Cas3 for DNA degradation<sup>9,12,13</sup>. The protospacer length of Cas9 and Cpf1 ranges from 20-24 base-pair (bp) and the PAM is 3-6 nucleotides (nt) long. Cascades have ~33 bp protospacer and 2-3 nt PAM. Cas9 from *Streptococcus Pyogenes* is the most widely used Cas9 and is thus the default Cas9 referenced here unless stated otherwise<sup>7,8</sup>.



**Figure 1.1 | Schematic of DNA targeting by CRISPR enzymes commonly used in genome engineering.**

(a) Complexation of Cas9 with guide-RNA results in Cas9-RNA which binds and cleaves DNA complementary to the sequence in guide-RNA (shown by red; protospacer) provided that the protospacer is adjacent to a protospacer adjacent motif (PAM). (b) Differences between Cpf1 and Cas9.

### 1.3 FLUORESCENCE and FORSTER RESONANCE ENERGY TRANSFER.

Energy in a molecule can be stored as potential energy or kinetic energy, in a variety of ways including and some of these can be due to: Rotational energy associated with rotational tumbling motion of molecules, vibration energy associated with oscillations of atoms within the molecule. Electronic energy, resulting from the potential energy of electrons in different electronic configurations/levels. These listed energies are quantized, meaning that they can only occupy discrete values/levels (Figure 1.2a) and investigation of transitions between these levels is called Spectroscopy and can be used to obtain useful information about the nature and environment of any molecule of interest.

One of the most commonly used techniques in spectroscopy is called Fluorescence which primarily involves the investigation of transitions between different electronic levels of a molecule. The fundamental principle of fluorescence is best described in a diagram known as the Jablonski diagram (Figure 1.2b). A molecule in a ground state electronic configuration, can be irradiated with a flash of energy (for example, laser). And if the photons from the energy source has energy of quanta equal to difference between ground state ( $S_0$ ) and a first excited electronic state ( $S_1$ ), the molecule can absorb these photons and jump to a higher energy excited electronic state ( $S_1$ ). The absorbance occurs on the order of  $10^{-15}$  seconds. As shown before in Figure 1.2b, the electronic energy levels are interspersed with the vibrational energy levels. One of the most common ways for the molecule to return to the ground state is to dissipate part of its absorbed energy via relaxation to the lowest vibrational energy level around  $S_1$ . From there, the molecule can relax back to  $S_0$ , while emitting a photon corresponding to the energy difference between these two levels. This process takes place on the order of



( $10^{-9}$ - $10^{-7}$  seconds) and is called as Fluorescence. Following this, the molecule can dissipate more of its energy to vibrational energy levels around  $S_0$  and finally relax to the lowest energy state of the  $S_0$  state. The transitions to lower vibrational energy levels is non-radiative i.e. no photons are emitted and energy is released in form of heat. This process is known as the relaxation and loss of part of energy originally obtained from the absorption of photon, results in photon emitted in fluorescence to have lower energy and thus higher wave-length. This shift in the higher value of wavelength is known as the Stokes shift.

There are many other pathways by which the molecule can relax back to the lowest energy state of the ground state ( $S_0$ ). Some of which are :

1. Collisional Quenching: Molecule in the excited state ( $S_1$ ) may collide with another molecule and transfer the energy in a non-radiative way.
2. Electron in the excited state ( $S_1$ ) of the molecule may undergo a spin transition and lead to conversion from singlet excited state  $S_1$  state to an excited triplet state ( $T_1$ ). Once in  $T_1$ , the molecule may then relax back to levels around  $S_0$ , in a radiative way emitting photons in the process, which is known as Phosphorescence.
3. Molecules have a net dipole. And dipole of fluorescent molecule (termed Donor) in excited energy state may get coupled in an interaction with the dipole of another fluorescent molecule (termed acceptor). This coupling results in a non-radiative energy transfer from donor to acceptor. The transferred energy in the acceptor (now in its excited energy state) relaxes back to its own ground state in a radiative way i.e. by emitting photons. This process of energy transfer is known as **Förster (or Fluorescence) Resonance Energy Transfer (FRET)**.

The efficiency of energy transfer, also known as FRET efficiency ( $E$ ), depends on:

- (1) Physical distance between the donor and the acceptor ( $r$ ).
- (2) Spectral overlap between the donor emission and acceptor absorption ( $J$ ).
- (3) Relative orientation of the dipoles of donor and acceptor molecules ( $\kappa$ ).
- (4) Quantum efficiency of the donor molecule ( $Q_d$ ).
- (5) Refractive index of the medium ( $n$ ).

And they are related by the following equations:

$$E = \frac{1}{1 + \left(\frac{r}{R_0}\right)^6}$$

**Equation 1.1**

$$R_0^6 = \frac{9 \ln 10}{128\pi^5 N_A} \frac{\kappa^2 Q_d}{n^4} J$$

**Equation 1.2**

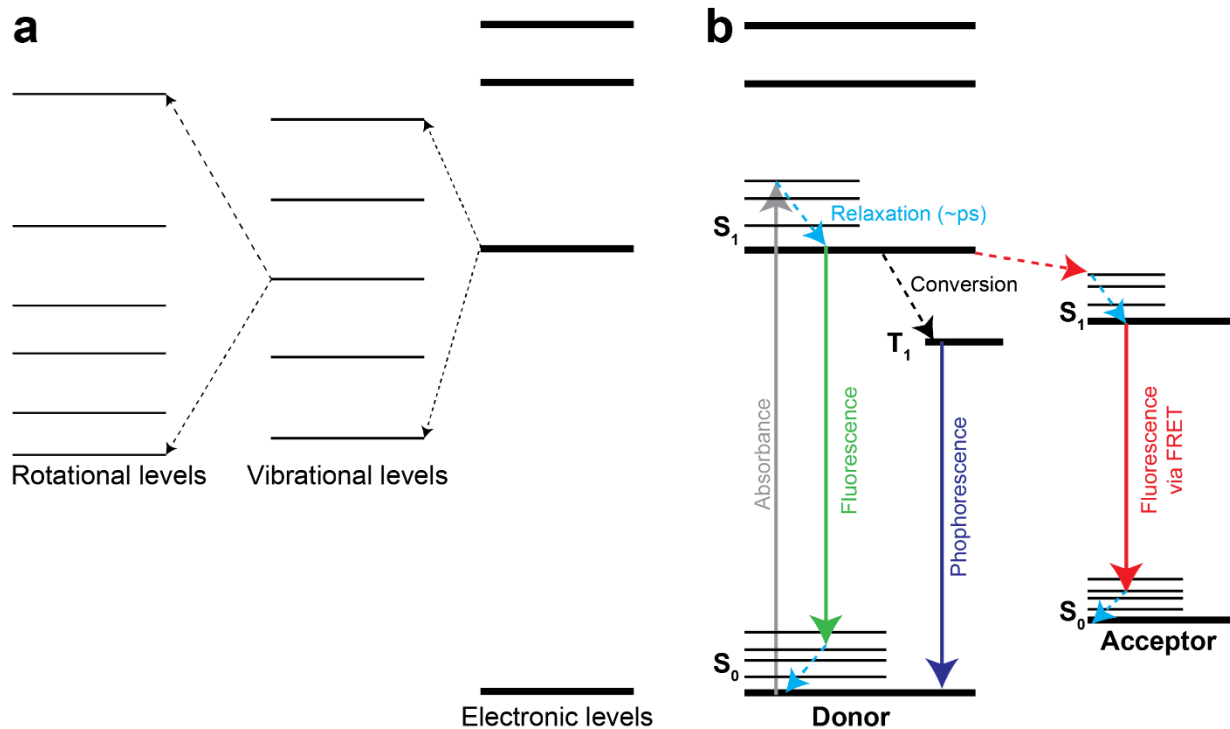
Where  $R_0$  is the distance at which the FRET efficiency is 50% (0.5), and  $N_A$  is Avogadro's number.

For practical applications, E can be simply estimated as

$$E = \frac{I_A}{I_D + I_A}$$

**Equation 1.3**

Where  $I_A$  is the intensity of the fluorescence from the acceptor (obtained via FRET) and  $I_D$  is the intensity of the fluorescence from the donor (fluorescence remaining after a fraction of it is FRET transferred to the acceptor).



**Figure 1.2 Molecular energy levels and transitions between them for fluorescence and FRET spectroscopy.**

(a) Quantized energy levels of a molecule owing to rotational and vibrational motion plus electronic states. (b) Transitions between different energy levels, giving rise to phenomenon of absorbance, fluorescence, phosphorescence and FRET.

#### 1.4 INTRODUCTION to SINGLE-MOLECULE TECHNOLOGIES.

As the name suggests, single-molecule techniques are those that enable investigation of properties of individual molecules as opposed to bulk techniques where readouts/measurements result from averaging of properties of multiple molecules. For the longest time, the biggest hurdle for single-molecule techniques was the ability to detect signal from a desired single molecule i.e. the signal was too weak, especially in a biological system. But since late 90s, advent of powerful cameras, new microscopy techniques, development and characterization of extremely bright dyes (fluorophores) have enabled detection from single molecules in biological systems. Why single-molecule?

### 1. Capture heterogeneity of a biological system:

Biological systems are extremely complex involving interaction between many components for e.g. DNA/RNA, proteins and metabolites. And even the weakest interactions can have critical outcomes. Therefore, to understand the system as a whole, it is imperative that we take all different interactions into account. Bulk measurements can only report on average of these interactions thus masking their intrinsic heterogeneity. Single-molecule measurements allow us to investigate each interaction separately thus helping us to unmask the intrinsic heterogeneity of a system.

### 2. Analyze molecular mechanism of important processes:

Biological processes are a combination of many stochastic events i.e. there is no synchronization between behaviors of multiple DNA/RNA or protein molecules (biomolecules) that make up the biological process. This makes investigating the process and its steps very difficult via bulk measurements, because they only report on the averaging of the multiple stochastic events. For e.g. if we relied solely on score lines to judge tennis players, we could make a fairly good assessment that Roger Federer is a great tennis player. But only the score lines obscure many critical aspects of his greatness, like his technique, foot work etc. which are critical information because they not only serve entertainment aspect of tennis, but is also used as a model for upcoming players and overall development of the sport. In a way, the scoreline can be thought of as a read of a bulk measurement i.e. average of multiple different shots (events) that Federer hit, with no information about quality and nature of the shots.

If we can watch single molecule(s) of that entire process or a system, we can understand how various parts of the system work individually and combinatorically. But capturing single biomolecules directly in action is almost impossible because their sizes are extremely small and are thus beyond fundamental optical limits of most microscopes. So single molecule detection for biological systems is typically achieved by attaching an extremely bright fluorescent dye on the biomolecule of interest which makes it 'fluorescently visualizable' as it emits light of a certain wavelength when illuminated with a light source

(laser) of particular wavelength (color). Apart from just being able to localize and track motion of the biomolecule via the fluorescent dye attached to it, there are many photophysical properties of these dyes that can be utilized to serve variety of purposes. One such example is FRET, which involves transfer of fluorescent 'energy' from one dye (donor) molecule to other (acceptor) molecule. So if a donor dye molecule is being illuminated with a bright laser for its fluorescent visualization via its emission of a certain wavelength or color. A part of its 'energy' can end up being donated to an acceptor dye molecule in its vicinity, which would emit the 'donated energy' via its emission of a different wavelength or color. The transfer of energy is dependent on distance between the dye molecules, so intensity and wavelength of the emission from these two dye molecules can act as a reporter of distance between them.

Implementation of FRET technique in a single-molecule setup is commonly referred to as single-molecule FRET (smFRET), which can be adapted to investigate any biological processes and its steps.

For e.g. if Federer had been extremely small like a protein molecule. Then we could watch the internal shape changes of the protein/Federer as it does its job/play tennis by attaching a donor and an acceptor molecule to specific locations (for e.g. hands of Federer) and illuminate them with a laser light of certain wavelength. The real-time output emission from the protein/Federer (donor + acceptor emission) will report on real-time changes in distance between two specific points in protein/ hands of Federer. This information will help you understand and analyze the molecular mechanism of the protein/Federer.

### 3. Localization of single molecules for super resolution imaging:

Biological structures (cells, tissue etc.) are made up of thousands of constituent biomolecules (protein, DNA, RNA etc.). Visualization and localization of each of the constituent biomolecules, *in singulo*, can be used to build up the entire biological structure giving us a high quality resolution image of it. Design and implementation of this idea was one of the subjects of 2014 Nobel prize in Chemistry.

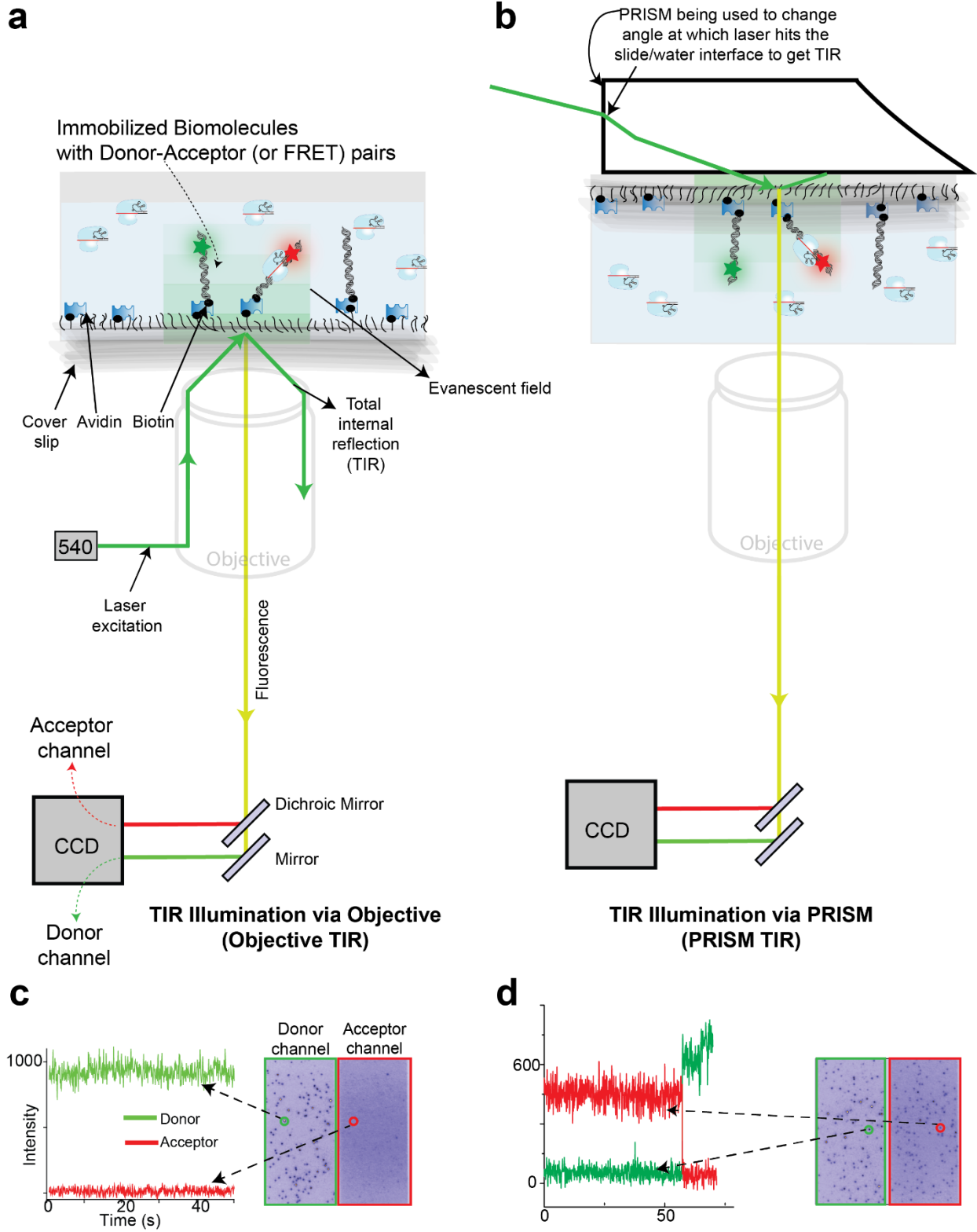
#### 4. Adaptation of single molecule technique in sequencing and other applications:

If multiple single molecules can be detected simultaneously, then multiple reactions can be performed simultaneously and investigated separately. This idea has been most successfully implemented in DNA sequencing applications. Previous methods of DNA sequencing relied on series of individual experiments that had to be painstakingly performed separately for each specific region in genome to be sequenced. Data was then collected from multiple such individual experiments for final analysis. But the ability to detect readout from single (or few) DNA molecule sequencing experiments, performed in parallel but analyzed in separate, has revolutionized biology. The point I want to emphasize here is that the ability of massively parallel single-molecule detection can be repurposed to separately analyze large number of reactions performed in parallel.

#### **1.5 SINGLE MOLECULE FRET (smFRET) and ITS IMPLEMENTATION**

One of the most commonly employed optical method of single molecule fluorescent detection is the Total internal reflection (TIR) microscopy. In it, an incident light is irradiated at the cover slip/slide and solution interface at an angle that would cause the phenomenon of total internal reflection instead of reflection. The total internal reflection creates a strong evanescent field, whose depth is limited to <100 nm into the solution. Thus only a small subset of molecules which lie within this evanescent volume will absorb incident photons and emit fluorescence or FRET depending on the presence and location of donors and acceptors. This strategy drastically improves the signal to noise as fluorescence from molecules outside the plane of interest is abolished since only molecules within 100 nm distance get excited.

One of the most common practices is to immobilize the biomolecules of interest (labeled with donor and/or acceptor) on the surface of the cover slip or the slide using Avidin-Biotin linkage which is one of the tightest non-covalent linkages). The most common optical strategies of achieving TIR and capturing subsequent fluorescence emission from single molecules is described in Figure 1.3.



**Figure 1.3 | Common strategies for single molecule fluorescence.**

**(a-d)** Biomolecules of interest, immobilized on a cover slip, are irradiated with a green laser (that can excite a particular donor molecule). The angle of incidence is managed by modulating the angles at which the laser beam enters the objective and hits the coverslip. Upon reaching the critical angle, the laser beam undergoes total internal reflection and creates a strong but limited evanescent field. The molecules within the evanescent field get excited by the energy of the evanescent field and undergo fluorescence or FRET depending on the geometry of donor and acceptor molecules within the biomolecules or biomolecular complexes. The emission from all the molecules within the evanescent field is captured by the objective. The total emission is then spectrally split into donor fluorescence and acceptor fluorescence (via FRET) and projected as two separate images (channels) of the same set of molecules at any given time point. The single fluorescent spots ideally represent the fluorescence from single biomolecules or biomolecular complexes. The intensity as a function of time of any spot in the donor channel ( $I_D(t)$ ) and acceptor channel ( $I_A(t)$ ) can be used to evaluate the changing FRET efficiency ( $E$ ) of the given single biomolecule as a function of time via Equation 1.3 we discussed before. Different values of  $E$  can report on different conformational states of a biomolecule or interaction between two biomolecules, depending on the labeling geometry of donor and acceptor. **(a)** and **(b)** are two primary strategies of achieving TIR. **(c)** and **(d)** are representative cases of **(c)** zero FRET and **(d)**  $E > 0$  FRET.



## **Chapter 2.**

### **Real-time observation of DNA recognition and rejection by the RNA-guided nuclease Cas9**

\*Contents of this chapter has been published and is available at:

Singh, D., Sternberg, S. H., Fei, J., Doudna, J. A. & Ha, T. Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. *Nature communications* **7**, 12778, doi:10.1038/ncomms12778 (2016).

## 2.1 ABSTRACT

Binding specificity of Cas9-guide RNA complexes to DNA is important for genome engineering applications, but how mismatches influence target recognition/rejection kinetics is not well-understood. We used single-molecule FRET to probe real-time interactions between Cas9-RNA and DNA targets. Bimolecular association rate is only weakly dependent on sequence, but the dissociation rate greatly increases from  $< 0.006 \text{ s}^{-1}$  to  $> 2 \text{ s}^{-1}$  upon introduction of mismatches proximal to protospacer adjacent motif (PAM), demonstrating that mismatches encountered early during heteroduplex formation induce rapid rejection of off-target DNA. In contrast, PAM-distal mismatches up to 11 base pairs in length, which prevent DNA cleavage, still allow formation of a stable complex (dissociation rate  $< 0.006 \text{ s}^{-1}$ ), suggesting that extremely slow rejection could sequester Cas9-RNA, increasing Cas9 expression level necessary for genome-editing thereby aggravating off-target effects. We also observed at least two different bound FRET states that may represent distinct steps in target-search and proofreading.

## 2.2 INTRODUCTION

CRISPR (clustered regularly interspaced short palindromic repeats)–Cas systems provide adaptive immunity against foreign genetic elements in bacteria and archaea<sup>14</sup>. In type II CRISPR-Cas systems, the Cas9 endonuclease functions together with a dual-guide RNA comprising CRISPR RNA (crRNA) and trans-activating crRNA (tracrRNA) to target 20 base pair (bp) DNA sequences (cognate sequence) for double-stranded cleavage<sup>15</sup>. Efficient targeting requires RNA-DNA complementarity as well as a specific motif flanking the target sequence called the PAM (protospacer adjacent motif, 5'-NGG-3' for *S. pyogenes* Cas9)<sup>15-17</sup>. Cas9-RNA complexes have proven to be extremely versatile tools for genome engineering applications<sup>18</sup>, and minimizing off-target effects<sup>19,20</sup> remains an active area of study.

Numerous studies have assessed off-target DNA binding and cleavage by the Cas9 RNA complex, both *in vitro* and *in vivo*<sup>15,17,21-56</sup>. While subtly different conclusions have been reached depending on the exact

method of analysis, these studies agreed about specificity being heavily influenced by the presence of a PAM, a 7-12 bp long seed sequence proximal to the PAM, and the concentration of Cas9 and guide-RNA. Most of previous studies lacked dynamic information on DNA targeting, yet in order to improve the efficacy of processing only the correct targets, we need such information on targeting dynamics. Single molecule methods are ideal for this task because they can detect wide-ranging interactions (transient to long-lived) and identify multiple states in real time<sup>57</sup>. Moreover, they can be used to obtain the dynamic information on short and specific DNA sequences, thus enabling the sequence specific estimation of various kinetic parameters<sup>58-60</sup>. Several single molecule studies have examined sequence specificity in CRISPR targeting<sup>53-56,61,62</sup>. Here, we report a systematic investigation of the binding and dissociation kinetics of Cas9-RNA as a function of sequence mismatches to determine how quickly cognate sequence is recognized and how quickly partially matching sequences are rejected.

## 2.3 RESULTS

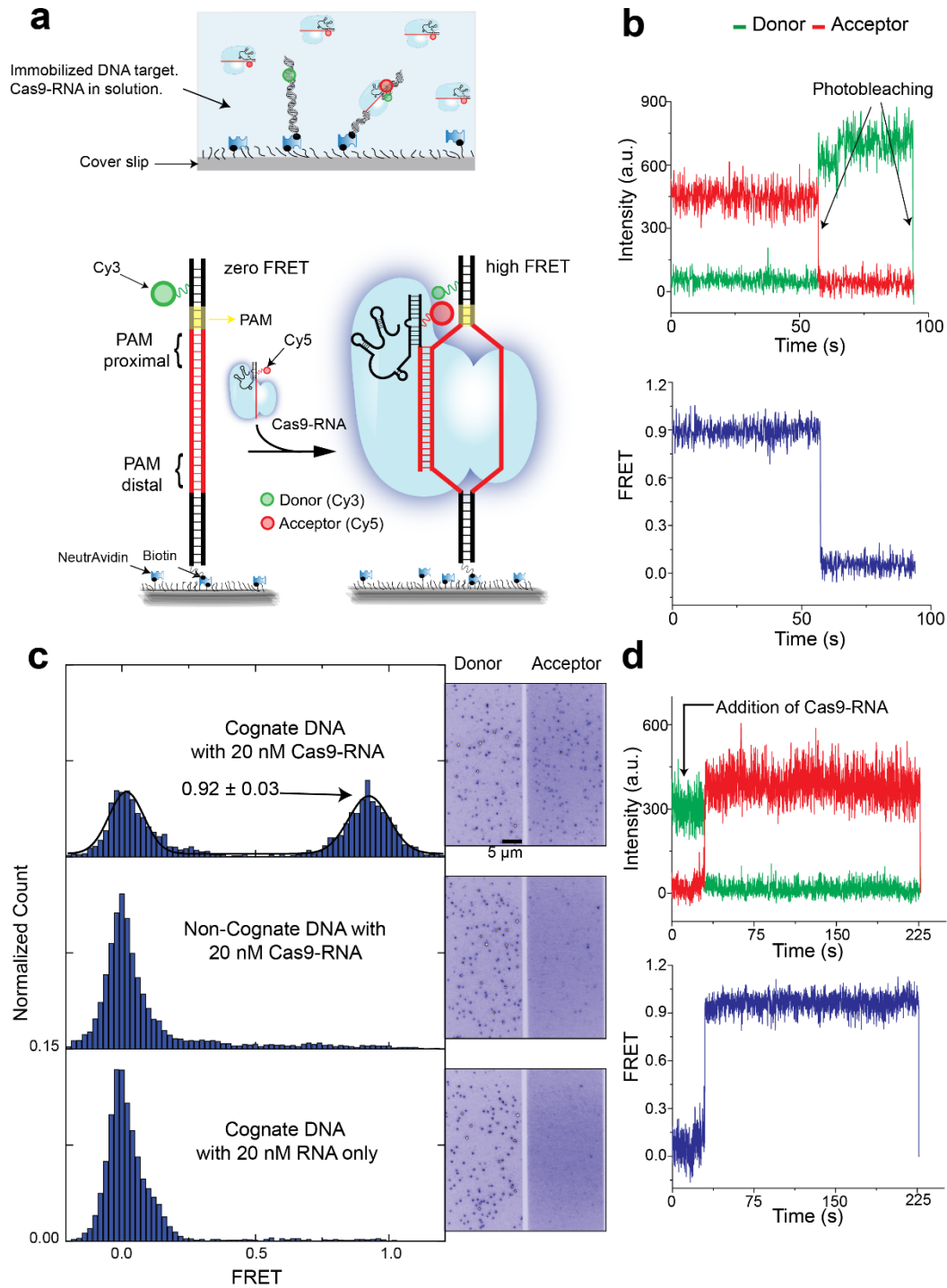
### 2.3.1 Real-time single molecule assay for DNA interrogation by Cas9-RNA

We used single-molecule fluorescence resonance energy transfer<sup>63,64</sup> (smFRET) to directly observe individual Cas9-RNA complexes binding to DNA targets in real time. Donor (Cy3) and acceptor (Cy5) fluorophores were conjugated to modified nucleotides in the DNA target and crRNA, respectively, so that FRET between them would report on Cas9-RNA binding to the DNA (Figure 2.1a and Figure 2.2). Fluorescence labeling did not compromise target cleavage (Figure 2.3). After introducing 20 nM Cas9-RNA complexes to cognate DNA target molecules immobilized on passivated microscope slides, two distinct populations were observed centered at FRET = 0.92 and 0, respectively (Figure 2.1b, c). The labeling sites are separated by 30 Å<sup>65</sup> (Figure 2.2), consistent with the observation of the high FRET value upon Cas9-RNA binding. In control experiments using a non-cognate (fully mismatched) DNA target with PAM (Table 2.1), or guide-RNA without Cas9, the 0.92 FRET state was not observed (Figure 2.1c). Therefore, we assigned the 0.92 FRET state to a stably formed Cas9-RNA-DNA complex. The high

FRET state was long-lived, with a lifetime ( $>3$  min) limited only by fluorophore photobleaching (Figure 2.4d). A catalytically dead Cas9 mutant (dCas9; D10A/H840A mutations<sup>15,16</sup>) showed signal indistinguishable from active Cas9 (Figure 2.4), indicating that DNA products remain tightly bound after cleavage as was observed previously<sup>17</sup> (Figure 2.3). To capture the moment of binding, we added Cas9-RNA into the flow cell during data acquisition. FRET efficiency increased from 0 to 0.92 in a single step (Figure 2.1d), suggesting that any intermediates on-path to target binding, if present, cannot be resolved at the time resolution of our measurements (0.1 s).

### **2.3.2 Effect of DNA target mismatches on Cas9-RNA binding**

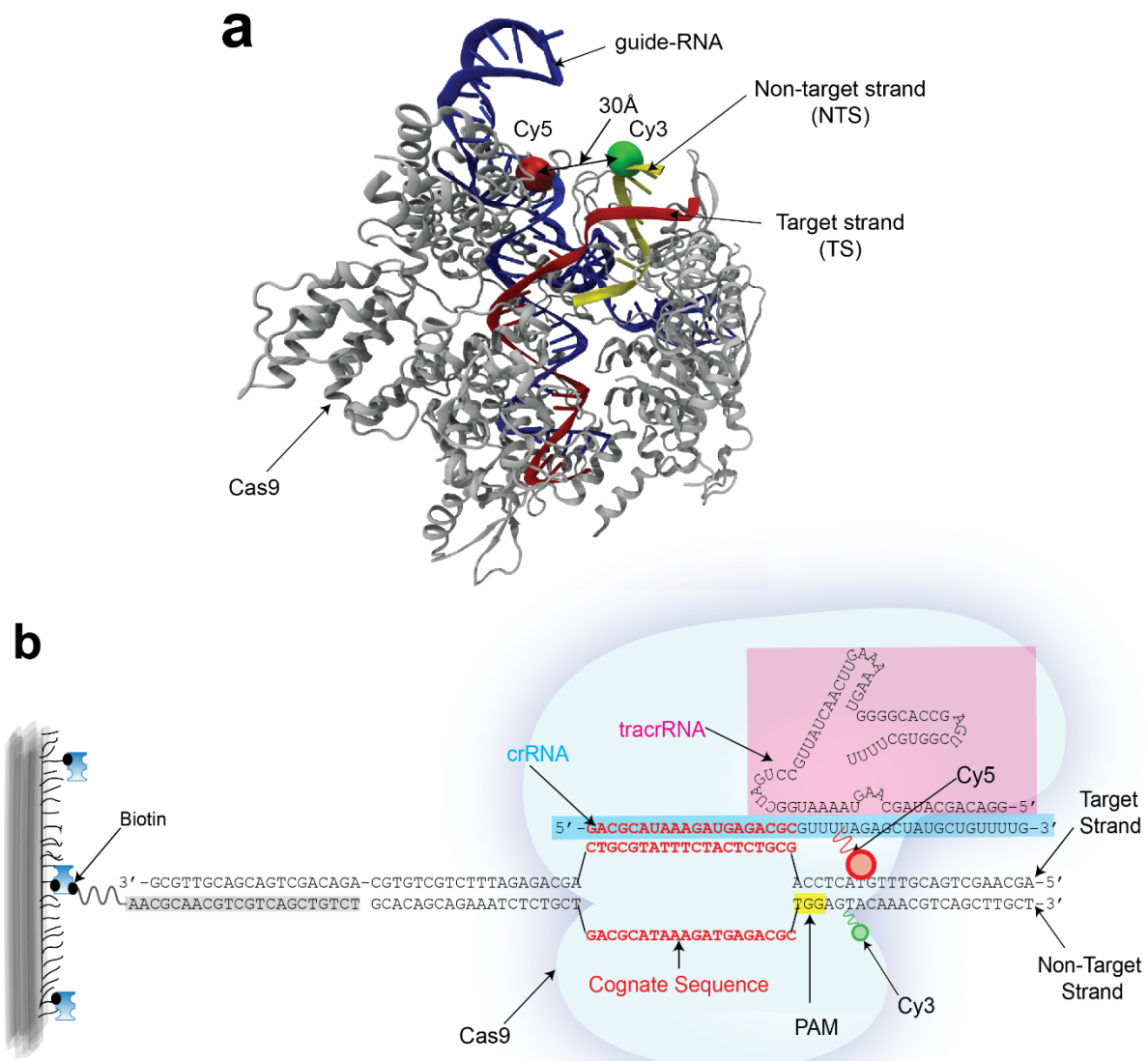
We next examined how DNA targets with imperfect RNA-DNA complementarity are discriminated against and rejected by Cas9-RNA. We prepared a series of donor-labeled, fully duplexed DNA containing mismatches relative to the guide-RNA (Table 2.1 and Figure 2.5a). The mismatches were introduced either from the PAM proximal side or from the PAM distal side, and are denoted using the naming convention  $x$ - $y_{\text{mm}}$  where  $x^{\text{th}}$  through  $y^{\text{th}}$  base pairs are mismatched. The fraction of DNA bound by Cas9-RNA (ratio between counts with FRET  $> 0.75$  and total counts in FRET histograms) remained identical to the cognate DNA up to 12 PAM-distal mismatches (17-20<sub>mm</sub>, 13-20<sub>mm</sub>, 12-20<sub>mm</sub>, 11-20<sub>mm</sub>, 10-20<sub>mm</sub>, 9-20<sub>mm</sub>) (Figure 2.5b, d). The bound state remained stable, with the observed lifetimes limited only by fluorophore photobleaching (Figure 2.4d). A large decrease in the bound fraction occurred only when the number of mismatches from the distal end exceeded 13 bp (7-20<sub>mm</sub>, 6-20<sub>mm</sub>, 5-20<sub>mm</sub>), corresponding to less than 7 matched bp from the PAM-proximal end. In contrast, even 2 bp mismatches from the PAM-proximal end (1-2<sub>mm</sub>) were deleterious for Cas9-RNA binding and binding to 4 bp PAM-proximal mismatches (1-4<sub>mm</sub>) was indistinguishable from binding to fully mismatched (1-20<sub>mm</sub>), underscoring the importance of the PAM-proximal seed region (Figure 2.5c, d).



**Figure 2.1 | Cas9-RNA binding to a cognate sequence.**

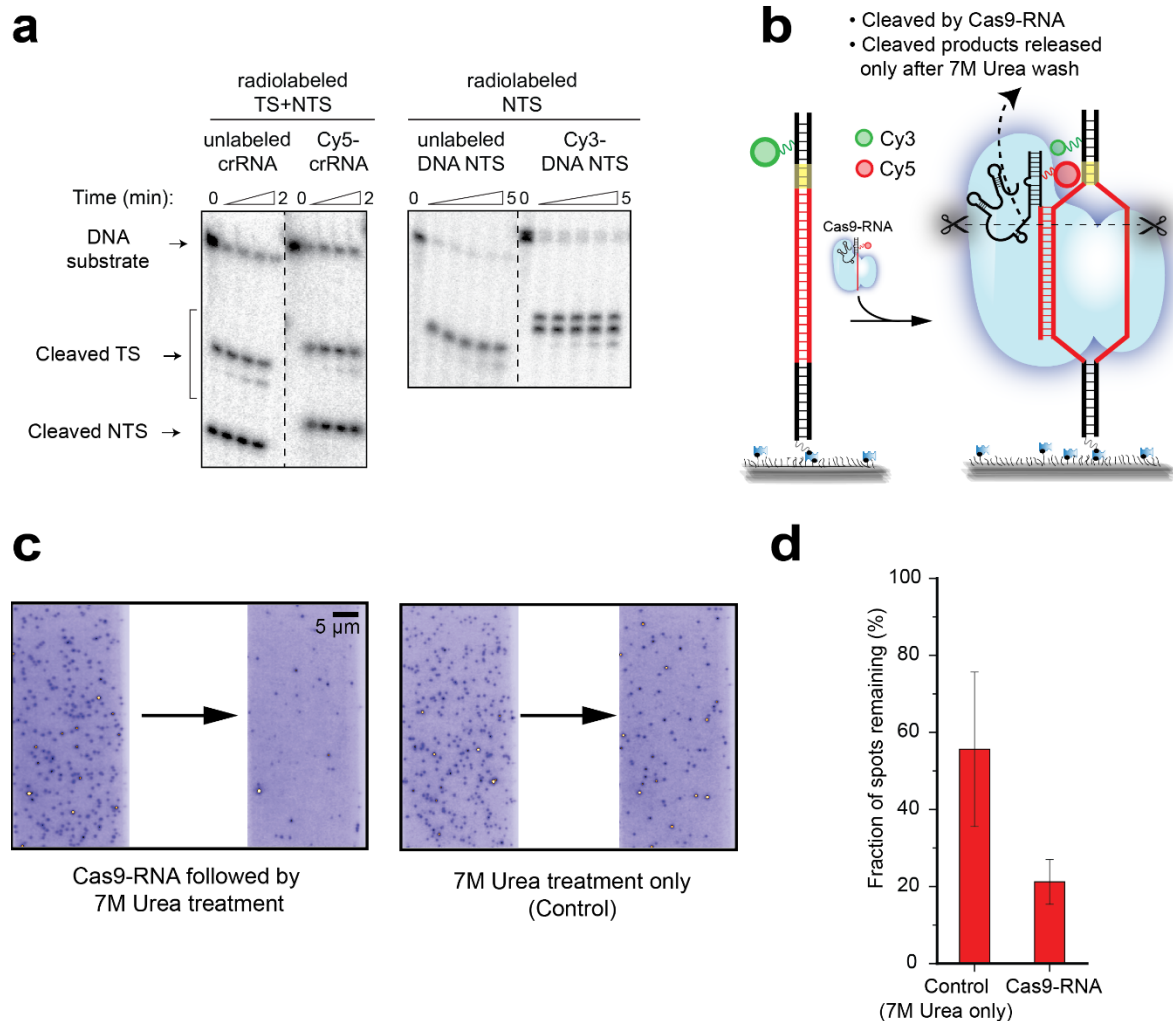
**(a)** Schematic of single-molecule FRET assay. High FRET signal resulted when Cas9 in complex with an acceptor (Cy5)-labeled guide RNA (Cas9-RNA) bound a surface-immobilized, donor (Cy3)-labeled target DNA that contains the cognate sequence (red DNA segment) and PAM (yellow segment). **(b)** A

representative smFRET time trajectory of a stably bound Cas9-RNA in the presence of 20 nM Cas9-RNA in solution. (c) FRET histograms obtained with cognate DNA (top) and negative controls with a non-cognate DNA (middle) and with RNA only (without Cas9) (bottom). The number of molecules included ranged from 568 to 1,314. Corresponding images of donor and acceptor channels are shown. (d) A representative smFRET time trajectory of real-time binding of Cas9-RNA in a single step after 20 nM Cas9-RNA is added at the time point indicated.



**Figure 2.2 | FRET probe labeling locations in the Cas9-RNA-DNA complex.**

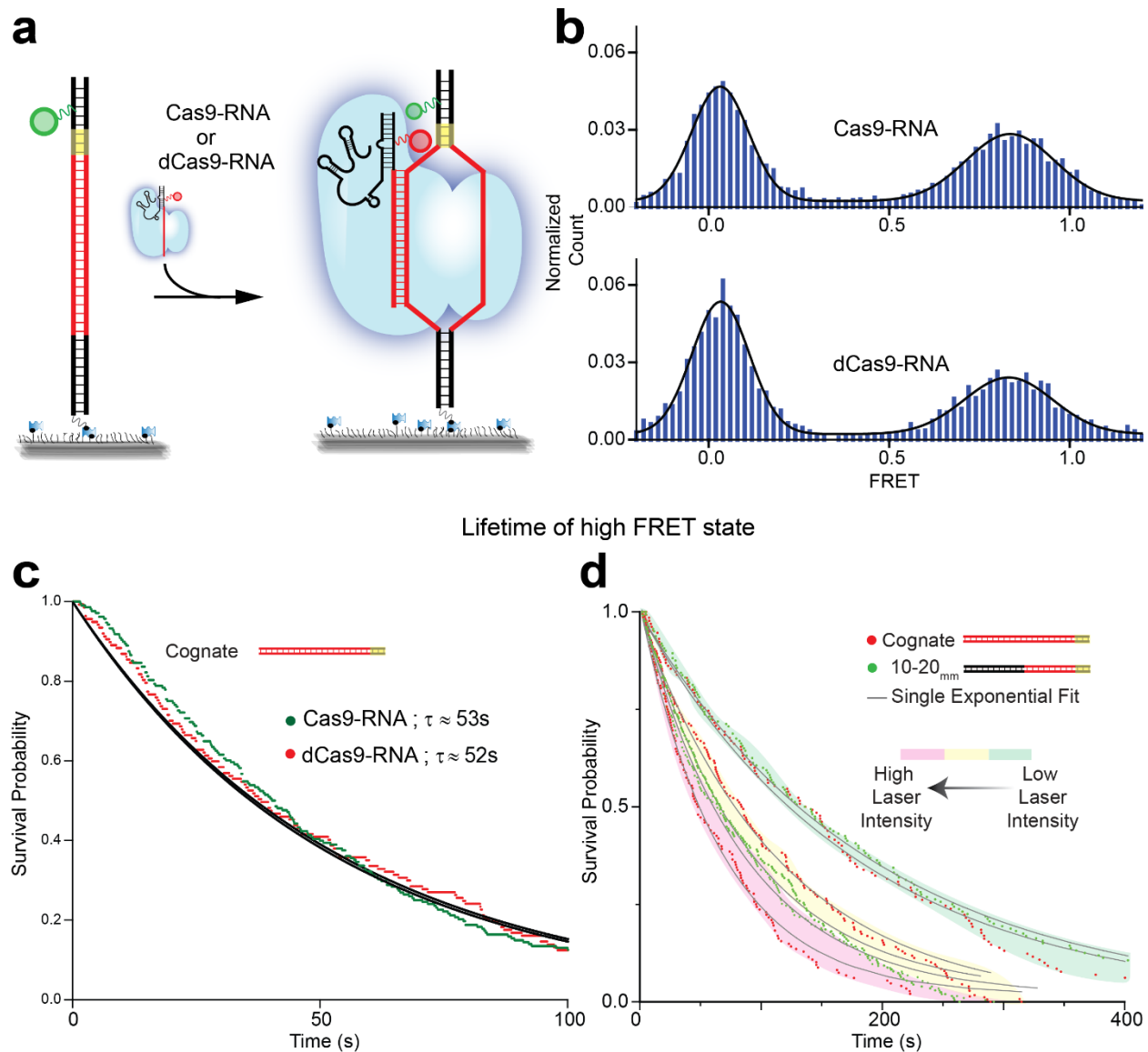
**(a)** Cy3 and Cy5 labeling locations shown in the crystal structure of Cas9-RNA bound to a cognate DNA target (PDB ID: 4UN3)<sup>65</sup>. The strand hybridized with the guide RNA to form the RNA-DNA heteroduplex is referred to as the target strand while the other strand, containing the PAM (5'-NGG-3'), is the non-target strand. **(b)** Schematic of a bound Cas9-RNA-DNA complex showing the base pairing between different components. The sequences shown in red denote the cognate sequence of the DNA target and the complementary guide sequence of the crRNA. The DNA sequence highlighted in light gray is a separate 22 nucleotide-long biotinylated adaptor used for surface immobilization of DNA target molecules.



**Figure 2.3 | Target cleavage activity is not impaired by fluorescent labeling of guide RNA or DNA target.**

**(a)** DNA cleavage assays using fluorescently labeled components, analyzed by denaturing PAGE. TS, target strand; NTS, non-target strand. Note that Cy3-DNA exhibits retarded gel mobility compared to unlabeled DNA, giving rise to two cleavage product bands and two substrate bands due to incomplete labeling. **(b)** Single-molecule assay to monitor DNA cleavage of the cognate DNA target. Because Cas9-RNA retains high-affinity binding to DNA cleavage products, 7 M urea was required to release Cas9-RNA and the DNA product containing Cy3<sup>17</sup>. **(c-d)** Disappearance of Cy3 spots resulted from Cas9-RNA mediated DNA cleavage, as shown by **(c)** fluorescence images and **(d)** quantification of Cy3 spot counts. About 20% of spots remained after Cas9-RNA reaction followed by urea wash compared to more than 50% in the control with urea treatment only. [Cas9-RNA] = 1  $\mu$ M. Error bars represent s.d. for  $n = 3$ .

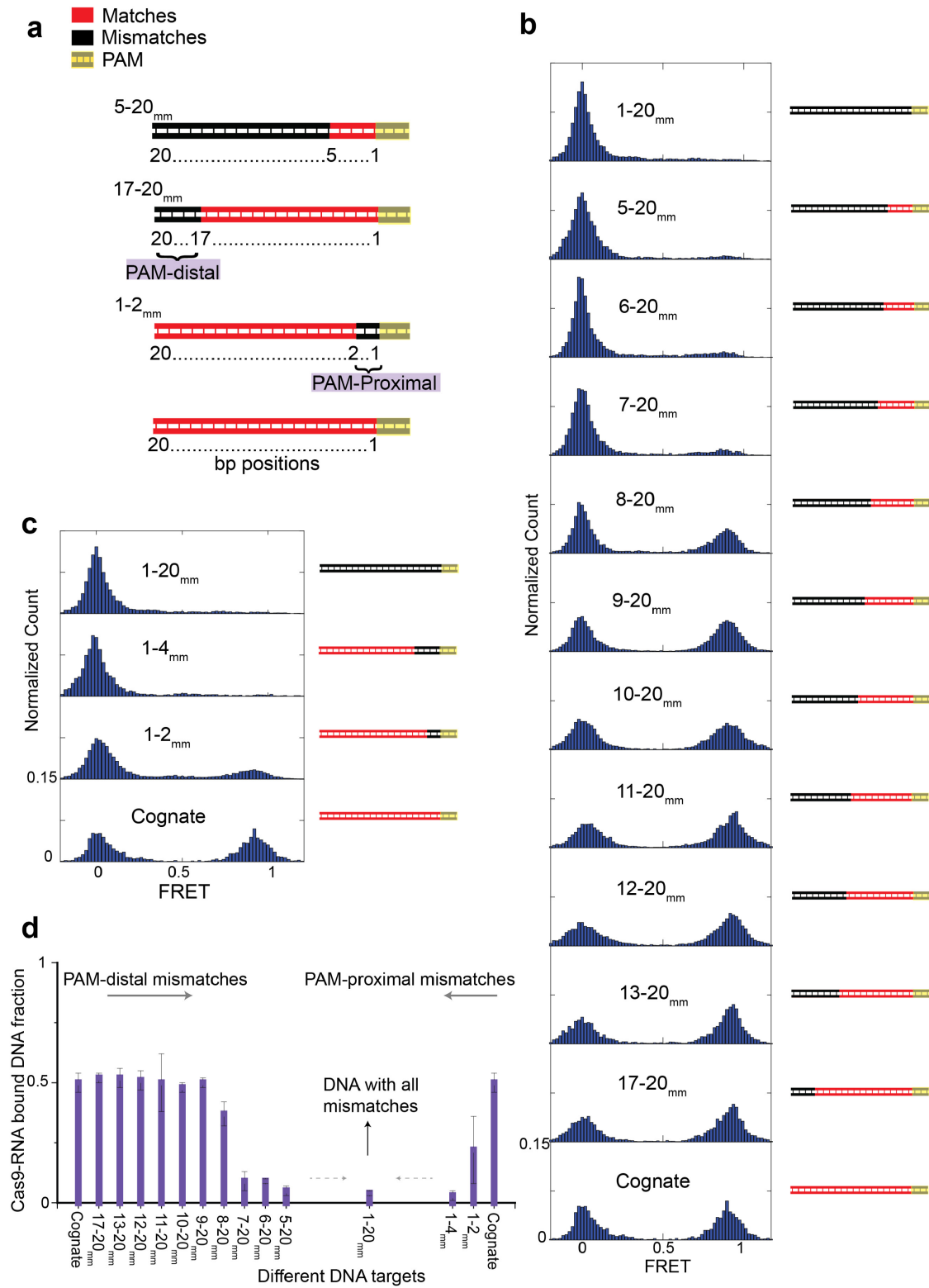




**Figure 2.4 | Comparison between Cas9 and dCas9, and high stability of Cas9-RNA-DNA for certain DNA targets.**

**(a)** Cas9 or dCas9 were assembled with acceptor-labeled guide RNA to monitor binding to a donor-labeled cognate DNA target. **(b-c)** Cas9 and dCas9 were indistinguishable in our binding assay, as shown by the similarity of **(b)** FRET histograms at 20 nM Cas9-RNA/dCas9-RNA and **(c)** lifetimes of the high FRET state at 10 nM Cas9-RNA/dCas9-RNA, fit with a single exponential decay. The number of molecules used for the FRET histograms was 346 (Cas9-RNA) and 297 (dCas9-RNA). **(d)** DNA targets with contiguous complementarity larger than 8 bps in the PAM-proximal end of the protospacer showed

long-lived binding as shown in (c). To confirm that the slow decay was caused by photobleaching instead of Cas9-RNA dissociation, we monitored the loss of signal at different laser intensities. As expected, the apparent lifetime of the high FRET state increased with decreasing laser intensity, reaching 3.3 min for the lowest intensity tested. [Cas9-RNA] = 20nM.



**(a)** A series of fully-duplexed DNA targets with a varying number of mismatches (blue segments) relative to the guide RNA. An  $x$ - $y_{\text{mm}}$  target has a contiguous mismatch running from position  $x$  to  $y$  relative to PAM. **(b, c)** FRET histograms of Cas9-RNA binding to DNA constructs carrying PAM-distal **(b)** and PAM-proximal **(c)** mismatches. The number of molecules for each histogram ranged from 568 to 3,053.  $[\text{Cas9-RNA}] = 20 \text{ nM}$ . **(d)** The fraction of Cas9-RNA bound DNA molecules for different DNA targets. All the data shown in the figure are from independent experiments and error bars represent s.d. for  $n = 3$  ( $n = 2$  for few sets).

### 2.3.3 Different DNA binding modes of Cas9-RNA

For DNA targets to which Cas9-RNA binds weakly, we observed a second bound state with a mid FRET peak at  $\sim 0.42$ , in addition to the 0.92 high FRET state. Single-molecule time trajectories (Figure 2.6a) and transition density plots reporting on the relative transition frequencies after hidden Markov modeling analysis<sup>66</sup> (Figure 2.6b, Figure 2.7, Figure 2.8, and Figure 2.9) revealed reversible transitions between the unbound (FRET $<0.2$ ) state and both mid and high FRET=2 bound states, and lifetime analysis as a function of Cas9-RNA concentration confirmed that transitions are due to Cas9-RNA association/dissociation events (Figure 2.10 and Figure 2.11). The mid FRET was more frequently observed as the number of mismatches increased (Figure 2.6b and Figure 2.7), and were also observed for DNA targets without PAM or without matching sequence, indicating that it does not require either (Figure 2.12 and Figure 2.13). The high FRET state was rarely observed without PAM (Figure 2.14). We propose that the Cas9-RNA has two binding modes (Figure 2.17). The mid FRET state (sampling mode) likely does not involve RNA-DNA heteroduplex formation and may represent a mode of PAM surveillance. It is possible that local diffusion may give rise to the time-averaged FRET value of the mid FRET state. Sequence-independent sampling of DNA target for PAM can occur at multiple locations in the DNA target, and the mid-FRET state, representative of this sampling, expectedly had a broad FRET distribution (Figure 2.7, Figure 2.8, and Figure 2.9). If PAM is recognized during transient binding in the sampling mode, RNA-DNA heteroduplex formation follows, resulting in the high FRET state.

Multimodal binding kinetics were also observed for Cascade in Type I CRISPR systems, but its shorter-lived binding mode plays a priming function which is absent for Cas9<sup>55</sup>.

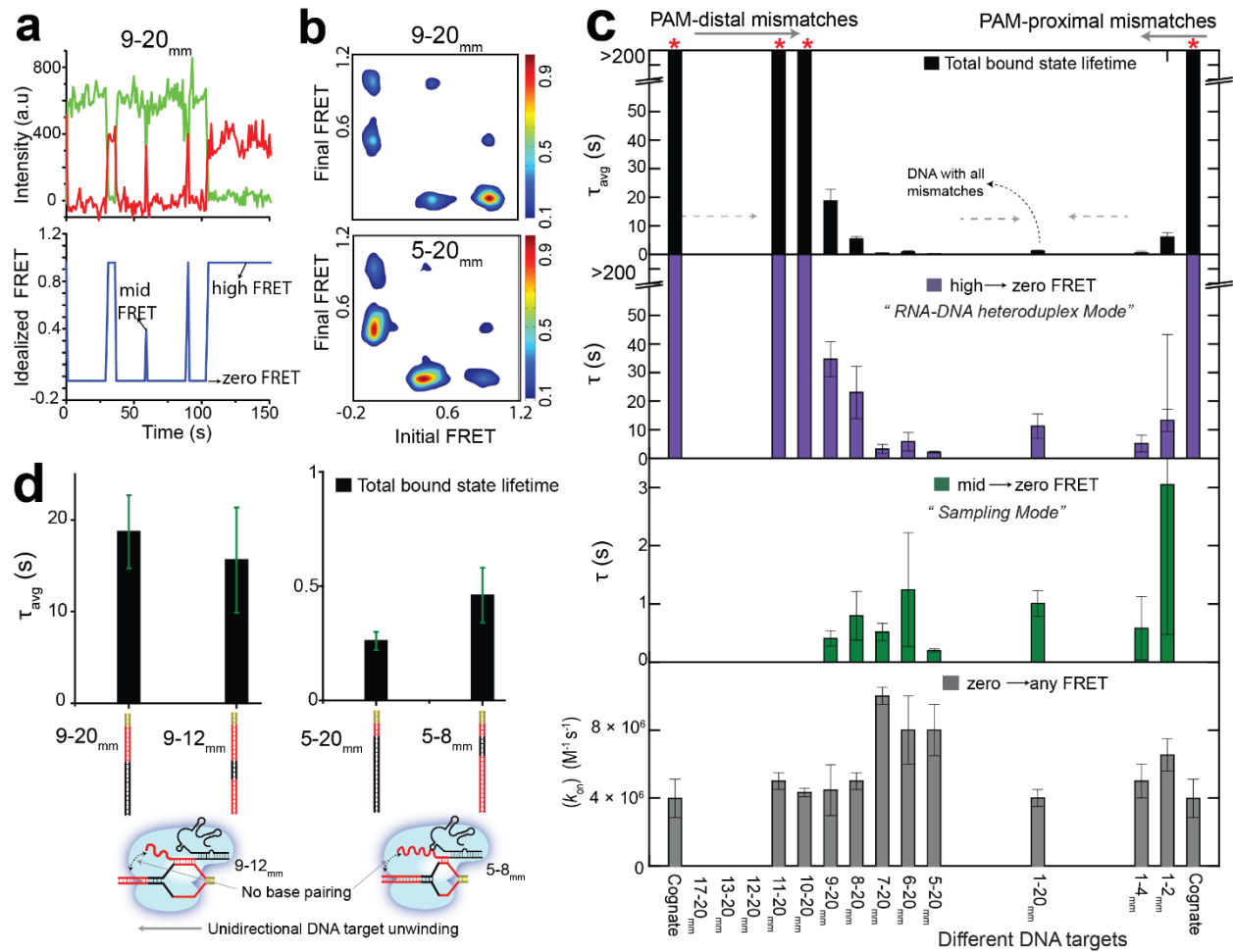
#### 2.3.4 Effect of DNA target mismatches on Cas9-RNA binding and dissociation kinetics

Survival probability distributions of dwell times in the bound state (FRET>0.2) before transitioning to the unbound state were best described by a double-exponential decay (Figure 2.14). The amplitude-weighted lifetime of the bound state ( $\tau_{\text{avg}}$ ) decreased precipitously even with just 2 bp PAM-proximal mismatches, likely because the R-loop failed to extend beyond the mismatches. In contrast, 12 bp mismatches were necessary from the distal end for any detectable decrease in  $\tau_{\text{avg}}$  (Figure 2.6c). Similar pattern was also observed for lifetime of transitions from high to zero FRET state whereas the lifetimes of the mid to zero FRET state remained short for all DNA targets tested, on average ~0.1 s (Figure 2.6c), supporting our proposal that the mid FRET state is a sampling mode that does not require sequence recognition.

In contrast to the bound state lifetimes, the lifetimes of the unbound state were only weakly dependent on sequence (Figure 2.6c and Figure 2.12), yielding the bimolecular association rate constant ( $k_{\text{on}}$ ) of  $\sim 6 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$  with some reduction for DNA targets without PAM. Overall, our kinetic analysis showed that mismatches affect Cas9-RNA binding mainly through changes in the dissociation rate. Complete kinetic model along with rates of transitions between different states for some DNA targets are in Figure 2.17 and Figure 2.15.

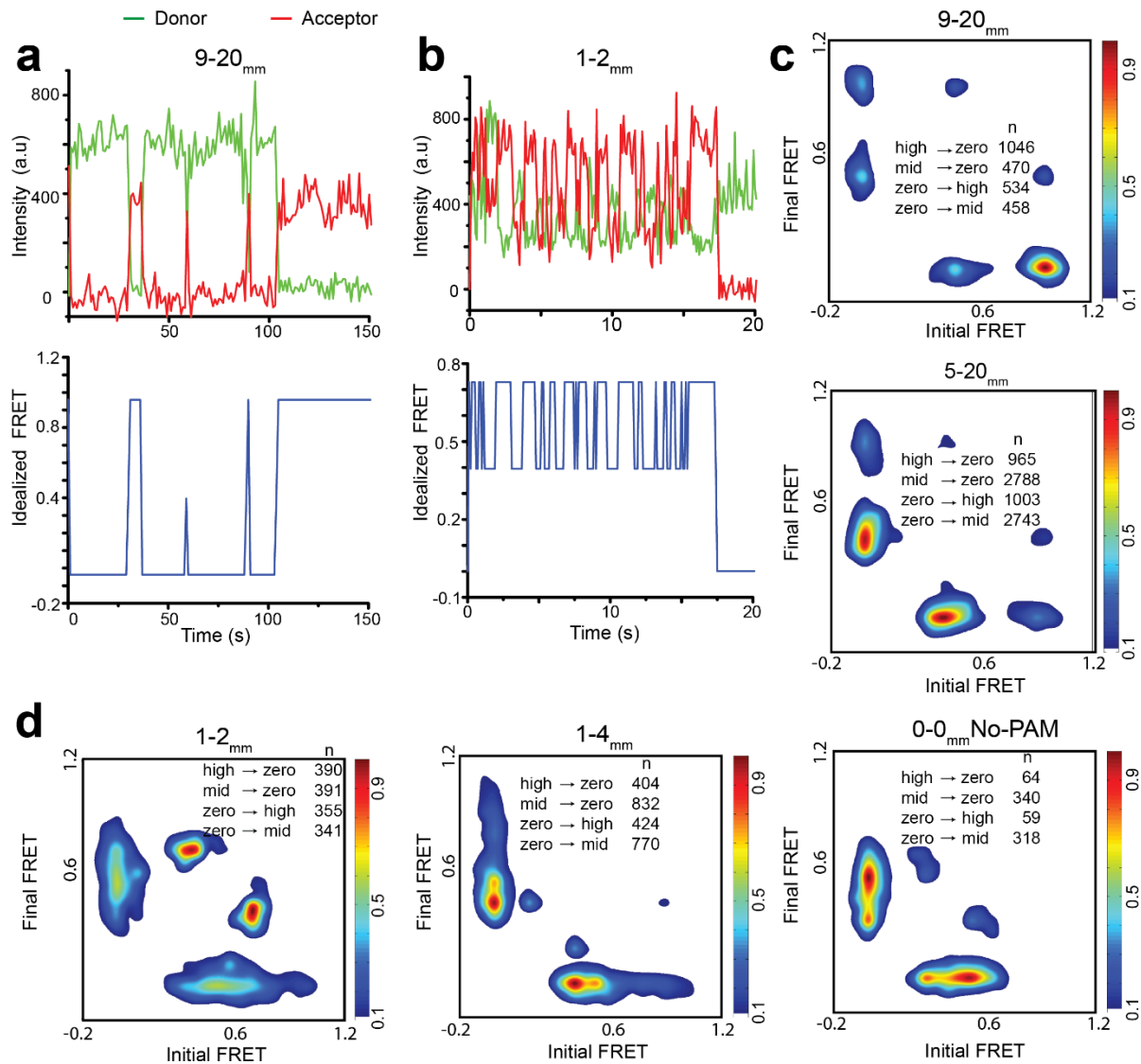
The relative importance of PAM-proximal base pairs over the PAM-distal base pairs (Figure 2.5d and Figure 2.6c) supports the model of unidirectional extension of the RNA-DNA heteroduplex starting from the PAM-proximal end<sup>17,61</sup>. For 5-20<sub>mm</sub> that has a maximum of 4 bp heteroduplex extension from PAM, the bound state lifetime was about 0.5 s, i.e. such potential targets are rapidly rejected. The bound state lifetime increased to about 8 s for 8-20<sub>mm</sub> with 7 bp heteroduplex, and to about 16 s for 9-20<sub>mm</sub> with 8 bp heteroduplex. For 9 bp or more heteroduplexes, the measured lifetime was limited by photobleaching

lifetime of about 3 min (Figure 2.4). Therefore, DNA sequences with 9 or more matching bp from the PAM proximal end have extremely long times, and Cas9-RNA would be unable to reject such sequences rapidly. A prediction of this model is that inserting a roadblock of mismatches near this boundary would prematurely terminate heteroduplex extension such that dissociation kinetics would be independent of the presence of a matched sequence beyond the block. To test this prediction, we created two “roadblock” targets, 9-12<sub>mm</sub> and 5-8<sub>mm</sub>. Indeed, the binding fraction and the lifetime of the bound state (Figure 2.16 and Figure 2.6d) for 9-12<sub>mm</sub> and 5-8<sub>mm</sub> were similar to those of 9-20<sub>mm</sub> and 5-20<sub>mm</sub>, respectively, confirming our prediction.



**Figure 2.6 | Cas9-RNA bound state lifetimes for different DNA targets.**

**(a)** smFRET time trajectory (donor and acceptor intensities, top, and idealized FRET via HMM analysis, bottom) for 9-20<sub>mm</sub> DNA target in the presence of 20 nM Cas9-RNA. Reversible Cas9-RNA association to high and mid FRET states and disassociation to zero FRET state are shown. **(b)** Transition density plots show relative transition frequencies between different FRET states for 9-20<sub>mm</sub> and 5-20<sub>mm</sub> DNA targets. [Cas9-RNA] = 20 nM. **(c)** The amplitude-weighted lifetime,  $\tau_{\text{avg}}$ , of the putative bound state, lifetime of high to zero and mid to zero FRET state transitions and biomolecular rate association constants for different DNA targets. Based on our model, the mid and high FRET states correspond to sampling and RNA-DNA heteroduplex modes respectively. **(d)** Lifetime comparison of DNA targets with the respective DNA targets containing mismatches after the roadblock. All the data shown in the figure are from independent experiments and error bars represent s.d. for  $n = 3$  ( $n = 2$  for few sets).

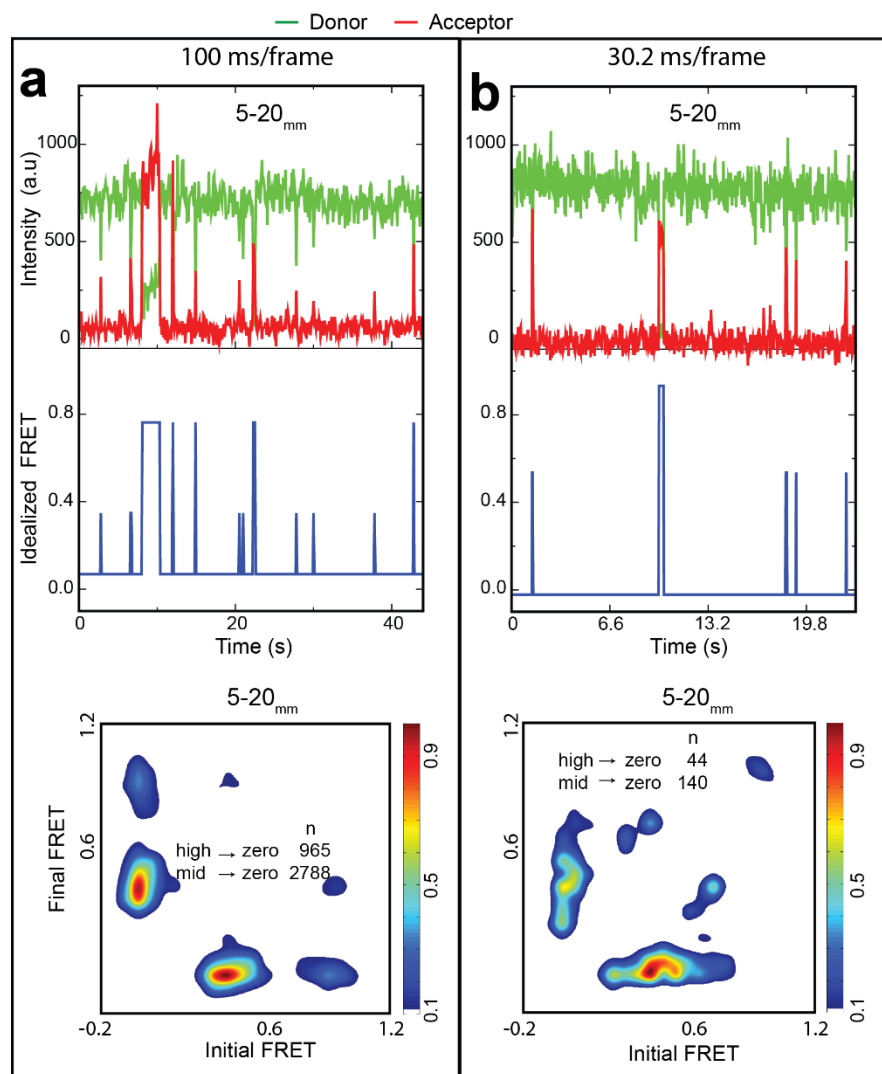


**Figure 2.7 | Hidden Markov model analysis and transition density plots reveal transitions between three different FRET states.**

**(a-b)** Representative single-molecule fluorescence time trajectories and their idealized FRET time trajectories, obtained using hidden Markov modeling<sup>66</sup>, for **(a)** the 9-20<sub>mm</sub> DNA target and **(b)** the 1-2<sub>mm</sub> DNA target. [Cas9-RNA] = 20 nM. **(c-d)** Transition density plots (TDP) reveal the relative frequencies of transitions between the initial FRET value and the final FRET value, as identified by hidden Markov modeling. The heat maps are separately scaled for each DNA target. A small proportion of rapid FRET fluctuations between high and mid FRET states were also observed. The relative population of transitions

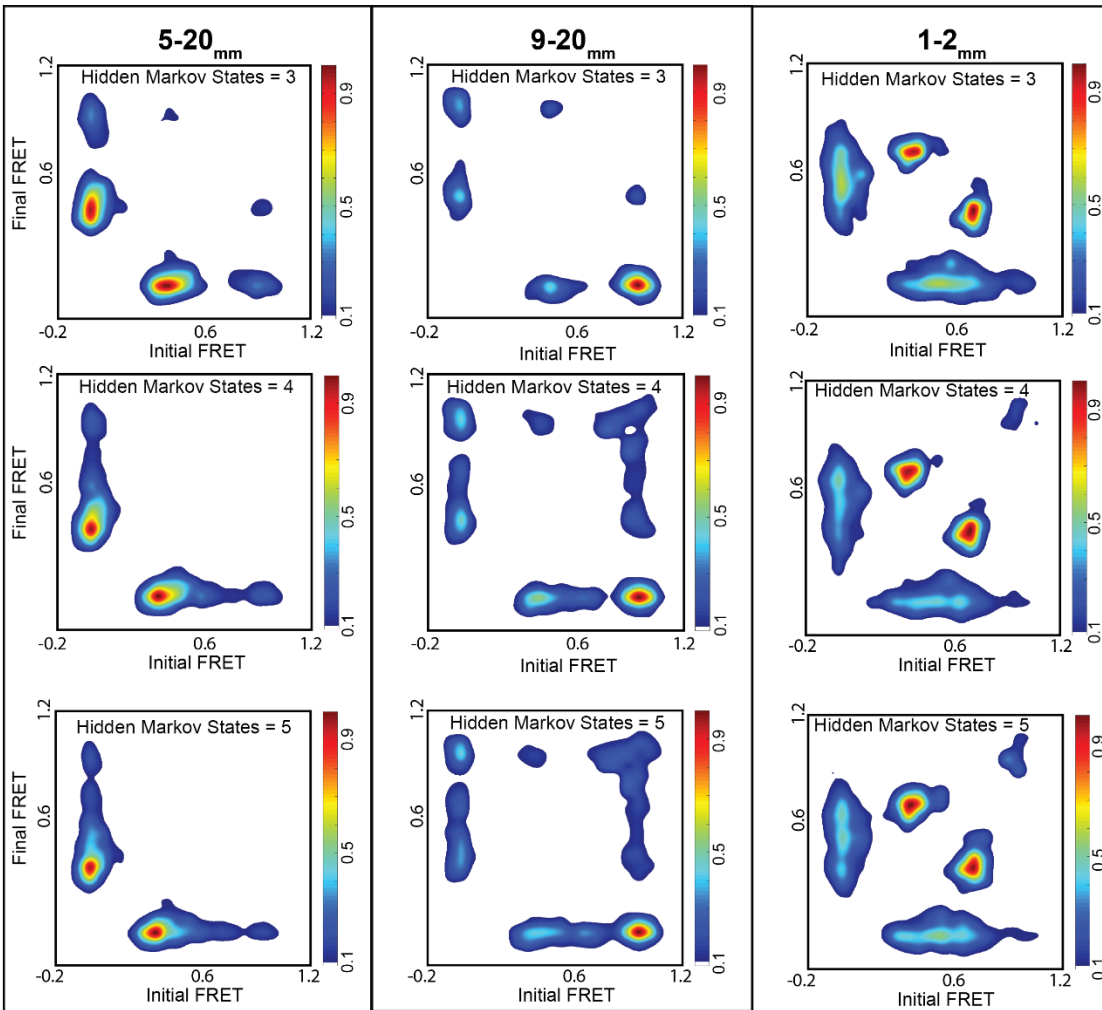


to/from the mid FRET state increased with increasing mismatches at both **(c)** PAM-distal and **(d)** PAM-proximal ends. The actual number of transitions from high to zero FRET state and mid to zero FRET state are shown within TDPs.



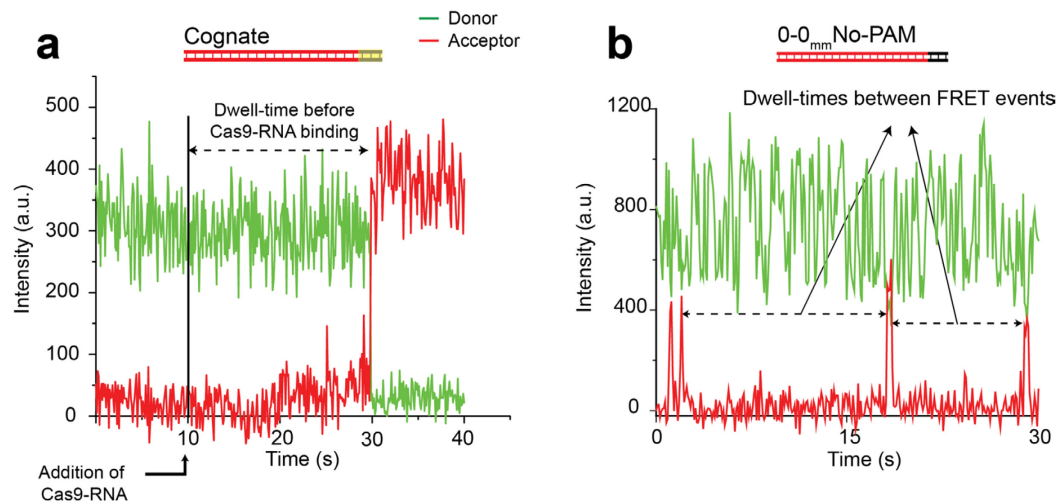
**Figure 2.8 | Transition density plots for 5-20<sub>mm</sub> DNA target at two different frame rates of image acquisition.**

In order to test if the short-lived mid FRET state observed at 100 ms time resolution was affected by time resolution, we performed additional measurements at 30.2 ms time resolution and compared the results for 5-20<sub>mm</sub> and found no appreciable difference. **(a)** A representative single-molecule fluorescence time trajectory showing Cas9-RNA binding at 100 ms/frame (top) and the associated transition density plot (bottom). [Cas9-RNA] = 20 nM. **(b)** A representative single-molecule fluorescence time trajectory showing Cas9-RNA binding at 30.2 ms/frame (top) and the associated transition density plot (bottom). [Cas9-RNA] = 20 nM.



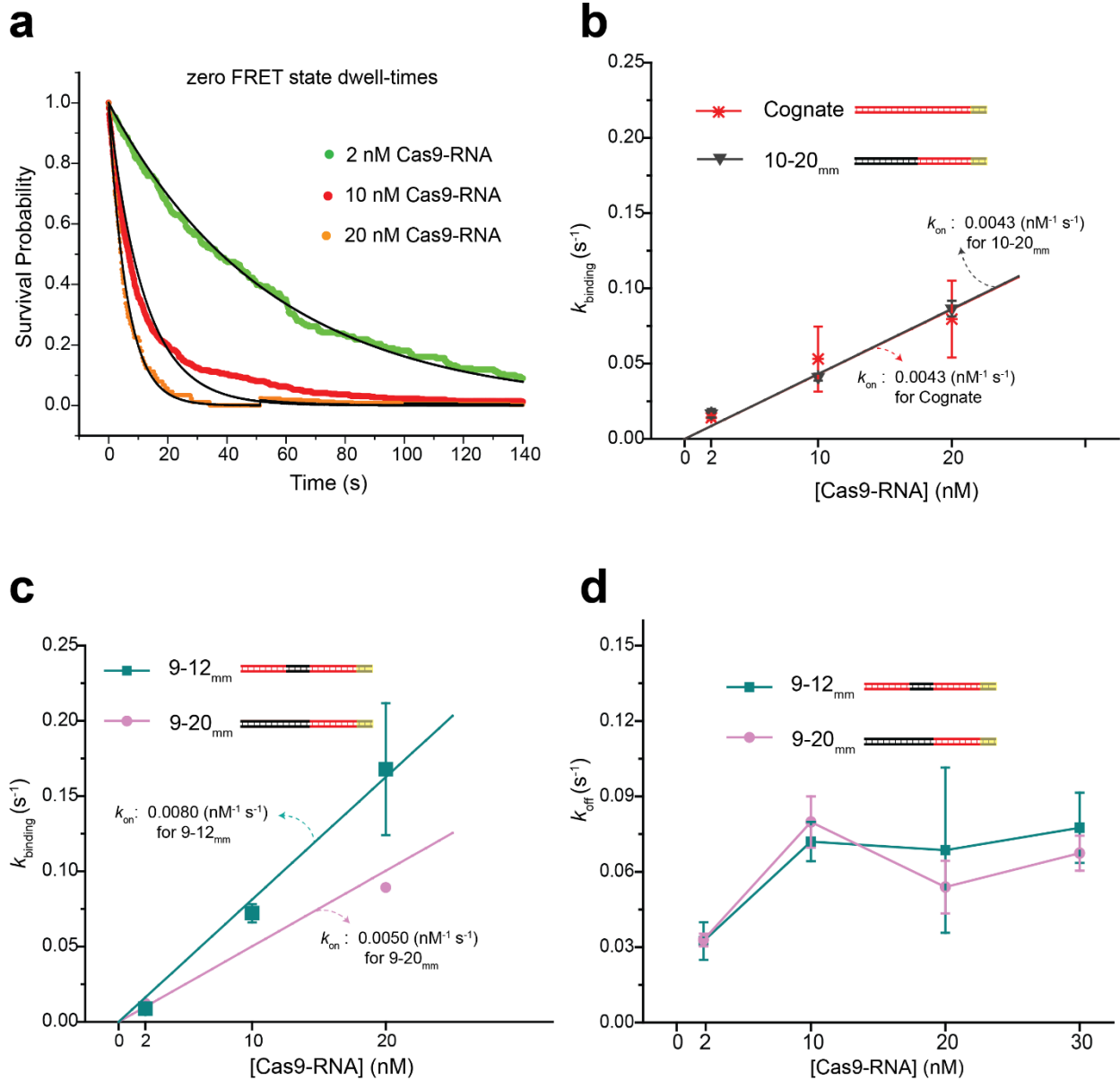
**Figure 2.9 | Transition density plots for 9-20<sub>mm</sub>, 5-20<sub>mm</sub> and 1-2<sub>mm</sub> DNA targets with different inputs for hidden Markov modeling.**

3 hidden Markov states were sufficient to capture the different FRET states of Cas9 targeting as any additional input state for hidden Markov modeling did not, evidently, result in any new discrete FRET state.



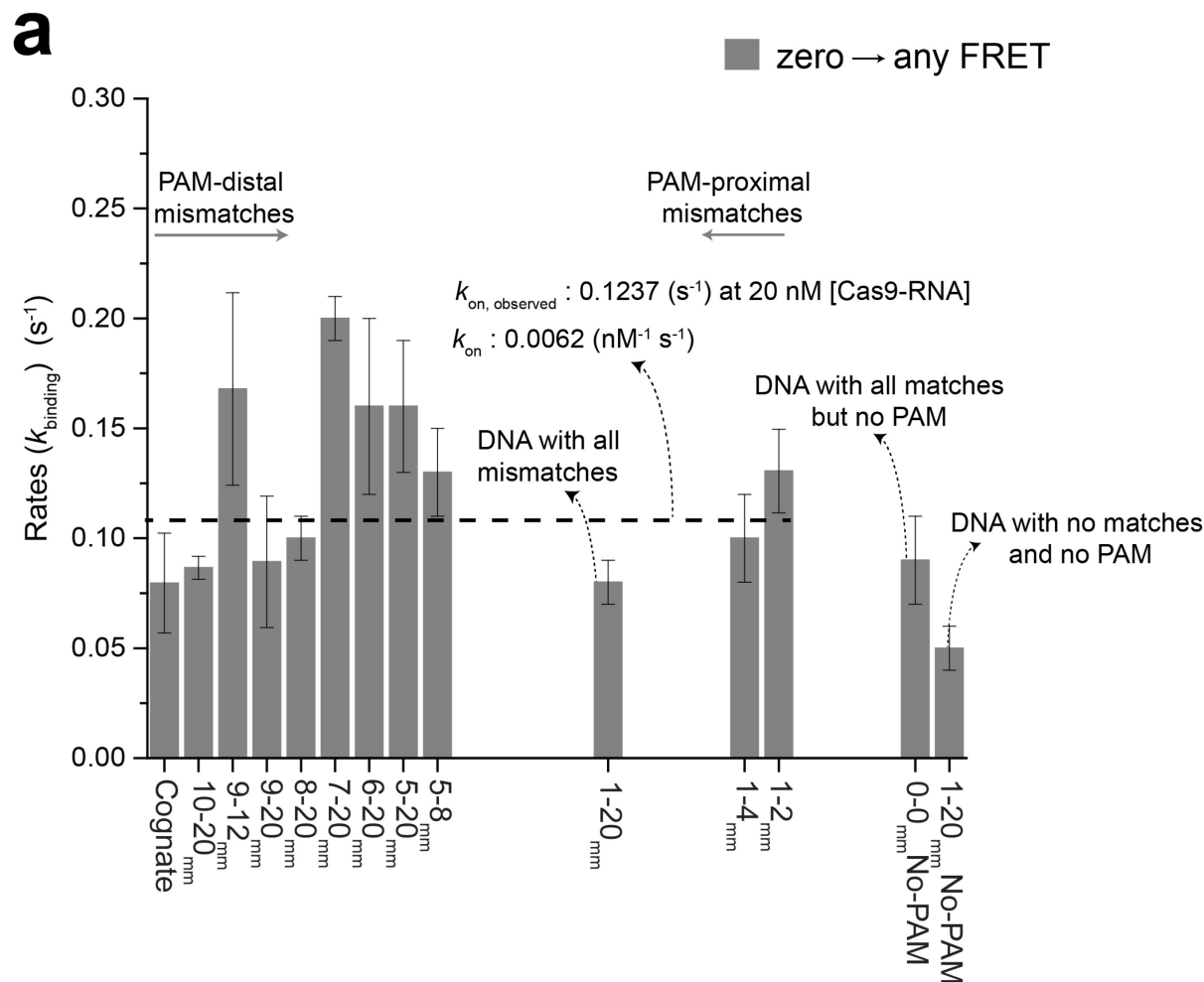
**Figure 2.10 | Determination of dwell times in the unbound state.**

**(a)** A representative smFRET time trajectory shows Cas9-RNA binding the cognate DNA target in real time. A 20 nM Cas9-RNA solution was added at the indicated time point and the dwell time in the zero FRET unbound state until the appearance of FRET was recorded. **(b)** A representative smFRET time trajectory for a DNA target with full sequence complementarity to guide RNA but without a PAM motif, obtained by imaging under steady state conditions with 20 nM Cas9-RNA in solution. Dwell times in the zero FRET unbound state between upward spikes in FRET due to transient binding events were recorded. See Figure 2.10 for the subsequent analysis.



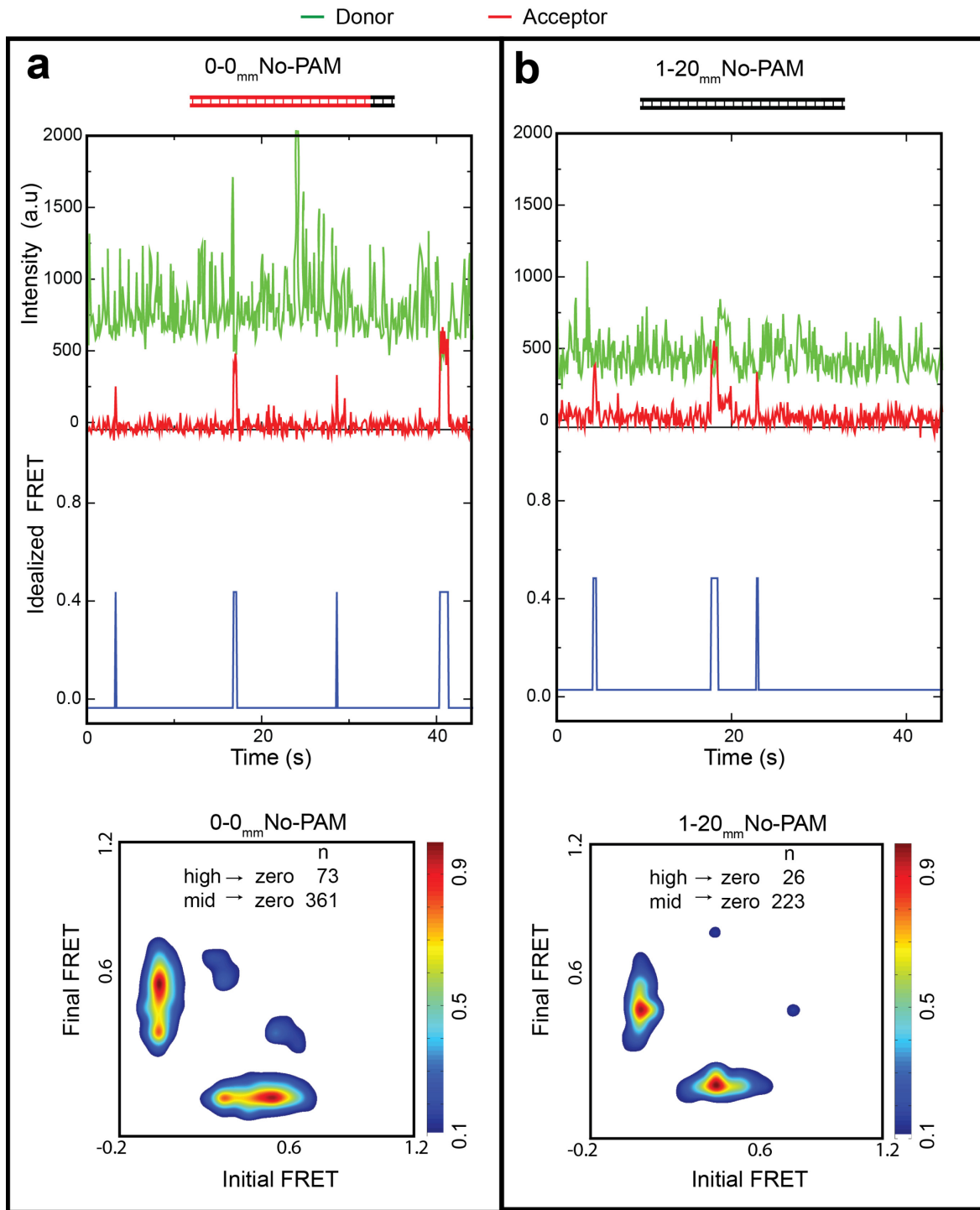
**Figure 2.11 | Determination of the rates of FRET appearance and disappearance.**

**(a)** Survival probability of dwell times of zero FRET state for 9-12<sub>mm</sub> DNA target (colored dots) fit with a single exponential decay (black line), to determine the observed binding rate,  $k_{\text{binding}}$  at three different Cas9-RNA concentrations. **(b, c)**  $k_{\text{binding}}$  vs.  $[\text{Cas9-RNA}]$  for different DNA constructs as noted. Linear fits were used to estimate the bimolecular association constant ( $k_{\text{on}}$ ). **(d)**  $k_{\text{off}}$  vs.  $[\text{Cas9-RNA}]$ . Error bars represent s.d. for  $n = 3$  ( $n = 2$  for few sets).



**Figure 2.12 | The binding rates.**

(a) Rate of Cas9-RNA DNA binding for a panel of DNA targets. For these calculations of the binding rate,  $k_{\text{binding}}$ , both the mid and high FRET state were taken together as a single state. A straight line fit to the observed binding rates was used to estimate the binding constant ( $k_{\text{on}}$ ) of Cas9-RNA averaged over all the DNA targets with PAM.  $[\text{Cas9-RNA}] = 20 \text{ nM}$ . Error bars represent s.d. for  $n = 3$  ( $n = 2$  for few sets).



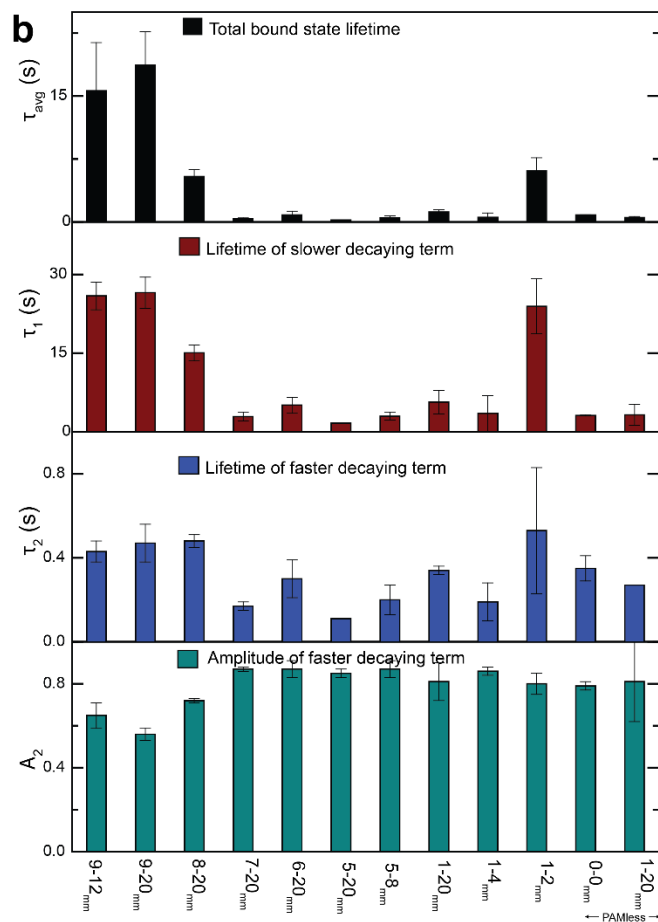
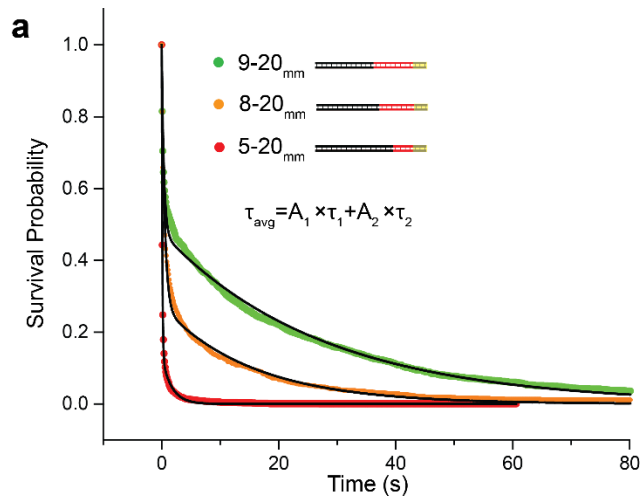
**Figure 2.13 | Binding dynamics for ‘PAM-less’ cognate DNA target and non-cognate target with PAM.**

Representative single-molecule fluorescence time trajectories (top) and associated transition density plots (bottom) of DNA target with full sequence complementarity (cognate) to guide-RNA but lacking a PAM

motif (**a**, 0-0<sub>mm</sub>No-PAM) and with no matches and no PAM (**b**, 1-20<sub>mm</sub>No-PAM). [Cas9-RNA] = 20 nM.

The number of transitions from high to zero FRET states and mid to zero FRET states are indicated within the plots.

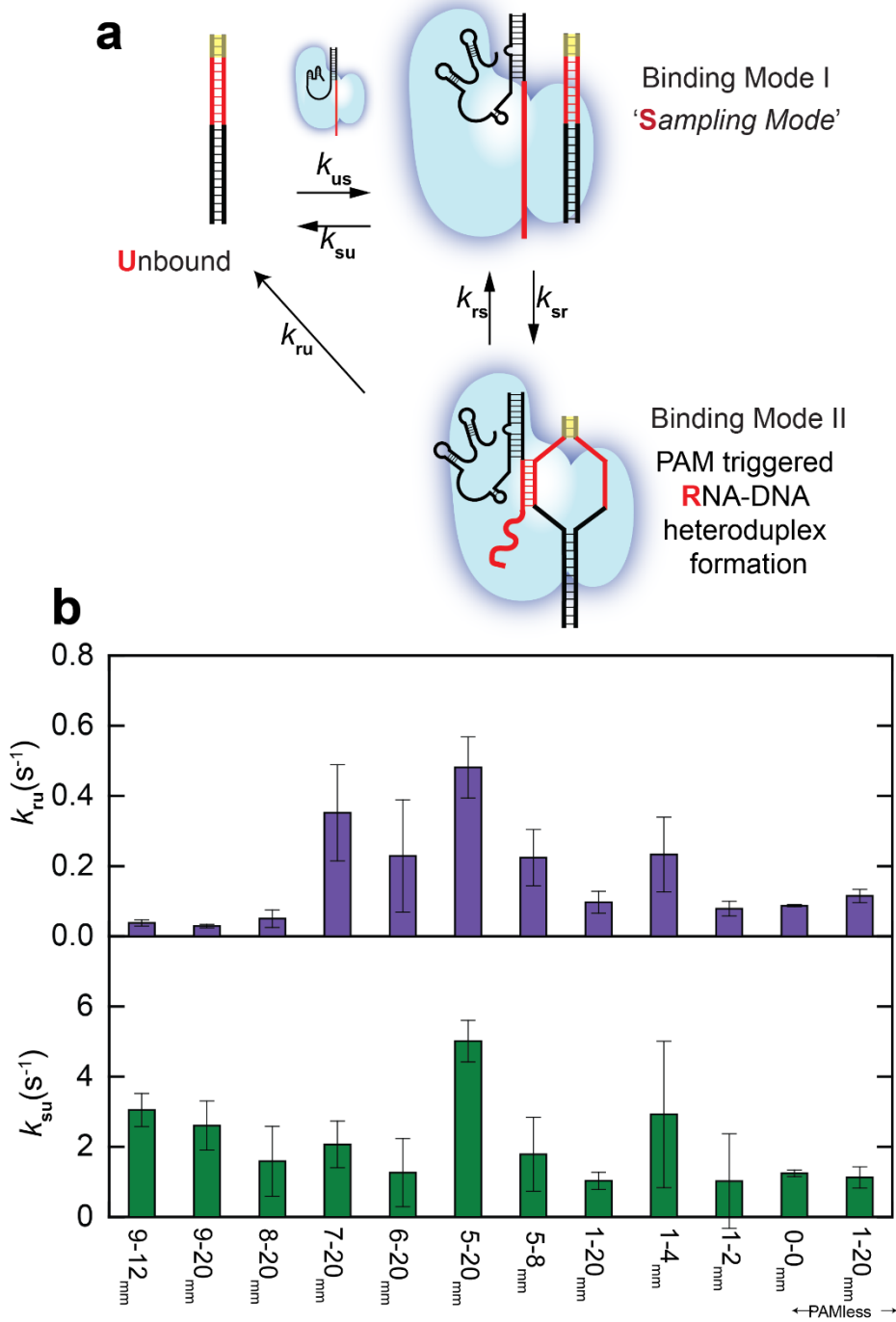




**Figure 2.14 | Fitting survival probability of Cas9-RNA bound states vs. time.**

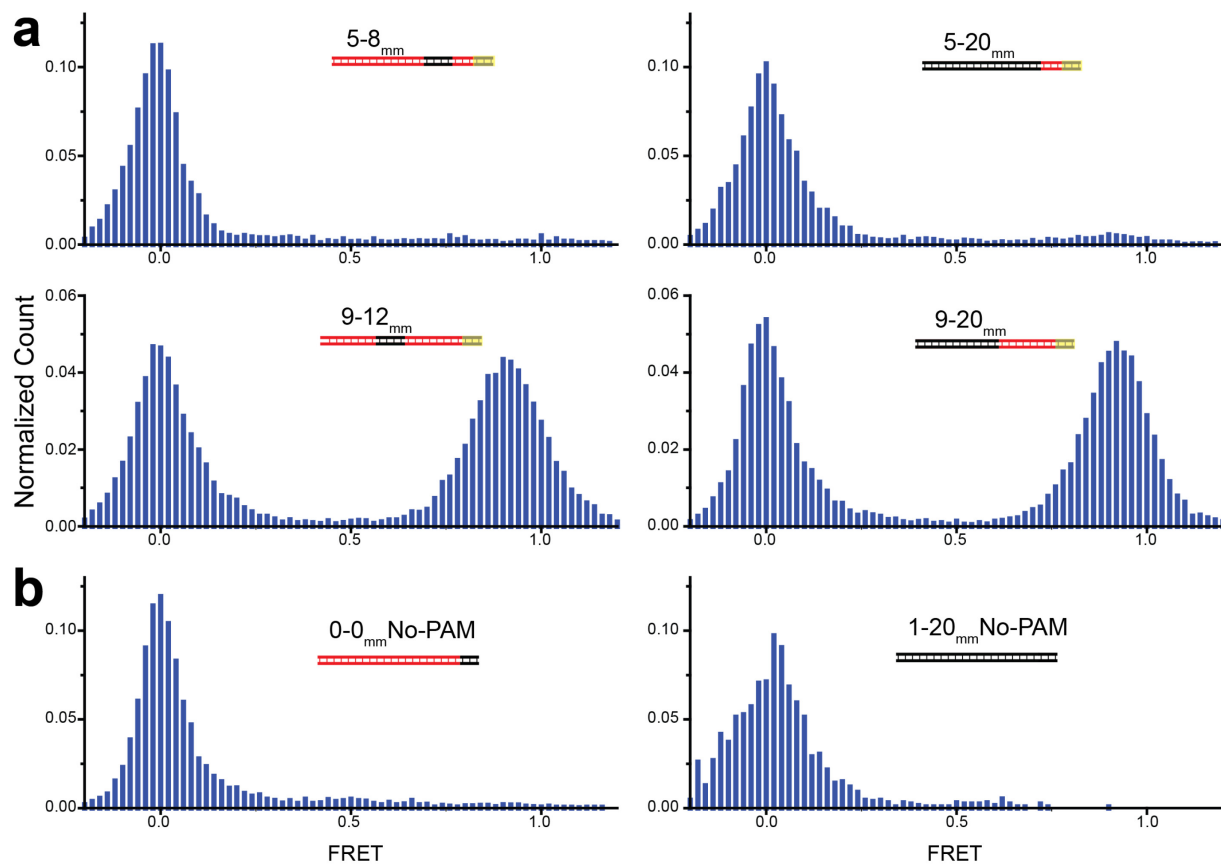
**(a)** The survival probability vs. time for different DNA targets (colored circles) of the putative bound states (FRET>0.2) (double exponential fit in black). **(b)** The parameters, of the double exponential fit

which results in two characteristic bound state lifetimes i.e. long lived and transient binding with their characteristic amplitudes. Error bars represent s.d. for  $n = 3$  ( $n = 2$  for few sets).



**Figure 2.15 | Kinetic model describing the transitions between various states of Cas9-RNA DNA targeting.**

(a) The kinetic model of Cas9-RNA and DNA interaction. Cas9-RNA has two binding modes i.e. the sampling mode for PAM surveillance and the RNA-DNA heteroduplex mode which are referred to as **s** and **r**. Unbound DNA state is **u**. (b) Rates of transitions from RNA-DNA heteroduplex and sampling mode to unbound state for certain DNA targets including the ones without PAM. Error bars represent s.d. for  $n = 3$  ( $n = 2$  for few sets).



**Figure 2.16 | FRET histograms for roadblock constructs and 'PAM-less' constructs.**

(a) FRET histograms for DNA targets with and without mismatches beyond 4 bp roadblock starting from 8<sup>th</sup> or 5<sup>th</sup> bp from PAM. (b) FRET histograms for DNA with full sequence complementarity but no PAM

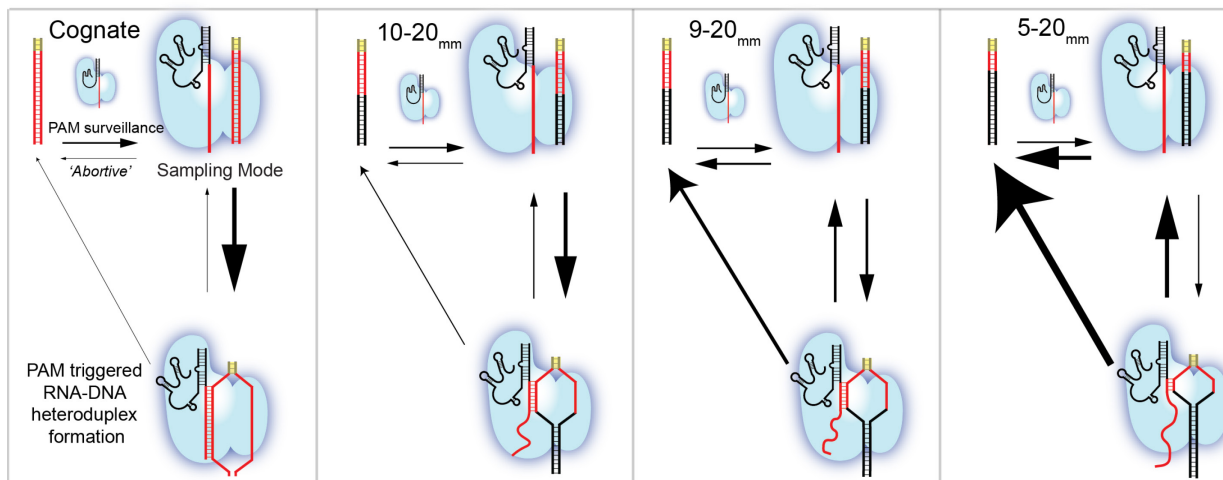
(left), and for DNA with no sequence complementarity and no PAM. The number of molecules per histogram ranged from 297 to 3,053.

## 2.4 DISCUSSION

A previous single-molecule<sup>61</sup> study that investigated the Cas9-RNA induced RNA-DNA heteroduplex formation via magnetic tweezers observed 11 PAM-proximal matches to be sufficient for stable RNA-DNA heteroduplex formation for StCas9 (the Cas9 ortholog from *Streptococcus thermophilus*). In our current study using SpCas9 (from *Streptococcus pyogenes*, simply referred to as Cas9), we found 9-10 PAM-proximal matches to be sufficient for ultra-stable Cas9-RNA binding. The stability of RNA guided CRISPR enzymes and DNA targets depends on energetic contributions of RNA-DNA heteroduplex and interactions between the DNA target and amino-acid residues of the CRISPR enzymes. The latter has been fine-tuned<sup>67,68</sup> through protein engineering to create more specific Cas9 variants and the small differences between different Cas9 orthologs may stem from the variations in the interactions between the DNA target and protein residues.

A two-step mechanism of Cas9-RNA binding involving PAM surveillance in the sampling mode and RNA-DNA heteroduplex formation upon PAM recognition (Figure 2.17) is also supported by structural analysis of Cas9 and Cas9-RNA-DNA ternary complexes, in which interactions between PAM-interacting amino acid motifs in Cas9 and the PAM of the DNA target precede and guide the further RNA-DNA heteroduplex formation<sup>32,65,69</sup>. Our observation that the heteroduplex lifetime increases greatly between 6 and 8 base pairs can be explained by the recently determined Cas9-RNA structure<sup>69</sup>, in which Watson-Crick faces of eight PAM-proximal nucleotides are solvent exposed, thus primed for heteroduplex formation. Once an RNA-DNA heteroduplex of 8 bp or more is formed, Cas9-RNA establishes a stable complex with the DNA, regardless of PAM-distal mismatches. Therefore, Cas9-RNA is unable to rapidly reject such off-target DNA, which it cannot cleave, and is sequestered by off-target DNA, limiting the

speed of genome editing. This effect would increase the minimal amount of Cas9-RNA required for genome editing, and may in turn lead to an increase in off-target cleavage. For applications requiring binding only, for example genome decoration or gene regulation, binding specificity will be almost entirely determined by the first 8 or 9 bp away from PAM, greatly reducing the ability to target well defined sequences in a large genome. For example, we found that 1,126 positions in the human reference genome match PAM plus 8 bp in the sequence we used (Table 2.2). For future improvements in Cas9 proteins, we suggest that one should focus on rapid rejection of such off targets. Our observations may also further inform the design of the guide-RNA and the DNA targets with minimal off-target effects<sup>70-81</sup>.



**Figure 2.17 | The proposed model of bimodal Cas9-RNA binding along with the kinetics of Cas9-RNA DNA targeting as a function of mismatches.**

## 2.5 MATERIALS AND METHODS

### 2.5.1 Preparation of DNA targets

All DNA oligonucleotides were purchased from Integrated DNA Technologies (IDT, Coralville, IA 52241). The Cy3 label in the DNA target is located 3 bp upstream of the protospacer adjacent motif (PAM 5'-NGG-3') and was achieved via conjugation of Cy3 N-hydroxysuccinimido (NHS) to an amino-

group attached to a modified thymine through a C6 linker (amino-dT). The entire panel of DNA targets used in our measurements is available in Table 2.1. A 22 nt long biotinylated adaptor strand was used for surface immobilization (Figure 2.2b). DNA targets were prepared by mixing all three component strands and heating to 90°C followed by cooling to room temperature over 3 hrs.

### **2.5.2 Expression and purification of Cas9 and dCas9**

The protein purification protocol was adapted from the methods described previously<sup>2,32</sup>. A fusion construct inserted into a custom pET-based expression vector was used for protein expression. The fusion construct consisted of the sequence encoding Cas9 (Cas9 residues 1-1368 from *S. pyogenes*) and an N-terminal decahistidine-maltose binding protein (His10-MBP) tag, followed by a peptide sequence containing a tobacco etch virus (TEV) protease cleavage site. The fusion protein was expressed in *E. coli* strain BL21 Rosetta 2 (DE3) (EMD Biosciences), grown in 2xYT medium at 18 °C for 16 h following induction with 0.5 mM IPTG. The harvested cells were lysed in 50 mM Tris pH 7.5, 500 mM NaCl, 5% glycerol, 1 mM TCEP, supplemented with protease inhibitor cocktail (Roche), and then homogenized (Avestin). Following ultra-centrifugation, the supernatant clarified cell lysate was separated from the cellular debris and bound in batch to Ni-NTA agarose (Qiagen). The resin was washed extensively with 50 mM Tris pH 7.5, 500 mM NaCl, 10 mM imidazole, 5% glycerol, 1 mM TCEP and the bound protein was eluted in a single-step with 50 mM Tris pH 7.5, 500 mM NaCl, 300 mM imidazole, 5% glycerol, 1 mM TCEP. TEV protease was added to the elutant and cleavage of the protein fusion was allowed to proceed overnight. Cas9 was then dialyzed into Buffer A (20 mM Tris-HCl pH 7.5, 125 mM KCl, 5% glycerol, 1 mM TCEP) for 3 h at 4°C, before being applied onto a 5 ml HiTrap SP HP sepharose column (GE Healthcare). After washing with Buffer A for three column volumes, Cas9 was eluted using a linear gradient from 0-100% Buffer B (20 mM Tris-HCl pH 7.5, 1 M KCl, 5% glycerol, 1 mM TCEP) over 20 column volumes. The protein was further purified by gel filtration chromatography on a Superdex 200 16/60 column (GE Healthcare) in Cas9 Storage Buffer (20 mM Tris-HCl pH 7.5, 200 mM KCl, 5%

glycerol, 1 mM TCEP). Cas9 was stored at -80°C. Catalytically dead Cas9 (dCas9; D10A/H840A mutations) was prepared with the same protocol.

### **2.5.3 Preparation of guide-RNA and Cas9-RNA**

The guide-RNA consists of CRISPR RNA (crRNA) and trans-activating crRNA (tracrRNA). The crRNA with an amino-dT was purchased from IDT and was labeled using Cy5-NHS. The tracrRNA was prepared using in vitro transcription as described previously<sup>17</sup>. The guide-RNA was assembled freshly for each experiment by mixing equimolar amounts of Cy5-labeled crRNA with tracrRNA, heated to 80°C followed by slow cooling to room temperature. The guide-RNA was then complexed with Cas9 (2-3 times the stoichiometric amount of guide-RNA) to form the Cas9-RNA complex for use in imaging experiments. RNA sequences are available in Table 2.1. A detailed schematic of the DNA and the Cas9-RNA design can be found in the Figure 2.2. The Cas9-RNA activity on the cognate sequence used in this study was characterized previously<sup>17</sup>. Our biochemical assays showed that fluorophore labeling in the DNA target or crRNA had not impaired DNA target cleavage. (Figure 2.3).

### **2.5.4 Single-molecule detection and data analysis**

Cy3-labeled DNA targets were immobilized on the PEG (Polyethylene glycol) passivated surface using neutravidin-biotin interaction. The DNA target molecules were then imaged in the presence of Cy5-labeled Cas9-RNA (referred to as Cas9-RNA for brevity here) using the total internal reflection fluorescence microscopy. Imaging was performed at room temperature in a buffer (20 mM Tris-HCl, 100 mM KCl, 5 mM MgCl<sub>2</sub>, 5 % (v/v) glycerol, 0.2 mg ml<sup>-1</sup> BSA, 1 mg ml<sup>-1</sup> glucose oxidase, 0.04 mg ml<sup>-1</sup> catalase, 0.8% dextrose and saturated Trolox (~3 mM)). The time resolution for all the experiments was 100 ms unless stated otherwise. Detailed methods of single-molecule FRET (smFRET) data acquisition and analysis were described previously<sup>64</sup>. The FRET efficiency of a single molecule was approximated as  $FRET = IA/(ID+IA)$ , where ID and IA are the background and leakage corrected emission intensities of the donor and acceptor, respectively.

### 2.5.5 FRET histograms and Cas9-RNA bound DNA fraction

The first five frames (100 ms each) of each of the molecule's FRET time trajectories were used as data points to construct the FRET histograms. The first ten frames were used for the FRET histograms in Figure 2.4. The Cas9-RNA bound DNA fraction was calculated as the fraction of data points with FRET > 0.75 and the total number of data points in the FRET histograms. For each DNA target, the single molecule FRET time trajectories from independent experiments were combined together to construct the FRET histograms as described.

### 2.5.6 Lifetime analysis of bound and unbound states via thresholding

To confirm that the FRET signal indeed reports on Cas9-RNA binding, the lifetimes of the zero FRET (FRET < 0.2) and the putative bound state (mid and high FRET states taken as a single state, FRET > 0.2) were determined as a function of Cas9-RNA concentration [Cas9-RNA]. Based on this cut-off of FRET=0.2, the survival probability of the zero FRET state vs. time could be fit well with a single exponential decay, and the decay rate increased linearly with [Cas9-RNA]. In contrast, the survival probability vs time for the bound state had to be fit with a double exponential decay and the decay rates did not depend on [Cas9-RNA]. (Figure 2.14). Therefore, a bimolecular association/disassociation kinetics was used for the analysis of DNA binding by Cas9-RNA.



$$k_{\text{binding}} (\text{s}^{-1}) = k_{\text{on}} (\text{M}^{-1}\text{s}^{-1}) \times [\text{Cas9-RNA}] (\text{M})$$

Lifetime of the bound state via thresholding. In order to perform unbiased analysis of apparently three-state FRET fluctuations observed from binding-challenged DNA targets, we employed hidden Markov model analysis and generated idealized FRET time trajectories<sup>66</sup>, assuming there are three distinct FRET



states (high, mid and zero FRET states). To estimate the lifetime of the putative bound states, the survival probability of all the bound state events (mid and high FRET states taken as a single state, FRET > 0.2) vs time was fit using a double exponential decay profile ( $A_1 \exp(-t/\tau_1) + A_2 \exp(-t/\tau_2)$ ) (Figure 2.14). The final bound state lifetime ( $\tau_{\text{avg, observed}}$ ), is an amplitude weighted average of two distinct lifetimes  $\tau_1$  and  $\tau_2$  i.e.

$$\tau_{\text{avg, observed}} = A_1 \tau_1 + A_2 \tau_2 \quad (k_{\text{observed}} = 1/\tau_{\text{avg, observed}}).$$

Association rates. We determined the observed rates of binding  $k_{\text{binding}}$  using two independent methods. First, the Cas9-RNA binding events were captured in real time by flowing Cas9-RNA into the sample chamber with immobilized DNA target molecules (Figure 2.10a). Second, for the binding challenged DNA targets that showed reversible association/disassociation, the smFRET time trajectories obtained under steady state conditions were used to extract the unbound state duration between adjacent binding events (Figure 2.10b). These dwell times in the unbound state were then used to get the rate of association by fitting their survival probability distribution to a single exponential decay (Figure 2.11a).

Rates of transitions between different states. Generation of idealized FRET time trajectories using the hidden Markov model<sup>66</sup> yielded three different FRET states (zero, mid and high) along with the probabilities of transitions between the various FRET states. The log of transitions probabilities between any two states was used to estimate the mean transition probability between the two given states which was then used to estimate the rate as following:

$$k_{A-B} (\text{s}^{-1}) = T_{p(A-B)} \times \text{Sampling rate of image acquisition}$$

where  $k_{A-B}$  is the rate of transition from state A to B

$T_{p(A-B)}$  is the mean probability of transition from A to B.

If each frame is acquired over 0.1s, then sampling rate (1/0.1) = 10 s<sup>-1</sup>

Correction factors. Because the high FRET state was very long-lived for certain DNA targets (i.e. 8-20<sub>nm</sub>, 9-20<sub>nm</sub>, 9-12<sub>nm</sub>, 1-2<sub>nm</sub>), their dwell times were not accurately captured due to photobleaching-induced truncation of smFRET time trajectories. The same is true for the dwell time of the unbound state. We made the following correction to obtain the actual rate.

$$k_{\text{actual}} = k_{\text{observed}} - k_{\text{photobleach (high/zero FRET state)}}$$

where  $k_{\text{observed}}$  is the rate calculated above and  $k_{\text{photobleach (high/zero FRET state)}}$  is the rate of photobleaching of the high or zero FRET state. Finally, we obtain  $\tau_{\text{avg}} = 1 / k_{\text{actual}}$ .

### 2.5.7 Counts of DNA target sequences in human genome.

The human genome assembly (GRCh38.p6) was analyzed using custom MATLAB scripts to calculate the total occurrences of DNA target sequences used in this study, which is referred to as the actual count (Table 2.2). The total number of occurrences expected for a sequence, assuming a random distribution of A, T, G and C nucleotides, is referred to as the probabilistic count and is calculated as follows:

$$\text{Probabilistic Count} = (1/4)^n \times \text{Total number of bp in human genome ( 3.2 billion)}$$

where  $1/4$  is the probability of occurrence of any given nucleotide at a position in the sequence, and  $n$  is the number of bp in the genome.

## 2.6 AUTHOR CONTRIBUTIONS

Digvijay Singh, Samuel H. Sternberg, Jingyi Fei, Taekjip Ha and Jennifer A. Doudna designed the experiments. Digvijay Singh conducted all the single-molecule experiments and synthesized guide-RNA. Samuel H. Sternberg prepared Cas9, dCas9, and guide RNAs, and conducted biochemical DNA cleavage assays. Digvijay Singh and Jingyi Fei performed the data analysis. All authors discussed the data; Digvijay Singh, Samuel H. Sternberg, Jingyi Fei, Taekjip Ha wrote the manuscript.

**Table 2.1 | Different DNA targets used in this study for Cas9-RNA binding and RNA sequences for constituting Cas9-RNA.**












Description	DNA Sequences
CognateSequence	<p>20 nucleotide biotinylated adaptor for surface immobilization</p> <p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGACGCATAAAGATGAGACGCTGGAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGACTGCGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
17-20 <sub>mm</sub>	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTCTGCCATAAAGATGAGACGCTGGAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGAGACGGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
13-20 <sub>mm</sub>	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTCTGCGTATAAGATGAGACGCTGGAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGAGACGCATATACTACTCTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
12-20 <sub>mm</sub>	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTCTGCGTATTAGATGAGACGCTGGAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGAGACGCATAAATCTACTCTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
11-20 <sub>mm</sub>	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTCTGCGTATTTGATGAGACGCTGGAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGAGACGCATAAATCTACTCTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
10-20 <sub>mm</sub>	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTCTGCGTATTTCTAGAGACGCTGGAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGAGACGCATAAAGTACTCTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
9-20 <sub>mm</sub>	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTCTGCGTATTTCTGAGACGCTGGAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGAGACGCATAAAGAATCTCTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
8-20 <sub>mm</sub>	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTCTGCGTATTTCTAGAGACGCTGGAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGAGACGCATAAAGATCTCTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
7-20 <sub>mm</sub>	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTCTGCGTATTTCTACAGACGCTGGAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGAGACGCATAAAGATGCTCTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
6-20 <sub>mm</sub>	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTCTGCGTATTTCTACTGACGCTGGAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGAGACGCATAAAGATGACTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
5-20 <sub>mm</sub>	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTCTGCGTATTTCTACTACGCTGGAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGAGACGCATAAAGATGAGTGGGACCTCATGTTTGCAGTCGAACGA-5'</p>
1-20 <sub>mm</sub>	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTCTGCGTATTTCTACTCTGCGTGGAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGAGACGCATAAAGATGAGACGACCTCATGTTTGCAGTCGAACGA-5'</p>
1-2 <sub>mm</sub>	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGACGCATAAAGATGAGACGCTGGAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGACTGCGTATTTCTACTCTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
1-4 <sub>mm</sub>	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGACGCATAAAGATGAGTGCCTGGAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGACTGCGTATTTCTACTCACGCACCTCATGTTTGCAGTCGAACGA-5'</p>
9-12 <sub>mm</sub>	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGACGCATAAAGATGAGACGCTGGAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGACTGCGTATAAAGACTCTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
5-8 <sub>mm</sub>	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGACGCATAAAGAACTACGCTGGAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGACTGCGTATTTCTGAGTGGGACCTCATGTTTGCAGTCGAACGA-5'</p>
0-0 <sub>mm</sub> No-PAM	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGACGCATAAAGATGAGACGCATAAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGACTGCGTATTTCTACTCTGCGTATTCATGTTTGCAGTCGAACGA-5'</p>
1-20 <sub>mm</sub> No-PAM	<p>5' <sup>●</sup>-<sup>24</sup>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTCTGCGTATTTCTACTCTGCGATAAG <sup>■</sup>ACAAAACGTCAGCTTGCT-3'</p> <p>3' -GCGTTGCAGCAGTCGACAGA -CGTGTGCTCTTTAGAGACGAGACGCATAAAGATGAGACGCTATTCATGTTTGCAGTCGAACGA-5'</p>
Description	RNA Sequences
crRNA	5' -GACGCAUAAAGAUGAGACGCGUUU <sup>■</sup> AGAGCUAUGCUGUUUUU-3'
tracrRNA	5' -GGACAGCAUAGCAAGUUAUAAAUAAGGCUAGUCCGUUAUCAUCUUGAAAAAGUGGCACCGAGUCGGUCUUUUU-3'

● Biotin    ■ Protospacer Adjacent Motif (PAM)    ■ Thymine modification for Cy3 and Cy5 labeling

DNA sequences complementary to guide RNA are shown in red (Cognate).

RNA sequences complementary to the protospacer in a cognate DNA target are shown in red (Cognate).

**Table 2.2 | The abundance count of the sequence of the different DNA targets in human genome (GRCh38.p6) along with its probabilistic count.**

Description	DNA Sequences	Actual Count	Probabilistic Count
CognateSequence		0	0
17-20 <sub>mm</sub>		0	0
13-20 <sub>mm</sub>		8	24
12-20 <sub>mm</sub>		35	96
11-20 <sub>mm</sub>		99	382
10-20 <sub>mm</sub>		211	763
9-20 <sub>mm</sub>		1146	1526
8-20 <sub>mm</sub>		5936	24414
7-20 <sub>mm</sub>		18370	48828
6-20 <sub>mm</sub>		59365	97656
5-20 <sub>mm</sub>		222480	1562500

# Chapter 3.

## **Mechanisms of improved specificity of engineered Cas9s revealed by single molecule analysis**

\*Contents of this chapter is available at:

Singh, D., Wang, Y., Mallon, J., Yang, O., Fei, J., Poddar, A., Ceylan, D., Bailey, S., and Ha, T. (2017)  
Mechanisms of improved specificity of engineered Cas9s revealed by single molecule analysis, *bioRxiv*.  
(2017)

### 3.1 ABSTRACT

In microbes, CRISPR-Cas systems provide adaptive immunity against invading genetic elements. Cas9 in complex with a guide-RNA targets complementary DNA for cleavage and has been repurposed for wide-ranging biological applications. New Cas9s have been engineered (eCas9 and Cas9-HF1) to improve specificity, but how they help reduce off-target cleavage is not known. Here, we developed single molecule DNA unwinding assay to show that sequence mismatches affect cleavage reactions through rebalancing the internal unwinding/rewinding equilibrium. Increasing PAM-distal mismatches facilitate rewinding, and the associated cleavage impairment shows that cleavage proceeds from the unwound state. Engineered Cas9s depopulate the unwound state more readily upon mismatch detection. Intrinsic cleavage rate is much lower for engineered Cas9s, preventing cleavage from transiently unwound off-targets. DNA interrogation experiments showed that engineered Cas9s require about one additional base pair match for stable binding, freeing them from sites that would otherwise sequester them. Therefore, engineered Cas9s achieve their improved specificity (1) by inhibiting stable DNA binding to partially matching sequences, (2) by making DNA unwinding more sensitive to mismatches, and (3) by slowing down intrinsic cleavage reaction.

### 3.2 INTRODUCTION

In bacteria and archaea, CRISPR (clustered regularly interspaced short palindromic repeats)–Cas systems impart adaptive defense against phages and plasmid<sup>1</sup>. In type II CRISPR-Cas systems, the Cas9 endonuclease in complex with two RNAs, a CRISPR guide-RNA (crRNA) and a trans-activating RNA (tracrRNA), targets complementary 20 base pair (bp) sequences (protospacers) in foreign DNA for double-stranded cleavage, with a requirement that they be followed by a motif called PAM (protospacer adjacent motif, 5'-NGG-3' for *S. pyogenes* Cas9)<sup>4,16,17</sup>. Programmable DNA binding and cleavage by Cas9 has revolutionized life sciences, where Cas9 cleavage is used for genome editing and Cas9 binding is used for tagging genomic sites with markers or effectors for wide-ranging applications<sup>7,8</sup>. Minimizing

off-target effects of both binding and cleavage<sup>20,82</sup> remains an active area of study. *In vitro* and *in vivo* investigations have shown that Cas9-RNA specificity is most affected by PAM plus an 8-10 PAM-proximal seed region and concentrations of both Cas9 and RNA<sup>82,83</sup>. Rationally designed and engineered *S. Pyogenes* Cas9s (EngCas9s), namely enhanced Cas9 (eCas9; K848A/K1003A/R1060A mutations) and high fidelity Cas9 (Cas9-HF1; N497A/R661A/Q695A/Q926A mutations), led to remarkable specificity improvements in unbiased genome wide CRISPR-Cas9 specificity measurements<sup>67,68</sup>.

The stability of Cas9-RNA-DNA complex is in part a function of the specific RNA-DNA base-pairing and sequence-independent interactions between Cas9 residues and DNA, the latter being a potential source of promiscuity. Cas9-RNA uses binding energy to initially melt DNA near PAM, and DNA unwinding can continue downstream in the presence of sequence complementary with guide-RNA. The motivation behind the mutations in EngCas9s was to destabilize the Cas9-RNA-DNA complex by diminishing the favorable sequence-independent interactions between Cas9 residues and the DNA backbone (Figure 3.1b). Single molecule FRET measurements showed that Cas9 conformational changes that bring a nuclease domain to the cleavage site can be disrupted by fewer mismatches for the EngCas9, leading to the proposal that the nuclease movement is the conformational checkpoint against off-target cleavage<sup>84,85</sup>. Here, we report a comparative analysis of DNA interrogation and rejection, DNA unwinding and rewinding dynamics, and cleavage activation among different Cas9s using single molecule imaging methods that can detect multiple conformations and transient intermediates<sup>57</sup> and have been used previously to study CRISPR systems *in vitro*<sup>6,53,54,62,83-88</sup>. We found evidence that DNA unwinding is the checkpoint that guard against off-target cleavage and could determine the equilibrium between unwinding and rewinding that allowed us to estimate the intrinsic cleavage rate from the unwound state. Based on our findings, we propose that engineered Cas9s achieve their improved specificity (1) by inhibiting stable DNA binding to partially matching sequences, (2) by making DNA unwinding more sensitive to mismatches, and (3) by slowing down intrinsic cleavage reaction.

### 3.3 RESULTS

#### 3.3.1 Real-time single molecule FRET assay for DNA interrogation

We employed a single-molecule fluorescence resonance energy transfer (smFRET) assay<sup>89</sup> to investigate DNA interrogation by EngCas9-RNA. DNA targets (donor-labeled, 82 bp long) were immobilized on polyethylene glycol (PEG)-passivated flow chambers and EngCas9 pre-complexed with acceptor-labeled guide-RNA (EngCas9-RNA) was added to observe their interactions in real time. Locations of donor (Cy3) and acceptor (Cy5) were chosen such that specific interaction between DNA target and EngCas9-RNA would lead to FRET (Figure 3.1a and Figure 3.2). Labeling at these locations do not affect Cas9 cleavage activity<sup>83</sup>. Unless mentioned otherwise, we used catalytically dead versions of all Cas9s (denoted with the prefix d) in order to focus on the properties prior to cleavage. Wild-type (WT) Cas9 and dCas9 showed similar behavior in DNA interrogation experiments<sup>83</sup>.

We examined different DNA targets containing mismatches relative to the guide-RNA. As the naming convention, we used  $n_{PD}$  (the number of PAM-distal mismatches) and  $n_{PP}$  (the number of PAM-proximal mismatches) (Figure 3.1b). The cognate DNA target gave two distinct populations centered at FRET efficiency ( $E$ ) of 0.9 and 0 in the presence of 20 nM EngCas9-RNA. The  $E=0.9$  population was negligible if only the labeled RNA was added without Cas9 or if a fully mismatched DNA without PAM was used. Therefore, we assigned the  $E=0.9$  species to a sequence-specific EngCas9-RNA-DNA complex (Figure 3.1c). The  $E=0$  population is a combination of unbound states and bound states with inactive or missing acceptor. Cas9-RNA titration gave the apparent dissociation constant ( $K_d$ ) of 0.50 nM (dCas9), 0.5 nM (deCas9) and 2.7 nM (dCas9-HF1) (Figure 3.3).

To quantify the impact of mismatches, we determined the apparent Cas9-RNA bound fraction  $f_{\text{bound}}$ , defined as the normalized fraction of DNA molecules with  $E > 0.75$  (20 nM Cas9-RNA) (Figure 3.1d and Figure 3.4).  $f_{\text{bound}}$  vs. mismatches for EngCas9s was similar to that observed previously with WT Cas9<sup>83</sup>;



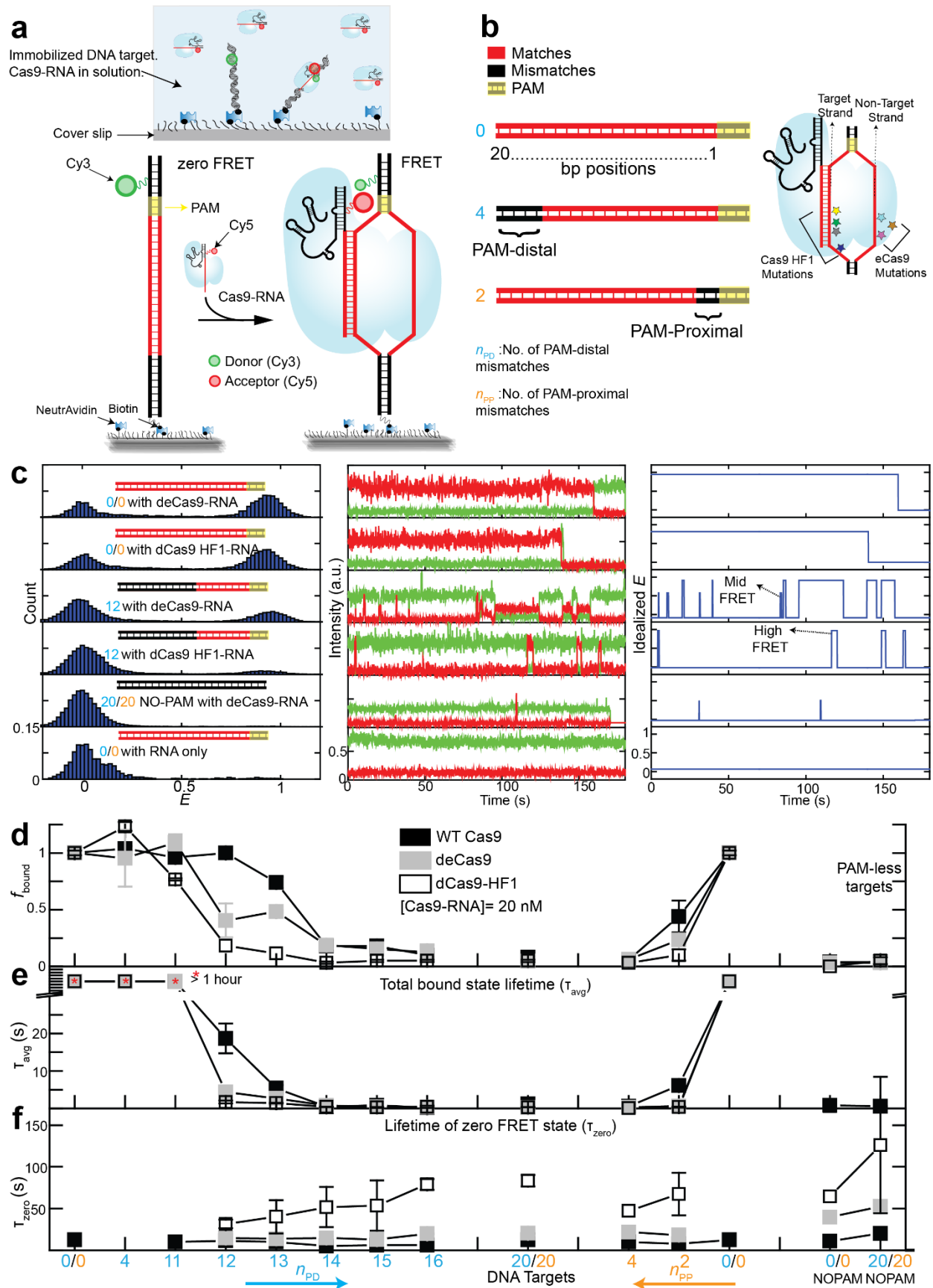
(i)  $f_{\text{bound}}$  remained unchanged when  $n_{\text{PD}}$  increased from 0 to 10 or 11, but precipitously decreased beyond, (ii) even  $n_{\text{PP}}$  of 2 or 4 caused a >50% or >95% drop in  $f_{\text{bound}}$ , respectively, (iii) binding is ultra-stable with >8-9 PAM-proximal matches as  $f_{\text{bound}}$  remained high even 1 hour after washing away free Cas9-RNA (Figure 3.5).

If there are enough mismatches to preclude stable binding, smFRET time-traces showed repetitive transitions between  $E=0$  and  $E=0.45$  states in addition to transitions between  $E=0$  and  $E=0.9$ , suggesting that there are multiple bound states distinguishable based on  $E$ <sup>83</sup> (Figure 3.1c). We used a hidden Markov modeling analysis<sup>90</sup> to determine  $\tau_{\text{avg}}$  as a fraction-weighted average of the high ( $E=0.9$ ) and mid ( $E=0.45$ ) FRET state lifetimes.  $\tau_{\text{avg}}$  was over 1 hour for DNA targets with at least  $m$  PAM-proximal matches ( $m=9$  for WT Cas9 and deCas9, 10 for dCas9-HF1) but decreased to 0.5 – 15 s with fewer than  $m$  PAM-proximal matches or any PAM-proximal mismatches (Figure 3.1e). Dwell times of the unbound state ( $E < 0.2$ ) were only weakly dependent on sequence (Figure 3.1f) (Figure 3.6 and Figure 3.7).

We list below qualitative features common between EngCas9s and WT Cas9 as well as quantitative differences:

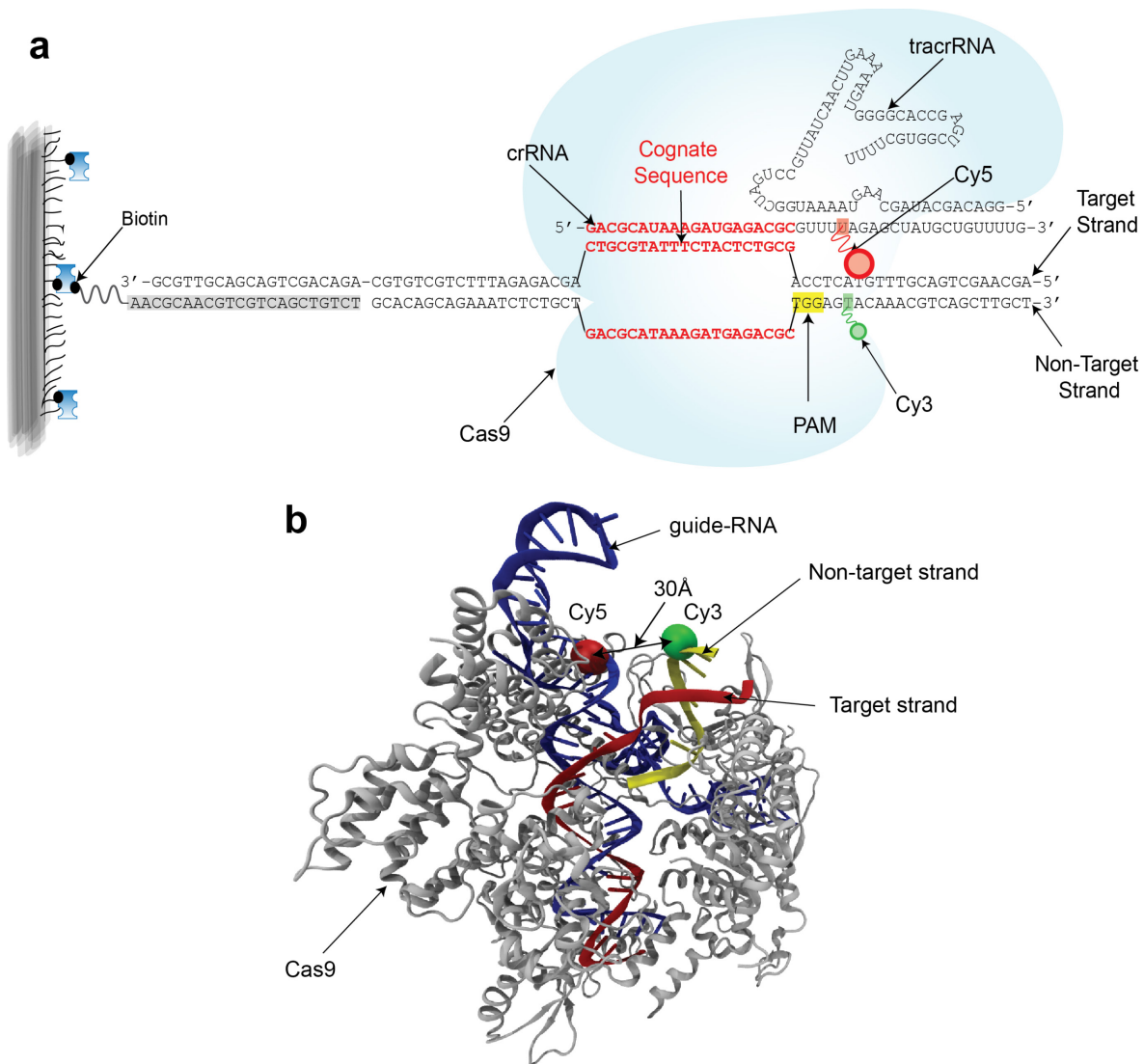
(1) All Cas9s interrogate and bind DNA in two distinct modes. Sequence-independent sampling of DNA target in search of PAM results in transient mid FRET events. The high FRET state results upon PAM detection and RNA-DNA heteroduplex formation, and its lifetime increased with increasing base-pairing between guide-RNA and DNA target. (2) The binding frequency is independent of sequence. The bimolecular association rate constant  $k_{\text{on}}$  decreased for EngCas9s ( $2.9 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$  for deCas9 and  $1 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$  for dCas9-HF1, compared to  $5.4 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$  for WT Cas9), possibly as a result of alteration in electrostatic and polar interaction upon removal of positively charged and polar residues (Figure 3.1f). (3) PAM-proximal mismatches are much more deleterious for stable binding compared to PAM-distal mismatches, and mismatches in the middle of the protospacer renders PAM-distal matches

inconsequential (Figure 3.4)<sup>83</sup>, suggesting that all Cas9s extend RNA-DNA heteroduplex unidirectionally from PAM-proximal to PAM-distal end. (4) PAM-proximal mismatches are more deleterious for EngCas9 binding than WT Cas9, and (5) deCas9 and WT Cas9<sup>83,91</sup> require 9 PAM-proximal matches for ultra-stable binding whereas dCas9-HF1 requires 10 (Figure 3.1e and Figure 3.5). Overall, EngCas9 are similar to WT Cas9 in their sequence-dependent binding properties except for the higher ability to reject both PAM-proximal and PAM-distal mismatches which may help reduce Cas9-RNA sequestration by partially matching targets.



**Figure 3.1 | smFRET assay to study DNA interrogation by engineered Cas9-RNA.**

(a) Schematic of smFRET assay. Cas9 in complex with an acceptor labeled guide-RNA binds a donor-labeled cognate DNA target. (b) DNA targets with mismatches in the protospacer region against the guide-RNA. The number of mismatches PAM-distal ( $n_{PD}$ ) and PAM-proximal ( $n_{PP}$ ) are shown in cyan and orange, respectively. Also shown are locations of EngCas9 mutations in dCas9-RNA-DNA complex (PDB ID: 4UN3). (c)  $E$  histograms (left) at 20 nM EngCas9-RNA or RNA only. Representative single molecule intensity time traces of donor (green) and acceptor (red) are shown (middle), along with  $E$  values idealized (right) by hidden Markov modeling. (d) Normalized fraction ( $f_{\text{bound}}$ ) of DNA molecules bound with Cas9-RNA at 20 nM Cas9-RNA. Fractions were normalized relative to the bound fraction of cognate DNA target. (e)  $\tau_{\text{avg}}$ , obtained from dwell times of  $E > 0.2$  states. (f) Unbound state lifetimes at [Cas9-RNA] = 20 nM. The mean of  $\tau_{\text{avg}}$  over various DNA targets was used to calculate  $k_{\text{on}}$ . Number of PAM-distal ( $n_{PD}$ ) and PAM-proximal mismatches ( $n_{PP}$ ) are shown in cyan and orange respectively. Error bars represent standard deviation (s.d.) from  $n = 2$  or  $3$ . Data for WT Cas9-RNA is taken from a previous study<sup>83</sup>.



**Figure 3.2 | FRET probe locations for DNA interrogation by Cas9-RNA.**

**(a)** Schematic of Cas9-RNA-DNA complex. The hybridized crRNA and tracrRNA are referred to as guide-RNA. Sequences in red denote guide sequence of the guide-RNA and the matching sequence of the DNA. The target strand is complementary to the guide-RNA. The non-target strand contains the PAM (5'-NGG-3'). A 22 nt biotin-labeled adaptor strand was used for surface immobilization of DNA and is highlighted in grey. **(b)** Cy3 and Cy5 labeling locations in Cas9-RNA-DNA complex (PDB ID: 4UN3)<sup>92</sup>.

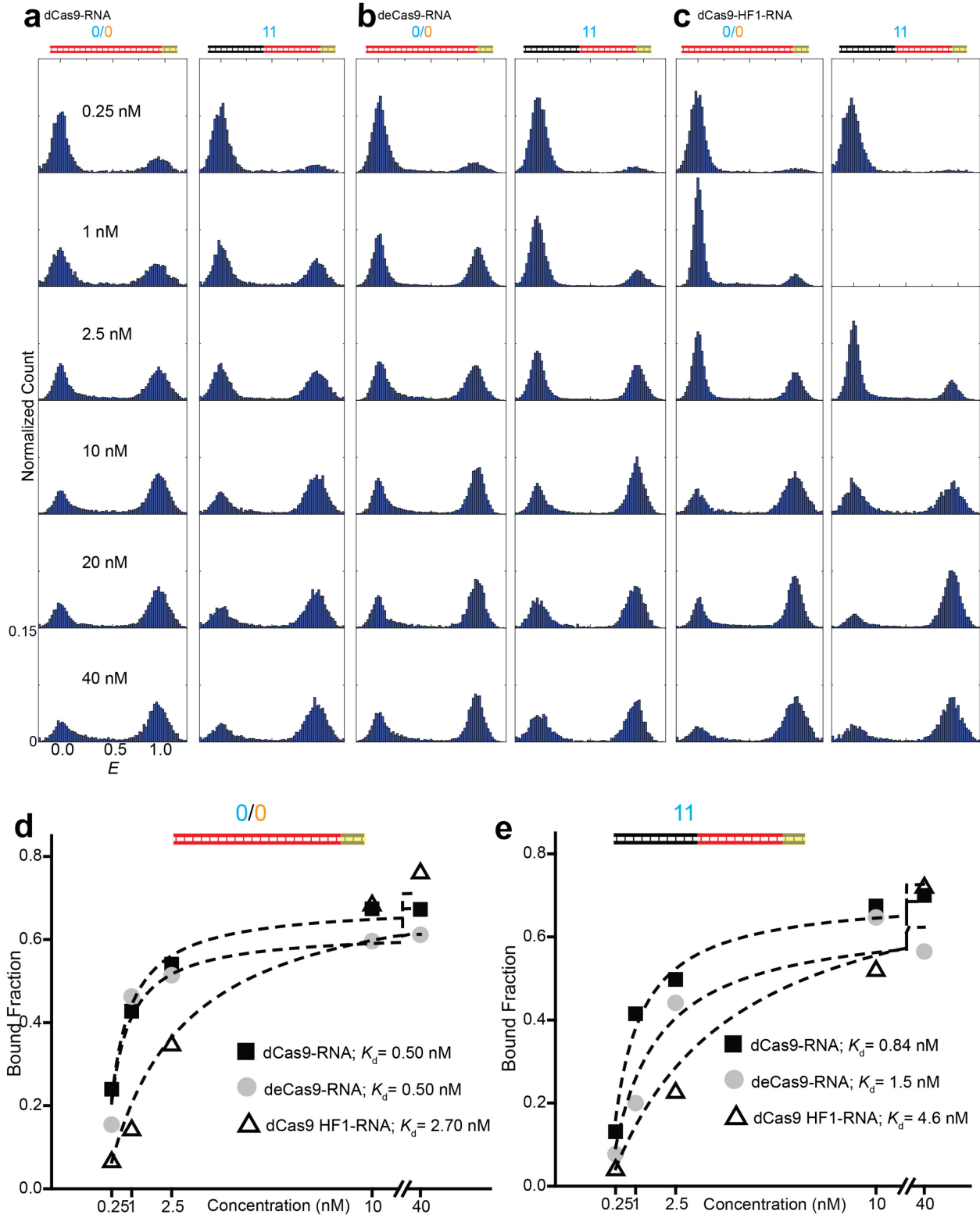
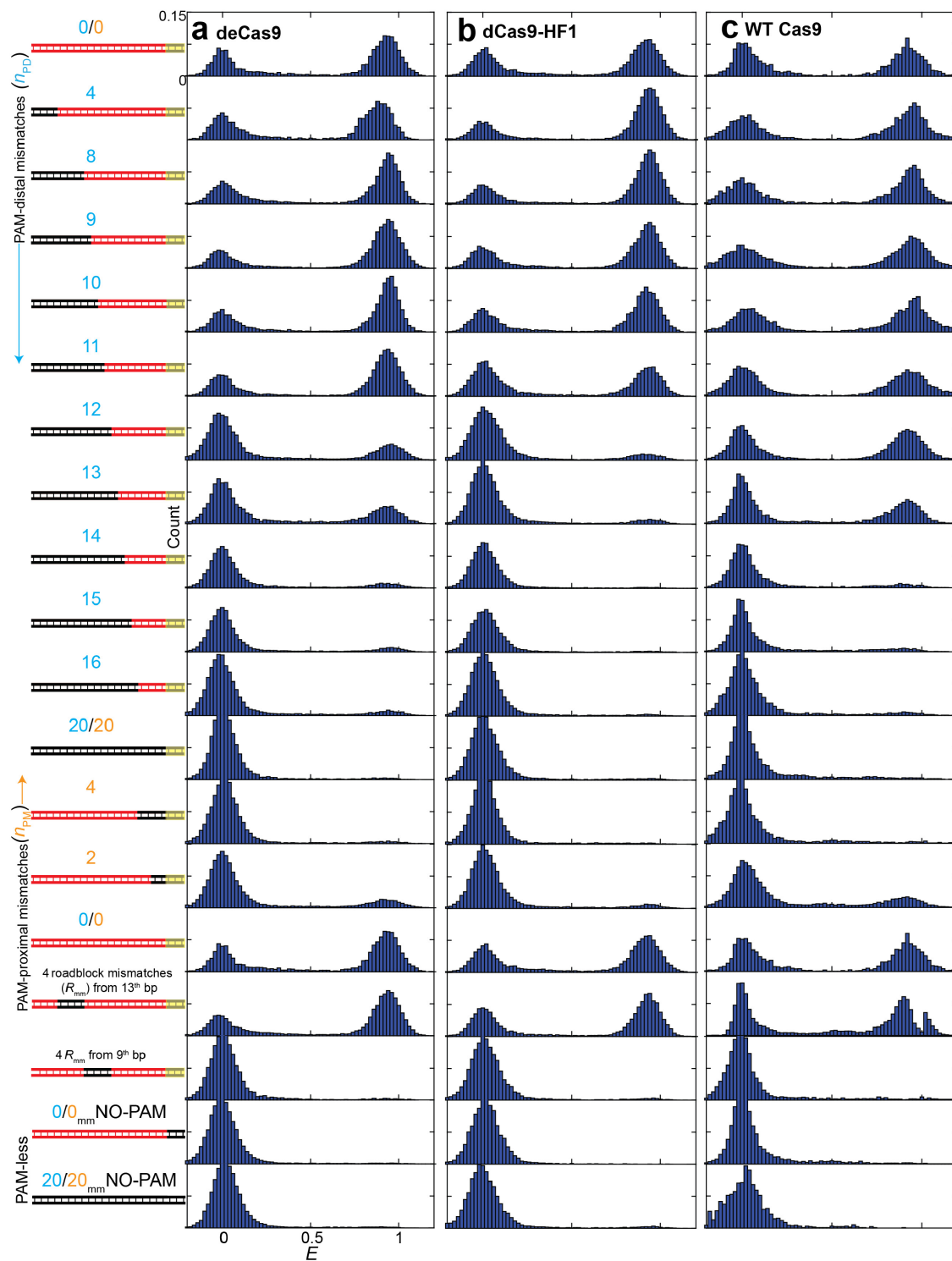


Figure 3.3 | Determination of  $K_d$  between Cas9-RNA and DNA.

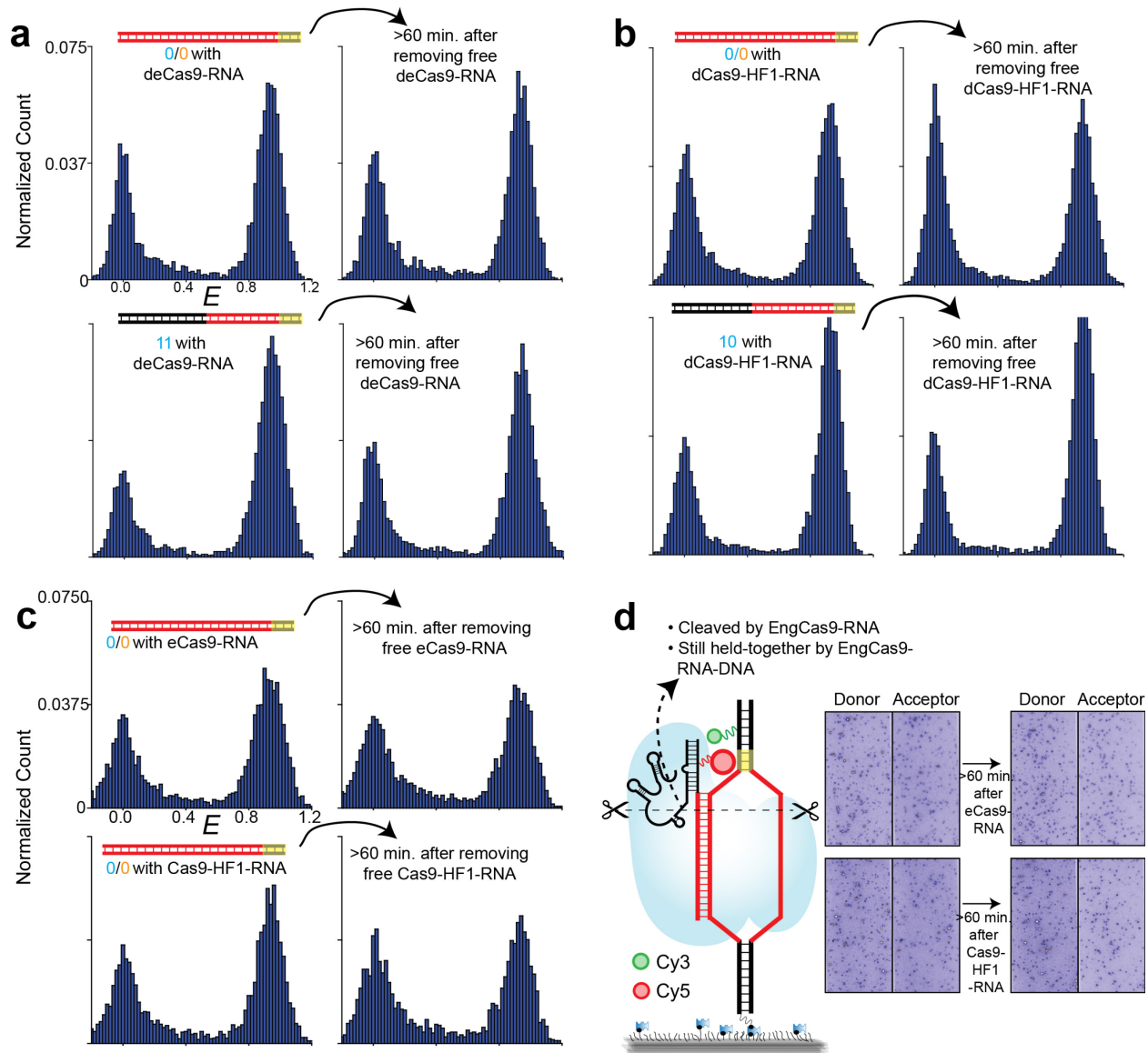
**(a-c)**  $E$  histograms for cognate DNA and DNA with  $n_{PD}=11$  for **(a)** dCas9 **(b)** deCas9 and **(c)** dCas9-HF1 vs [Cas9-RNA] obtained in smFRET DNA interrogation experiments. **(d-e)** The apparent bound fraction vs. [Cas9-RNA] and fits for  $K_d$  estimation. The number of PAM-distal mismatches ( $n_{PD}$ ) and PAM proximal mismatches ( $n_{PP}$ ) are shown in cyan and orange respectively.



**Figure 3.4 |  $E$  histograms for different DNA targets obtained smFRET DNA interrogation experiments at 20 nM Cas9-RNA.**



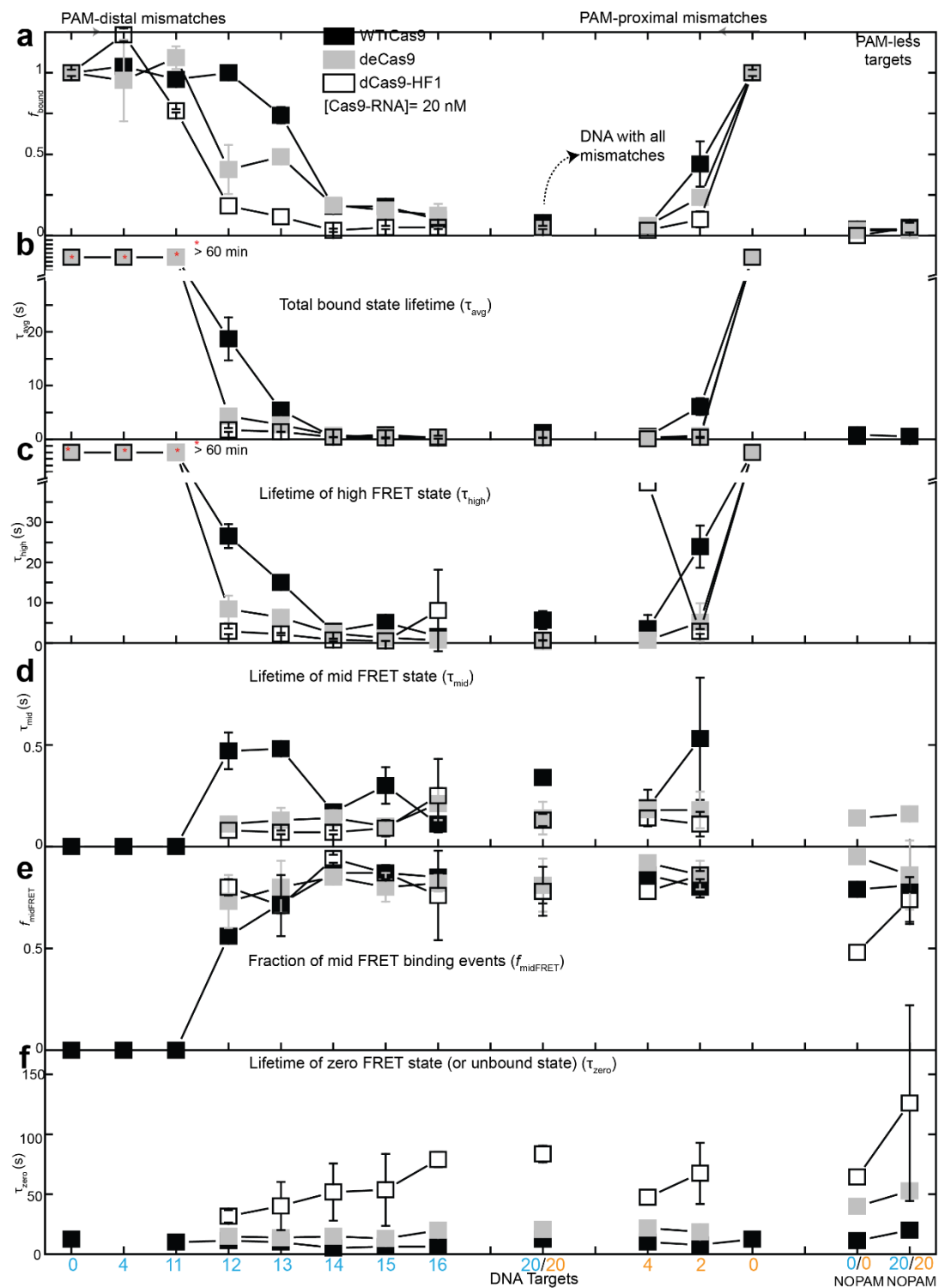
**(a)** deCas9. **(b)** dCas9-HF1. **(c)** WT Cas9. The number of PAM-distal mismatches ( $n_{PD}$ ) and PAM-proximal mismatches ( $n_{PP}$ ) are shown in cyan and orange, respectively.



**Figure 3.5 | Ultrastable binding of EngCas9-RNA to DNA.**

$E$  histograms before and >60 min after washing away free Cas9-RNA in solution obtained from smFRET DNA interrogation experiments. Like WT Cas9-RNA<sup>6,83</sup>, EngCas9-RNA remains near-irreversibly bound to the DNA targets if there are >8-9 PAM-proximal matches. **(a)** deCas9 for cognate DNA and DNA with  $n_{PD} = 11$ . **(b)** dCas9-HF1-RNA for cognate DNA and DNA with  $n_{PD} = 10$ . **(c-d)** Like WT Cas9-RNA<sup>6,83</sup>,

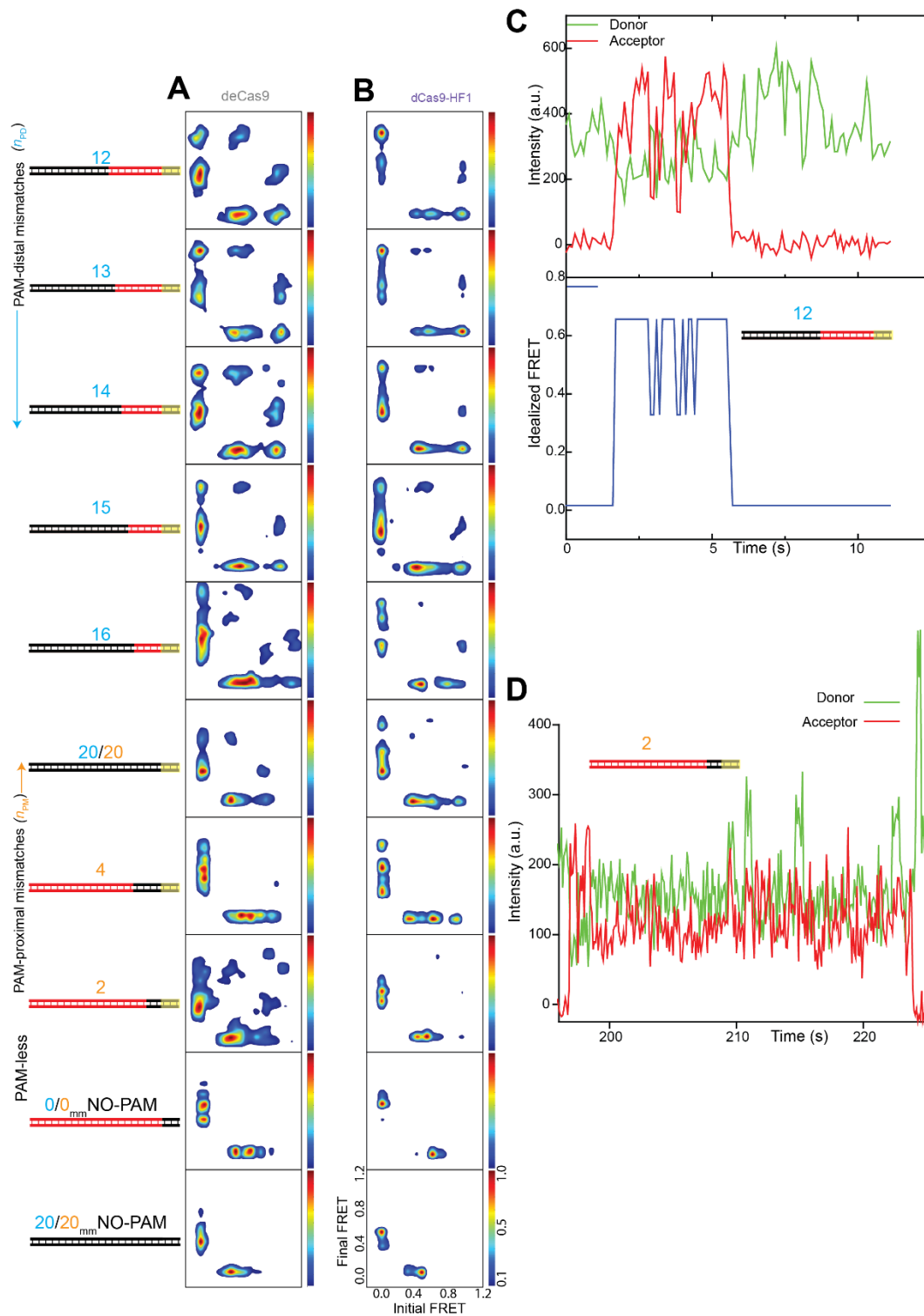
catalytically capable EngCas9s do not release cleaved DNA products. The release of cleaved products would have resulted in disappearance of fluorescent/FRET spots. **(c)** *E* histograms before and after >60 minutes after removing free eCas9-RNA and Cas9-HF1-RNA for cognate DNA. **(d)** Representative images showing no reduction in the number of spots. The number of PAM-distal mismatches ( $n_{PD}$ ) is shown in cyan.



**Figure 3.6 | Binding and kinetic parameters of DNA interrogation (20 nM Cas9-RNA).**

(a) Normalized fraction of DNA molecules bound with Cas9-RNA. Fractions were normalized relative to the bound fraction of cognate DNA for each Cas9-RNA. (b) Average bound state lifetime obtained from

dwelt times of  $E > 0.2$  states. **(c)** Average lifetime of high FRET ( $E > 0.6$ ) binding events ( $\tau_{\text{high}}$ ), obtained from the dwell times of  $E > 0.6$  states. **(d)** Average lifetime of mid FRET binding events ( $\tau_{\text{mid}}$ ) obtained from dwell times of mid FRET ( $E > 0.2$  &  $E < 0.6$ ) states. **(e)** Fraction of binding events with mid FRET. **(f)** Average unbound state lifetime obtained from the single-exponential fits to the distributions of dwell times of  $E < 0.2$  state. The number of PAM-distal mismatches ( $n_{\text{PD}}$ ) and PAM-proximal mismatches ( $n_{\text{PP}}$ ) are shown in cyan and orange, respectively. Data for WT Cas9-RNA is taken from our previous study<sup>83</sup>.



**Figure 3.7 | Transition density plots for smFRET DNA interrogation experiments.**

**(a-b)** Transition density plots pictorialize the relative number of transitions between different FRET

states, as identified by hidden Markov modeling. **(a;** deCas9 and **b;** dCas9-HF1) **(c-d)** A small fraction of

binding events showed rapid fluctuations between high and mid FRET states. The source of these fluctuations, also observed for WT Cas9<sup>83</sup>, is unknown and dwells in these states were not included in the lifetime calculation. The number of PAM-distal mismatches ( $n_{PD}$ ) and PAM-proximal mismatches ( $n_{PP}$ ) are shown in cyan and orange, respectively.

### 3.3.2 Cas9-RNA induced DNA unwinding

Genome wide characterization of EngCas9 cleavage specificity showed marked improvements over WT Cas9<sup>67,68</sup>. Yet, our binding study shows that the EngCas9s still stably bind to DNA with  $n_{PD}$  as large as 10. We reasoned that EngCas9s may differ from WT Cas9 in their mismatch dependence of DNA unwinding which, if promoted by annealing to guide-RNA, may be a determinant of cleavage action<sup>93</sup>. Therefore, we probed the internal unwinding of PAM-distal region of the protospacer by labeling the target and non-target strands with a donor and an acceptor, respectively, with 9 bp spacing (Figure 3.8). Labeling at these locations did not perturb cleavage (Figure 3.9). The DNA by itself showed high FRET ( $E \sim 0.75$ ) (Figure 3.10a-b), and upon addition of 100 nM WT dCas9-RNA to cognate DNA, we observed a shift to a stable low FRET state ( $E \sim 0.30$ ), likely because of DNA unwinding (Figure 3.10a-b). A similar FRET change was observed when the locations of the donor and acceptor were swapped (Figure 3.8b). A DNA with  $n_{PD} = 1$  showed a similarly stable unwinding but with a slightly higher  $E$  of  $\sim 0.35$  and occasional short-lived transitions to the high FRET state, likely due to one bp fewer unwinding and increased frequency to rewind, respectively. The trend continued upon increasing  $n_{PD}$  to 2 and 3;  $E$  value for the unwound state increased, and the relative population of the rewind state increased. With  $n_{PD} = 4$ , the unwound state is rarely populated. deCas9 and dCas9-HF1 showed a similar behavior but their unwound state population and lifetime were smaller and decreased more quickly with increasing  $n_{PD}$  (Figure 3.10b). Therefore, EngCas9s are more sensitive to PAM-distal mismatches in their ability to unwind DNA. The rewind state must still have at least 8 PAM-proximal bp unwound because we did not see stable binding with fewer PAM-proximal matches (Figure 3.1), and can have up to 16 PAM-proximal bp unwound.

We observed two distinct  $E$  populations even at a higher time resolution (35 ms) (Figure 3.11), indicating that Cas9-RNA induces primarily two states, unwound and rewound, without spending appreciable time in between. We used the Hidden Markov modeling analysis to segment single molecule time traces into two states (Figure 3.12). All Cas9s showed a reduction in the relative population and lifetime of the unwound state with increasing  $n_{PD}$  (Figure 3.10c-e). Combined with the gradual increase of  $E$  values for the unwound state (Figure 3.10b), we can conclude that PAM-distal mismatches reduce the time spent in the unwound state in addition to reducing the maximal extent of unwinding. For a given  $n_{PD}$ , EngCas9s showed lower occupancy and shorter lifetime of the unwound state (Figure 3.10c-e), suggesting that the mutations indeed destabilize the maximally unwound states. We observed a similar unwinding behavior from catalytically active Cas9 but  $E$  distribution was broader and state transitions were less frequent (Figure 3.13), suggesting a change in unwinding dynamics after cleavage, possibly due to disordered non-target strand<sup>93,94</sup>.

In order to capture the initial DNA unwinding event, we added labeled DNA to surface-immobilized Cas9-RNA molecules (Figure 3.10f) (Figure 3.14 and Figure 3.15). DNA binding is detected as a sudden appearance of fluorescence signal in a high  $E$  state, followed by a single step change to a low  $E$  (unwound state) (Figure 3.10g). Subsequently, we observed unwinding/rewinding dynamics similar to what we observed in the steady state (Figure 3.10g) (Figure 3.16 and Figure 3.17). The average dwell time of the initial high  $E$  state which we attribute to the time it takes to unwind DNA for the first time ( $\tau_{unwinding}$ ) remained constant ( $\sim 1$  s) for dCas9 when  $n_{PD}$  changed from 0 to 2. In contrast,  $\tau_{unwinding}$  increased  $\sim 9$  folds for EngCas9s when  $n_{PD}$  changed from 0 to 2 (Figure 3.10h), suggesting that EngCas9s take longer to unwind in the presence of PAM-distal matches.

### 3.3.3 DNA cleavage vs. mismatches

Next, we probed how mismatches influence target cleavage through modulating DNA unwinding/rewinding dynamics. Cleavage was more specific for EngCas9s because they cleaved only up to  $n_{PD} = 3$  compared to  $n_{PD} = 4$  for WT Cas9 (Figure 3.18a and Figure 3.9). EngCas9s did not release cleavage products under physiological conditions (Figure 3.5), as was the case for WT Cas9<sup>83</sup>. In terms of the cleavage reaction, EngCas9s were much slower than WT Cas9 for cognate or PAM-distal mismatched DNA (Figure 3.18b-c) but were faster for PAM-proximal mismatched DNA (Figure 3.18a). Overall, we observed a clear correlation between the cleavage time scale,  $\tau_{\text{cleavage}}$ , and DNA unwinding signatures;  $\tau_{\text{cleavage}}$  decreased with increasing population (Figure 3.18d) and lifetime (Figure 3.18e) of the unwound state. Therefore, cleavage must proceed from the unwound state.

HNH and RuvC nuclease domains cleave target and non-target strand respectively. Additional 3'-5' exonuclease activity of RuvC<sup>4,93,95</sup>, which normally causes multiple bands of cleaved non-target strand, decreased with increasing  $n_{PD}$ , and more efficiently so for EngCas9s (Figure 3.9).

Armed with the kinetic information on DNA binding, interrogation, unwinding/rewinding and cleavage, we can clarify in which reaction steps sequence mismatches and Cas9 mutations exert their influence. Because cleavage likely occurs from the unwound state, we estimated the lower limits of intrinsic rate of cleavage ( $k_{c,int}$ ) from the apparent cleavage time  $\tau_{\text{cleavage}}$  and unwound fraction by fitting  $\tau_{\text{cleavage}}$  vs unwound fraction using  $\tau_{\text{cleavage}} = 1/([\text{unwound fraction}] * k_{c,int} + C)$ . Because a single value of  $k_{c,int}$  was able to fit the data for each Cas9, we conclude that changes in  $\tau_{\text{cleavage}}$  caused by mismatches can be attributed to changes in the unwound fraction (Figure 3.18d). Therefore, although DNA molecules with different  $n_{PD}$  values show different extents of unwinding, their unwound states have similar intrinsic cleavage rates.  $k_{c,int}$  was  $\sim 0.67 \text{ s}^{-1}$  for WT Cas9 but decreased to 0.025-0.015  $\text{s}^{-1}$  for EngCas9s.



Inserting two single bp mismatches in the PAM-distal region gave a predominantly rewound state with only brief excursions to the unwound state for all Cas9s (Figure 3.19a), indicating that there is no difference in unwound fraction among them. Nevertheless, WT Cas9 cleaved the DNA whereas EngCas9s did not, further confirming the much higher  $k_{c,int}$  for WT Cas9 which in turn allows cleavage even from a transiently populated unwound state, likely contributing to higher off-target cleavage yields (Figure 3.19b).

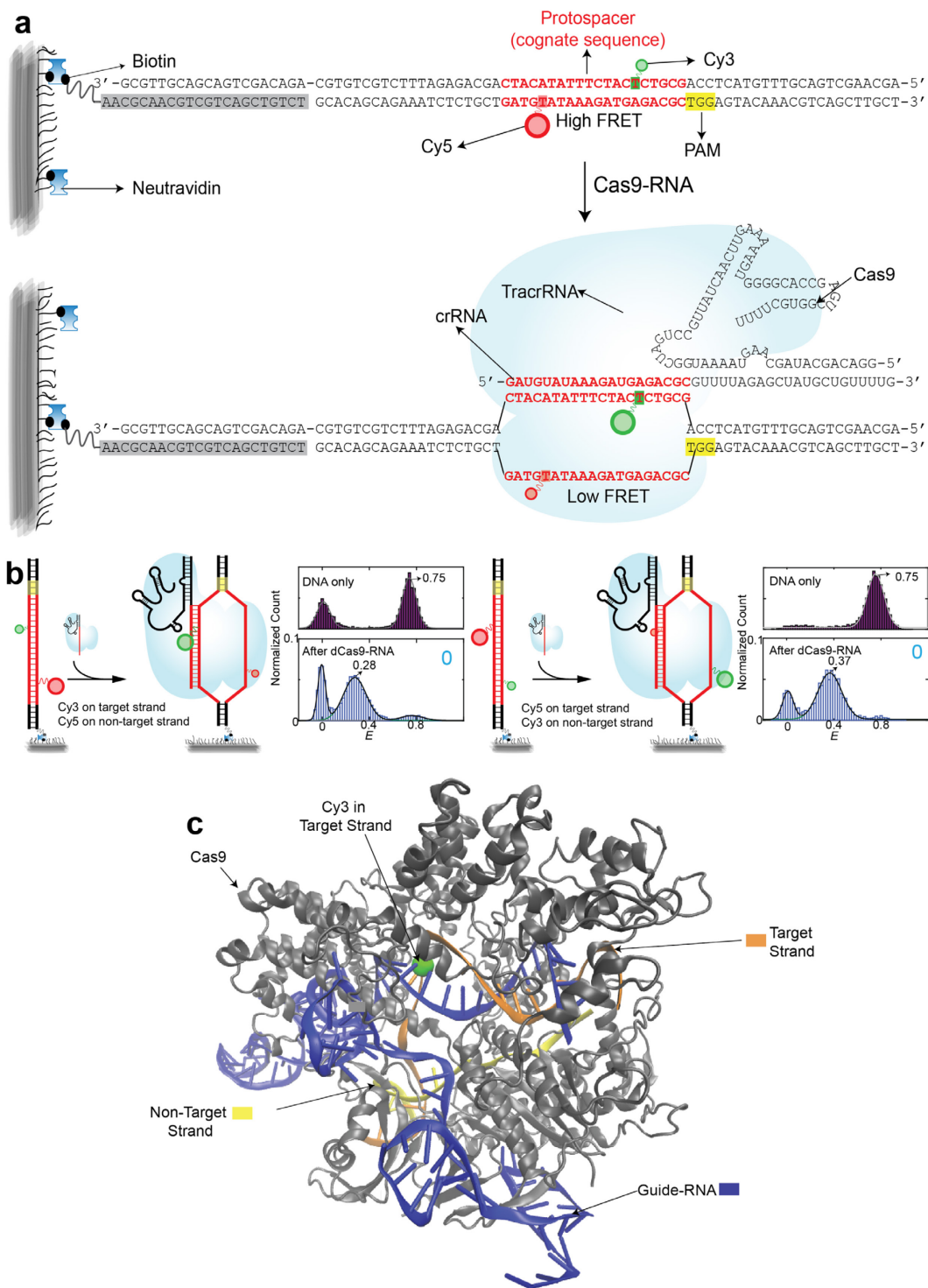
### 3.3.4 Mechanism of mismatch sensitivity of EngCas9s

The changing balance between unwinding and rewinding caused by mismatches reflect the energetic competition between target strand annealing with RNA vs with non-target strand. Mismatches would favor rewinding to recover the parental duplex by disrupting RNA/DNA duplex. We found that mutations in EngCas9s further promote rewinding, helping improve cleavage specificity. In order to gain additional insights into how EngCas9s more readily rewind in the presence of mismatches, we disrupted the parental DNA duplex by pre-unwinding the PAM-distal mismatched portion (Figure 3.19c). Pre-unwinding shifted the balance toward unwinding for WT Cas9 and Ca9-HF1 (Figure 3.19d), possibly because the residues that sequester the non-target strand, mutated in eCas9, are still intact. The result also explains why pre-unwinding allows rapid cleavage of mismatched DNA<sup>96</sup>. In contrast, eCas9 failed to shift  $E$  distribution toward a low value in the presence of PAM-distal mismatches (Figure 3.19d and Figure 3.20), likely because the residues that help sequester the unwound non-template strand are mutated (Figure 3.1b and Figure 3.21).

### 3.3.5 PAM-proximal DNA unwinding

To investigate if EngCas9-RNA mutations also affect PAM-proximal DNA unwinding we designed a DNA construct with a donor and an acceptor at the PAM proximal site separated by 12 bp (Figure 3.19e and Figure 3.22). Cas9-RNA binding to cognate DNA shifted  $E$  from 0.55 to 0.35 for all Cas9s, indicating a stable DNA unwinding at PAM-proximal site. With  $n_{pp} = 2$ , the fraction of unwound

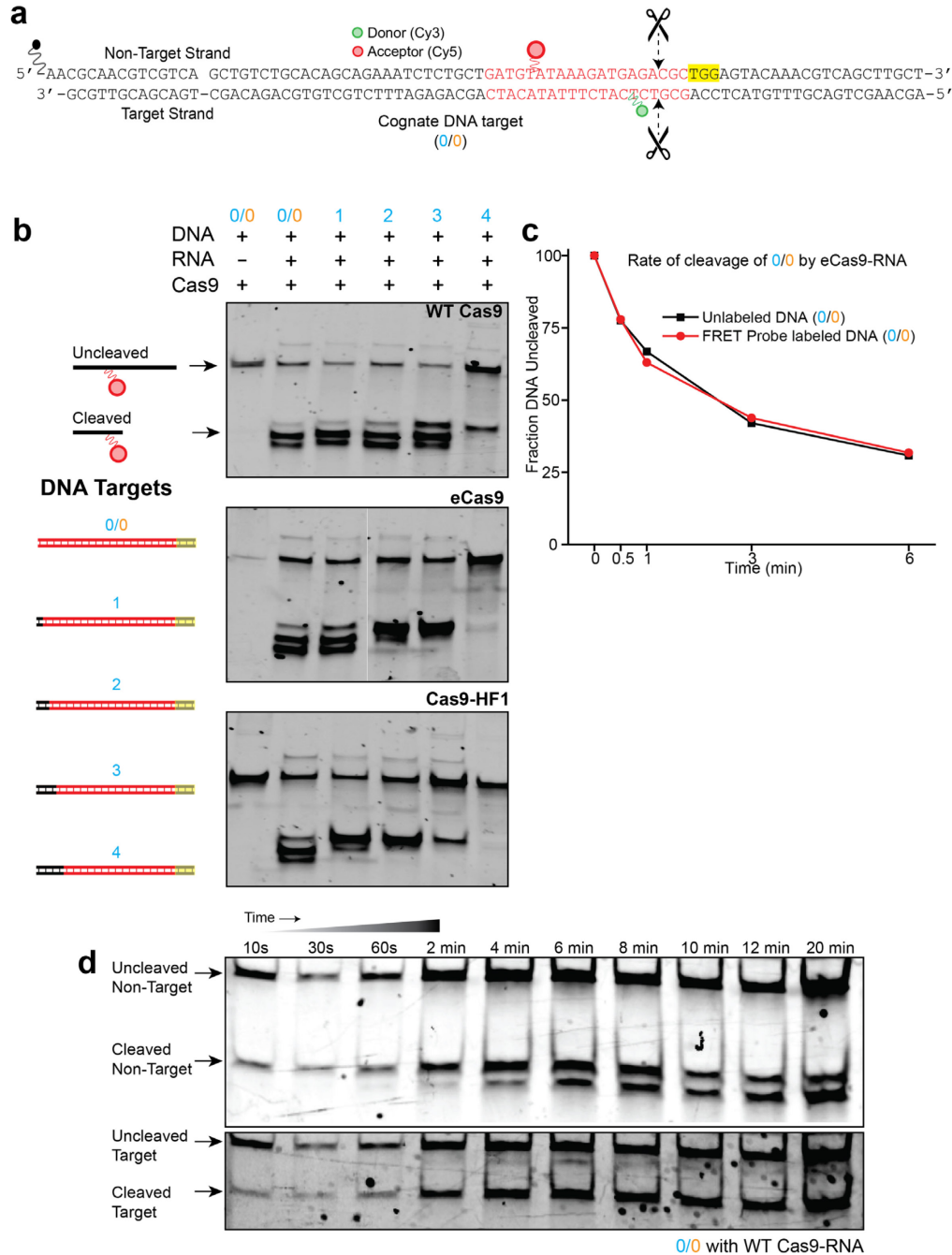
population decreased to 50% for dCas9, 40% for deCas9 and < 5% for dCas9-HF1 (Figure 3.19f), mirroring the decreases in  $f_{\text{bound}}$ . However, this decrease did not correlate with cleavage efficiency of DNA target with  $n_{\text{pp}} = 2$ , where EngCas9 were more efficient than WT Cas9 (Figure 3.18a). Therefore, EngCas9s can cleave DNA better once the initial barrier caused by PAM-proximal mismatches is overcome.



**Figure 3.8 | FRET probe locations for smFRET DNA unwinding experiments.**

(a) Schematic of Cas9-RNA-DNA complex. The hybridized crRNA and tracrRNA are referred to as guide-RNA. Sequences in red denote guide sequence of the guide-RNA and the matching sequence of the

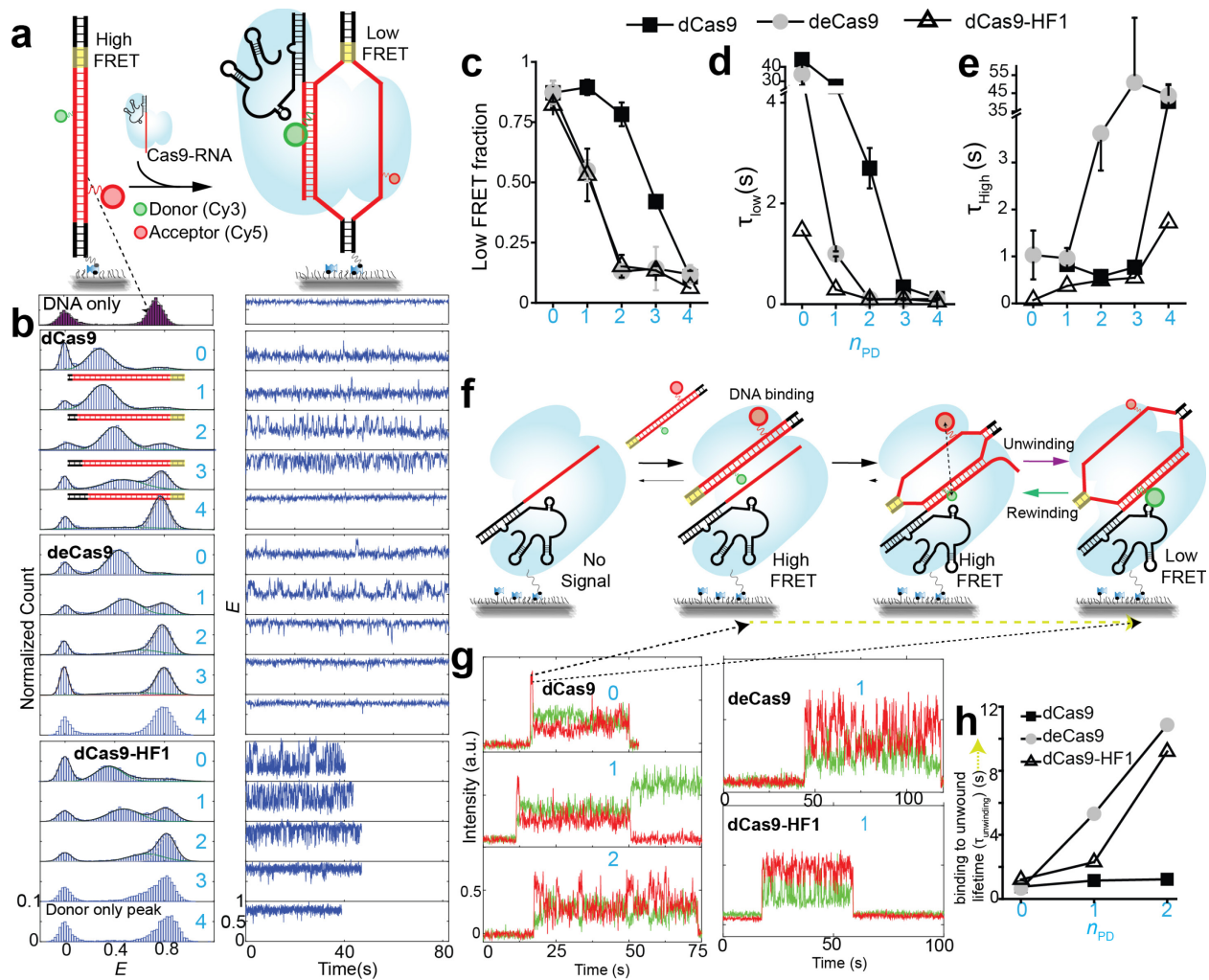
DNA. The target strand is complementary to the guide-RNA. The non-target strand contains the PAM (5'-NGG-3'). A 22 nt biotin-labeled adaptor strand was used for surface immobilization of DNA and is highlighted in grey. The donor is attached to the target strand in the 6<sup>th</sup> position from PAM and the acceptor is attached to the non-target strand in the 16<sup>th</sup> position from PAM. Formation of Cas9-RNA-DNA complex leads to the unwinding of DNA target, causing a decrease in *E*. **(b)** Similar FRET changes upon Cas9-RNA induced unwinding was observed when the donor and acceptor positions were swapped. **(c)** Probe locations in the Cas9-RNA-DNA complex (PDB ID: 5F9R<sup>93</sup>).



**Figure 3.9 | Fluorescent labeling for smFRET DNA unwinding experiments does not affect cleavage.**

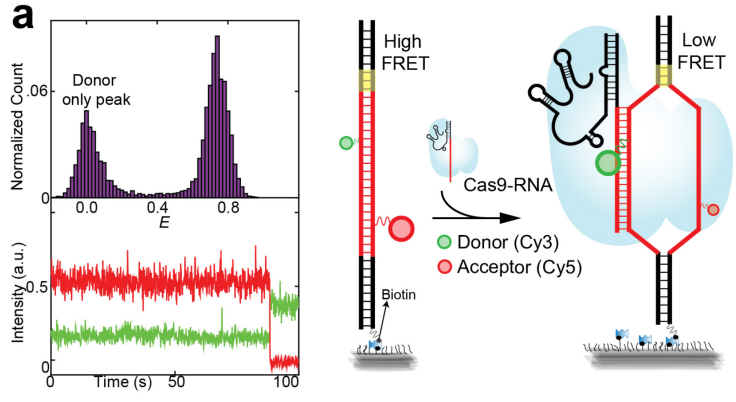
(a) Schematic of cognate DNA target used in smFRET DNA unwinding experiments. Protospacer is highlighted in red and PAM in yellow. Also indicated are Cas9-RNA induced cleavage positions. (b) Cleavage of the DNA targets analyzed by denaturing polyacrylamide gel electrophoresis and Cy5

imaging. Multiple cleaved products for the non-target strand likely arises from 3'-5' additional exonuclease activity of RuvC domain<sup>4</sup>. [DNA] =5 nM and [Cas9-RNA] =100 nM in 10  $\mu$ l reaction. DNA was incubated with Cas9-RNA for ~75 minutes before being denatured by formamide loading buffer and heating at 95 °C for 10 minutes. The denatured sample products were then resolved using 15% polyacrylamide denaturing gel electrophoresis and imaged via Cy5 fluorescence. **(c)** Time courses of cognate DNA cleavage by eCas9-RNA with and without fluorescent probes (Cy3 and Cy5 pair) on the DNA show that labeling does not affect cleavage kinetics. Aliquots from a running reaction were taken at different time points and analyzed by PAGE. DNA targets were radio-labeled at the 5' end of the target strand with <sup>32</sup>P via a T4 polynucleotide kinase reaction for visualization. **(d)** Time evolution of multiple cleavage products of non-target strand due to additional 3'-5' exonuclease activity of RuvC. HNH has no exonuclease activity and consequently does not result in cleaved products of different sizes. The number of PAM-distal mismatches ( $n_{PD}$ ) is shown in cyan.

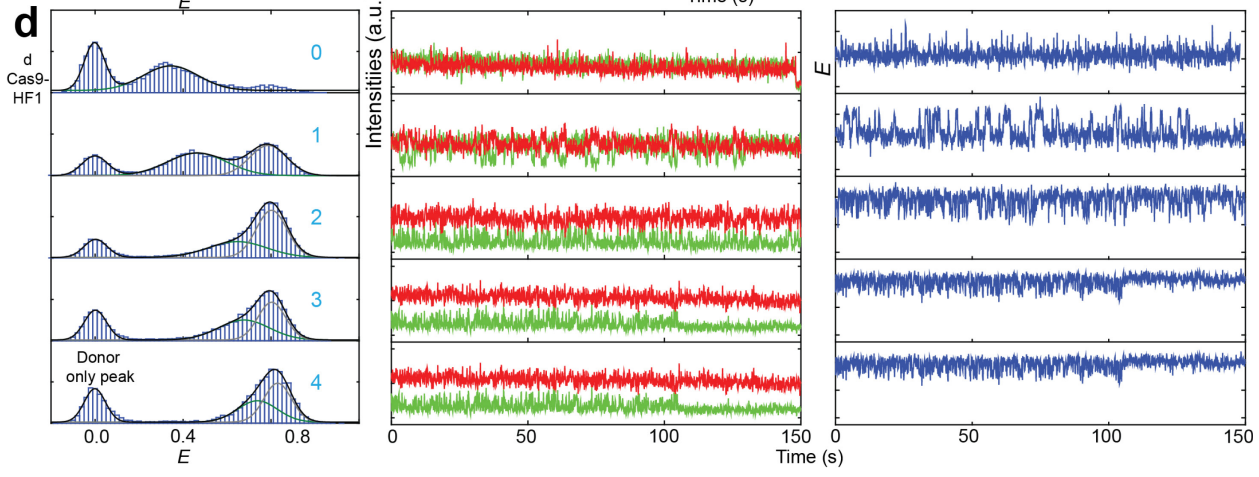
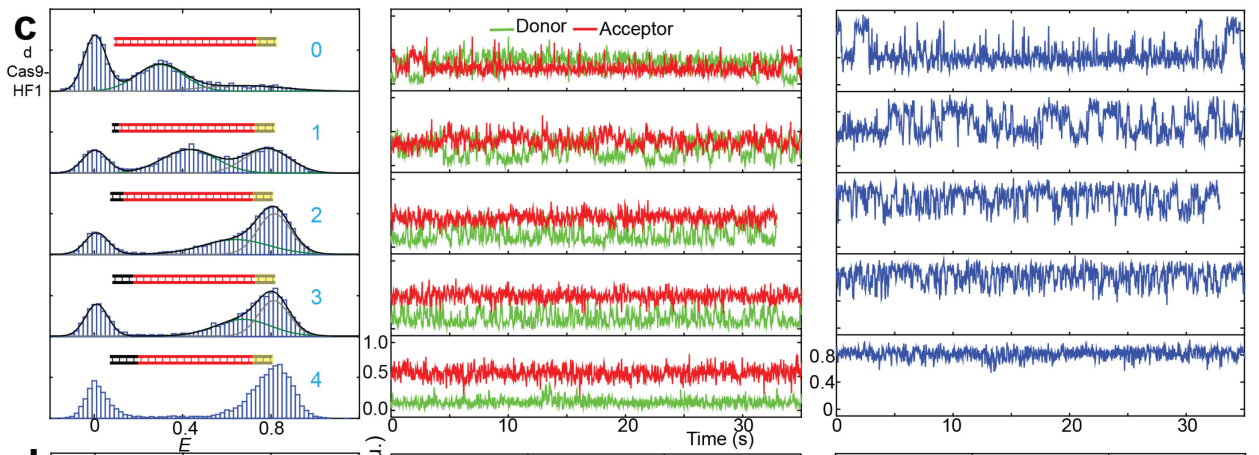
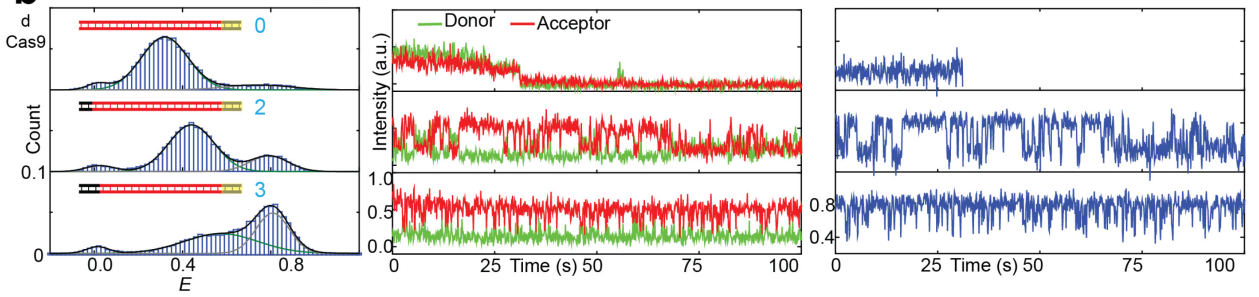


**Figure 3.10 | Internal DNA unwinding/rewinding dynamics modulated by mismatches and Cas9 mutations.**

(a) Schematic of smFRET assay for Cas9-RNA induced unwinding of surface-tethered DNA. (b)  $E$  histograms for  $n_{PD} = 0, 1, 2, 3$  and 4 (cyan numbering) and representative time traces. (c-e) Relative population of the unwound state, average lifetime of the unwound state and average lifetime of the rewound state vs.  $n_{PD}$ . (f) Schematic of smFRET assay for DNA unwinding by surface-tethered Cas9-RNA. (g) Representative single molecule fluorescence intensity time traces of donor (green) and acceptor (red) show abrupt signal increase upon DNA binding followed by FRET changes. (h) Time taken to go from initial binding to its first unwound state configuration ( $\tau_{unwinding}$ ) as denoted by a dark yellow line (f). Error bars represent standard deviation (s.d.) from  $n=2$ ,  $n=1$  in absence of error bar.



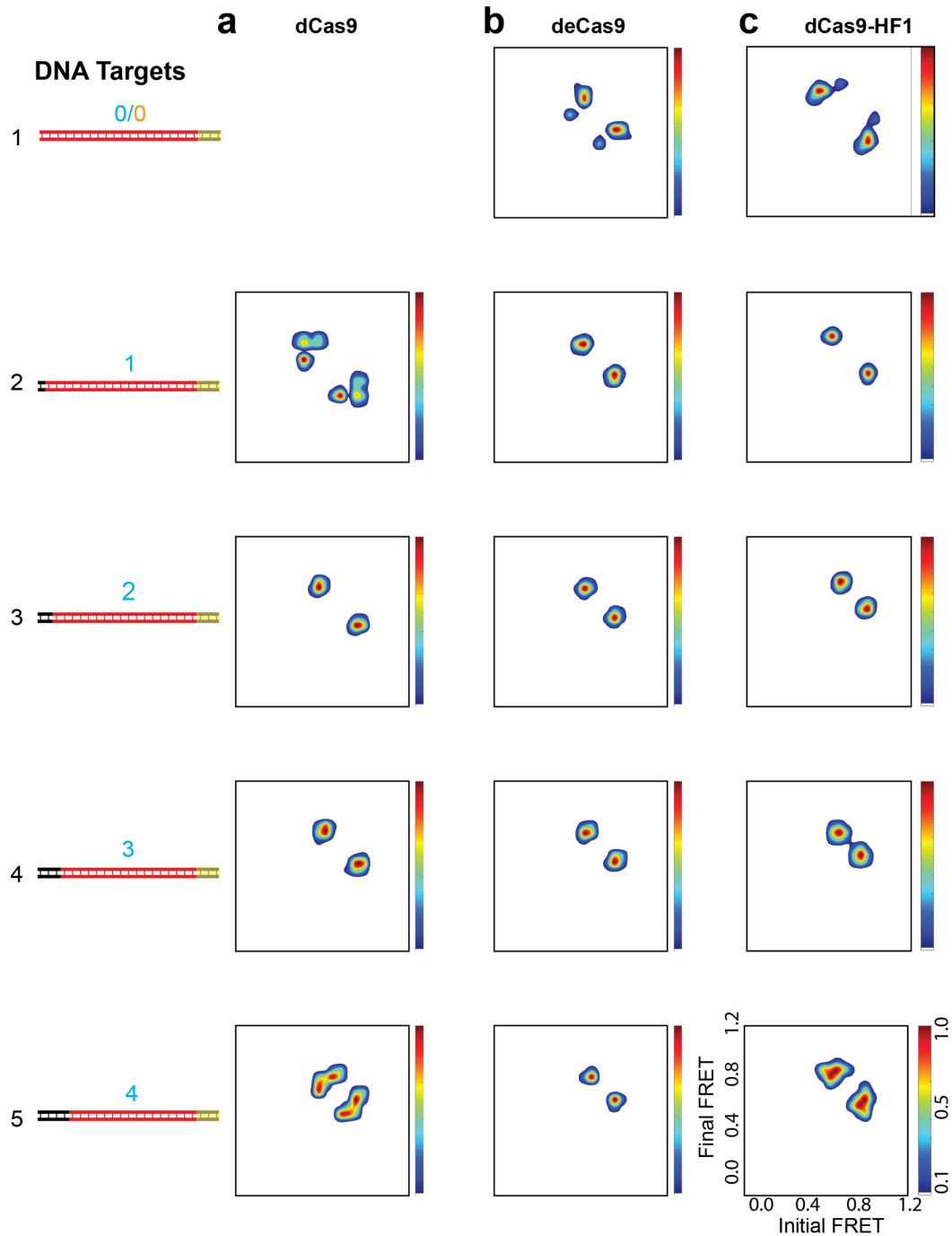
**b** 5 nM of dCas9-RNA for these experiments. 100 nM Cas9-RNA is used for all other DNA unwinding smFRET experiments





**Figure 3.11 | smFRET DNA unwinding experiments at different Cas9-RNA concentrations and different frame rates of image acquisition.**

**(a)** smFRET assay to investigate the Cas9-RNA induced DNA unwinding as described in Figure 3.10a. **(b-d)**  $E$  histograms (left) of DNA targets vs  $n_{PD}$  and their representative single molecule time traces of donor and acceptor intensities (middle) and  $E$  values (right). **(b)** DNA unwinding smFRET experiments performed at 5 nM of dCas9-RNA produce results similar to those observed with 100 nM Cas9-RNA. We used 100 nM for all unwinding experiments reported elsewhere in this study. **(c-d)** Frame rate of image acquisition for these experiments = 35 ms **(c)** and 100 ms **(d)**. Both experiments performed with dCas9-HF1 show similar  $E$  distributions. Although the overall signal was noisier for 35 ms data, fast transitions are better resolved.



**Figure 3.12 | Transition density plots for smFRET DNA unwinding experiments.**

Transition density plots pictorialize the relative number of transitions between different FRET states, as identified by hidden Markov modeling<sup>90</sup>. **(a)** dCas9 **(b)** deCas9 **(c)** dCas9-HF1. Number of PAM-distal mismatches ( $n_{PD}$ ) is shown in cyan.

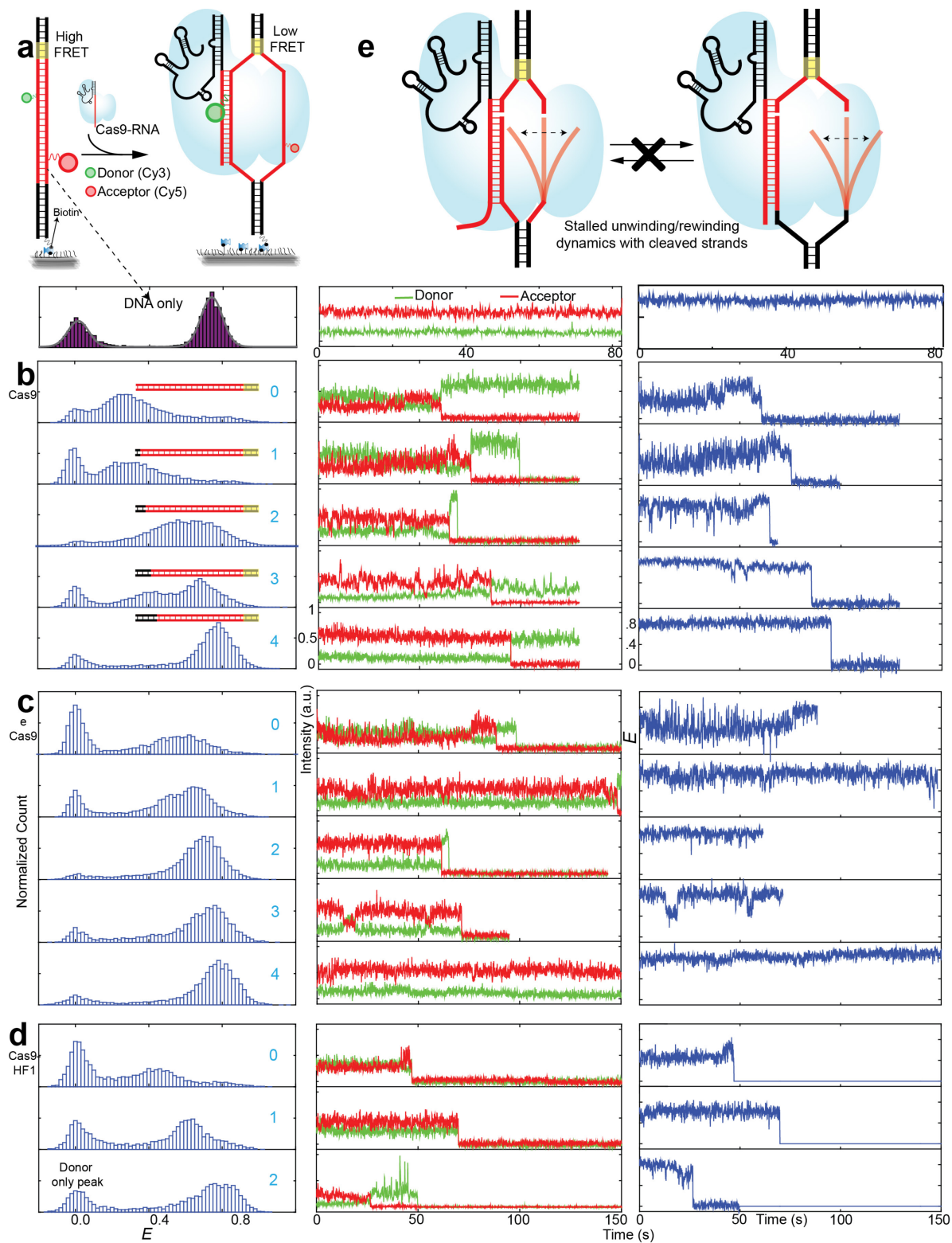


Figure 3.13 | smFRET DNA unwinding experiments using catalytically active Cas9-RNA.

**(a)** smFRET assay to investigate the Cas9-RNA induced DNA unwinding as described in Figure 3.10a. smFRET data of DNA target only are shown. **(b-d)**  $E$  histograms vs  $n_{PD}$  (left) and their representative time traces of donor and acceptor intensities (middle) and  $E$  values (right) for **(b)** dCas9 **(c)** deCas9 **(d)** dCas9-HF1. **(e)** Possible scenario of altered DNA unwinding/rewinding dynamics with cleaved strands in Cas9-RNA-DNA. The number of PAM-distal mismatches ( $n_{PD}$ ) is shown in cyan.

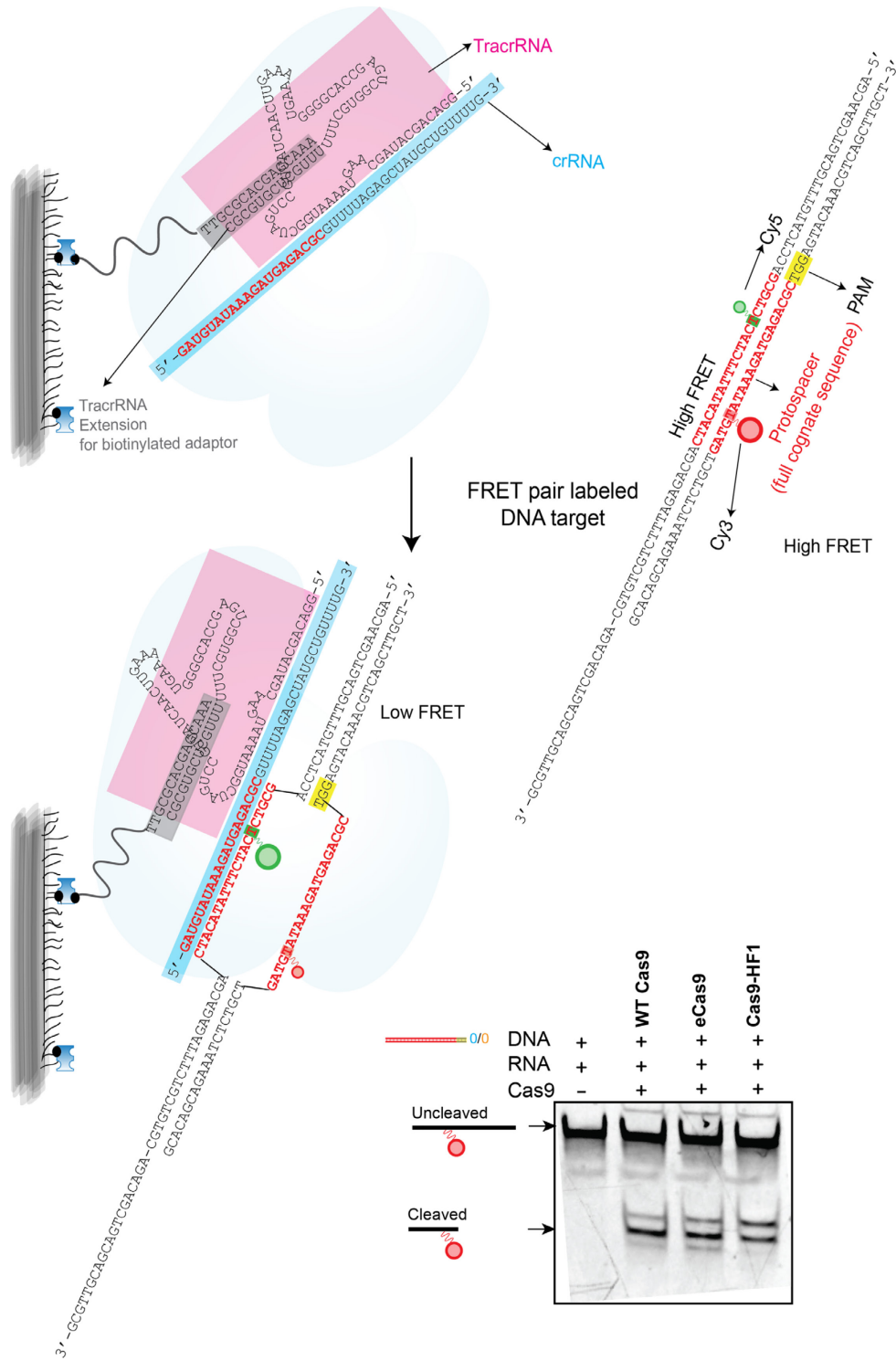
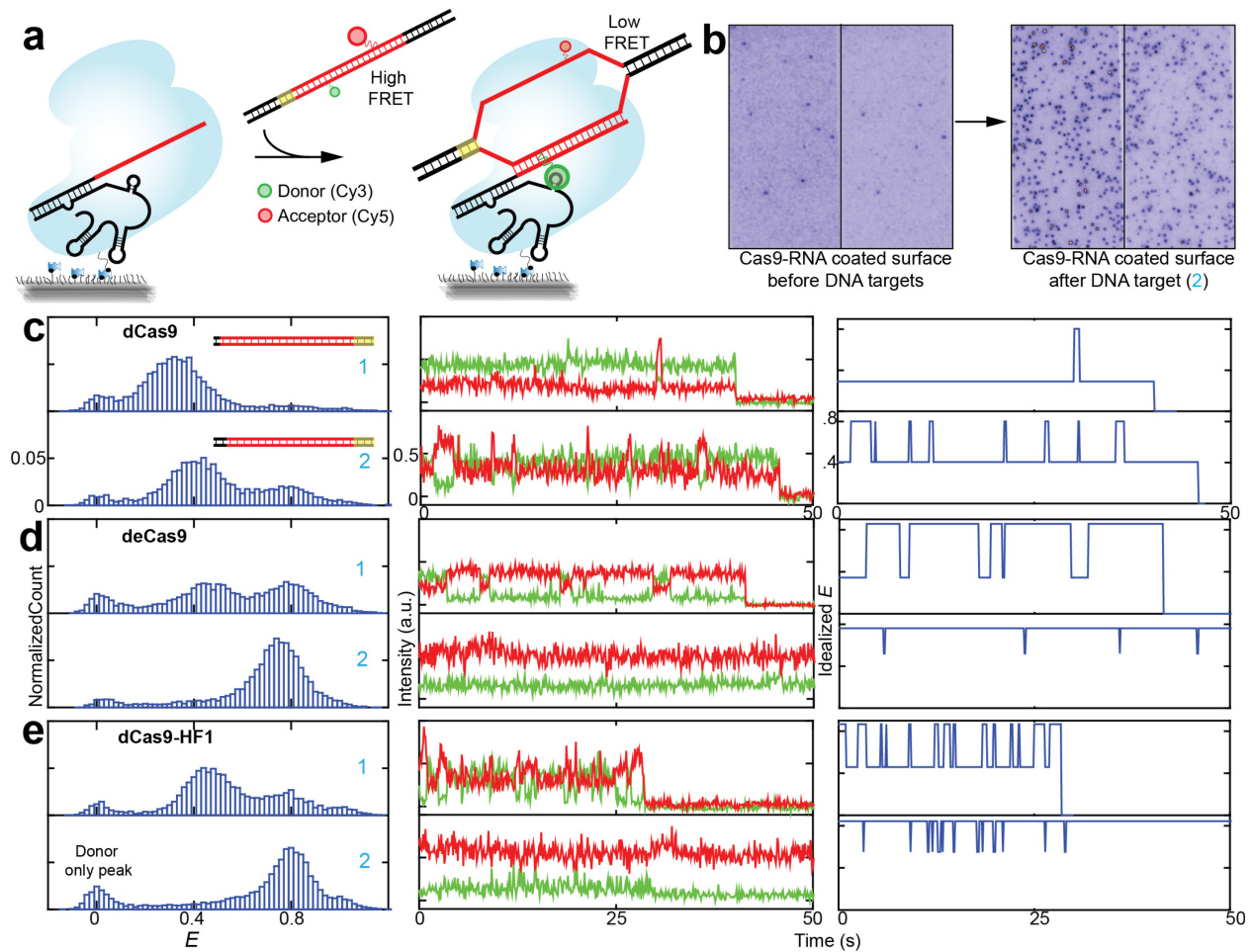


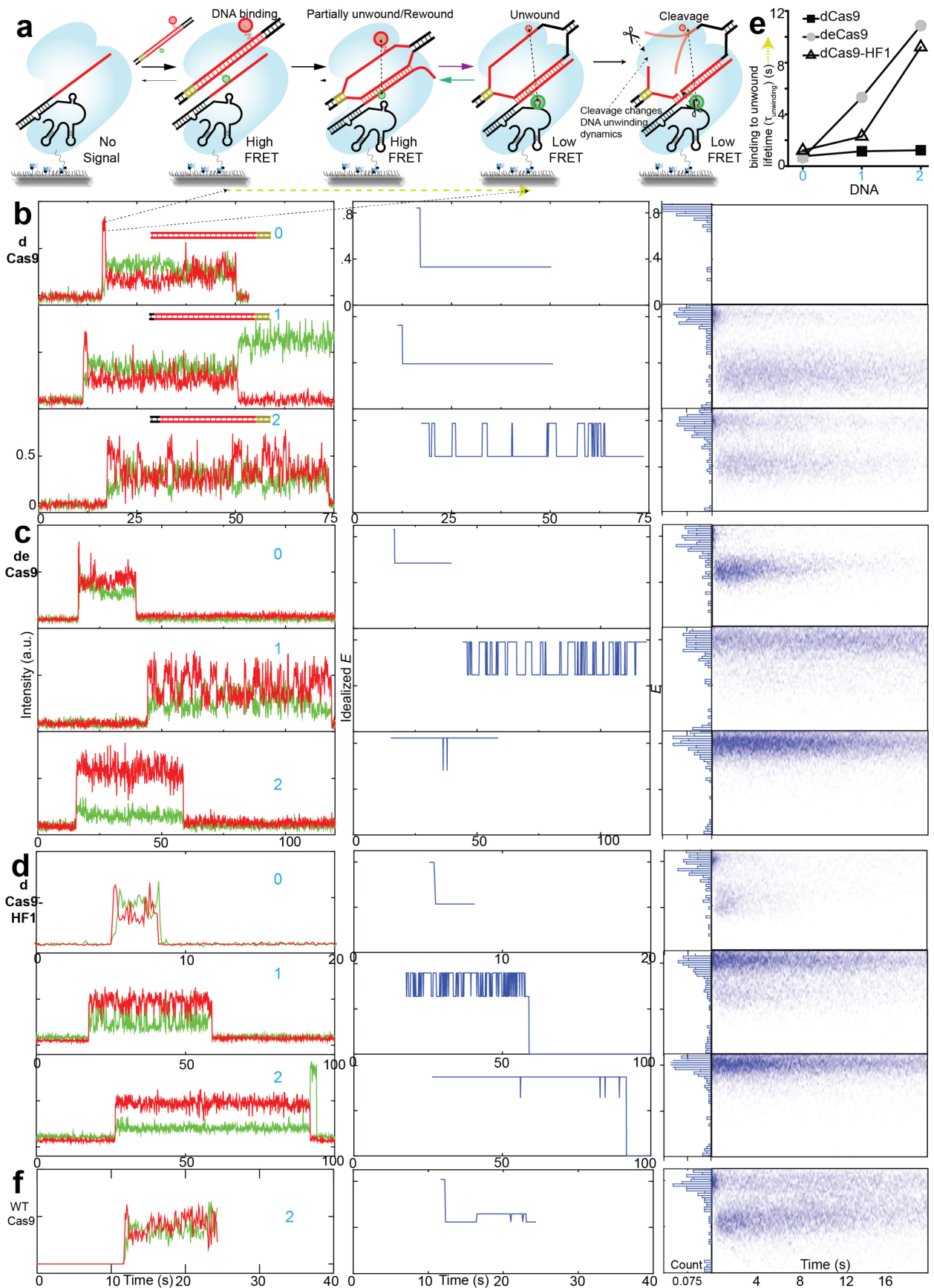
Figure 3.14 | Schematic of smFRET DNA unwinding experiments using surface-tethered Cas9-RNA.

The tracrRNA with the indicated 3' extension for the biotinylated DNA adaptor did not affect Cas9-RNA activity as shown by cleavage products analyzed by 15% denaturing polyacrylamide gel electrophoresis and Cy5 imaging.



**Figure 3.15 | DNA unwinding dynamics with surface immobilized Cas9-RNA.**

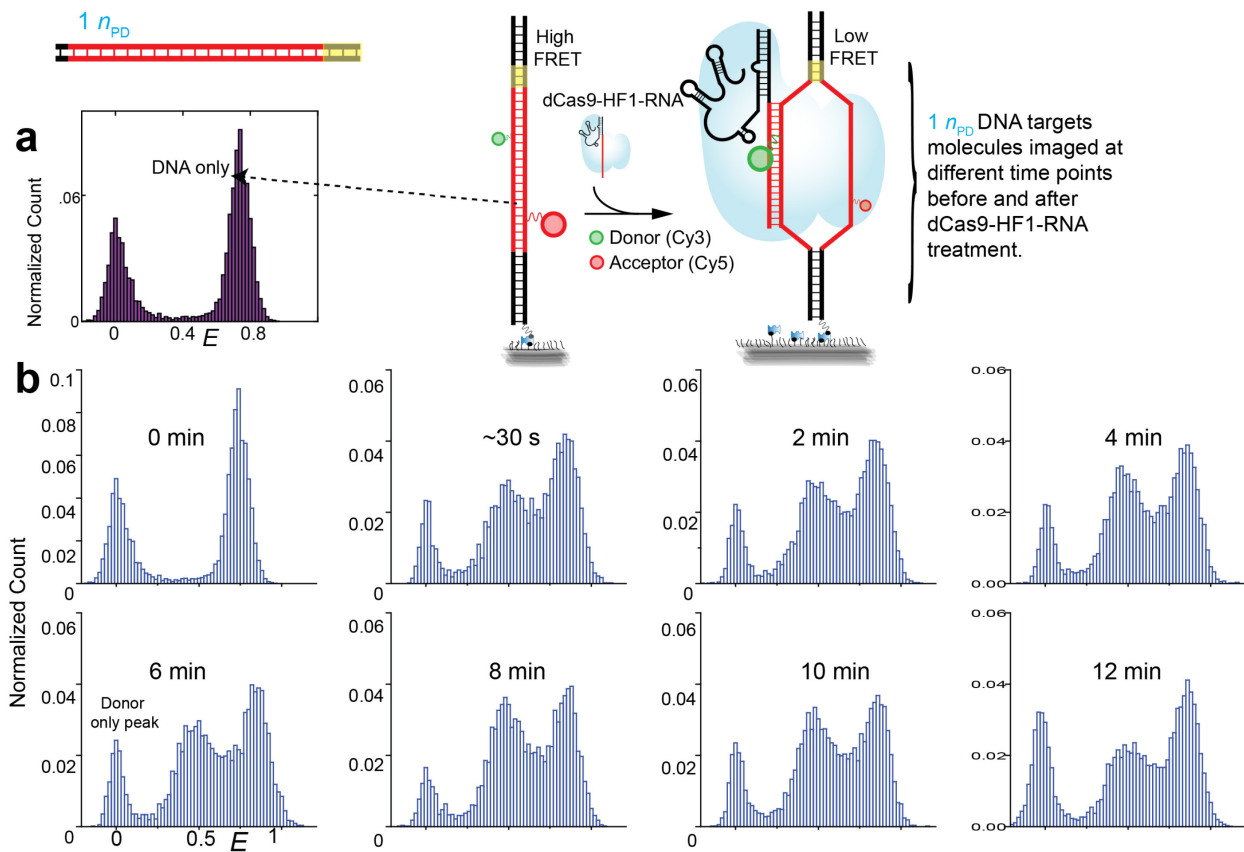
(a) smFRET DNA unwinding experiments as described in Figure 3.10f but with the Cas9-RNA tethered to the surface instead of DNA. (b) Representative images of the donor and acceptor imaging channels before and after addition of labeled DNA with  $n_{PD}=2$ . (c, d, e)  $E$  histograms vs  $n_{PD}$  obtained  $>\sim 10$  minutes after addition of labeled DNA (left), and the corresponding representative single molecule time traces of donor and acceptor intensities (middle) and their idealized  $E$  values (right).



**Figure 3.16 | DNA unwinding dynamics upon initial formation of Cas9-RNA-DNA complex.**

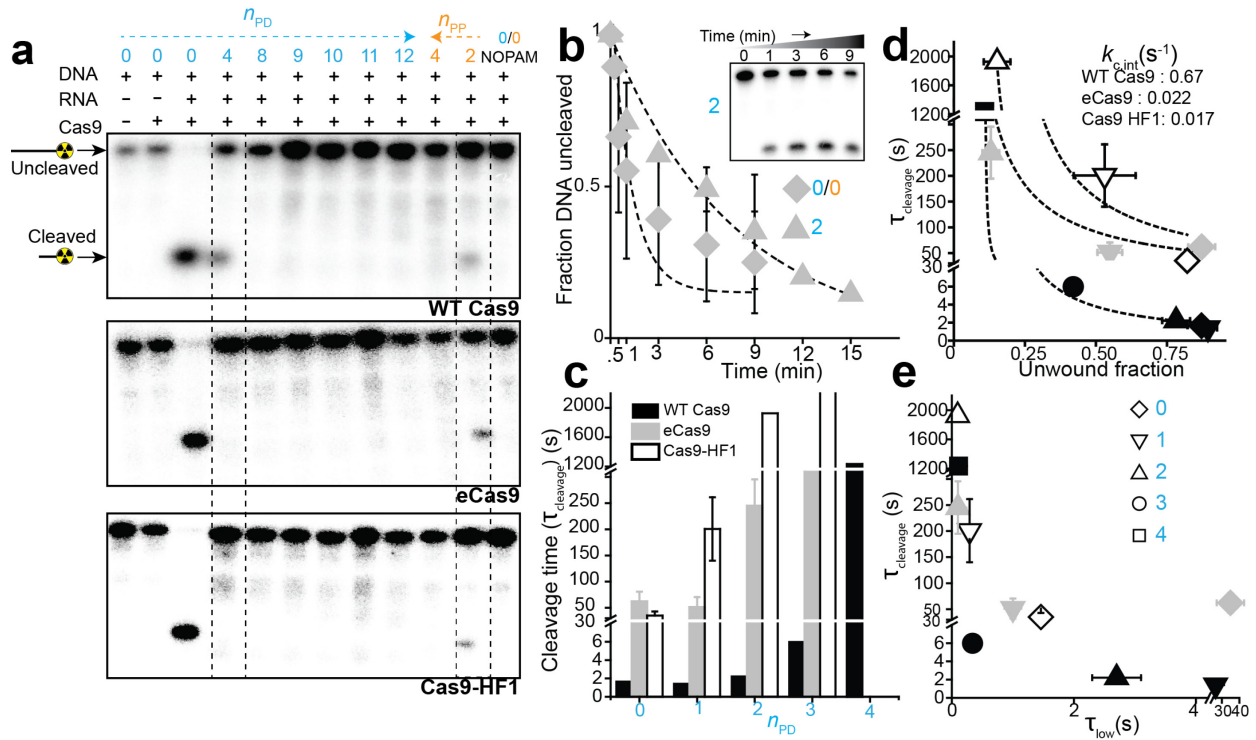
(a) Schematic of experiment to observe the DNA unwinding dynamics during very first moments of Cas9-RNA-DNA complex formation. The Cas9-RNA immobilized flow chamber surface was imaged during the addition of the FRET pair labeled DNA target to the flow chamber. Results with catalytically dead Cas9s are in: (b) dCas9 (c) deCas9 (d) dCas9-HF1. Representative single molecule time traces of donor and acceptor intensities (left) and their idealized  $E$  values (middle).  $E$  density maps (right) show the real-time evolution of different FRET states synchronized to the moment of DNA binding to Cas9-RNA. Histograms (in border blue bars) show  $E$  distribution at the moment of binding (time = 0). (e) Average time it takes for catalytically dead Cas9-RNA to go from initial binding (no unwinding) to its first unwound state configuration ( $\tau_{\text{unwinding}}$ ) is described by a dark yellow line in the schematic shown in (a). (f) Data obtained using catalytically active WT Cas9. The number of PAM-distal mismatches ( $n_{\text{PD}}$ ) is shown in cyan.





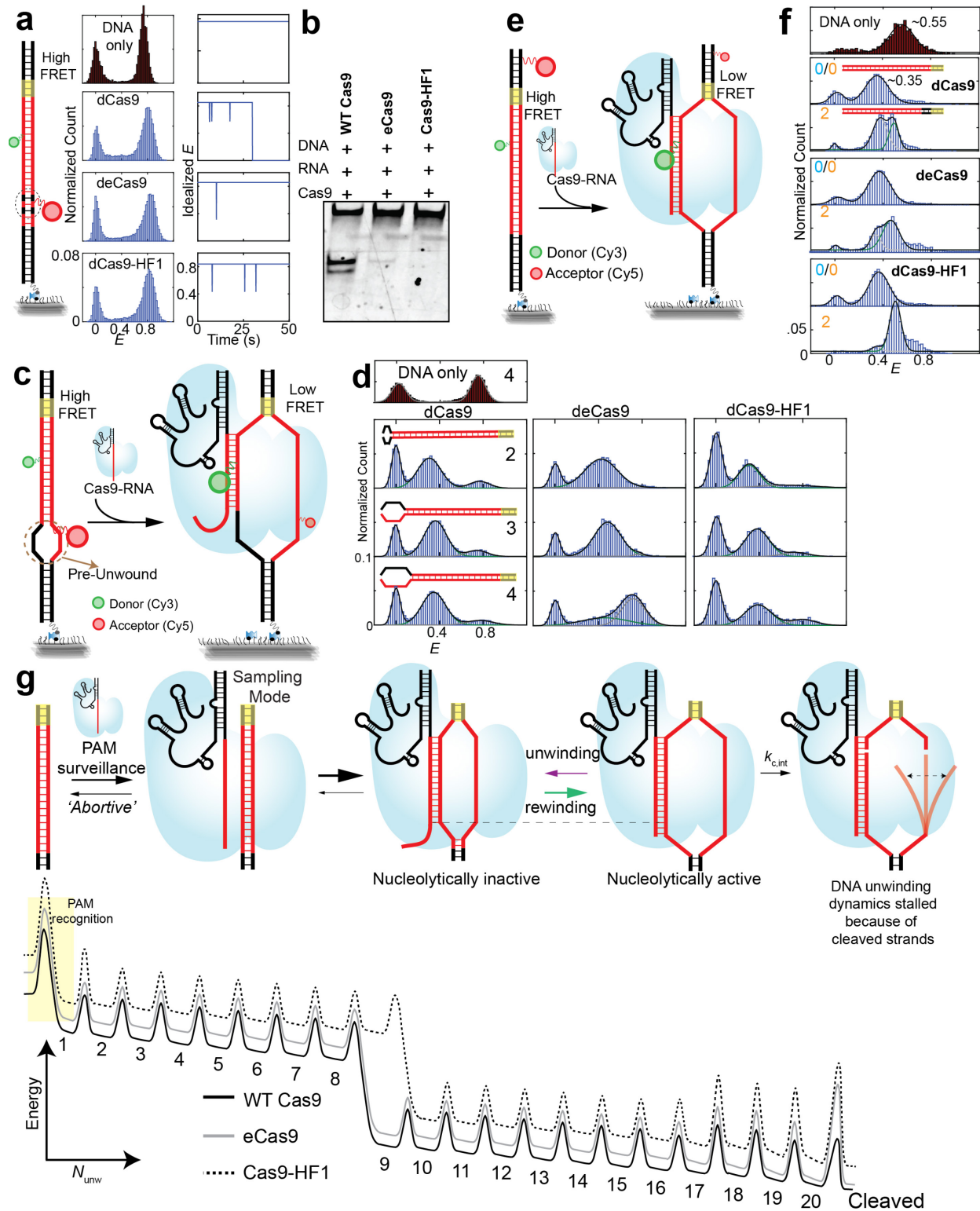
**Figure 3.17 | Appearance of unwound state over time (dCas9-HF1-RNA on DNA with  $n_{PD}=1$ ).**

**(a)** Schematics of smFRET DNA unwinding experiments as described in Figure 3.10a.  $E$  Histogram of DNA target only ( $n_{PD}=1$ ) is shown. **(b)**  $E$  histograms of 1  $n_{PD}$  DNA target before and at different time points after adding 100 nM dCas9-HF1-RNA. Rapid appearance of a peak near  $E=0.5$  indicates that Cas9-RNA induced DNA unwinding occurs as soon as Cas9-RNA-DNA complex forms. The number of PAM-distal mismatches ( $n_{PD}$ ) is shown in cyan.



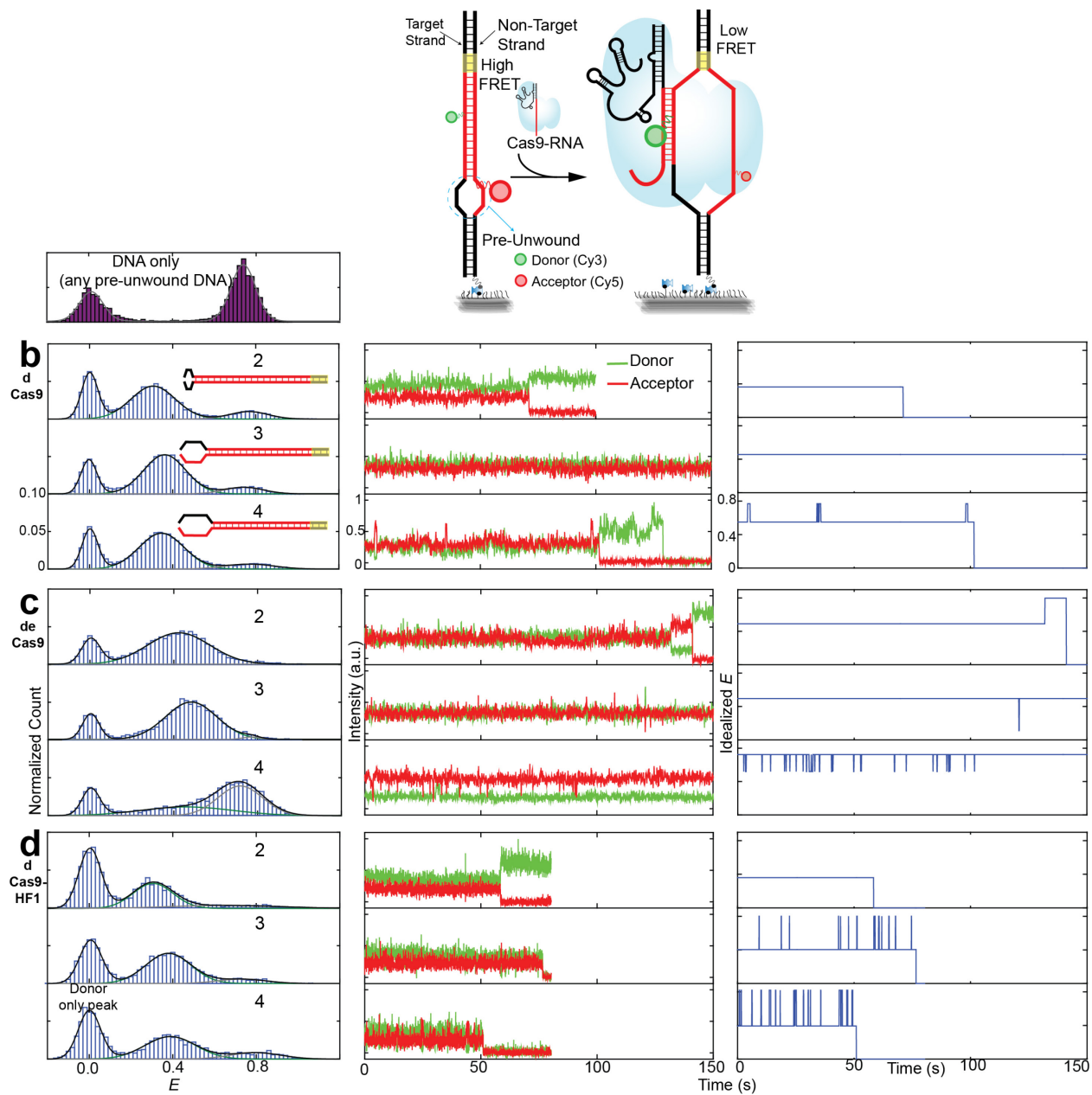
**Figure 3.18 | Cleavage vs mismatches and its relation with DNA unwinding.**

(a) Cas9-RNA induced cleavage pattern analyzed by 15% denaturing polyacrylamide gel electrophoresis of target strand 5'-labeled with  $^{32}\text{P}$ . Dashed boxes highlight DNA targets with  $n_{PD}=4$  and  $n_{PP}=2$  that show differences between EngCas9s and WT Cas9. (b) Fraction of uncleaved DNA vs time for eCas9 as a function of time and single exponential decay fits to obtain cleavage time  $\tau_{\text{cleavage}}$ . Inset shows a representative gel image.  $[\text{DNA}] = 1 \text{ nM}$ .  $[\text{Cas9-RNA}] = 100 \text{ nM}$ . (c)  $\tau_{\text{cleavage}}$  vs  $n_{PD}$ . Data for WT Cas9-RNA is taken from a previous study<sup>96</sup>. (d)  $\tau_{\text{cleavage}}$  vs unwound fraction as shown in Figure 3.10c. Fits are made to determine  $k_{c,\text{int}}$  (see methods). (e)  $\tau_{\text{cleavage}}$  vs. unwound state lifetime as shown in Figure 3.10d.  $n_{PD}$  and  $n_{PP}$  are shown in cyan and orange, respectively. Error bars represent standard deviation (s.d.) from  $n=2$  or 3.  $n=1$  in absence of error bar.



**Figure 3.19 | Cas9-RNA induced unwinding of various DNA and mechanisms of increased specificity by EngCas9s.**

**(a)** Unwinding and cleavage of DNA with internal single base mismatches at positions 16<sup>th</sup> and 18<sup>th</sup>. *E* histograms were obtained at [Cas9-RNA] =100 nM or in its absence. Idealized *E* traces show transient unwinding observed in a subset of molecules. **(b)** Gel image of non-target strand shows that WT Cas9, but not EngCas9s, cleave the DNA. **(c)** Schematics of Cas9-RNA induced unwinding of pre-unwound DNA. **(d)** *E* histograms obtained at [Cas9-RNA] =100 nM or in its absence. Black numbers denote the number of base pairs pre-unwound and are mismatched. **(e)** Schematics of Cas9-RNA induced unwinding of PAM-proximal region. **(f)** *E* histograms obtained at [Cas9-RNA] =100 nM or in its absence.  $n_{PP}$  is shown in orange digits. **(g)** Proposed model of DNA targeting, unwinding, rewinding, and cleavage (top). Energy diagram as a function of  $N_{unw}$ , the number of unwound base pairs.



**Figure 3.20 | smFRET unwinding experiments using pre-unwound DNA.**

(a) Schematics of smFRET DNA unwinding experiments utilizing pre-unwound DNA targets as described in Figure 3.19c. (b-d)  $E$  histograms vs  $n_{PD}$  (left) and representative single molecule time traces of donor and acceptor intensities (middle) and idealized  $E$  values (right).

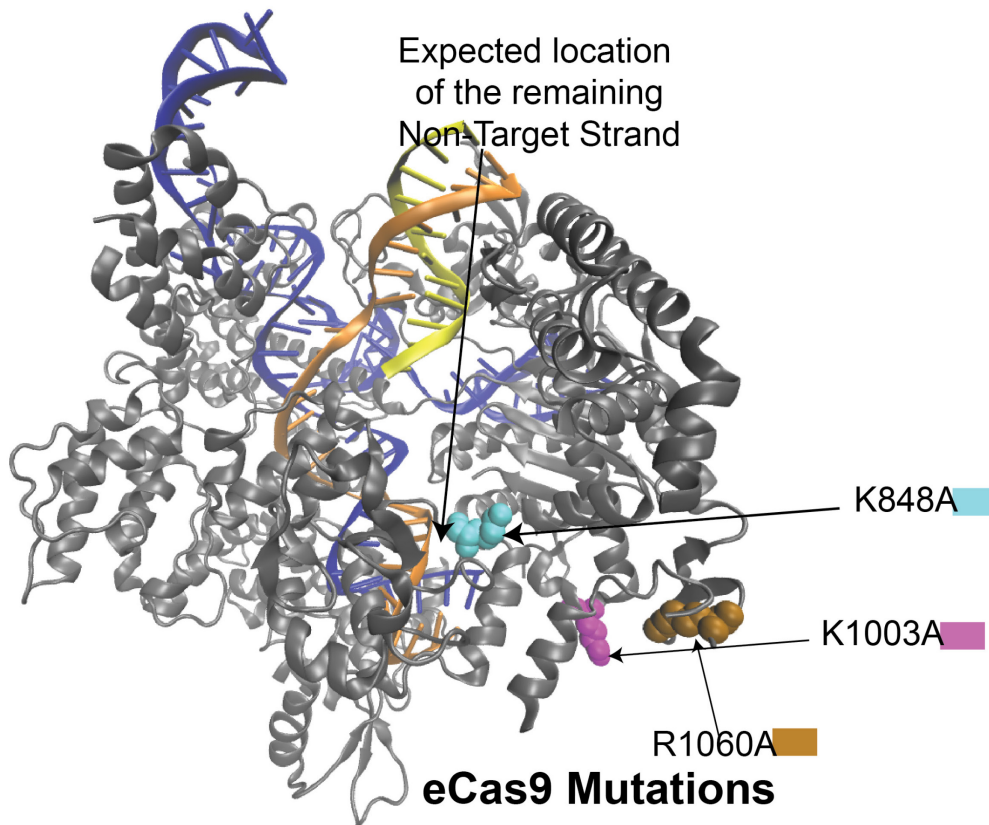
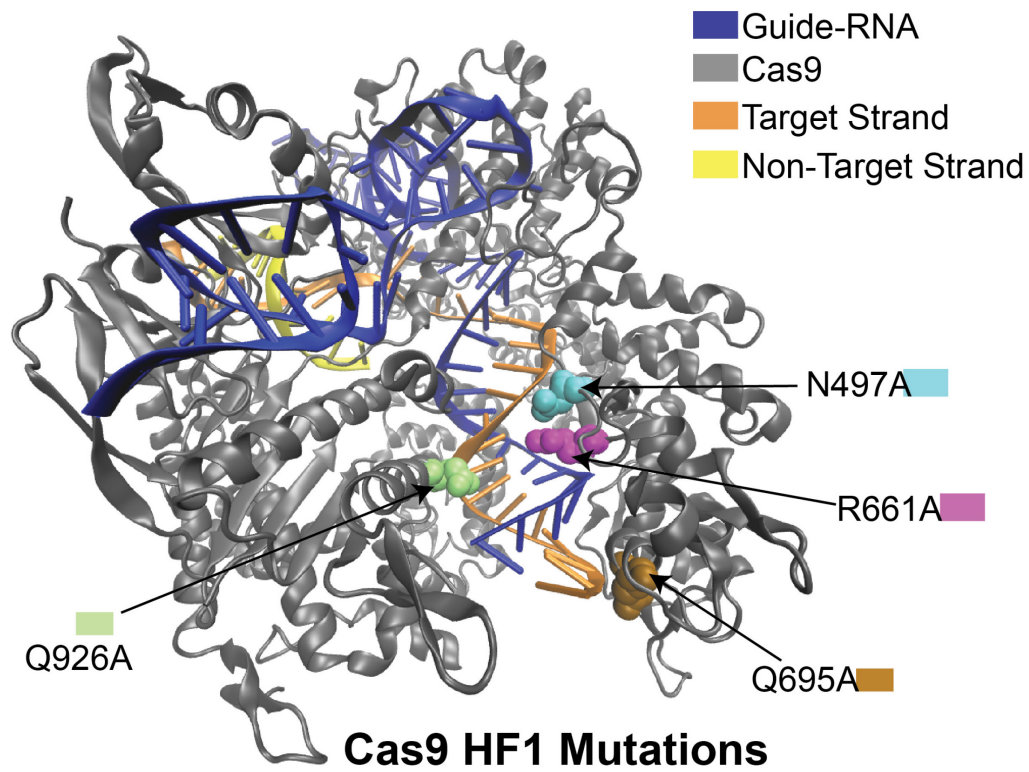
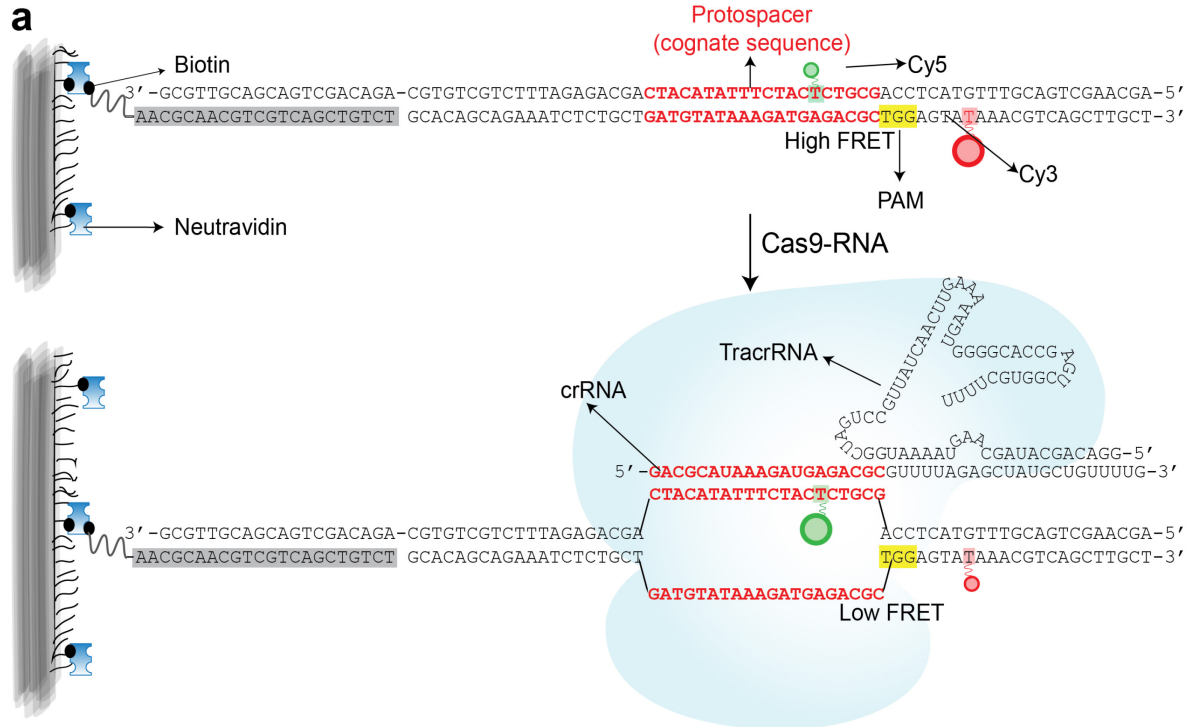
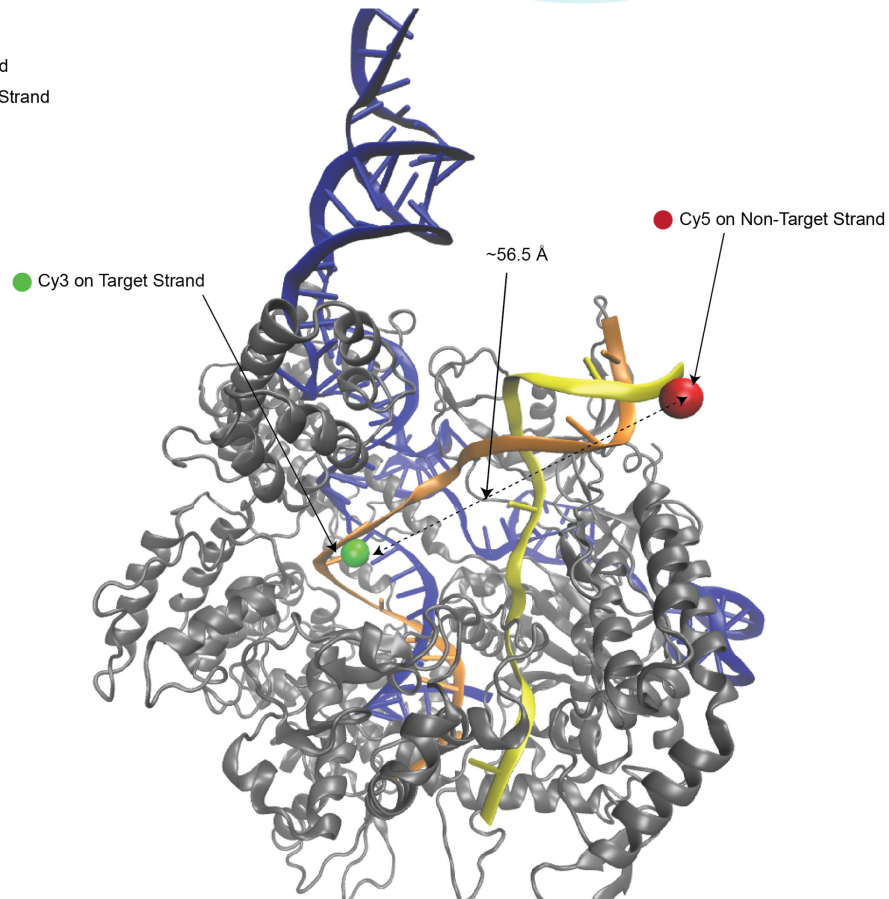


Figure 3.21 | Locations of EngCas9 mutations in dCas9-RNA-DNA complex (PDB ID: 4UN3).



- b**
- Guide-RNA
  - Target Strand
  - Non-Target Strand
  - Cas9



**Figure 3.22 | Probe locations for smFRET assay to investigate Cas9-RNA induced DNA unwinding in PAM-proximal site.**

(a) Schematic of a cognate dsDNA target before and after Cas9-RNA binding. (b) Probe locations mapped to the structure of Cas9-RNA-DNA complex (PDB ID: 5F9R<sup>93</sup>).

### 3.4 DISCUSSION

Like WT Cas9, EngCas9s also require only 9-10 PAM-proximal matches for ultra-stable binding and thus will likely not confer significant specificity advances for applications utilizing Cas9 binding, for example transcription regulation<sup>97</sup> or imaging<sup>98</sup>. EngCas9s, however, do show a modest increase in binding specificity. For example, Cas9-HF1 requires one additional basepair match for ultra-stable binding compared to WT Cas9. Ultra-stable binding to partially matching sequences can sequester Cas9-RNA, limiting the speed of genome editing. To overcome sequestration, one would need higher Cas9-RNA concentrations which may in turn lead to an increase in off-target cleavage. One base pair difference would mean about four-fold reduction in such off-target sinks for Cas9-HF1, and even such a modest improvement in binding specificity may improve cleavage specificity by reducing the total concentration of Cas9-RNA required.

Activation and dynamics of proofreading (REC2) and HNH domains in response to PAM-distal mismatches and Cas9 mutations has been studied using single molecule FRET<sup>84,85</sup>. But what changes in the DNA target conformation trigger these movements of Cas9 domains were not known. Cas9's cleavage activity requires many more matches than stable binding<sup>4,17</sup> and we found that we found that DNA unwinding, instead of DNA binding, is strongly correlated with the cleavage rate when we vary the DNA sequence or Cas9 mutations. Unwound DNA configuration is likely verified by REC3 domain and linker connecting HNH and RuvC nuclease domains in Cas9<sup>85,99</sup>, guiding HNH nuclease movement for cleavage. Here, we show that PAM-distal mismatches impair DNA unwinding, consequently this also



results in impairments in REC2 (coupled to REC3) proofreading and HNH movements<sup>84,85</sup>. In fact, Chen *et al* showed that the cleavage active conformation is more readily depopulated with mismatches for EngCas9s compared to WT Cas9, raising the possibility that HNH movement and DNA unwinding are concurrent. To test this possibility, we utilized the observation that HNH movement requires divalent cations and is inhibited by EDTA<sup>84</sup>. We found that DNA unwinding occurs even in the absence of divalent cations (Figure 3.23), suggesting that DNA unwinding is upstream of HNH movement and that HNH movement does not occur with or because of HNH movement as had been suggested in a simulation study<sup>99</sup>. All these results suggest that DNA unwinding is the critical step that verifies sequence matching, which then subsequently triggers a series of steps toward cleavage including HNH movement.

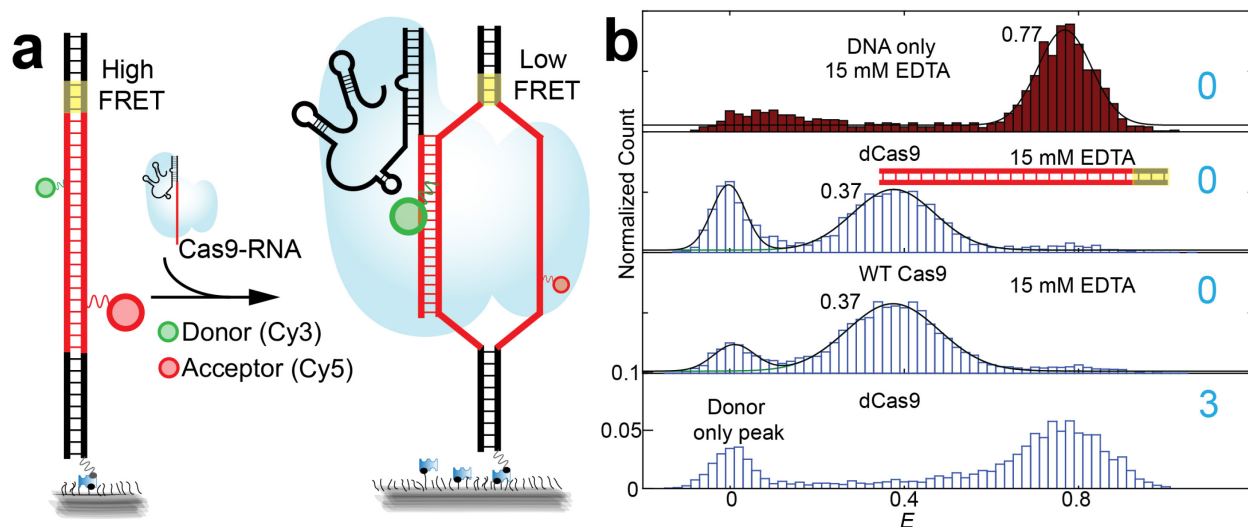
The impact of PAM-distal matches in opposing DNA unwinding is even greater for EngCas9 because their mutations destabilize the unwound DNA configuration, and because cleavage likely proceeds from the unwound state, the increased sensitivity to mismatches in DNA unwinding is likely to contribute to the increased specificity of cleavage by EngCas9s. Any perturbation that destabilizes unwound DNA or fully extended R-loop will delay or fully impair the cleavage action and is also likely the reason why Cas9 with truncated guide-RNAs are more specific in cleavage<sup>100</sup>.

DNA targets with only a few interspersed single base-pair mismatches at PAM-distal site can prevent stably unwound states by all Cas9 but even transiently unwound states appear to allow cleavage by WT Cas9, but not by EngCas9s which we showed here to have much lower intrinsic cleavage rate. Therefore, the large enhancement in cleavage specificity for EngCas9s may in part be due to the lower intrinsic cleavage rate that prevents cleavage from transiently unwound states that are populated in partially matching sequences.

Figure 3.19g summarizes our findings on the three possible mechanisms of specificity increase for EngCas9 in the form of free energy diagram along the axis of  $N_{\text{unw}}$ , the number of bp unwound, leading

toward cleavage. Once the initial barrier of PAM detection is overcome, DNA unwinding proceeds in the direction from PAM-proximal to PAM-distal. There is a sudden drop in energy at  $N_{\text{unw}}=9$  for WT Cas9 and eCas9, marking a threshold for stable binding. Because the threshold is shifted to  $N_{\text{unw}}=10$  for Cas9-HF1, Cas9-HF1 can achieve higher cleavage specificity by reducing the number of off-target sites which can otherwise sequester Cas9-RNA (Mechanism 1). EngCas9s also tilt the energetic balance toward rewinding in the PAM-distal region such that PAM-distal mismatches readily depopulate the fully unwound state needed for cleavage (Mechanism 2). Finally, EngCas9s have lower intrinsic cleavage rates from the unwound state, represented as higher energetic barrier against cleavage, preventing cleavage from transiently unwound off-targets (Mechanism 3).

EngCas9s still cleave DNA targets with  $\sim 3$  PAM-distal mismatches thus there is room for further improvement. Cas9-HF1 and eCas9 mutations eliminated some of the favorable sequence-independent interactions for target and non-target strands separately and may act in an independent manner and thus, their mutations may be combined. EngCas9 mutations may also be combined with truncated RNA to improve the specificity even further. But, combining a number of such strategies that destabilize unwound state and decrease the intrinsic cleavage rate from the unwound state as we showed here could lead to substantially reduced on-target cleavage as well, as has been observed when Cas9-HF1 was used with the truncated RNA<sup>68</sup>.



**Figure 3.23 | DNA unwinding occurs in the absence of divalent cations.**

Without divalent cations, the unwound fraction is unchanged for cognate DNA although final  $E$  value is slightly higher. But unwound fraction is much lower for DNA with  $n_{PD}=3$  with its smFRET traces showing rare cases of transient excursions to low FRET state. These results indicate that divalent cations are not required for DNA unwinding but may help in unwinding. A fraction of high FRET population with  $n_{PD}=3$  could also be due to deleterious binding of Cas9 in absence of divalent cations.

### 3.5 MATERIALS AND METHODS

#### 3.5.1 Preparation of DNA targets

DNA targets used in smFRET assay for DNA interrogation by Cas9-RNA are the same as what were used previously for WT Cas9 studies<sup>83</sup> and Figure 3.2 shows the overall schematics. The schematic of DNA targets used for smFRET assay of Cas9-RNA induced DNA unwinding is shown in Figure 3.8 and Figure 3.14. All DNA oligonucleotides were purchased from Integrated DNA Technologies (IDT, Coralville, IA 52241). A thymine modified with an amine group through a C6 linker (amino-dT) was used to label DNA with Cy3 or Cy5 N-hydroxysuccinimido (NHS). Non-target strand, target strand and a 22 nt biotinylated adaptor strand were assembled by mixing them in buffered solution with 10 mM Tris-HCl, pH 8 and 50 mM NaCl and heating to 90 °C followed by cooling to room temperature over 3 hrs. For DNA unwinding

by surface-tethered Cas9-RNA, the biotin adaptor strand was omitted. Full sequence and modifications of DNA targets used in smFRET assays for DNA interrogation and DNA unwinding are shown in Table 3.1 and respectively. For radio-labeled gel electrophoresis cleavage experiments, the target strand was phosphorylated with P32 using T4 polynucleotide kinase reaction and was annealed with the non-target strand as described above. Sequences are the same as in Table 3.1 and Table 3.2 but without the biotinylated adaptor strand. For fluorescently-labeled gel electrophoresis cleavage experiments, the DNA targets used are the same as those used in smFRET assays.

### **3.5.2 Expression and purification of Cas9**

All Cas9 were expressed and purified as described previously<sup>2,32</sup>. A pET-based expression vector was used for protein expression which consisted of sequence encoding Cas9 (Cas9 residues 1-1368 from *S. pyogenes*) and an N-terminal decahistidine-maltose binding protein (His10-MBP) tag, followed by a peptide sequence containing a tobacco etch virus (TEV) protease cleavage site.

Mutations for cleavage impairment, enhanced mutations (eCas9), HF1 mutations or the desired combinations of them were introduced by site-directed mutagenesis (QuickChange Lightning; Agilent Technologies, Santa Clara, CA 95050). Proteins were expressed in *E. coli* strain BL21 Rosetta 2 (DE3) (EMD Biosciences), grown in TB (Terrific Broth) or 2YT medium (higher expression obtained for TB) at 37 °C for a few hours. When the optical density at 600 nm (OD<sub>600</sub>) reached 0.6, protein expression was induced with 0.5 mM IPTG and the temperature was lowered to 18 °C. The induction was then continued for 12-16 h. The medium was then discarded and cells were harvested. The harvested cells were lysed in 50 mM Tris pH 7.5, 500 mM NaCl, 5% glycerol, 1 mM TCEP, supplemented with protease inhibitor cocktail (Roche) and with/without Lysozyme (Sigma Aldrich), and then homogenized in a microfluidizer (Avestin) or homogenized with Fisher Model 500 Sonic Dismembrator (Thermo Fisher Scientific) at 30% amplitude in 3 one minute cycles, each consisting of series of 2 s sonicate-2 s repetitions. The lysed solution was then ultra-centrifuged at 15,000 *g* for 30-45 minutes, supernatant of lysate was collected and cellular debris was discarded. The supernatant was added to Ni-NTA agarose

resin (Qiagen). The resin was washed extensively with 50 mM Tris pH 7.5, 500 mM NaCl, 10 mM imidazole, 5% glycerol, 1 mM TCEP and the bound protein was eluted in a single-step with 50 mM Tris pH 7.5, 500 mM NaCl, 300 mM imidazole, 5% glycerol, 1 mM TCEP. Dialysis of Cas9 into Buffer A (20 mM Tris-HCl pH 7.5, 125 mM KCl, 5% glycerol, 1 mM TCEP) and cleavage of TEV-protease site by TEV protease was simultaneously carried out overnight at 4 °C. Deconstitution of 10-His-MBP-TEV-Cas9 by TEV protease resulted in 10-His-MBP and Cas9 constituents in the solution. Another round of Ni-NTA agarose column was performed to arrest 10-His-MBP out of the solution and obtain free Cas9. Cas9 was then further purified by size-exclusion chromatography on a Superdex 200 16/60 column (GE Healthcare) in Cas9 Storage Buffer (20 mM Tris-HCl pH 7.5, 100 mM KCl, 5 % glycerol and 5 mM MgCl<sub>2</sub>) and stored at -80 °C. All the purification steps were performed at 4 °C. In some preparations, TEV protease was first added to the elutant and cleavage of the protein fusion was carried out overnight. Following TEV protease cleavage, Cas9 was then dialyzed into Buffer A (20 mM Tris-HCl pH 7.5, 125 mM KCl, 5% glycerol, 1 mM TCEP) for 3 h at 4 °C, before being applied onto a 5 ml HiTrap SP HP sepharose column (GE Healthcare). After washing with Buffer A for three column volumes, Cas9 was eluted using a linear gradient from 0-100% Buffer B (20 mM Tris-HCl pH 7.5, 1 M KCl, 5% glycerol, 1 mM TCEP) over 20 column volumes. The protein was further purified by gel filtration chromatography on a Superdex 200 16/60 column (GE Healthcare) in Cas9 Storage Buffer (20 mM Tris-HCl pH 7.5, 200 mM KCl, 5% glycerol, 1 mM TCEP). Cas9 was stored at -80 °C.

### 3.5.3 Preparation of guide-RNA and Cas9-RNA

Guide-RNA for Cas9 is a combination of CRISPR RNA (crRNA) and trans-activating crRNA (tracrRNA). For smFRET assay for DNA interrogation by Cas9-RNA, crRNA with an amino-dT was purchased from IDT and labeled with Cy5-NHS. Location of the Cy5 in the crRNA is shown in Figure 3.2. The tracrRNA was *in vitro* transcribed as described previously<sup>17</sup>. For DNA unwinding smFRET assay, both crRNA and tracrRNA were unlabeled and transcribed *in vitro*. Guide-RNA was assembled by mixing crRNA and tracrRNA at 1:1.2 ratio in buffer containing 10 mM Tris HCl (pH 8) and 50 mM

NaCl, heated to 90 °C and slowly cooled to room temperature. Cas9-RNA was assembled by mixing guide-RNA and Cas9 at a ratio of 1:3 in Cas9-RNA activity buffer (20 mM Tris HCl (pH 8), 100 mM KCl, 5 mM MgCl<sub>2</sub>, 5% v/v glycerol). Cas9-RNA cleavage activity on cognate sequence used for smFRET assay for DNA interrogation was characterized previously<sup>17</sup>. We used a slightly difference sequence for optical placement of amino-dTs for DNA unwinding smFRET assay and Cas9-RNA cleavage activity on that sequence was also confirmed (Figure 3.9). RNA sequences are available in Table 3.1 and Table 3.2. For smFRET assay of DNA unwinding by surface-tethered Cas9-RNA, a biotin adaptor DNA strand was annealed to a complementary extension on guide-RNA (Figure 3.14)

### 3.5.4 Single-molecule fluorescence imaging and data analysis

DNA targets were immobilized on the polyethylene glycol-passivated flow chamber surface (purchased from Johns Hopkins University Microscope Supplies Core or prepared following protocols reported previously<sup>89</sup> using neutrAvidin-biotin interaction and imaged in the presence of Cas9-RNA at the stated concentration using two-color total internal reflection fluorescence microscopy. For DNA unwinding by surface-tethered Cas9-RNA smFRET assay, 20 nM of biotin-labeled Cas9-RNA was immobilized on surface before adding FRET pair labeled DNA targets. All the imaging experiments were done at room temperature in a Cas9-RNA activity buffer with oxygen scavenging for extending photostability (20 mM Tris-HCl, 100 mM KCl, 5 mM MgCl<sub>2</sub>, 5% (v/v) glycerol, 0.2 mg ml<sup>-1</sup> BSA, 1 mg ml<sup>-1</sup> glucose oxidase, 0.04 mg ml<sup>-1</sup> catalase, 0.8% dextrose and saturated Trolox (>5 mM)<sup>101</sup>. Time resolution was 100 ms unless stated otherwise. Detailed methods of single-molecule FRET (smFRET) data acquisition and analysis have been described previously<sup>89</sup>. Video recordings obtained using EMCCD camera (Andor) were processed to extract single molecule fluorescence intensities at each frame and custom written scripts were used to calculate FRET efficiencies. Data acquisition and analysis software can be downloaded from <https://cplc.illinois.edu/software/>. FRET efficiency ( $E$ ) of the detected spot was approximated as  $E = I_A / (I_D + I_A)$ , where  $I_D$  and  $I_A$  are background and leakage corrected emission intensities of the donor and acceptor, respectively.

### 3.5.5 FRET histograms, Cas9-RNA bound DNA fraction and unwound fraction

Unless stated otherwise, first five frames of each molecule's  $E$  value time traces were used as data points to construct  $E$  histograms. At least 2,500 molecules were included in each histogram. Cas9-RNA bound DNA fraction was calculated as a fraction of data points with  $E > 0.75$  and was then normalized relative to Cas9-RNA bound DNA fraction for the cognate DNA target, which is not 100% because missing or inactive acceptor fluorophore, and is referred to as  $f_{\text{bound}}$ . To estimate  $K_d$ ,  $f_{\text{bound}}$  vs Cas9-RNA concentration ( $c$ ) was fit using

$f_{\text{bound}} = M \times c / (K_d + c)$  where  $M$  is the maximum observable  $f_{\text{bound}}$ .  $M$  is typically less than 1 because inactive or missing acceptors or because not all of the DNA on the surface are capable of binding Cas9-RNA.

For DNA unwinding smFRET assay, unwound fraction was calculated as the ratio of data points with  $0.2 < E < 0.6$  and total number of data points with  $E > 0.2$ .

### 3.5.6 Kinetic analysis of DNA interrogation by Cas9-RNA

A bimolecular association/dissociation kinetics was used for the analysis of DNA binding by Cas9-RNA.



$$k_{\text{binding}} (\text{s}^{-1}) = k_{\text{on}} (\text{M}^{-1}\text{s}^{-1}) \times [\text{Cas9-RNA}] (\text{M})$$

A hidden Markov model analysis of smFRET traces of DNA targets that showed real-time reversible association/dissociation of Cas9-RNA yielded three pre-dominant FRET states, of zero, mid and high  $E$  values. Dwell times of high FRET ( $E > 0.6$ ) states before their transition to zero FRET ( $E < 0.2$ ) states were used to calculate lifetime of high FRET ( $\tau_{\text{high}}$ ) (Figure 3.6c) binding events by fitting their distribution using a double-exponential decay. Dwell times of mid FRET ( $0.2 < E < 0.6$ ) states before their transition to

zero FRET ( $E < 0.2$ ) states were used to calculate lifetime of mid FRET ( $\tau_{\text{mid}}$ ) (Figure 3.6d) binding events by fitting their distribution using a single exponential decay.

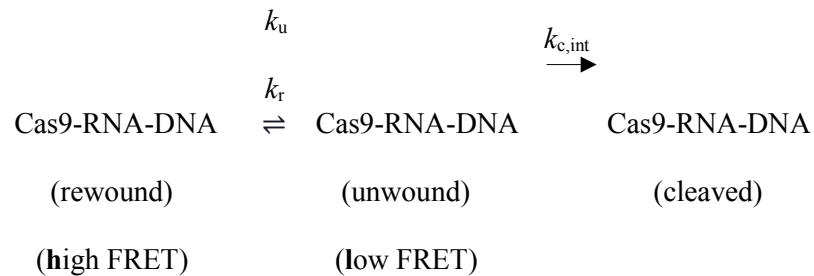
To calculate the total bound state lifetime ( $\tau_{\text{avg}}$ ) (Figure 3.1d), mid and high FRET states were taken as bound states and zero FRET state as the unbound state. The survival probability of all the bound state events (mid and high FRET states taken as a single state,  $\text{FRET} > 0.2$ ) vs time was fit using a double exponential decay profile ( $A_1 \exp(-t/\tau_1) + A_2 \exp(-t/\tau_2)$ ).  $\tau_{\text{avg}}$  is an amplitude weighted average of two distinct lifetimes  $\tau_1$  and  $\tau_2$ .

$$\tau_{\text{avg}} = A_1 \tau_1 + A_2 \tau_2.$$

$A_1$  values are also used as the fraction of mid FRET state binding events ( $f_{\text{mid FRET}}$ ). Dwell time distribution of zero FRET state ( $E < 0.2$ ) was fit to a single exponential decay to calculate lifetime of unbound state ( $\tau_{\text{unbound}}$ ) (Figure 3.1f). Cy5 labeling efficiency of guide-RNA was 88.7% and thus  $\tau_{\text{unbound}}$  was appropriately corrected to account for it. Inverse of  $\tau_{\text{unbound}}$  was used to calculate  $k_{\text{binding}}$  and  $k_{\text{on}}$ . Since Cas9-RNA DNA association were sequence independent, a mean of  $k_{\text{on}}$  for different DNA target was used to determine  $k_{\text{on}}$  for each Cas9 (Figure 3.1f)

### 3.5.7 Kinetic analysis of Cas9-RNA induced DNA unwinding and rewinding

Cas9-RNA induced DNA unwinding was modeled as a two-state system (Figure 3.10b):



A hidden Markov model analysis segmented single molecule time traces into unwound and rewind states (low and high FRET states). Dwell time histograms were fitted using single exponential decay to



determine the average lifetime of the high ( $\tau_{\text{high}}$ ) and low FRET ( $\tau_{\text{low}}$ ) states respectively (Figure 3.10d-e). A small fraction of its smFRET time trajectories had a constant high FRET without any fluctuations, likely representing DNA molecules unable to bind Cas9-RNA and were excluded from kinetic analysis. Unwinding rate ( $k_u$ ) was calculated as inverse of  $\tau_{\text{high}}$  and rewinding rate ( $k_r$ ) was calculated as inverse of  $\tau_{\text{low}}$ . Intrinsic cleavage rate  $k_{c,\text{int}}$  was determined by fitting cleavage lifetime vs unwound fraction using  $\tau_{\text{cleavage}} = 1/([\text{unwound fraction}] * k_{c,\text{int}} + C)$ , where  $C=0$  for eCas9 but an extremely small value of  $C$  had to be used for WT Cas9 and Cas9-HF1.

In smFRET experiments to capture the initial DNA unwinding event, distribution of dwell times of the initial high FRET state before the first transition to the low FRET state was fit with a single exponential to obtain  $\tau_{\text{unwinding}}$ .

### **3.5.8 Gel-electrophoresis to investigate Cas9-RNA induced cleavage**

FRET pair labeled DNA targets used in the smFRET DNA unwinding assays were also used in fluorescence-based gel electrophoresis experiments. The DNA targets were incubated with Cas9-RNA in Cas9-RNA activity buffer (20 mM Tris HCl (pH 8), 100 mM KCl, 5 mM MgCl<sub>2</sub>, 5% v/v glycerol) at the specified concentrations and for specified durations. The reaction samples were then denatured by formamide loading buffer (95% formamide, 5 mM EDTA) and resolved via polyacrylamide gel electrophoresis (PAGE) using 15% Polyacrylamide TBE Urea pre-cast gels (Bio-rad laboratories) and imaged via Cy5/Cy3 excitation (GE Amersham Imager 600). Sequences of DNA targets and guide-RNA are shown in Table 3.2. For radio-labeled PAGE experiments, P<sup>32</sup>-labeled DNA targets at 1 nM concentration were incubated with 100 nM Cas9-RNA in Cas9-RNA activity buffer (20 mM Tris HCl (pH 8), 100 mM KCl, 5 mM MgCl<sub>2</sub>, 5% v/v glycerol) at for 10 minutes. The reaction samples were then denatured by formamide loading buffer (95% formamide, 5 mM EDTA), resolved via PAGE and imaged via phosphorimaging (GE Healthcare). For the rate of cleavage experiments, aliquots of Cas9-RNA DNA running reaction were taken at different time points for the PAGE analysis. Sequences of DNA targets

and guide-RNA are the same as in Table 3.2 except that the biotinylated adaptor strand is omitted.  $\tau_{\text{cleavage}}$  for WT Cas9 is taken from a study<sup>96</sup> which used a slightly different protospacer, shown in Table 3.1. In a control, the  $\tau_{\text{cleavage}}$  for WT Cas9 were similar between the two protospacers and much lower than  $\tau_{\text{cleavage}}$  for EngCas9. For all experiments reported in the manuscript, error bars represent s.d. from n=2 or 3. n=1 in absence of error bar.

**Table 3.1 | Guide-RNA and DNA targets used for the smFRET assay for DNA interrogation by Cas9-RNA and gel electrophoresis cleavage assay.**

Description	DNA Sequences
	20 nucleotide biotinylated adaptor for surface immobilization
0/0	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGACGCATAAAGATGAGACGC TGGAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGACTGCGTATTTCTACTCTGCGACCTCATGTTGCAGTCGAACGA-5'
4	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGCCATAAAGATGAGACGC TGGAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGAGACGATGTTTCTACTCTGCGACCTCATGTTGCAGTCGAACGA-5'
8	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGCGTATAAGATGAGACGC TGGAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGAGACGCATACTACTCTGCGACCTCATGTTGCAGTCGAACGA-5'
9	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGCGTATTAGATGAGACGC TGGAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGAGACGCATAATCTACTCTGCGACCTCATGTTGCAGTCGAACGA-5'
10	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGCGTATTTGATGAGACGC TGGAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGAGACGCATAAACTACTCTGCGACCTCATGTTGCAGTCGAACGA-5'
11	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGCGTATTTCTAGAGACGC TGGAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGAGACGCATAAAGTACTCTGCGACCTCATGTTGCAGTCGAACGA-5'
12	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGCGTATTTCTTGAGACGC TGGAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGAGACGCATAAAGACTCTGCGACCTCATGTTGCAGTCGAACGA-5'
13	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGCGTATTTCTAGAGACGC TGGAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGAGACGCATAAAGATCTCTGCGACCTCATGTTGCAGTCGAACGA-5'
14	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGCGTATTTCTACAGACGC TGGAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGAGACGCATAAAGATGCTCTGCGACCTCATGTTGCAGTCGAACGA-5'
15	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGCGTATTTCTACTGACGC TGGAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGAGACGCATAAAGATGACTGCGACCTCATGTTGCAGTCGAACGA-5'
16	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGCGTATTTCTACTACGC TGGAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGAGACGCATAAAGATGAGTCTGCGACCTCATGTTGCAGTCGAACGA-5'
20/20	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGCGTATTTCTACTCTGCG TGGAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGAGACGCATAAAGATGAGACGCGACCTCATGTTGCAGTCGAACGA-5'
2	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGACGCATAAAGATGAGACGC TGGAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGACTGCGTATTTCTACTCTGCGACCTCATGTTGCAGTCGAACGA-5'
4	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGACGCATAAAGATGAGTGGC TGGAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGACTGCGTATTTCTACTACGCGACCTCATGTTGCAGTCGAACGA-5'
4 roadblock mismatches ( $R_{mm}$ ) from 9 <sup>th</sup> bp	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGACGCATAATTTCTTGAGACGC TGGAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGACTGCGTATAAAGACTCTGCGACCTCATGTTGCAGTCGAACGA-5'
4 $R_{mm}$ from 5 <sup>th</sup> bp	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGACGCATAAAGAACTACGC TGGAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGACTGCGTATTTGAGTGGCAGCTCATGTTGCAGTCGAACGA-5'
0/0 <sub>mm</sub> NO-PAM	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGACGCATAAAGATGAGACGCATAAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGACTGCGTATTTCTACTCTGCGTATTTCATGTTGCAGTCGAACGA-5'
20/20 <sub>mm</sub> NO-PAM	5' <sup>•</sup> -AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGCGTATTTCTACTCTGCGATAAG ACAACGTCAGCTTGCT-3' 3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGCTTTAGAGACGAGACGCATAAAGATGAGACGCTATTTCATGTTGCAGTCGAACGA-5'
Description	RNA Sequences
crRNA	5' -GACGCAUAAAGAUGAGACGCGUUUAGAGCUAUGCUGUUUUU-3'
tracrRNA	5' -GGACAGCAUAGCAAGUUUUUUUUAGGUCUAGUCCGUUUUCAACUUGAAAAAGUGGCACCGAGUCGGUCUUUUU-3'

<sup>•</sup> Biotin    ■ Protospacer Adjacent Motif (PAM)    ■ Thymine modification for Cy3 and Cy5 labeling

DNA sequences complementary to guide RNA are shown in red (Cognate). RNA sequences complementary to the protospacer in a cognate DNA target are shown in red (Cognate).

**Table 3.2 | Guide-RNA and DNA targets used for Cas9-RNA induced DNA unwinding smFRET assay and gel electrophoresis cleavage assays.**

Description	DNA Sequences
0/0	<p>20 nucleotide biotinylated adaptor for surface immobilization</p> <p>5' <sup>●</sup>-*<sub>2</sub>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGATG<sup>■</sup>ATAAAGATGAGACGCTGGAGTACAAACGTCAGCTTGCT-3'            3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGTCTTTAGAGACGACTACATATTTCTAC<sup>■</sup>CTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
1	<p>5' <sup>●</sup>-*<sub>2</sub>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTCATG<sup>■</sup>ATAAAGATGAGACGCTGGAGTACAAACGTCAGCTTGCT-3'            3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGTCTTTAGAGACGAGTACATATTTCTAC<sup>■</sup>CTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
2	<p>5' <sup>●</sup>-*<sub>2</sub>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTCTG<sup>■</sup>ATAAAGATGAGACGCTGGAGTACAAACGTCAGCTTGCT-3'            3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGTCTTTAGAGACGAGAACATATTTCTAC<sup>■</sup>CTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
3	<p>5' <sup>●</sup>-*<sub>2</sub>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTCTAG<sup>■</sup>ATAAAGATGAGACGCTGGAGTACAAACGTCAGCTTGCT-3'            3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGTCTTTAGAGACGAGATCATATTTCTAC<sup>■</sup>CTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
4	<p>5' <sup>●</sup>-*<sub>2</sub>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTCTAC<sup>■</sup>ATAAAGATGAGACGCTGGAGTACAAACGTCAGCTTGCT-3'            3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGTCTTTAGAGACGAGATGATATTTCTAC<sup>■</sup>CTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
20-20 <sub>mm</sub> 20-20 <sub>uw</sub>	<p>5' <sup>●</sup>-*<sub>2</sub>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGATG<sup>■</sup>ATAAAGATGAGACGCTGGAGTACAAACGTCAGCTTGCT-3'            3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGTCTTTAGAGACGAGTACATATTTCTAC<sup>■</sup>CTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
18-20 <sub>mm</sub> 18-20 <sub>uw</sub>	<p>5' <sup>●</sup>-*<sub>2</sub>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGATG<sup>■</sup>ATAAAGATGAGACGCTGGAGTACAAACGTCAGCTTGCT-3'            3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGTCTTTAGAGACGAGATCATATTTCTAC<sup>■</sup>CTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
17-20 <sub>mm</sub> 17-20 <sub>uw</sub>	<p>5' <sup>●</sup>-*<sub>2</sub>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGATG<sup>■</sup>ATAAAGATGAGACGCTGGAGTACAAACGTCAGCTTGCT-3'            3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGTCTTTAGAGACGAGATGATATTTCTAC<sup>■</sup>CTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
0-0 <sub>mm</sub> (PAM-proximal unwinding)	<p>5' <sup>●</sup>-*<sub>2</sub>-AACGCAACGTCGTCAGCTGTCT GCACAGCAGAAATCTCTGCTGATG<sup>■</sup>TATAAAGATGAGACGCTGGAG<sup>■</sup>ACAAACGTCAGCTTGCT-3'            3' -GCGTTGCAGCAGTCGACAGA-CGTGTCGTCTTTAGAGACGACTACATATTTCTAC<sup>■</sup>CTGCGACCTCATGTTTGCAGTCGAACGA-5'</p>
Description	RNA Sequences
crRNA	5' -GAUGUAUAAAGAUAGACGCGUUUUUAGAGCUAUGCUGUUUUU-3'
tracrRNA	5' -GGACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUU-3'
tracrRNA with 3' extension for biotinylated adaptor	5' -GGACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUUGCUCGUGCGC-3'
Biotinylated adaptor	5' <sup>●</sup> -* <sub>2</sub> -TTGCGCAGCAGCAAA-3'
2 n <sub>pp</sub> -crRNA	5' -GAUGUAUAAAGAUAGACCGUUUUUAGAGCUAUGCUGUUUUU-3'
Internal roadblock mismatches crRNA	5' -GACGCAUAAAGAUAGACGCGUUUUUAGAGCUAUGCUGUUUUU-3'

<sup>●</sup> Biotin    ■ Protospacer Adjacent Motif (PAM)    ■ Thymine modification for Cy3 and Cy5 labeling

DNA sequences complementary to guide RNA are shown in red (Cognate). RNA sequences complementary to the protospacer in a cognate DNA target are shown in red (Cognate).

For experiments in Fig.6a-b and 6c-f, the mismatches were in the guide-RNA, against the cognate DNA (0-0mm) target. i.e. 2 n<sub>pp</sub>-crRNA and internal roadblock mismatches crRNA respectively.

### **3.6 AUTHOR CONTRIBUTIONS and ACKNOWLEDGEMENTS**

Digvijay Singh and Taekjip Ha designed the experiments. Digvijay Singh and Yanbo Wang performed smFRET DNA interrogation experiments. Digvijay Singh performed smFRET DNA unwinding experiments. Digvijay Singh and John Mallon performed gel-based experiments. Digvijay Singh, John Mallon expressed and purified Cas9s. Digvijay Singh, Yanbo Wang, Olivia Yang, Jingyi Fei, Anustup Poddar, and Damon Ceylan performed or helped with the data analysis. Olivia Yang assisted with some experiments. Anustup Poddar assisted with PEG passivation of some slides.

We would like to thank Jennifer A. Doudna and Samuel H. Sternberg for useful early discussions about the design of experiments. We would also like to thank Samuel H. Sternberg and Janice S. Chen of Doudna lab for generously providing some Cas9 stocks and EngCas9 expression plasmids, respectively.

# Chapter 4.

## **Real-time observation of DNA target interrogation and product release by the RNA-guided endonuclease CRISPR Cpf1**

\*Contents of this chapter is available at:

Singh, D., Mallon, J., Poddar, A., Wang, Y., Tipanna, R., Yang, O., Bailey, S., and Ha, T. (2017) Real-time observation of DNA target interrogation and product release by the RNA-guided endonuclease CRISPR Cpf1, *bioRxiv* (2017).

## 4.1 ABSTRACT

CRISPR-Cas9, which imparts adaptive immunity against foreign genomic invaders in certain prokaryotes, has been repurposed for genome-engineering applications. More recently, another RNA-guided CRISPR endonuclease called Cpf1 was identified and is also being repurposed. Little is known about the kinetics and mechanism of Cpf1 DNA interaction and how sequence mismatches between the DNA target and guide-RNA influence this interaction. We have used single-molecule fluorescence imaging and biochemical assays to characterize DNA interrogation, cleavage, and product release by three Cpf1 orthologues. Like Cas9, Cpf1 initially binds DNA in search of PAM (protospacer-adjacent motif) sequences, verifies the target sequence unidirectionally from the PAM-proximal end and rapidly rejects any targets that lack a PAM or that are poorly matched with the guide-RNA. Cpf1 requires ~ 17 bp sequence match for both stable binding and cleavage, contrasting it with Cas9 which requires 9 bp for stable binding and ~16 bp for cleavage. Unlike Cas9, which does not release the DNA cleavage products, Cpf1 rapidly releases the PAM-distal cleavage product, but not the PAM-proximal product. Our findings have important implications on Cpf1-based genome engineering and manipulation applications.

## 4.2 INTRODUCTION

In prokaryotes, CRISPR (clustered regularly interspaced short palindromic repeats)–Cas (CRISPR-associated) acts as an adaptive defense system against foreign genetic elements<sup>102</sup>. The system achieves adaptive immunity by storing short sequences of invader DNA into the host genome, which get transcribed and processed into small CRISPR RNA (crRNA). These crRNAs form a complex with a CRISPR nuclease to guide the nuclease to complementary foreign nucleic acids (protospacers) for cleavage. Binding and cleavage also require that the protospacer be adjacent to a protospacer adjacent motif (PAM)<sup>4,103</sup>. CRISPR-Cas9, chiefly the Cas9 from *Streptococcus pyogenes* (*SpCas9*), has been repurposed to create an RNA-programmable endonuclease for gene knockout and editing<sup>7,8,104</sup>. Nuclease deficient Cas9 has also been used for tagging genomic sites in wide-ranging applications<sup>7,8,104</sup>. This

repurposing has revolutionized biology and sparked a search for other novel CRISPR-Cas enzymes<sup>105,106</sup>. One such search led to the discovery of the Cas protein Cpf1, with some of its orthologues reporting highly specific cleavage activities in mammalian cells<sup>9,11,107,108</sup>.

Compared to Cas9, Cpf1 has an AT rich PAM (5'-YTTN-3' vs. 5'-NGG-3' for SpCas9), a longer protospacer (24 bp vs. 20 bp for Cas9), creates staggered cuts distal to the PAM vs. blunt cuts proximal to the PAM by Cas9<sup>11</sup>, and is an even simpler system than Cas9 because it does not require a trans-activating RNA for nuclease activity or guide-RNA maturation<sup>109</sup>. Off-target effects remain one of the top concerns for CRISPR-based applications but Cpf1 is reportedly more specific than Cas9<sup>107,108</sup>. However, its kinetics and mechanism of DNA recognition, rejection, cleavage and product release as a function of mismatches between the guide-RNA and target DNA remain unknown. Precise characterization of differences amongst different CRISPR enzymes should help in expanding the functionalities of the CRISPR toolbox.

Here, we have used single-molecule imaging and biochemical assays to understand how mismatches between the guide-RNA and DNA target modulate the activity of three Cpf1 orthologues from *Acidaminococcus sp. (AsCpf1)*, *Lachnospiraceae bacterium (LbCpf1)* and *Francisella novicida (FnCpf1)*<sup>11</sup>. Single-molecule methods have been helpful in the study of CRISPR mechanisms<sup>6,53,62,83,86,88,110-113</sup> because they allow real-time detection of multiple and distinct steps of varying time lengths i.e. transient to long-lived<sup>57</sup>.

## 4.3 RESULTS

### 4.3.1 Real-time DNA interrogation by Cpf1-RNA

We employed a single-molecule fluorescence resonance energy transfer (smFRET) binding assay<sup>63,89</sup>. DNA targets (donor-labeled, 82 bp long) were immobilized on a polyethylene glycol (PEG) passivated surface and Cpf1 pre-complexed with acceptor-labeled guide-RNA (Cpf1-RNA) was added. Cognate



DNA and guide-RNA sequences are identical to the Cpf1 orthologue-specific sequences that were previously characterized biochemically<sup>11</sup> with the exception that we used canonical guide-RNA of AsCpf1 for FnCpf1 analysis because guide-RNAs of AsCpf1 and FnCpf1 are interchangeable<sup>11</sup> (Figure 4.1). Locations of donor (Cy3) and acceptor (Cy5) fluorophores were chosen such that FRET would report on interaction between the DNA target and Cpf1-RNA<sup>114</sup> (Figure 4.2a and Figure 4.1). Fluorescent labeling did not affect cleavage activity of Cpf1-RNA (Figure 4.3). We used a series of DNA targets containing different degrees of mismatches relative to the guide-RNA referred to here with  $n_{PD}$  (the number of PAM-distal mismatches) or  $n_{PP}$  (the number of PAM-proximal mismatches) (Figure 4.1b). Cognate DNA target in the presence of 50 nM Cpf1-RNA gave two distinct populations with FRET efficiency  $E$  centered at 0.4 and 0. Using instead a non-cognate DNA target ( $n_{PD}$  of 24 and without PAM) or guide-RNA only without Cpf1 gave a negligible  $E=0.4$  population, allowing us to assign  $E\sim 0.4$  to a sequence-specific Cpf1-RNA-DNA complex where the labeling sites are separated by 54 Å<sup>114</sup> (Figure 4.2c and Figure 4.1). The  $E=0$  population is a combination of unbound states and bound states but with an inactive or missing acceptor. smFRET time trajectories of the cognate DNA target showed a constant  $E\sim 0.4$  value within measurement noise (Figure 4.2c).

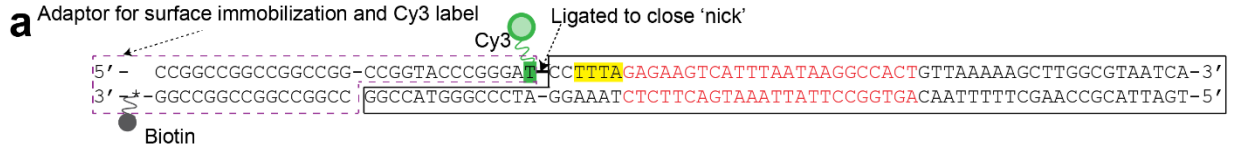
Cpf1-RNA titration experiments yielded dissociation constants ( $K_d$ ) of 0.27 nM (FnCpf1), 0.1 nM (AsCpf1), 3.9 nM (LbCpf1) in our standard imaging condition and 0.13 nM (LbCpf1) in a reducing condition (Figure 4.4). Binding is much tighter than the 50 nM  $K_d$  previously reported for FnCpf1<sup>109</sup>. We performed purification and biochemical experiments in buffer containing dithiothreitol (DTT) as per previous protocols<sup>11</sup> but did not include DTT for standard imaging condition because of severe fluorescence intermittency of Cy5 caused by DTT<sup>115</sup>. DTT did not affect FnCpf1 or AsCpf1 DNA binding but made binding >20-fold tighter for LbCpf1 (Figure 4.4). Cleavage by AsCpf1 is most effective at pH 6.5-7.0 (Figure 4.5). Therefore, we used pH 7.0 for AsCpf1 and standard pH 8.0 for FnCpf1 and LbCpf1.

$E$  histograms obtained at 50 nM Cpf1-RNA show the impact of mismatches on DNA binding (Figure 4.6). The apparent bound fraction  $f_{\text{bound}}$ , defined as the fraction of DNA molecules with  $E > 0.2$ , remained unchanged when  $n_{\text{PD}}$  increased from 0 to 7 (0 to 6 for LbCpf1 in non-reducing conditions) (Figure 4.6 and Figure 4.7d). Binding was ultra-stable for  $n_{\text{PD}} \leq 7$  because  $f_{\text{bound}}$  did not change even 1 hour after washing away free Cpf1-RNA (Figure 4.7a).  $f_{\text{bound}}$  decreased steeply when  $n_{\text{PD}}$  exceeded 7 for FnCpf1 and LbCpf1 but the decrease was gradual for AsCpf1 and for LbCpf1 in the reducing condition (Figure 4.6 and Figure 4.7d). For all Cpf1 orthologues, ultra-stable binding required  $n_{\text{PD}} \leq 7$ , corresponding to a 17 bp PAM-proximal sequence match. This is much larger than the 9 bp PAM-proximal sequence match required for ultra-stable binding of Cas9<sup>83</sup>. PAM-proximal mismatches are highly deleterious for Cpf1 binding because  $f_{\text{bound}}$  dropped by more than 95% if  $n_{\text{PP}} \geq 2$  (Figure 4.6 and Figure 4.7d). In comparison, Cas9 showed a more modest ~50% drop for  $n_{\text{PP}} = 2$ <sup>83</sup>. Overall, Cpf1 is much better than Cas9 in discriminating against both PAM-distal and PAM-proximal mismatches for stable binding.

Single molecule time-trajectories of all Cpf1 orthologues for  $n_{\text{PD}} \leq 7$  showed a constant  $E \sim 0.4$  value within noise, limited only by photobleaching. For  $n_{\text{PD}} > 7$ , we observed reversible transitions in  $E$  likely due to transient binding (Figure 4.8 and Figure 4.9 and Figure 4.10). Dwell time analysis as a function of Cpf1-RNA concentration confirmed that  $E$  fluctuations are due to binding and dissociation, not conformational changes (Figure 4.7b, Figure 4.7c, and Figure 4.4). We used hidden Markov modeling analysis<sup>90</sup> to segment the time traces to bound and unbound states. Average lifetime of the bound state,  $\tau_{\text{avg}}$ , was  $> 1$  hour for  $n_{\text{PD}} \leq 7$  but decreased to a few seconds with  $n_{\text{PD}} > 7$  or any PAM-proximal mismatches (Figure 4.7e). The unbound state lifetime differed between orthologues but was nearly the same among most DNA targets, indicating that initial binding has little sequence dependence. The bimolecular association rate  $k_{\text{on}}$  was  $2.37 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$  (FnCpf1),  $0.87 \times 10^6 \text{ M}^{-1} \text{ s}^{-1}$  (LbCpf1) and  $1.33 \times 10^7 \text{ M}^{-1} \text{ s}^{-1}$  (LbCpf1 in reducing conditions) (Figure 4.7c and Figure 4.7f). Much longer apparent unbound state lifetimes with PAM-proximal mismatches or DNA targets without PAM are likely due to binding

events shorter than the time resolution (0.1 s).

These results show that Cpf1-RNA has dual binding modes. It first binds DNA non-specifically (mode I) in search of PAM and upon detection of PAM, RNA-DNA heteroduplex formation ensues (mode II) and if it extends  $\geq 17$  bp, Cpf1-RNA remains ultra-stably bound to the DNA. Some reversible transitions in  $E$  were observed even for DNA with  $n_{PD} = 7$ , indicating that multiple short-lived binding events take place before the one resulting in ultra-stable binding (Figure 4.8, Figure 4.9 and Figure 4.10). RNA-DNA heteroduplex extension is likely unidirectional from PAM-proximal to PAM-distal end because any PAM-proximal mismatch prevented stable binding. Consistent with dual binding modes, survival probability distributions of bound and unbound state were best described by a double and single exponential decay, respectively (Figure 4.11).



Oligonucleotides used to constitute the above DNA target :

smFRET Cy3 oligo

5' -CCGGGCCGGCCGGCCGGCCGGTACCCGGGA-3'

Target strand

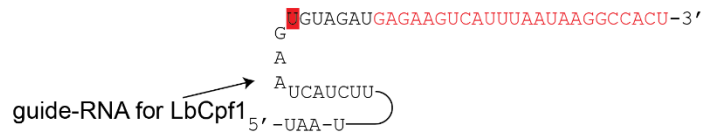
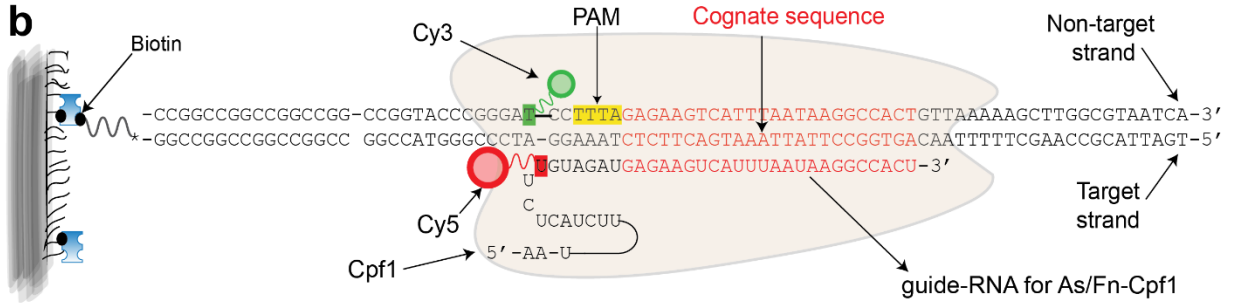
3' -GGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTATTCGGGTGACAATTTTTCGAACCGCATTAGT-5'

Non-target strand

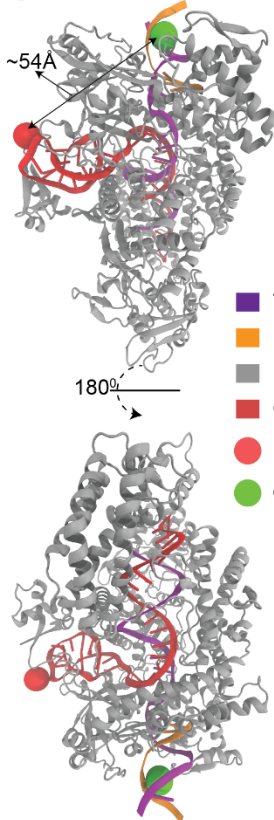
5' -CCTTTAGAGAAGTCATTTAATAAGGCCACTGTTAAAAGCTTGGCGTAATCA-3'

Biotin oligo

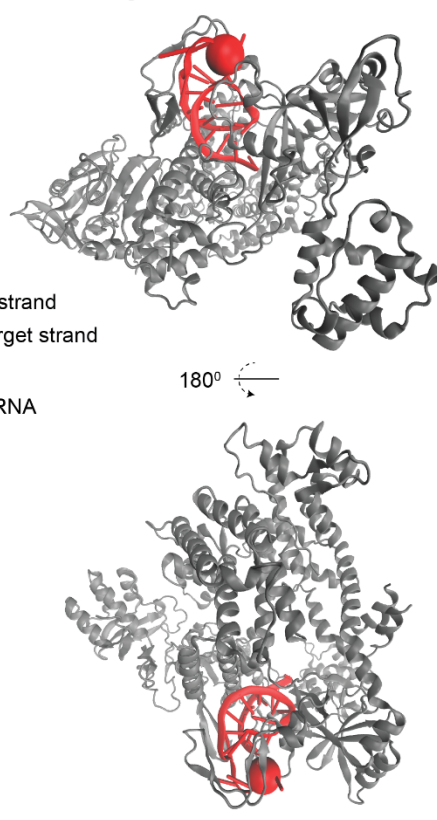
3' -GGCCGGCCGGCCGGCC-5'



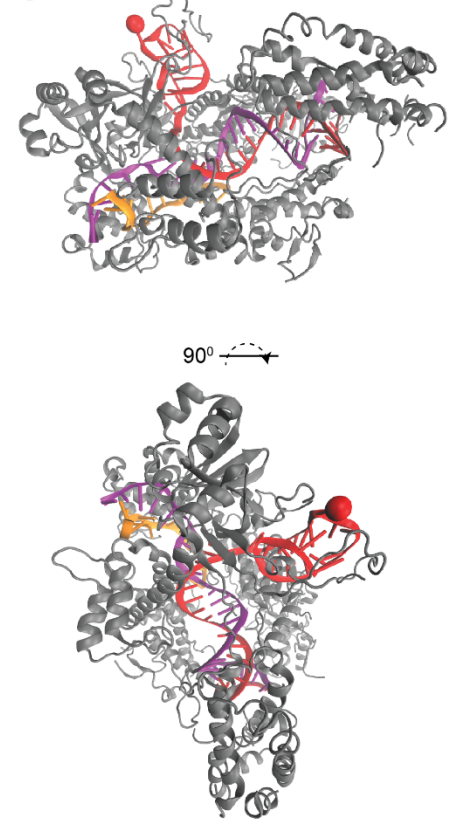
**c** AsCpf1-RNA-DNA



**d** LbCpf1-RNA

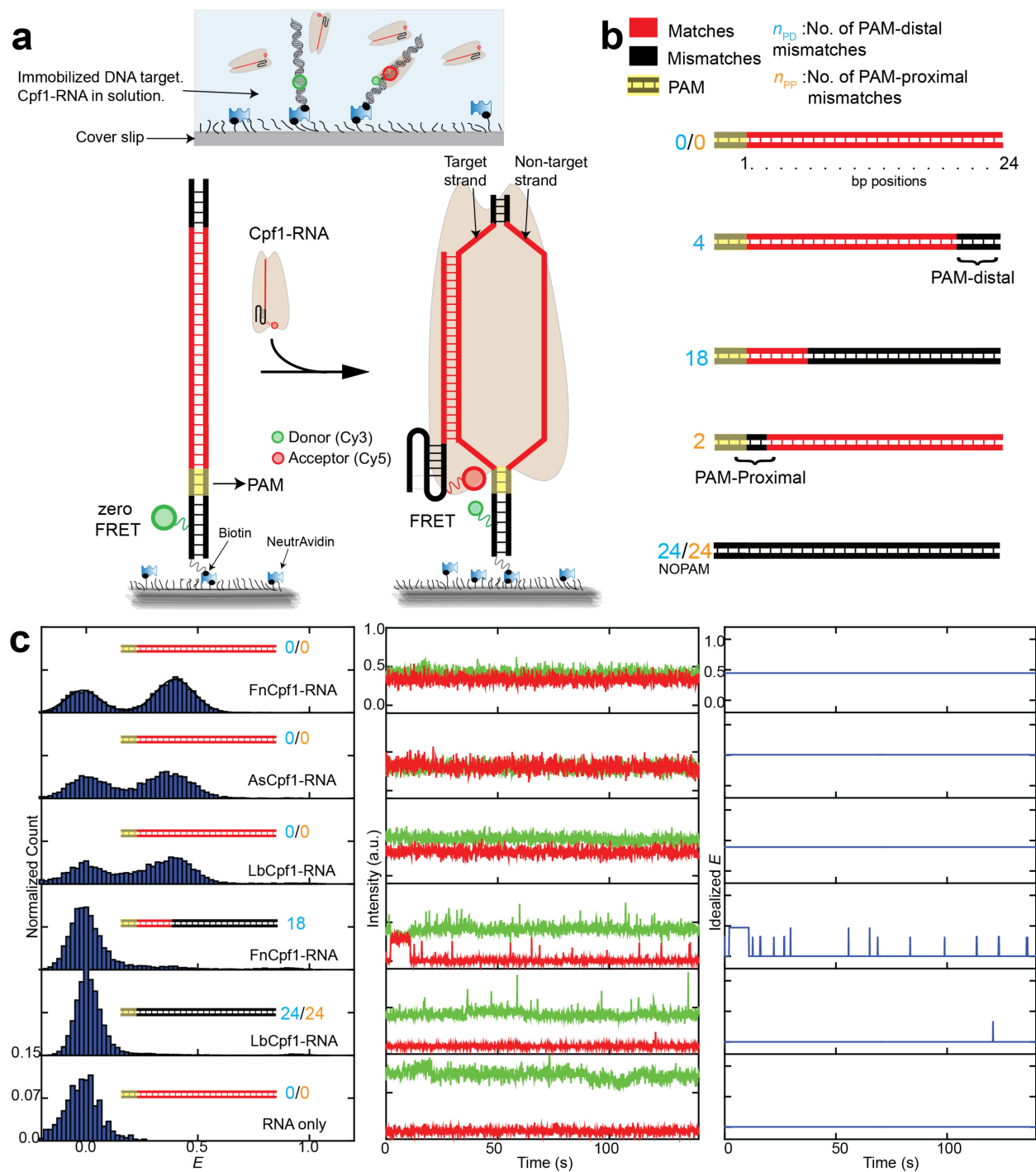


**e** FnCpf1-RNA-DNA



**Figure 4.1 | Design of DNA targets and guide-RNA along with FRET probes labeling locations in the Cpf1-RNA-DNA complex for smFRET assay for DNA interrogation by Cpf1-RNA.**

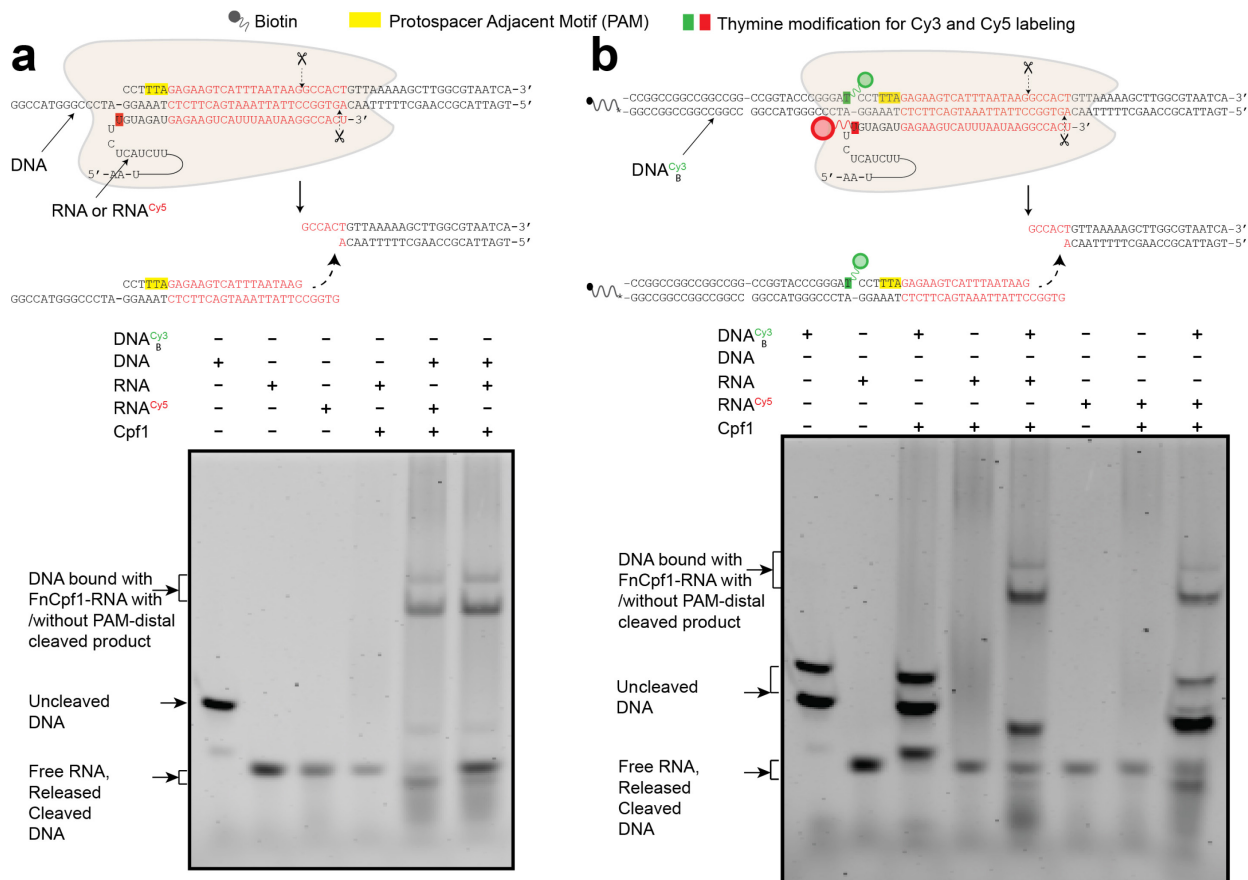
**(a)** Description of single-stranded DNA oligonucleotides with appropriate modifications for constitution of a fully duplexed DNA target for use in the smFRET assay. Oligonucleotides referred to as smFRET Cy3 oligo and Biotin oligo provide donor label for the smFRET assay and anchor for surface immobilization of the fully duplexed DNA target respectively. These oligonucleotides were same for all DNA targets. Other two strands were unmodified oligonucleotides, and strand that hybridizes with the guide-RNA (of Cpf1-RNA) is referred to as target strand and strand complementary to the target strand is called non-target strand. Base sequence of the target and non-target strands were changed to create DNA targets with mismatches against fixed guide-RNA sequence. These smFRET experiments were done both with DNA targets with ‘nick’ close to PAM at the indicated location and with DNA targets where this nick was ligated to close it. **(b)** Illustrated schematic of a complete Cpf1-RNA-DNA (b; As/FnCpf1-RNA-DNA and LbCpf1-RNA-DNA) complex showing base-pairing between different components. Sequences written in red denote cognate sequence in the DNA target and complementary sequence in the guide-RNA. **(c)** Fluorescent (Cy3 and Cy5) labeling locations shown in structure of AsCpf1-RNA bound to cognate DNA target (PDB ID: 5B43)<sup>114</sup>. As mentioned, strand which hybridizes with guide-RNA to form RNA-DNA heteroduplex is referred to as the target strand while the other strand, containing the PAM (5'-YTN-3'), is the non-target strand. **(d)** Cy5 labeling location shown in structure of LbCpf1-RNA complex (PDB ID: 5ID6)<sup>116</sup>. **(e)** Cy5 labeling location shown in structure of FnCpf1-RNA-DNA complex (PDB ID: 5MGA)<sup>117</sup>.



**Figure 4.2 | smFRET assay to study DNA interrogation by Cpf1-RNA.**

(a) Schematic of single-molecule FRET assay. Cy3-labeled DNA immobilized on a passivated surface is targeted by a Cy5-labeled guide-RNA in complex with Cpf1, referred to as Cpf1-RNA. (b) DNA targets

with mismatches in the protospacer region against the guide-RNA. The number of mismatches PAM-distal ( $n_{PD}$ ) and PAM-proximal ( $n_{PP}$ ) are shown in cyan and orange, respectively. (c)  $E$  histograms (left) at 50 nM Cpf1-RNA or 50 nM RNA only. Representative single molecule intensity time traces of donor (green) and acceptor (red) are shown (middle), along with  $E$  values idealized (right) by hidden Markov modeling<sup>90</sup>.

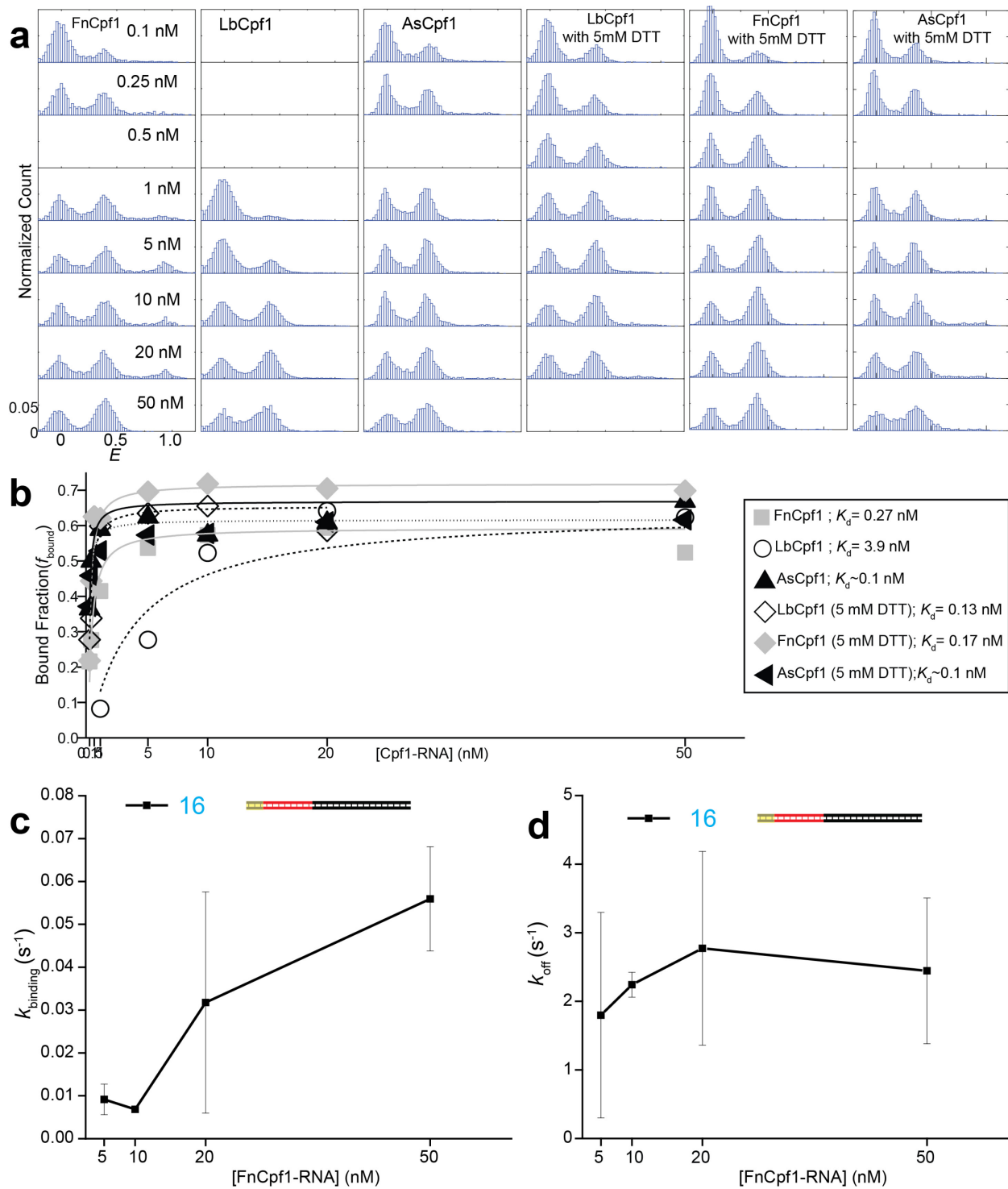


**Figure 4.3 | Fncpf1-RNA activity is not impaired by fluorescent labeling of guide-RNA and DNA target.**

Fncpf1-RNA induced DNA cleavage & binding analyzed by 4% native agarose gel electrophoresis and visualization of SYBR Gold II stained nucleic acids. (a) Activity of unmodified RNA and Cy5 labeled RNA for Fncpf1-RNA targeting of cognate DNA target without Cy3 or Biotin label. (b) Activity of unmodified RNA and Cy5 labeled RNA for Fncpf1-RNA targeting of cognate DNA target labeled with

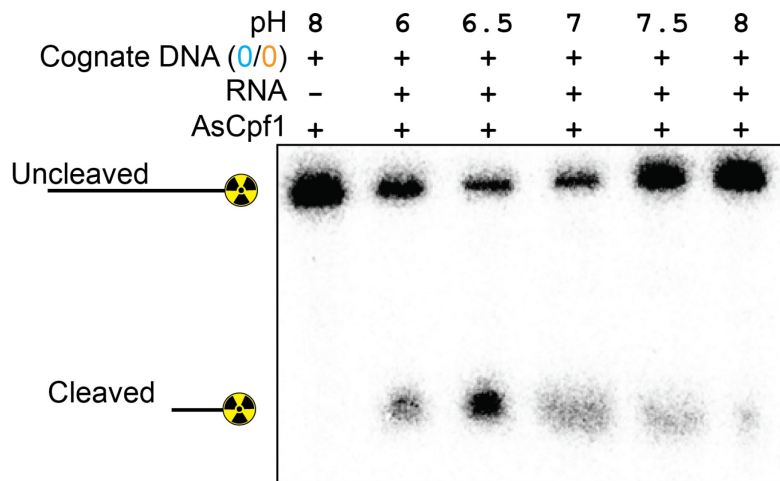
Cy3 and Biotin. It must be noted that the strands used to constitute dsDNA target with and without Cy3 and Biotin label were mixed with excess of non-target strand to ensure a near 100% hybridization of target strand with non-target strand, which results in multiple bands for the DNA substrate.





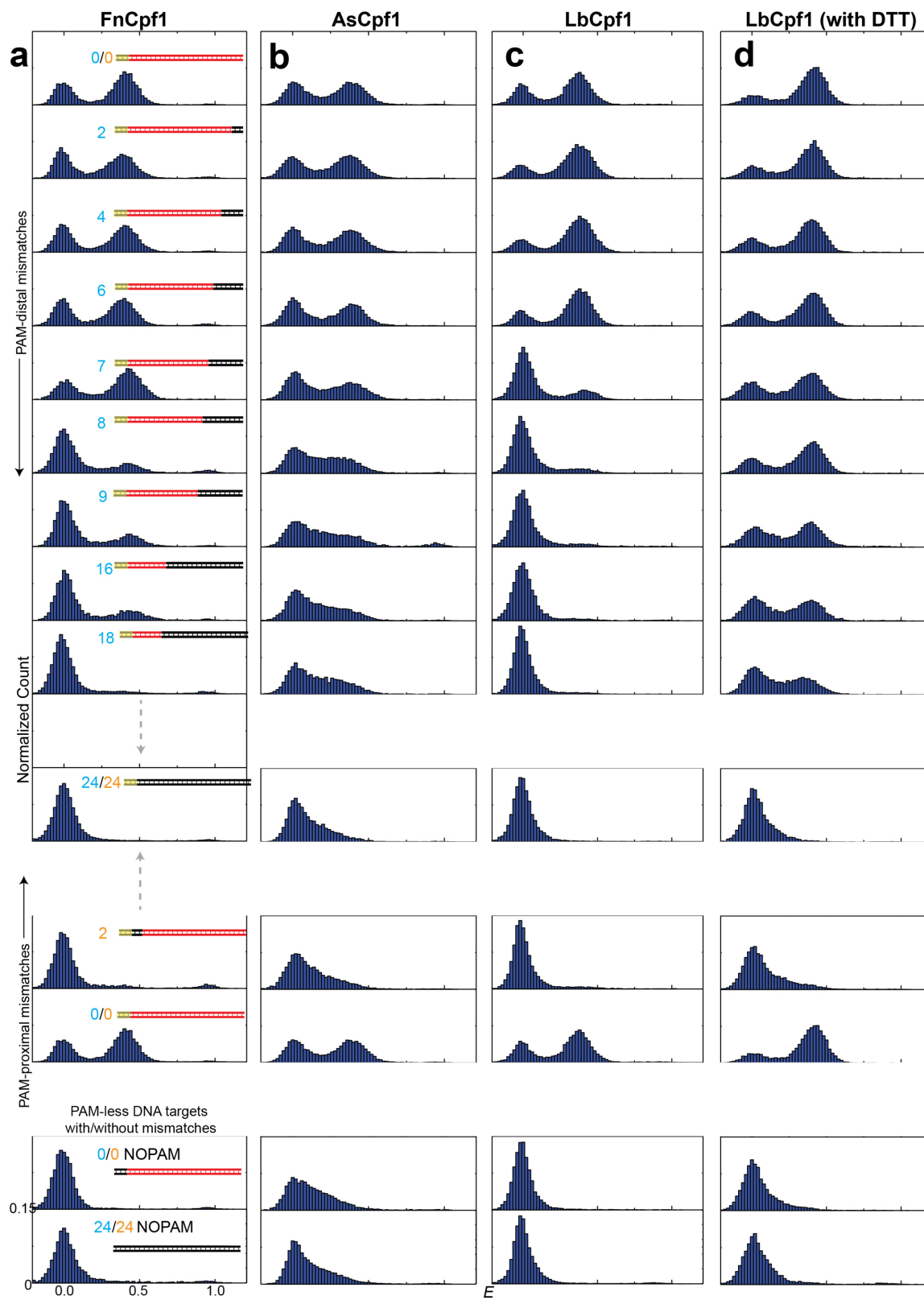
**Figure 4.4 | Bound fraction and rates of FRET appearance and disappearance with increasing Cpf1-RNA concentration.**

(a) *E* histograms demonstrating increase in Cpf1-RNA bound state population of cognate DNA targets with increasing Cpf1-RNA concentration for different Cpf1 orthologs. (b) Cpf1-RNA bound fraction defined as a fraction of population of DNA target molecules with FRET ( $>0.2$  &  $<0.6$ ) vs. Cpf1-RNA concentration. Trend was fit to obtain disassociation constant ( $K_d$ ) of Cpf1-RNA and DNA interaction. (C, D) After a hidden Markov analysis<sup>90</sup> on smFRET time-trajectories, bound and unbound states were divided based on the FRET values via thresholding at FRET= 0.2. The FRET states  $> 0.2$  were taken as putative bounds states. Dwell-times of bound states were used to estimate the overall bound state lifetime, inverse of which was taken as FnCpf1-RNA-DNA dissociation rate ( $k_{off}$ ). The dwell-times of the unbound states were used to estimate rate of FnCpf1-RNA and DNA association ( $k_{binding}$ ). (c) The rate of binding for DNA target with  $n_{PD}=16$  showing a linear increase with the increasing concentration of FnCpf1-RNA. (d) The rate of FnCpf1-RNA-DNA dissociation for DNA target with  $n_{PD}=16$  remained largely unchanged at different FnCpf1-RNA concentrations. Error bars represent s.d for replicate experiments. Number of PAM-distal mismatches ( $n_{PD}$ ) is shown in cyan.



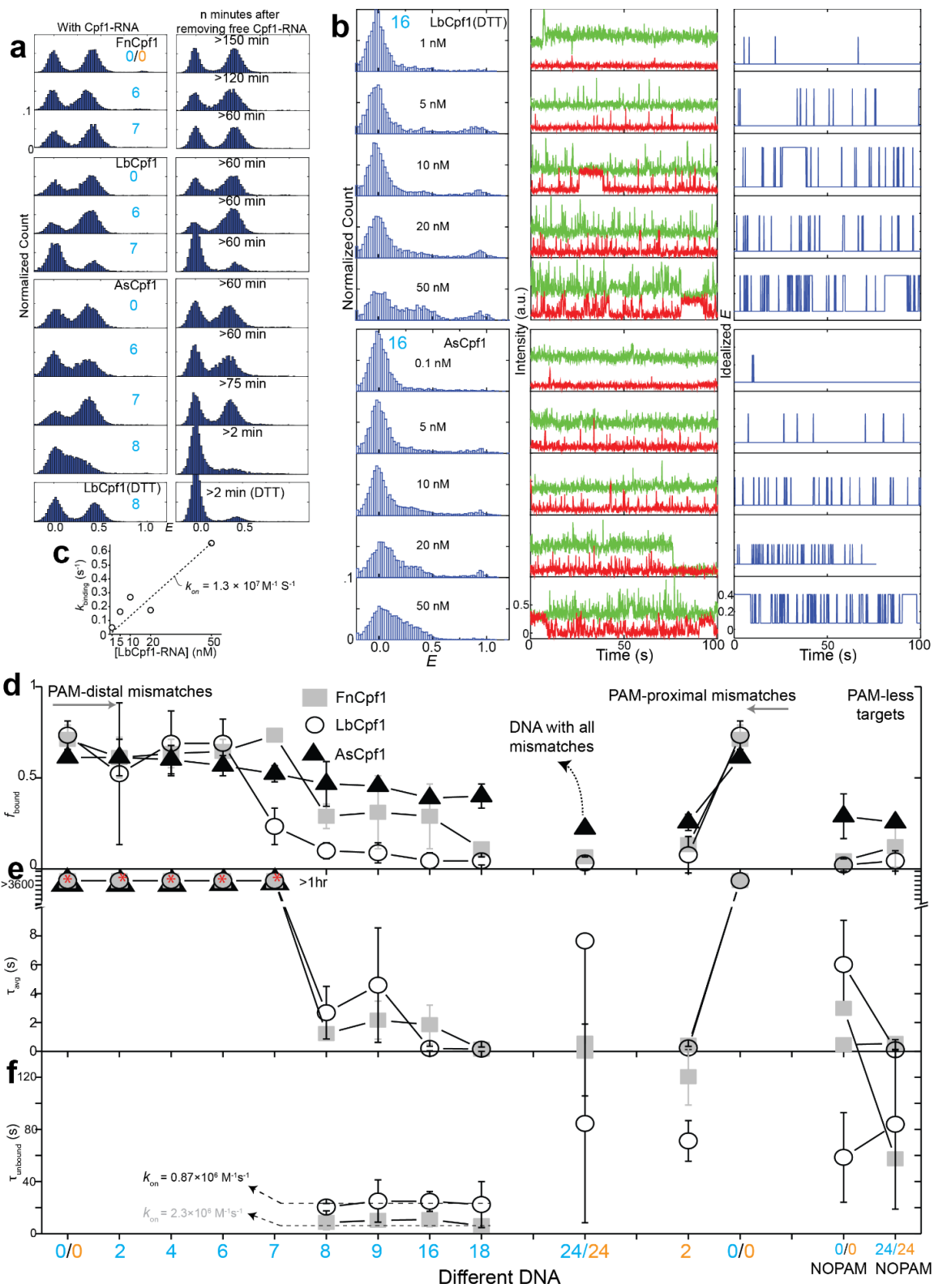
**Figure 4.5 | Cleavage activity of AsCpf1 at different pH conditions.**

Number of PAM-distal mismatches ( $n_{PD}$ ) is shown in cyan.



**Figure 4.6 | *E* histograms during DNA interrogation by Cpf1-RNA.**

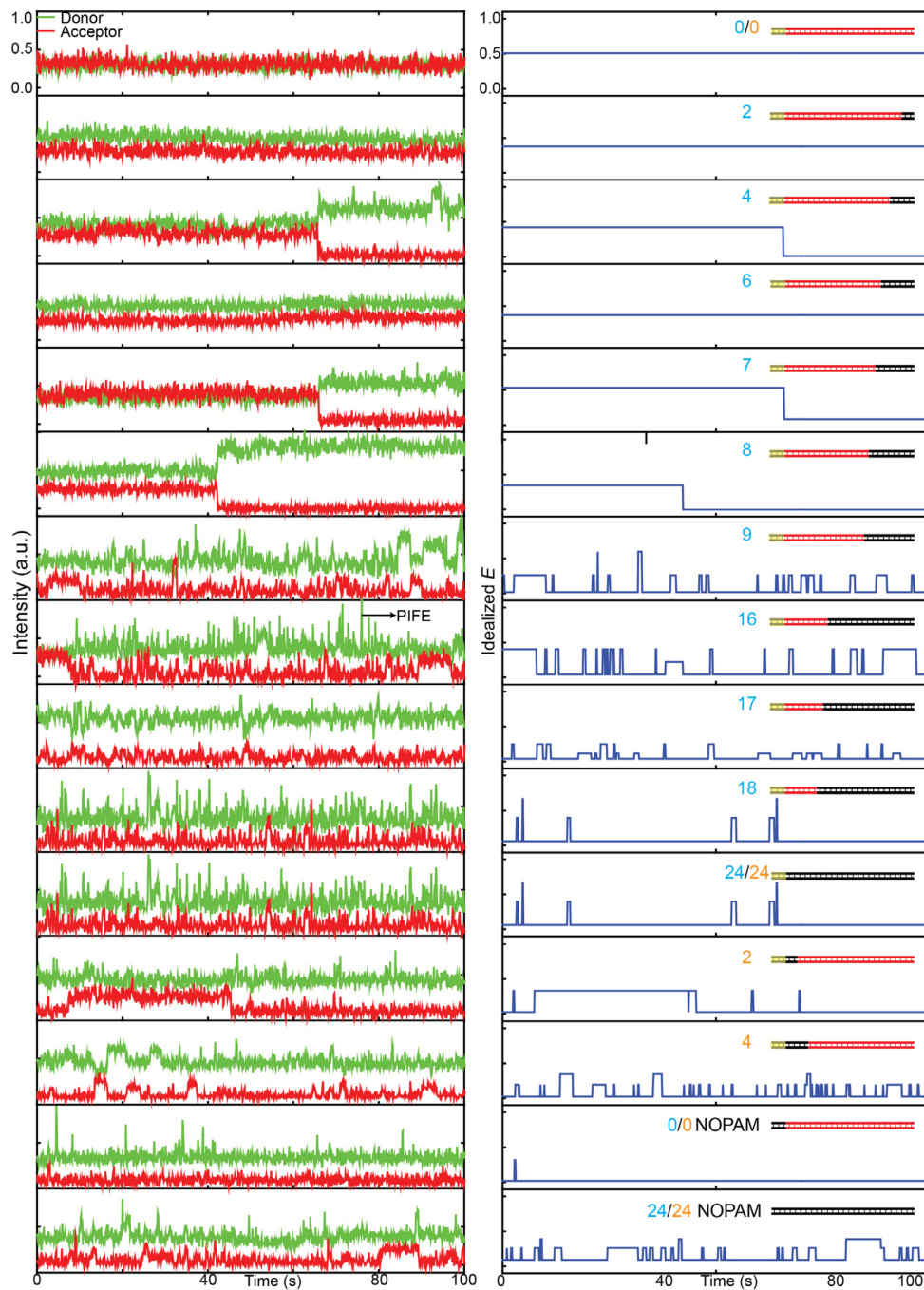
**(a)** FnCpf1. **(b)** AsCpf1. **(c)** LbCpf1. **(d)** LbCpf1 (in reducing conditions of 5 mM DTT). Number of PAM-distal ( $n_{PD}$ ) and PAM-proximal mismatches ( $n_{PP}$ ) are shown in cyan and orange respectively. [Cpf1-RNA] = 50 nM.



**(a)**  $E$  histograms for various  $n_{PD}$  with 50 nM Cpf1-RNA (left) and indicated minutes after free Cpf1-RNA was washed out (right) for FnCpf1, LbCpf1, AsCpf1 and LbCpf1 in reducing condition of 5 mM DTT.

**(b)**  $E$  histograms (left) and representative smFRET time-trajectories (middle) with their idealized  $E$  values (right) for  $n_{PD} = 16$  at various concentrations of LbCpf1-RNA in reducing condition and AsCpf1-RNA.

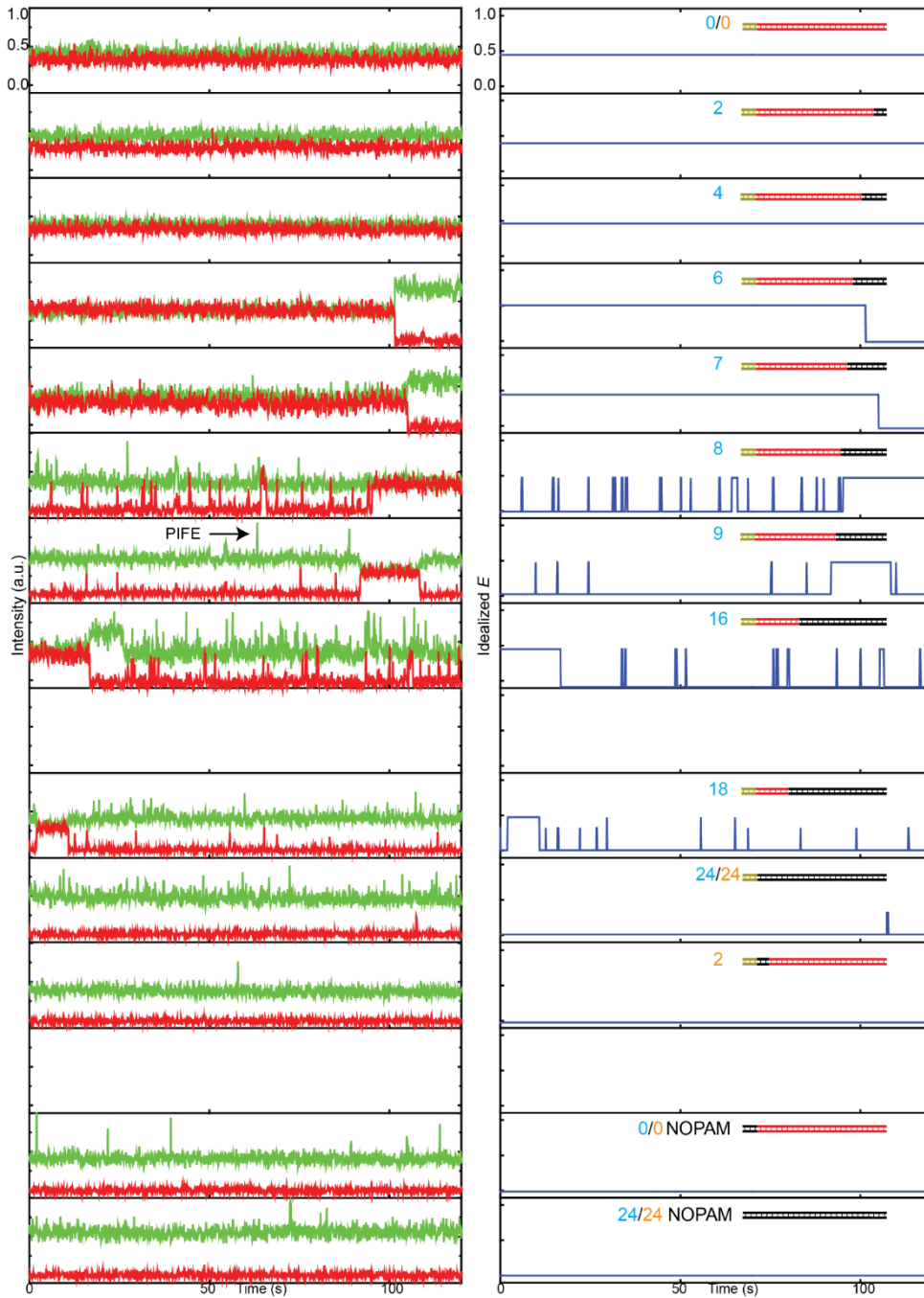
**(c)** Rate of LbCpf1-RNA and DNA association ( $k_{\text{binding}}$ ) at different LbCpf1-RNA concentration.  $E > 0.2$  and  $E < 0.2$  states were taken as putative bound and unbound states. Dwell-times of the unbound states were used to calculate  $k_{\text{binding}}$ . **(d)**  $f_{\text{bound}}$ , **(e)** bound state lifetime, **(f)** unbound state lifetime for various mismatches at 50 nM Cpf1-RNA. Average of rates of binding ( $\tau_{\text{inbound}}^{-1}$ ) of DNA with  $n_{PD} = 8-18$  were used to calculate  $k_{\text{on}}$  for FnCpf1 and LbCpf1.  $n_{PD}$  and  $n_{PP}$  are shown in cyan and orange, respectively.



**Figure 4.8 | Representative smFRET time-trajectories from smFRET experiments to study DNA interrogation by AsCpf1-RNA.**

These representative smFRET time- trajectories (left) along with their idealized FRET values (right) are taken from representative DNA targets with/without nick near the PAM (Figure 4.1) and no differences were observed between them. The indicated protein induced fluorescence enhancement (PIFE) on the

donor only signal is likely resulting from non-specific interaction of free-excess of Apo AsCpf1 with the DNA and was observed for all the DNA targets. Number of PAM-distal ( $n_{PD}$ ) and PAM-proximal mismatches ( $n_{PP}$ ) are shown in cyan and orange respectively.

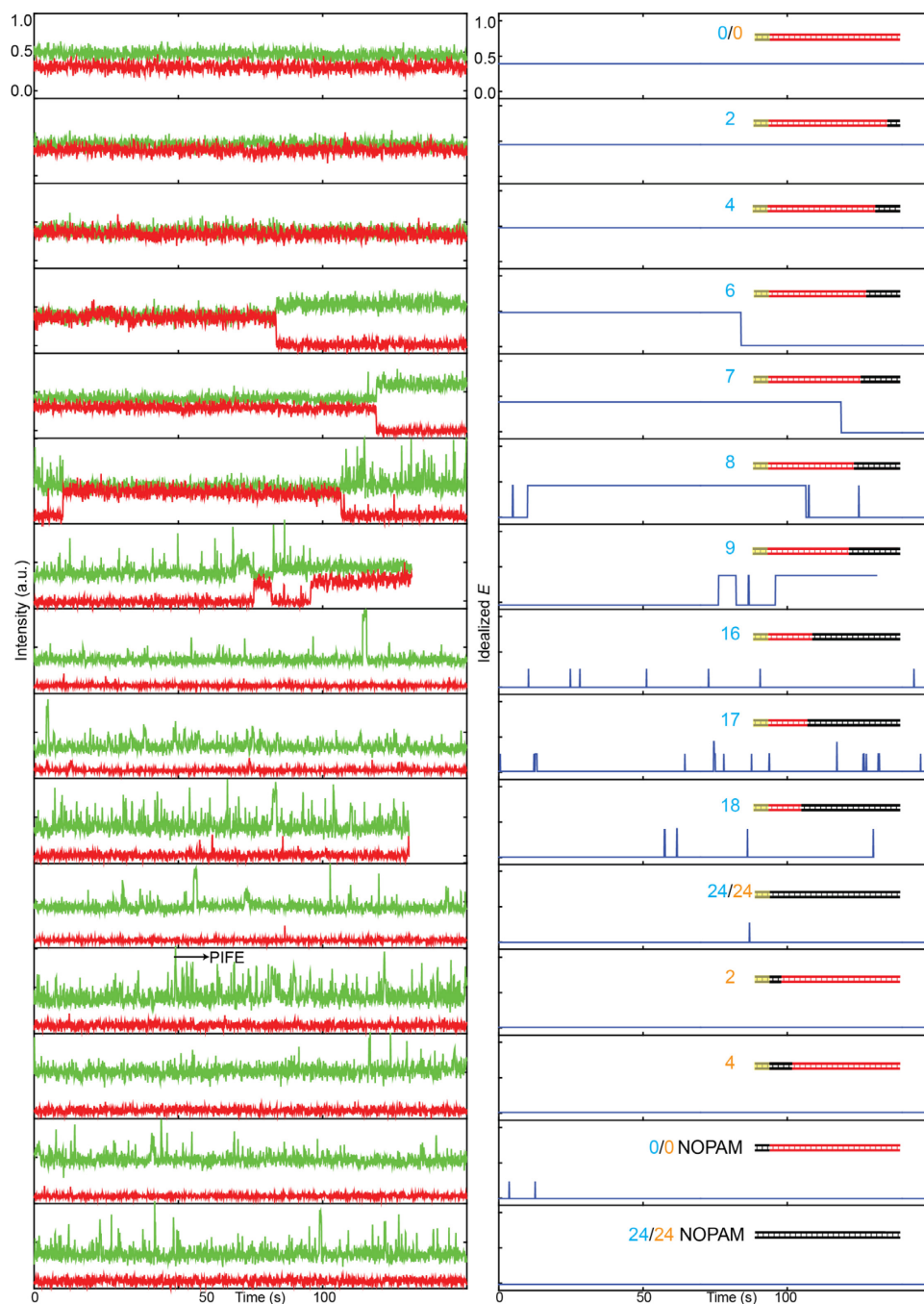




**Figure 4.9 | Representative smFRET time-trajectories from smFRET experiments to study DNA interrogation by FnCpf1-RNA.**

These representative smFRET time- trajectories (left) along with their idealized FRET values (right) are taken from representative DNA targets with/without nick near the PAM (Figure 4.1) and no differences were observed between them. The indicated protein induced fluorescence enhancement (PIFE) on the donor only signal is likely resulting from non-specific interaction of the free-excess of Apo FnCpf1 with the DNA and was observed for all the DNA targets. DNA targets with nick (Figure 4.1) exhibit stronger anti-correlation between donor and acceptor signal for transient FRET binding events. PIFE upon the FnCpf1-RNA binding can influence this anti-correlation and slightly different levels of PIFE on Cy3 in DNA target with and without nick i.e. different flexibilities could be causing these variations.

Photophysics or isomerization of Cy3 resulting PIFE is guided by local DNA flexibility<sup>118</sup>. Number of PAM-distal ( $n_{PD}$ ) and PAM-proximal mismatches ( $n_{PP}$ ) are shown in cyan and orange respectively.

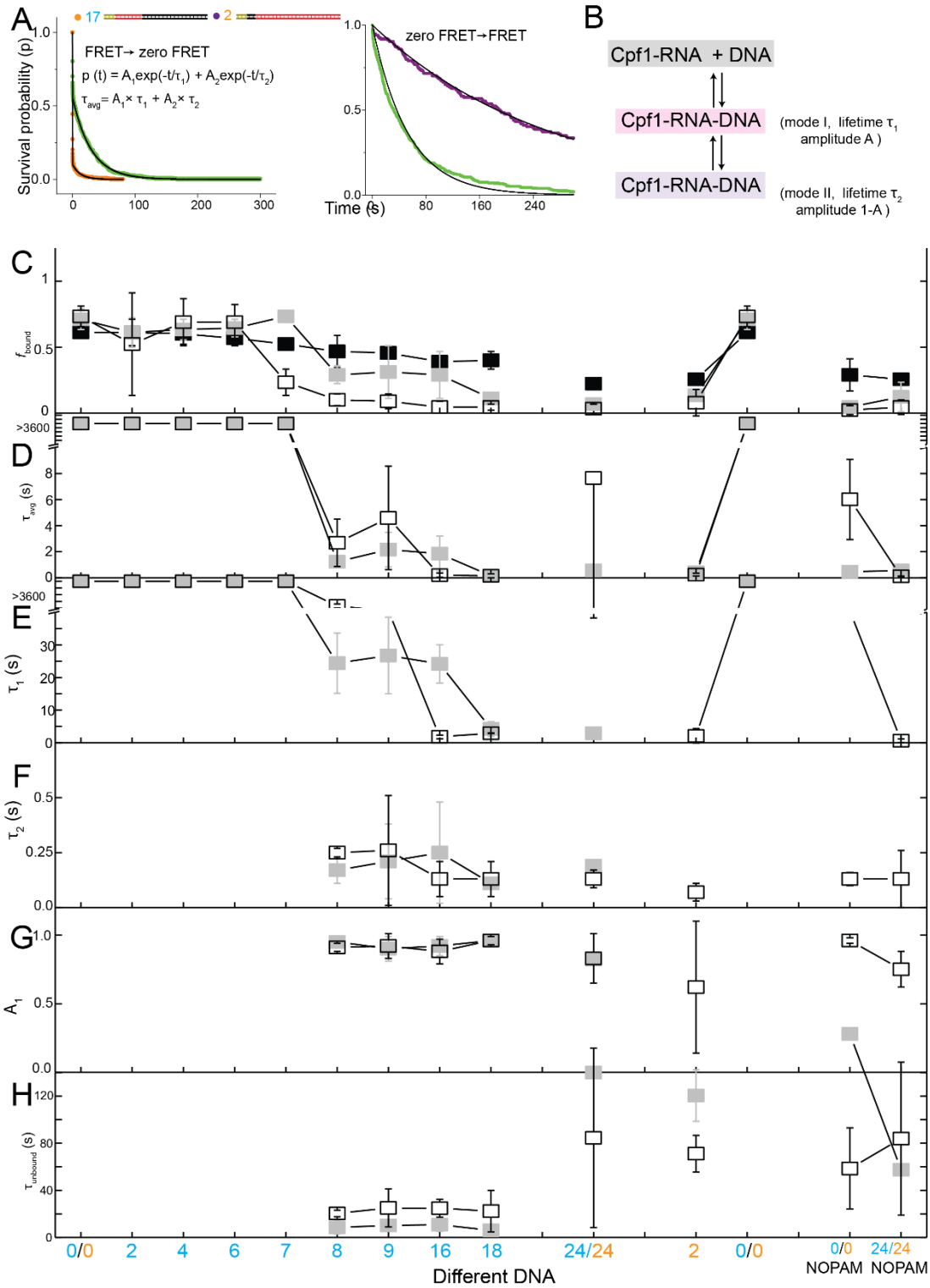


**Figure 4.10 | Representative smFRET time-trajectories from smFRET experiments to study DNA interrogation by LbCpf1-RNA.**

These representative smFRET time- trajectories (left) along with their idealized FRET values (right) are taken from representative DNA targets with/without nick near the PAM (Figure 4.1) and no differences were observed between them. The indicated protein induced fluorescence enhancement (PIFE) on the

donor only signal is likely resulting from non-specific interaction of the free-excess of Apo LbCpf1 with the DNA and was observed for all the DNA targets. DNA targets with nick (Figure 4.1) exhibit stronger anti-correlation between donor and acceptor signal for transient FRET binding events. PIFE upon the LbCpf1-RNA binding can influence this anti-correlation and slightly different levels of PIFE on Cy3 in DNA target with and without nick i.e. different flexibilities could be causing these variations.

Photophysics or isomerization of Cy3 resulting PIFE is guided by local DNA flexibility<sup>118</sup>. Number of PAM-distal ( $n_{PD}$ ) and PAM-proximal mismatches ( $n_{PP}$ ) are shown in cyan and orange respectively.



**Figure 4.11 | Bimodal nature of DNA interrogation by Cpf1-RNA and its parameters (50 nM Cpf1-RNA).**

**(a)** Survival probability of FRET state ( $E > 0.2$ ; putative bound states) and zero FRET state ( $E < 0.2$ ; unbound states) dwell-times vs. time, fit with double-exponential and single exponential decay to obtain lifetime of bound state ( $\tau_{\text{avg}}$ ) and unbound state ( $\tau_{\text{unbound}}$ ) respectively.

**(b)** Model describing a bimodal binding nature of Cpf1-RNA. **(c)** Fraction of DNA molecules bound with Cpf1-RNA ( $f_{\text{bound}}$ ). **(d)** Amplitude-weighted lifetime,  $\tau_{\text{avg}}$ , of the Cpf1-RNA bound state. **(e)** Lifetime of long-lived binding mode ( $\tau_1$ ). **(f)** Lifetime of transient ( $\tau_2$ ) binding mode which was similar for all the DNA targets ( $\sim < 0.5$  s). **(g)** Amplitude of the transient binding mode only. **(h)** unbound state lifetime. Error bars represent s.d. for replicate experiments. Large error bars were observed due to the under-sampling (i.e. below detection limit of 0.1s) of extremely transient binding events for certain DNA targets, especially with PAM-proximal mismatches or without PAM. This indicates that Cpf1-RNA rejected DNA faster if there were PAM-proximal mismatches or in absence of PAM. Number of PAM-distal ( $n_{\text{PD}}$ ) and PAM-proximal mismatches ( $n_{\text{PP}}$ ) are shown in cyan and orange respectively.

### 4.3.2 DNA cleavage by Cpf1 as a function of mismatches

Next, we performed gel-based experiments using the same set of DNA targets to measure cleavage by Cpf1. Cleavage was observed at a wide range of temperatures (4-37 °C), required divalent ions ( $\text{Ca}^{2+}$  could substitute for  $\text{Mg}^{2+}$ ), and showed a pH dependence. AsCpf1 is most active only at slightly acidic to neutral pH (6.5-7.0) whereas FnCpf1 has more activity at pH 8.5 than pH 8.0 (Figure 4.12 and Figure 4.13 and Figure 4.14). Cleavage required 17 PAM-proximal matches, corresponding to  $n_{\text{PD}} \leq 7$ , (Figure 4.15a, Figure 4.12 and Figure 4.13 ) which is identical to the threshold for stable binding (Figure 4.6 and Figure 4.7). This contrasts with Cas9, which requires only 9 PAM-proximal matches for stable binding<sup>83</sup> but 16-18 PAM-proximal matches for cleavage<sup>4,6</sup>.

We measured the time it takes to cleave DNA,  $\tau_{\text{cleavage}}$  (Figure 4.16).  $\tau_{\text{cleavage}}$  remained approximately the same among DNA with  $0 \leq n_{\text{PD}} \leq 6$  for FnCpf1 (30-60 s) but steeply increased upon increasing  $n_{\text{PD}}$  to 7

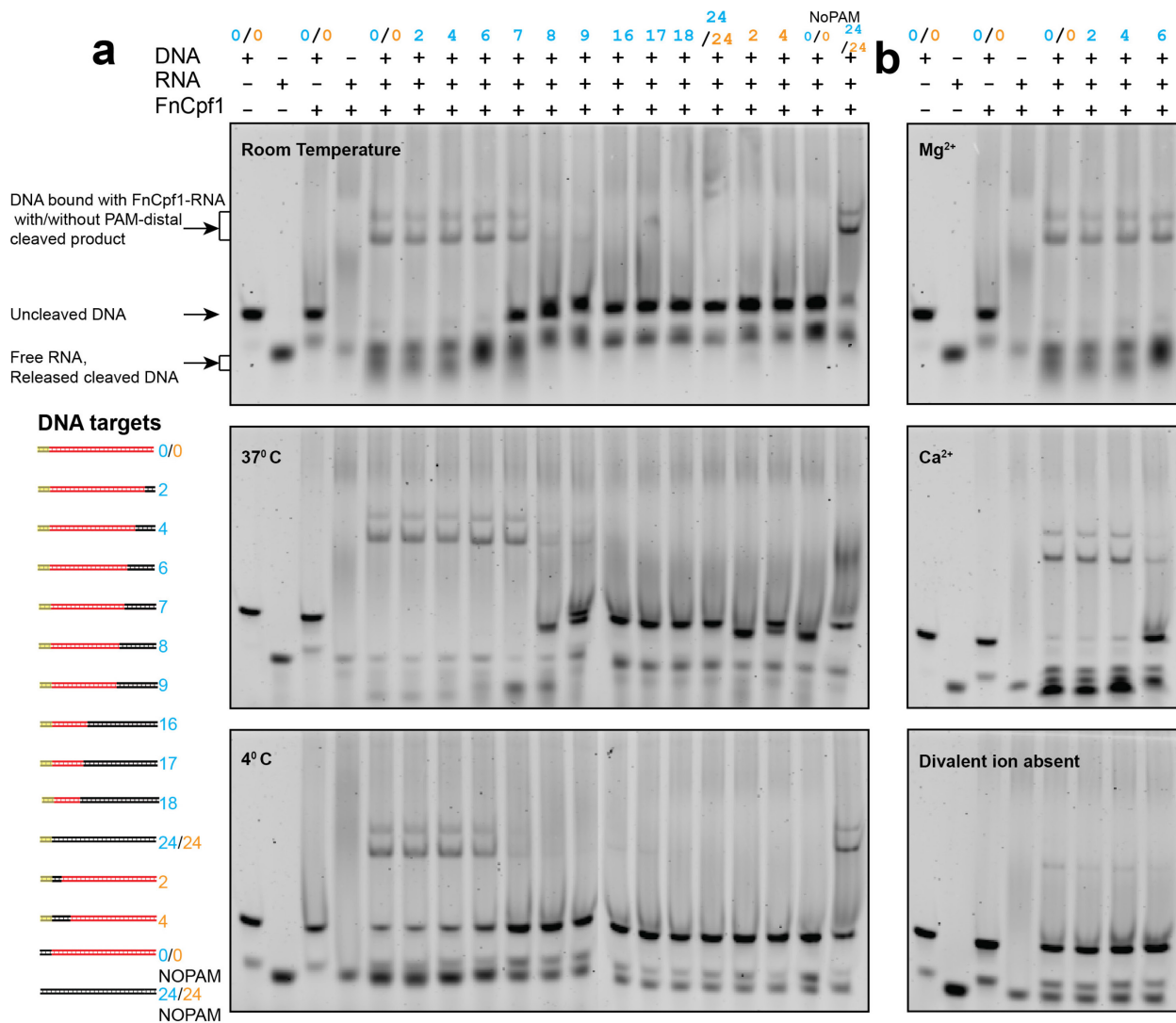
(Figure 4.15b and Figure 4.15c). AsCpf1 showed a more complex  $n_{PD}$  dependence with a minimal  $\tau_{\text{cleavage}}$  value of 8 minutes for  $n_{PD} = 6$ . (Figure 4.15c).  $\tau_{\text{cleavage}}$  is much longer than the 1 to 15 seconds it takes Cpf1-RNA to bind the DNA at the same Cpf1-RNA concentration, suggesting that Cpf1-RNA-DNA undergoes additional rate-limiting steps after DNA binding and before cleavage. These additional steps are likely the conformational rearrangement of Cpf1-RNA-DNA complex that position the nuclease domains and DNA strands for cleavage, as has been described in structural analysis of Cpf1-RNA-DNA complex<sup>114,117</sup>.

Because of the finite  $\tau_{\text{cleavage}}$  we can infer that the ultra-stable binding (lifetime > 1 hr) for  $n_{PD} \leq 7$  is that of Cpf1-RNA binding to the cleaved product, and it is in principle possible that cleavage stabilizes Cpf1-RNA binding. In order to test this possibility, we purified catalytically dead FnCpf1 (dFnCpf1) and performed DNA interrogation experiments. dFnCpf1 binding was ultra-stable for cognate DNA but showed a substantial dissociation after 5-10 min for  $n_{PD}=6$  or 7 (Figure 4.17). Therefore, cleavage can further stabilize Cpf1-RNA binding to DNA. Cleavage was negligible for DNA targets that showed transient binding. Therefore, transient binding and dissociation we observed is not to and from a cleaved DNA product.

### 4.3.3 Fate of cleaved DNA.

For an efficient addition of a new piece of DNA at a cleaved site, the cleaved site needs to be exposed<sup>94</sup>. To investigate the fate of DNA targets post cleavage, we relocated the Cy3 label to a PAM-distal DNA segment that would depart the imaging surface if the Cpf1 releases cleavage product(s) (Figure 4.15d and Figure 4.18). The number of fluorescent spots decreased over time (Figure 4.15e), suggesting the cleavage product is released under physiological conditions, which is in stark contrast to Cas9, which holds onto the cleaved DNA and does not release except in denaturing condition<sup>6,83</sup>. Cpf1 releases only the PAM-distal cleavage product, however, because when Cy3 is attached to a site on the PAM-proximal

cleavage product, the number of fluorescence spots did not decrease over time (Figure 4.2, Figure 4.6 and Figure 4.7). The average time for fluorescence signal disappearance ranged from ~30 s to 30 min depending on the PAM-distal mismatches and Cpf1 orthologues. By subtracting the time it takes to bind and cleave, we estimated the product release time scale ( $\tau_{\text{release}}$ ) (Figure 4.15f), which showed a dependence on  $n_{\text{PD}}$ . Therefore, PAM-distal mismatches can also affect product release.

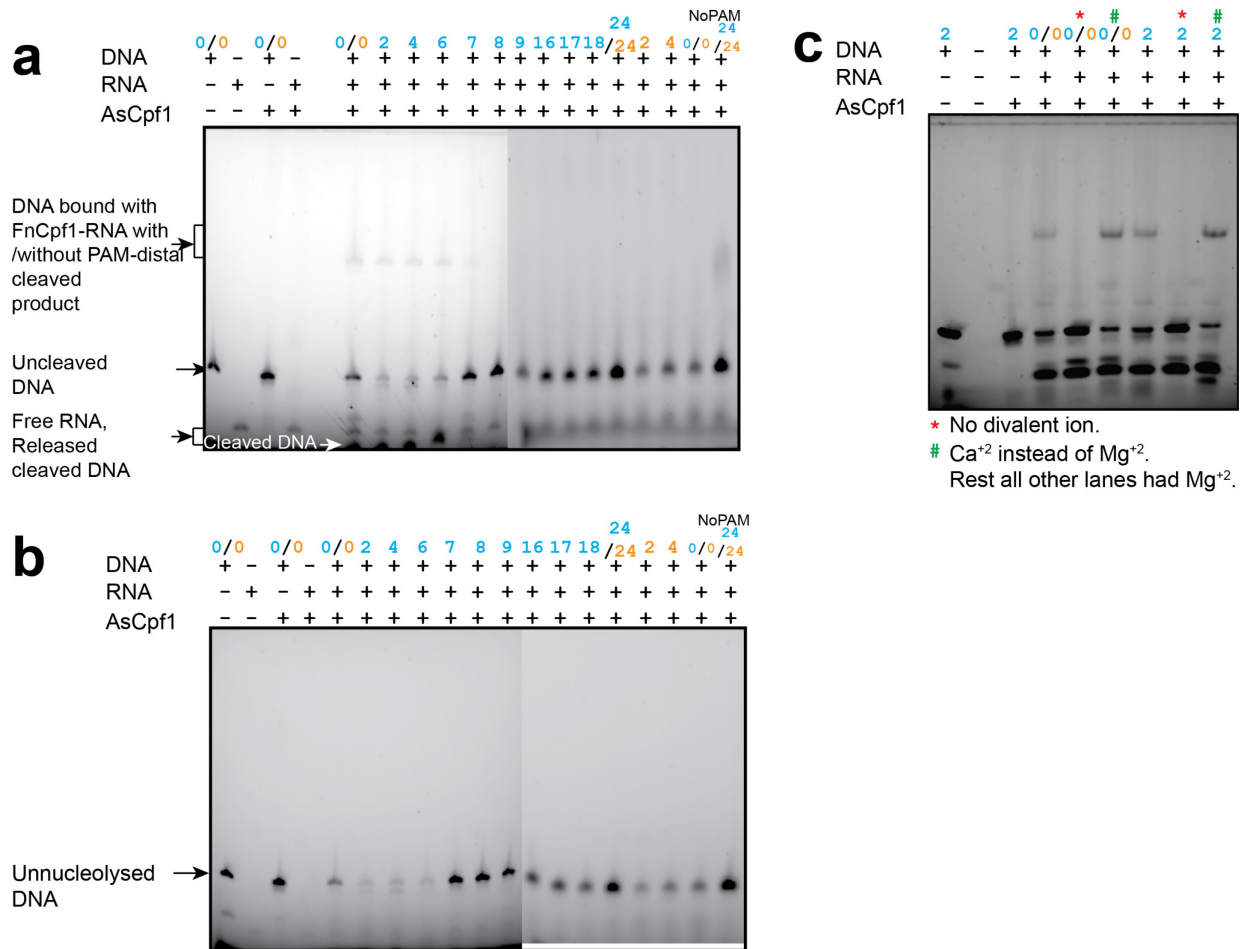


**Figure 4.12 | FnCpf1 activity at different temperatures and divalent cation conditions.**

DNA cleavage and binding by FnCpf1 at **(a)** different temperature and **(b)** divalent cation conditions analyzed by 4% native agarose gel electrophoresis and SYBR Gold II staining of nucleic acids. **(a)** DNA

targets with  $n_{PD}$  ranging from 0 to 7 were stably bound and cleaved by FnCpf1, as seen from depletion of uncleaved DNA band for these DNA targets. Binding and cleavage cut-off was observed at  $n_{PD} = 7$  (substantially reduced cleavage activity), beyond which all the DNA targets remained unbound and uncleaved. FnCpf1 remained active across different temperatures but its efficiency was higher at 37 °C and lower at 4 °C compared to room temperature. Also, differences in cleavage efficiency  $n_{PD} = 7$  were markedly different across different temperature conditions. **(b)** Divalent cation is crucial for DNA targeting and cleavage by FnCpf1 as shown by the fully intact uncleaved DNA target band (bottom) for experiments without divalent cation, the binding was also significantly impaired for such cases with a faint band corresponding to the FnCpf1-RNA-DNA complex. Both  $Ca^{2+}$  and  $Mg^{2+}$  supported the FnCpf1 binding and cleavage activity (top, middle). All these experiments were performed with short dsDNA targets with only 6 bp flanking the protospacer (4 of which is for PAM). These results also underlie FnCpf1's ability to affect interference on short dsDNA targets. Sequences of DNA targets and guide-RNA used for these experiments is in Table 4.3 and Table 4.5. Number of PAM-distal ( $n_{PD}$ ) and PAM-proximal mismatches ( $n_{PP}$ ) are shown in cyan and orange respectively.

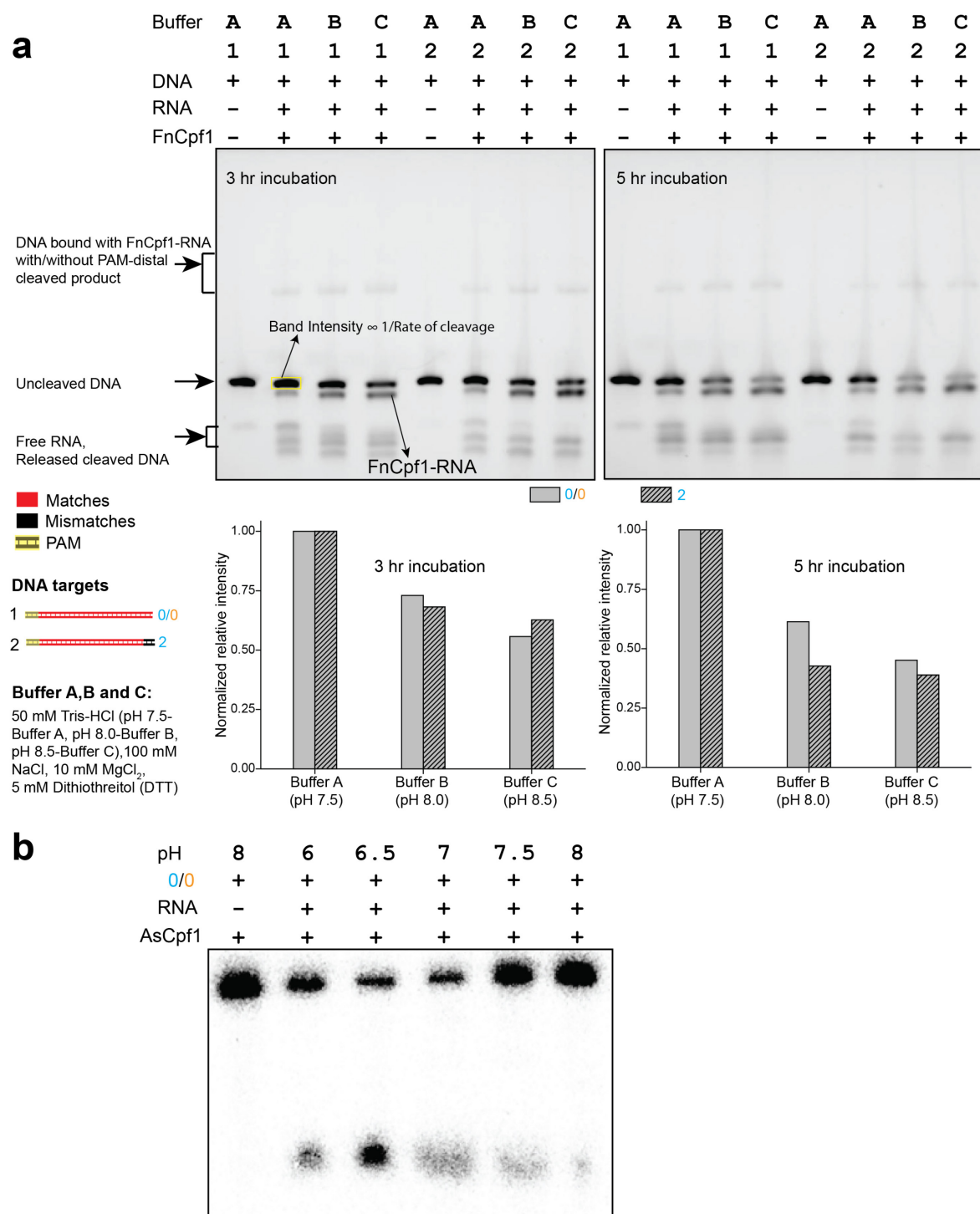




**Figure 4.13 | AsCpf1-RNA activity and effect of divalent cation conditions.**

(a) DNA cleavage & binding by AsCpf1 at room temperature analyzed by 4% (a) native and (b) denaturing agarose gel electrophoresis and SYBR Gold II staining of nucleic acids. (a-b) DNA targets with  $n_{PD}$  ranging from 0 to 7 were stably bound and cleaved by FnCpf1, as seen from depletion of uncleaved DNA band for these DNA targets. Binding and cleavage cut-off was observed at  $n_{PD} = 7$  (substantially reduced cleavage activity), beyond which all the DNA targets remained unbound and uncleaved. (c) Divalent cation is crucial for DNA targeting and cleavage by AsCpf1 as shown by the fully intact uncleaved DNA target band (bottom) for experiments without divalent cation, the binding was also significantly impaired for such cases with a faint band corresponding to the AsCpf1-RNA-DNA complex. Both Ca<sup>2+</sup> and Mg<sup>2+</sup> supported the AsCpf1 binding and cleavage activity (top, middle). All these experiments were performed with short dsDNA targets with only 6 bp flanking the protospacer (4 of

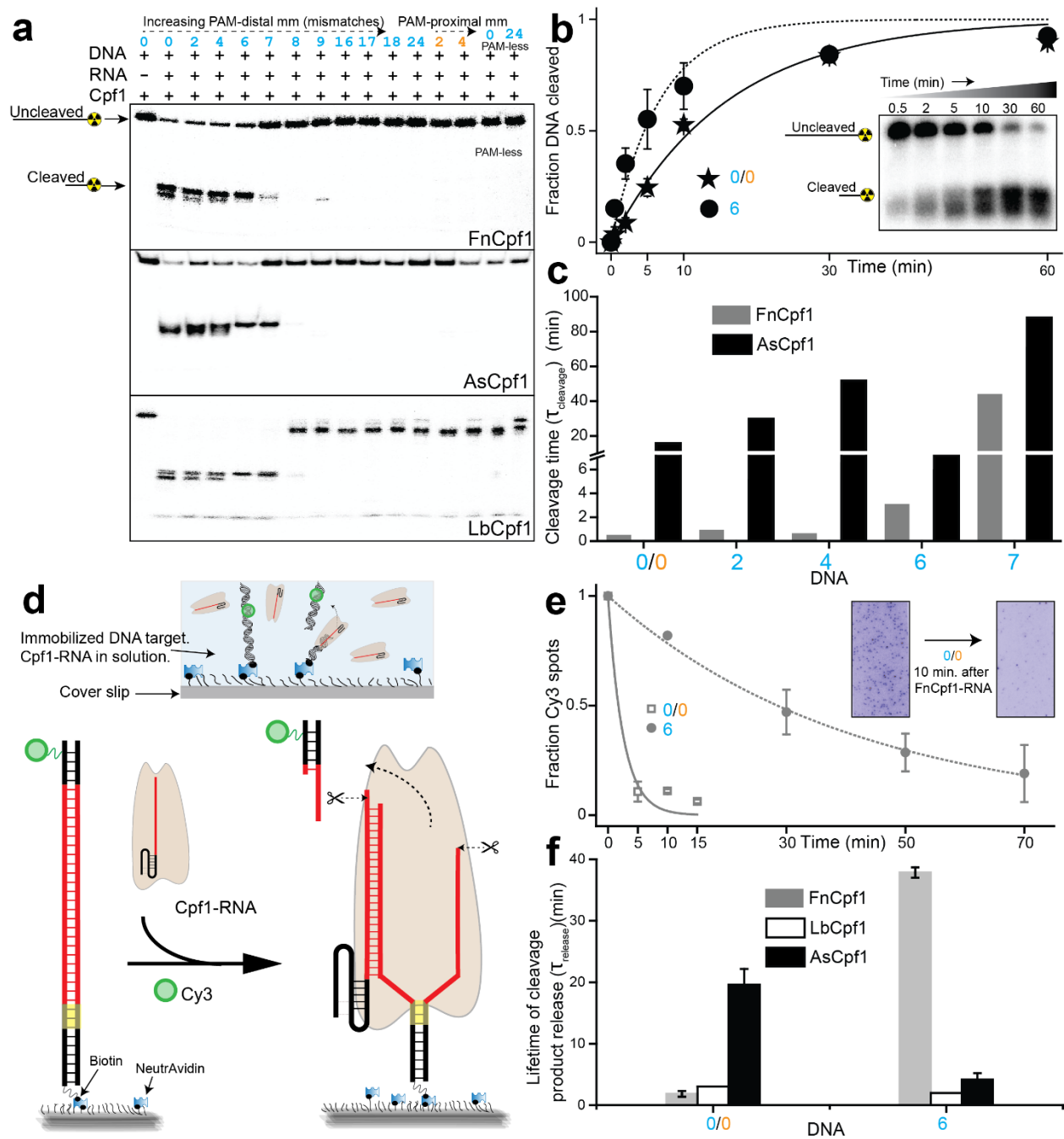
which is for PAM). These results also underlie AsCpf1's ability to affect interference on short dsDNA targets. Sequences of DNA targets and guide-RNA used for these experiments is in Table 4.3 and Table 4.5. All these experiments were performed in pH 8.0 conditions, and it was later discovered (Figure 4.5) that AsCpf1 shows strong pH dependence and works efficiently only at pH 6.5-7.0. Therefore, all subsequent experiments were performed at pH 7.0. Lower activity of the experiments shown here can be attributed to its high pH conditions. Number of PAM-distal ( $n_{PD}$ ) and PAM-proximal mismatches ( $n_{PP}$ ) are shown in cyan and orange respectively.



**Figure 4.14 | Cpf1-RNA activity at different pH conditions.**

**(a)** FnCpf1-RNA DNA cleavage and binding reactions incubated for 3 and 5 hr at different pH conditions analyzed by 4% native agarose gel electrophoresis and SYBR Gold II staining of nucleic acids. 60 nM of

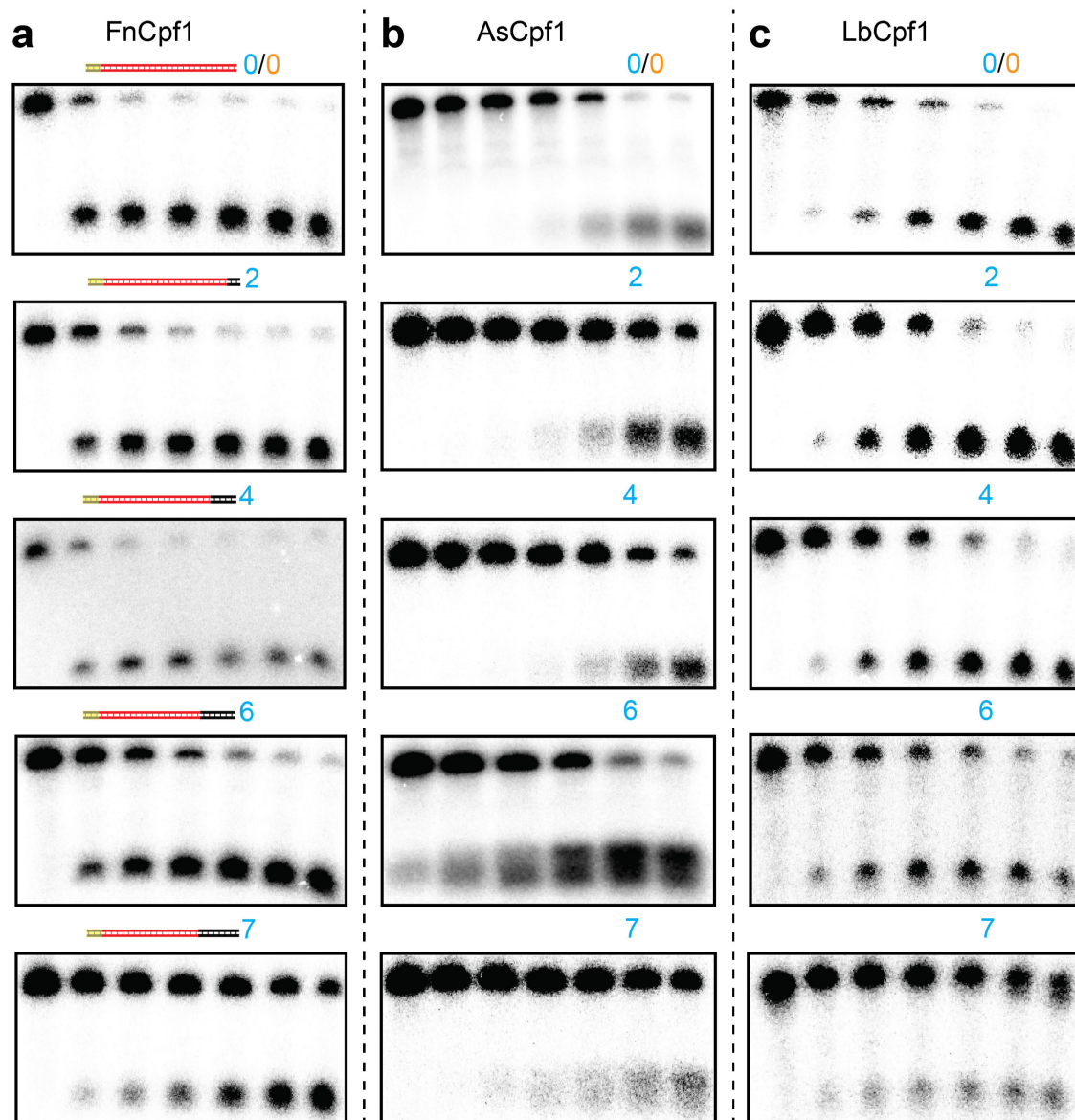
DNA targets was mixed with the ~300 nM of FnCpf1-RNA in reaction buffer volume of 20  $\mu$ L. 10  $\mu$ L of the reaction was analyzed at 3 hr time point and the remaining 10  $\mu$ L of the reaction was analyzed at 5 hr time point. For the reactions incubated for 3 hr, intensity of the uncleaved DNA target band was taken as being inversely proportional to rate of DNA interference by FnCpf1 and was ~37% and ~79% higher when pH of the Tris-HCl component used in the reaction buffer was 8.0 and 8.5 respectively as compared to 7.5. For the reactions incubated for 5 hr, the rate of cleavage was ~46% and ~59% higher when pH of the Tris-HCl component used in the reaction buffer was 8.0 and 8.5 respectively as compared to 7.5. **(b)** Effect of pH on AsCpf1 activity. Sequences of DNA targets and guide-RNA used for these experiments is in Table 4.3 and Table 4.5. Number of PAM-distal ( $n_{PD}$ ) and PAM-proximal mismatches ( $n_{PP}$ ) are shown in cyan and orange respectively.



**Figure 4.15 | DNA cleavage and product release.**

(a) Cpf1 induced DNA cleavage at room temperature analyzed by 10% denaturing polyacrylamide gel electrophoresis of radio-labeled DNA targets. (b) Fraction of DNA cleaved by AsCpf1 vs time for cognate and DNA with  $n_{PD}=6$ , and single exponential fits. A representative gel image is shown in inset. (c) Cleavage time ( $\tau_{\text{cleavage}}$ ) determined from cleavage time courses as shown in (b). (d) Schematic of single-

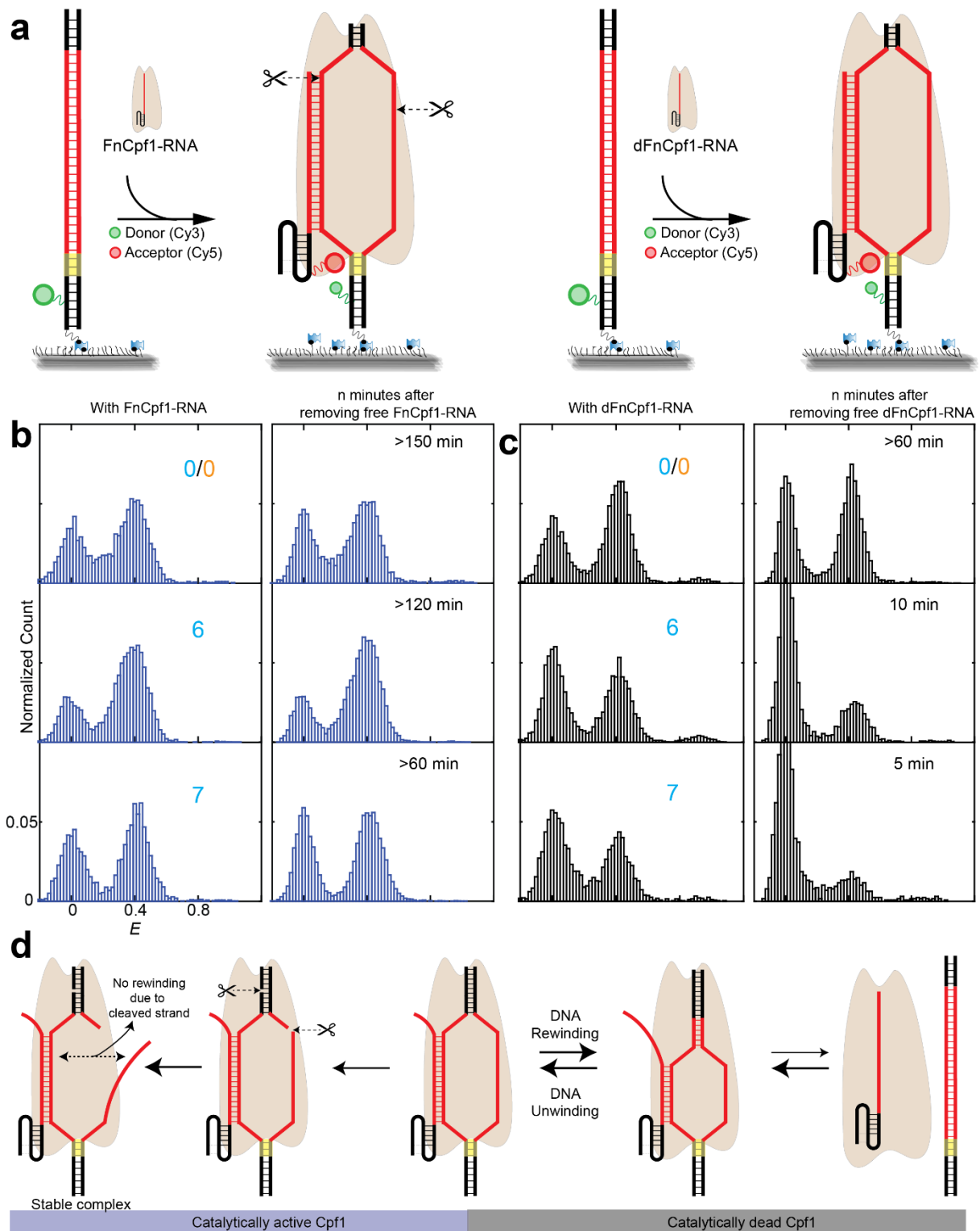
molecule cleavage product release assay. PAM-distal cleavage product release can be detected as disappearance of fluorescence signal from Cy3 attached to the PAM-distal product. **(e)** Average fraction of Cy3 spots remaining vs time for FnCpf1-RNA (50 nM). Inset shows images before and after 10 min reaction. **(f)** Average time of cleavage product release ( $\tau_{\text{release}}$ ).



**Figure 4.16 | Time-lapse of cleavage by Cpf1.**

Representative images of cleavage by Cpf1 at room temperature for different DNA targets as a function of time, from experiments utilizing 10% Polyacrylamide denaturing gel electrophoresis. Target strand in

DNA targets was radio-labeled. [Cpf1-RNA] =50 nM. [DNA] =0.5 nM. From 3 replicate experiments, uncleaved and cleaved DNA intensities were quantified and their decay fit to single exponential profile to obtain lifetime of cleavage ( $t_{\text{cleavage}}$ ) as shown and quantified in Figure 4.15b and Figure 4.15c Sequences of DNA targets and guide-RNA used for these experiments is in Table 4.2 and Table 4.5. Number of PAM-distal ( $n_{\text{PD}}$ ) and PAM-proximal mismatches ( $n_{\text{PP}}$ ) are shown in cyan and orange respectively.

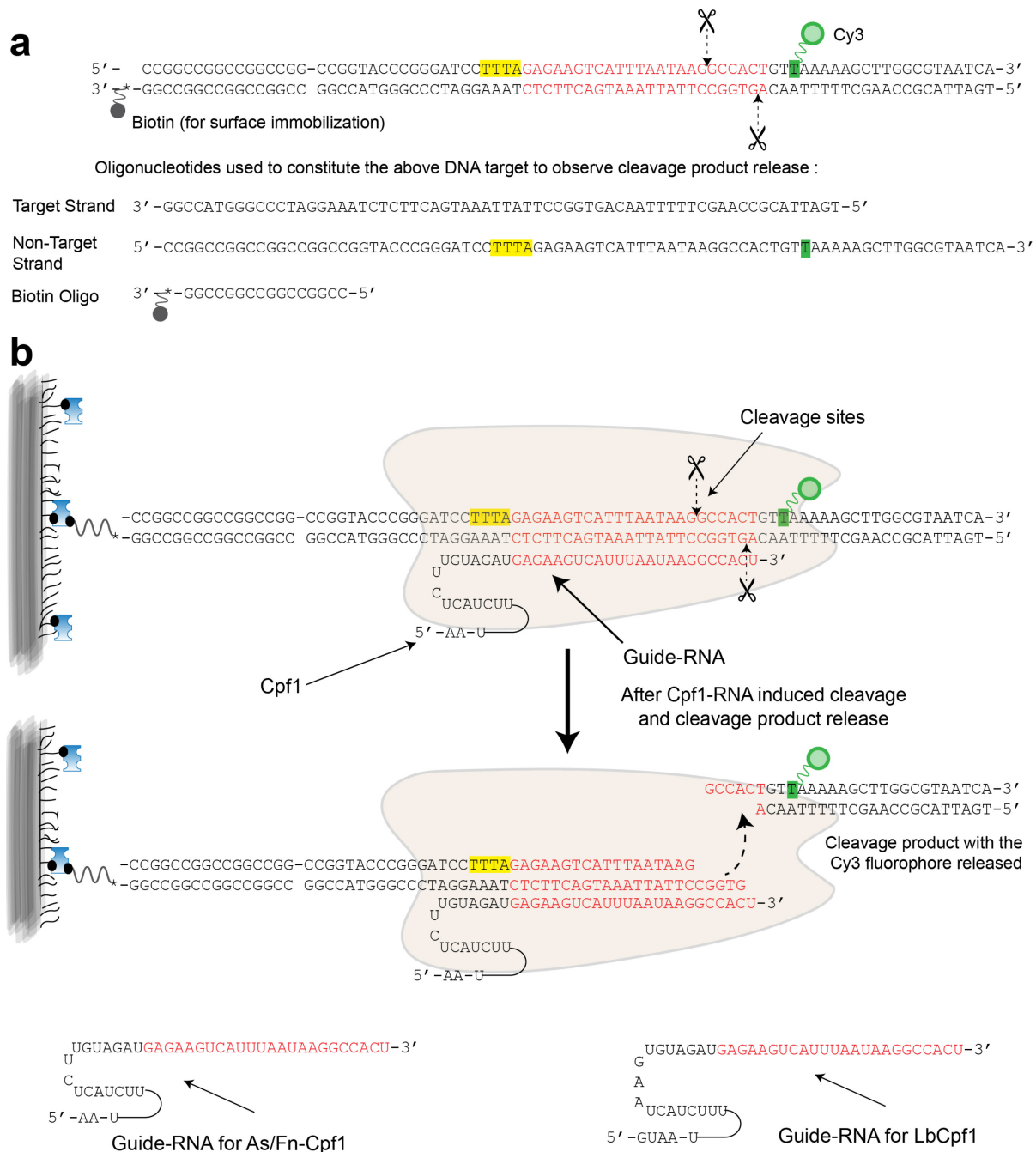


**Figure 4.17 | Catalytic activity of Cpf1 increases stability of Cpf1-RNA-DNA.**

(a) Schematic of single-molecule FRET assay to investigate DNA binding by catalytically active FnCpf1 (left) and catalytically dead FnCpf1 (right). DNA cleavage sites are indicated.



**(b)** *E* histograms of different DNA targets with 50 nM FnCpf1-RNA and *n* minutes post removal of free FnCpf1-RNA from the imaging chamber. **(c)** *E* histograms of different DNA targets with 50 nM dFnCpf1-RNA and *n* minutes (as indicated) post removal of free FnCpf1-RNA from the imaging chamber. Number of PAM-distal ( $n_{PD}$ ) in DNA targets are shown in cyan. **(d)** Internal DNA unwinding and rewinding dynamics in Cpf1-RNA-DNA likely causes eventual disassociation of Cpf1-RNA from DNA. Cleavage stalls the unwinding/rewinding dynamics thus preventing Cpf1-RNA dissociation leading to a stable Cpf1-RNA-DNA.



**Figure 4.18 | Design of DNA targets and guide-RNA for single-molecule cleavage product release assay to study Cpf1-RNA induced cleavage product release of DNA targets.**

(a) Description of single-stranded DNA oligonucleotides with appropriate modifications for constitution of a fully duplexed DNA target for use in the single-molecule cleavage product release assay.

Oligonucleotides referred to as Biotin oligo provide anchor for surface immobilization of the fully

duplexed DNA target. These oligonucleotides were same for all the DNA targets. The other two strands being, target strand that hybridizes with the guide-RNA (of Cpf1-RNA) and non-target strand, complementary to the target strand. The non-target strand was labeled with Cy3 at the indicated position. Base sequence of the target and non-target strands were changed to create DNA targets with mismatches against the fixed guide-RNA sequence. **(b)** Illustrated schematic of a complete Cpf1-RNA-DNA complex showing the base-pairing between different components. Sequence written in red denote the cognate sequence of the DNA target and the complementary sequence in the guide-RNA. Also shown are the guide-RNA sequences used for these experiments.

#### 4.4 DISCUSSION

The two-step mechanism of sampling for PAM followed by unidirectional RNA-DNA heteroduplex extension (Figure 4.19) is shared between Cas9 and Cpf1, suggesting this to be a general target identification mechanism of these CRISPR systems. Ultra-stable binding of Cpf1 requires the same extent of sequence match (17 bp PAM-proximal matches) as target cleavage. This contrasts with Cas9, which requires only 9 bp and 16 bp PAM-proximal matches for ultra-stable binding and cleavage respectively<sup>83,96,119</sup>. Therefore, Cpf1 can be more sequence specific in experiments involving the use of catalytically dead CRISPR for imaging, tracking and transcription regulation purposes<sup>18</sup>. The binding specificity of engineered Cas9s (eCas9<sup>67</sup> & Cas9-HF1<sup>68</sup>) is still much lower than that of Cpf1<sup>119</sup>. Therefore, Cpf1 has the potential to be a better alternative to all current Cas9 variants.

Cleavage rate is reduced with increasing PAM-distal mismatches (Figure 4.15c) even when the mismatches do not affect stable binding (Figure 4.7), suggesting that shorter RNA-DNA heteroduplexes result in slower conformational changes required for cleavage activation. Previous studies on Cas9

revealed that mismatches alter the kinetics of DNA unwinding, RNA-DNA heteroduplex extension, and nuclease and proof-reading domain movements<sup>88,96,112,119</sup>.

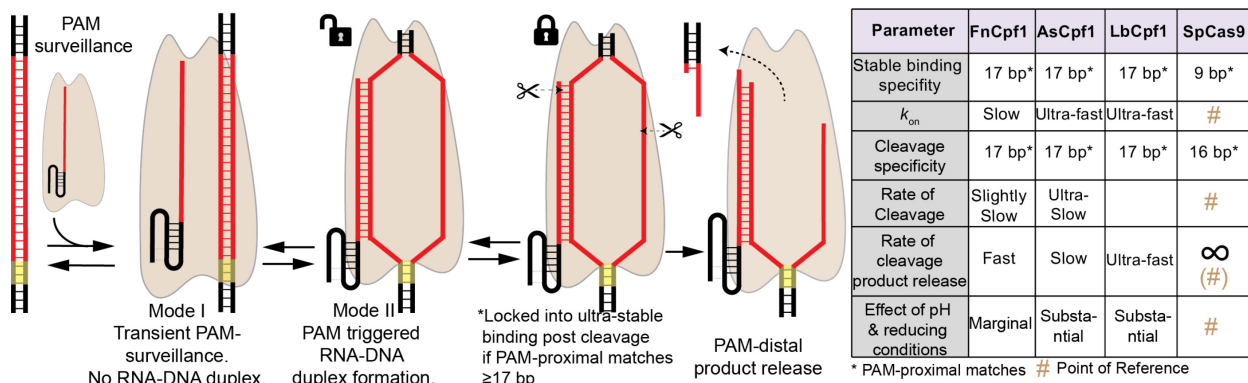
For cognate DNA target, RNA-DNA heteroduplex extension would require unwinding of the parental DNA duplex. In the crystal structure of AsCpf1-RNA-DNA complex, four PAM-distal base pairs are unwound but not involved in RNA-DNA heteroduplex<sup>114</sup>, hinting that DNA unwinding does not necessarily cause a concomitant annealing with the RNA. We performed cleavage experiments using DNA with PAM-distal mismatched region pre-unwound in order to test the relative importance of parental DNA duplex unwinding and annealing with RNA in cleavage activation. Cpf1 needed much fewer PAM-proximal matches to cleave if the mismatched region is pre-unwound (Figure 4.20) indicating indeed DNA unwinding is likely more important than RNA-DNA heteroduplex in activating cleavage. Accordingly, ssDNA can also be cleaved by Cpf1 (Figure 4.20). Therefore, the role of RNA may primarily be in keeping the DNA unwound through annealing with the target strand.

CRISPR enzymes bend DNA to cause a local kink near the PAM, which acts as a seed for unwinding and heteroduplex extension<sup>92,93,114</sup>. Perturbation of DNA bending by introducing a nick near the PAM slowed down cleavage, underscoring the importance of DNA bending for Cpf1 induced cleavage (Figure 4.21). Cas9 causes a larger DNA bend than Cpf1<sup>93,114</sup>, possibly contributing to its higher tolerance of PAM-proximal mismatches in binding and cleavage activity.

Shorter and simpler guide-RNA<sup>11</sup> for Cpf1 could potentially be deleterious for its engineering or extension, as is done for Cas9's guide-RNA<sup>120</sup>. For e.g., an extra 5' guanine in the guide-RNA was extremely deleterious for LbCpf1 (Figure 4.22). This feature could affect applications where guide-RNAs are transcribed using U6/T7 RNA polymerase systems that require first nucleotide in transcribed RNA to be the guanine<sup>121,122</sup>.

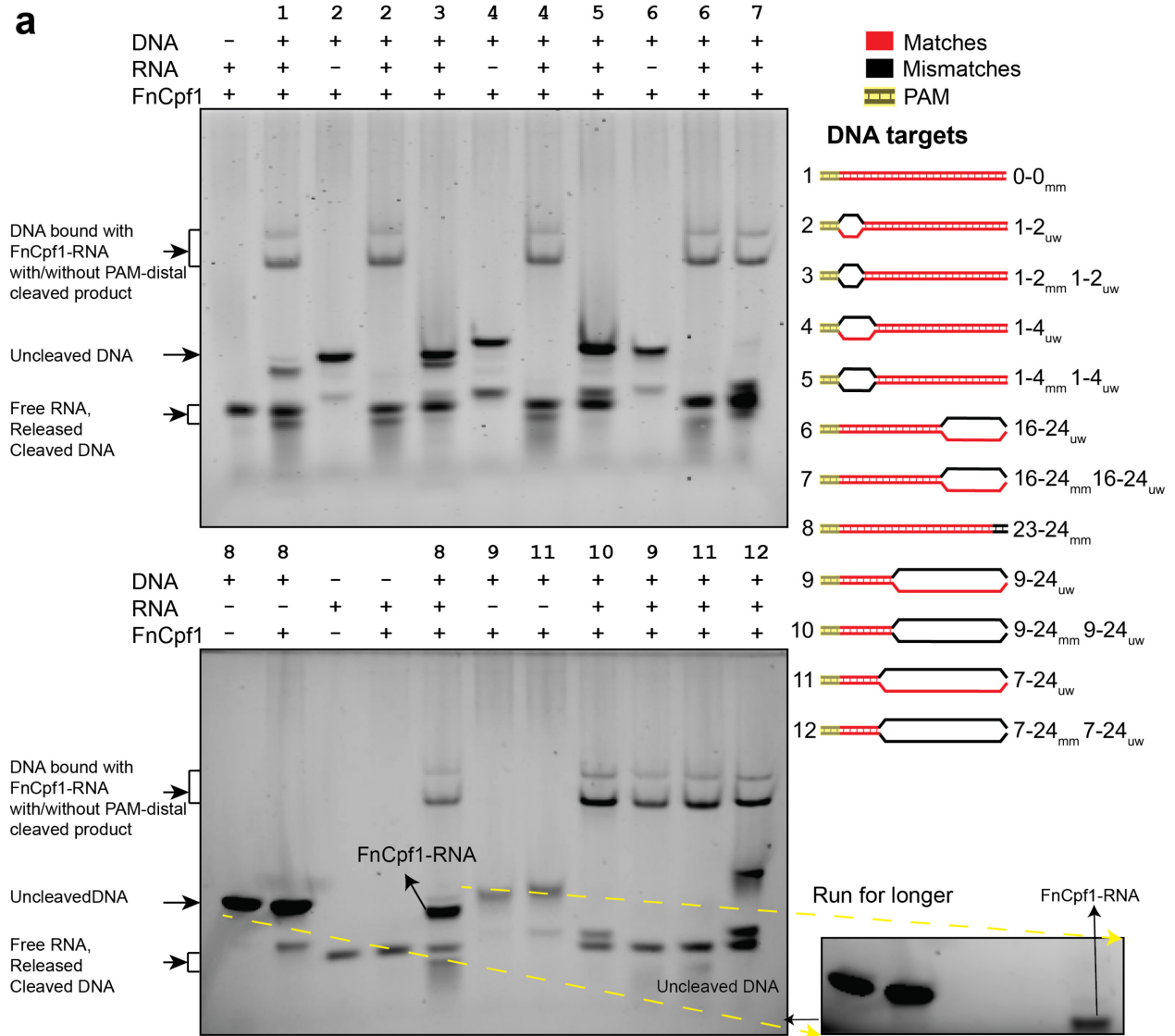
Cas9 has provided a highly efficient and versatile platform for DNA targeting, but the efficiency of gene knock-in is low<sup>123</sup>. Amongst the possible reasons is the inability of Cas9 to release and expose cleaved DNA ends. In contrast, the ability of Cpf1 to release a cleavage product readily, combined with staggered cuts it generates, could in principle increase the knock-in efficiency. Although it remains to be seen how this property affects the downstream processing *in vivo*, we can also envision a scenario where product release by Cpf1 can be detrimental to genome engineering applications. Applying positive twist to the DNA in a Cas9-RNA-DNA complex can release Cas9-RNA from DNA by promoting rewinding of parental DNA duplex<sup>86</sup>. Positive supercoiling is generated ahead of a transcribing RNA polymerase<sup>124</sup> and Cas9 holding onto the double strand break product may help build the torsional strain required to eject Cas9-RNA. If the PAM-distal cleavage product is released prematurely as in the case of Cpf1, transcription-induced positive supercoiling cannot build up and the Cpf1-RNA would remain bound stably to the PAM-proximal cleavage product, hiding the cleaved end and preventing efficient knock-in.

High specificity of adaptive immunity by Cpf1 against hypervariable genetic invaders is a little paradoxical. But Cpf1 and Cas9 systems co-exist in many species and thus they likely provide immunity suited to their features, effectively broadening the scope of immunity. Overall, our results establish major different and common features between Cpf1 and Cas9 which can be useful for the broadening of genome engineering applications as well.

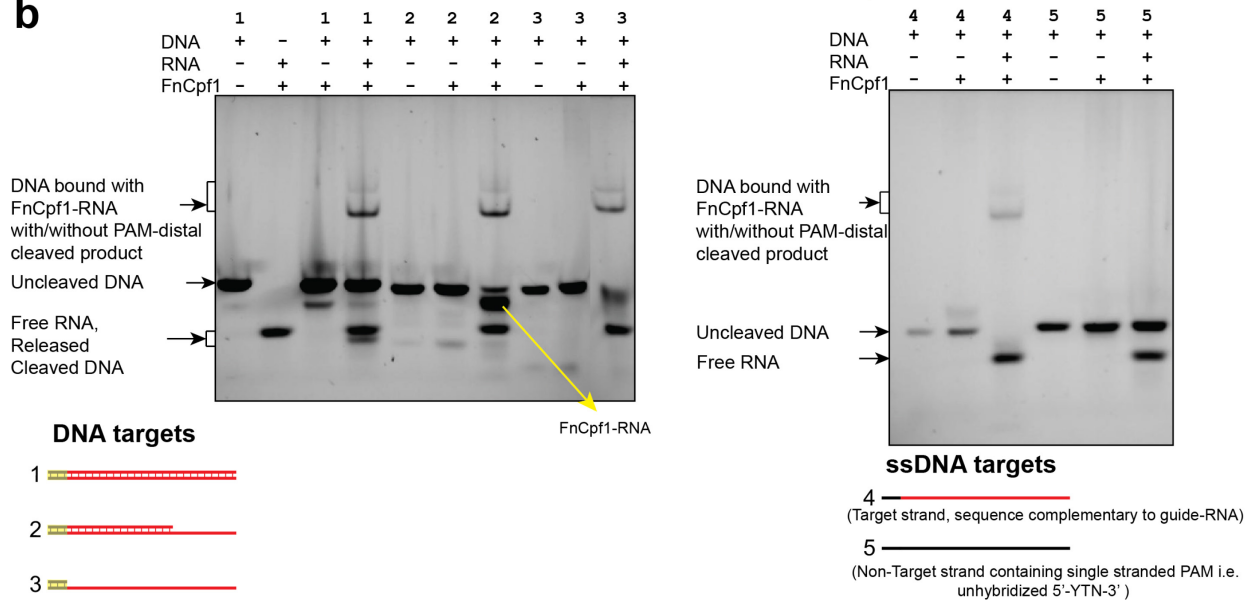


**Figure 4.19 | Model of Cpf1-RNA DNA targeting, cleavage and product release.**

**a**



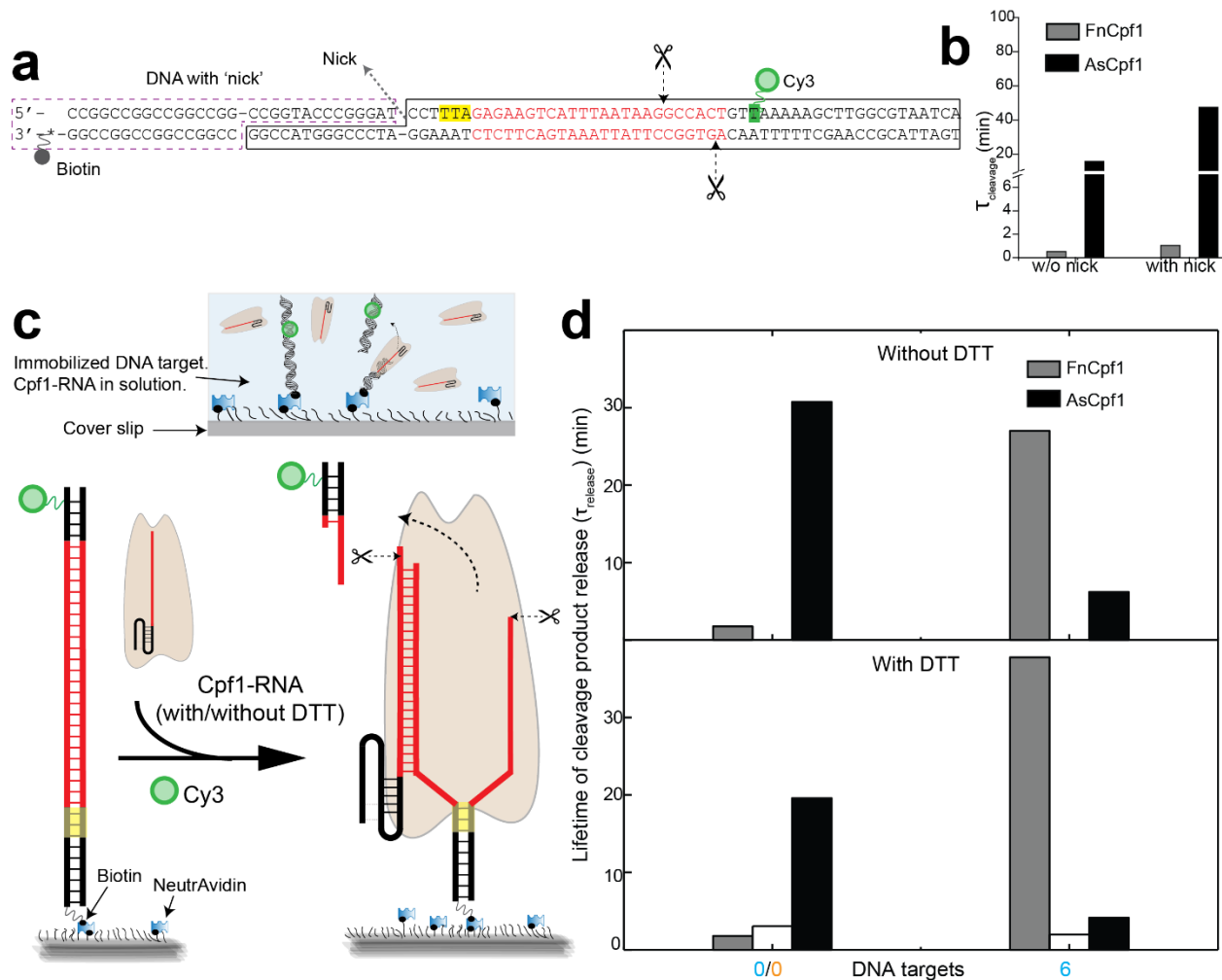
**b**



**Figure 4.20 | FnCpf1 activity for DNA targets in pre-unwound configuration and DNA targets with partially or completely single stranded target strand analyzed by 4% native gel electrophoresis and SYBR Gold II staining of nucleic acids.**

**(a)** Pre-unwound DNA targets with mismatches i.e. DNA targets with no base-pairing between target and non-target strands at the indicated bulged portion along with mismatches in the target strand. These were bound and cleaved by FnCpf1-RNA even with high extent of PAM-distal mismatches. This indicates that if DNA unwinding was facilitated by using pre-unwound DNA configuration, then the FnCpf1 was more likely to cause DNA cleavage. But even for pre-unwound DNA targets, PAM-proximal complementarity between guide-RNA and the target strand of the DNA targets was still a necessary condition for the FnCpf1 activity.  $x-y_{mm}$  refers to a contiguous mismatch (between target strand and the guide-RNA) running from position  $x$  to  $y$  relative to PAM, whereas  $x-y_{uw}$  refers to a contiguous stretch (relative to PAM) in the DNA target where the bases between the target and the non-target strands are not base-paired. FnCpf1-RNA band was sometimes observed as a faint band right below the uncleaved DNA target band. Possibly, complexation of guide-RNA with FnCpf1 prevents efficient staining of RNA in the FnCpf1-RNA. As shown, if the gel-electrophoresis was run longer, then a clearer resolution between uncleaved DNA target and FnCpf1-RNA band could be observed. **(b)** Partially duplex DNA targets, i.e., DNA targets with a portion of target strand in single-stranded confirmation were also bound and cleaved by FnCpf1 even if the PAM was single-stranded. Sequences of DNA targets and guide-RNA used for all these experiments is in Table 4.3, Table 4.4 and Table 4.5.

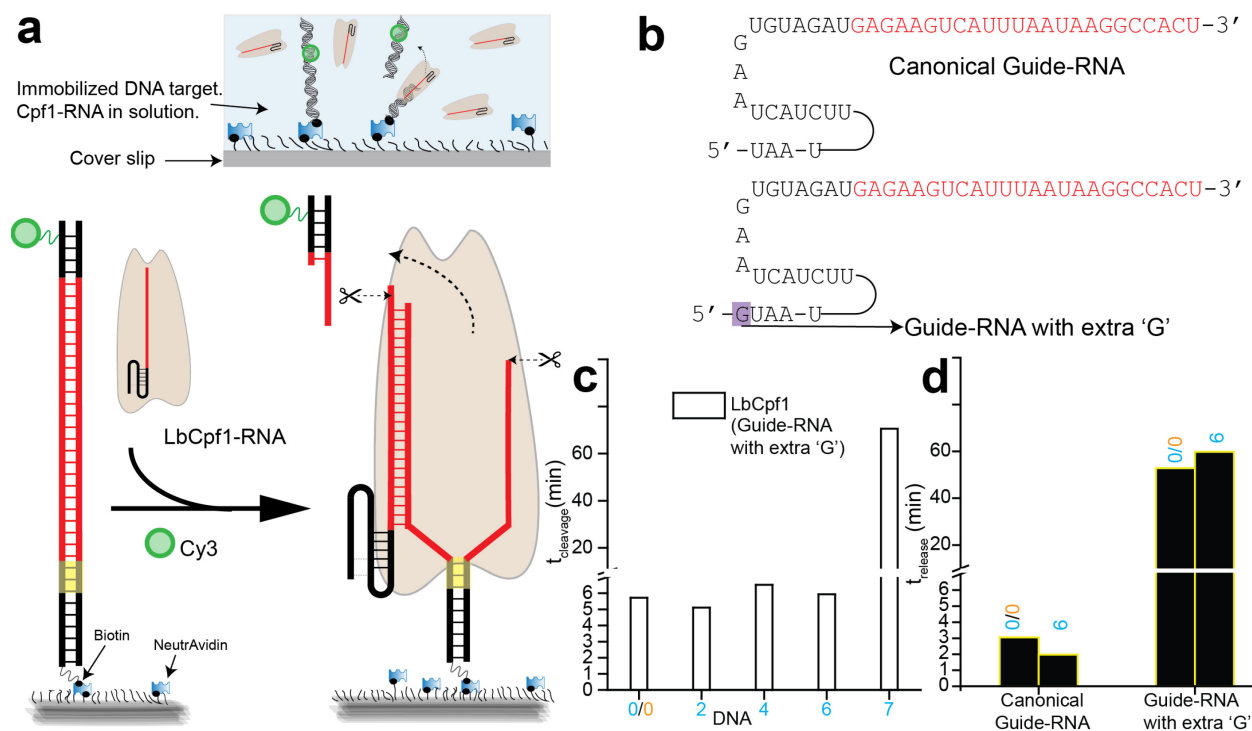




**Figure 4.21 | Effect of 'nick' and reducing conditions on the Cpf1 induced DNA cleavage and release of cleavage products.**

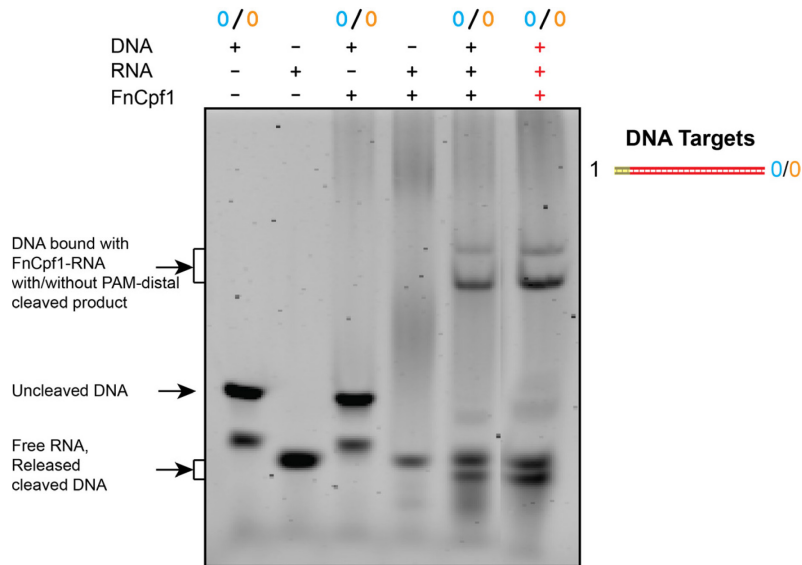
(a) Presence of nick close to PAM can perturb DNA bending, which has been structurally characterized<sup>93</sup> to be important for Cas9 induced DNA unwinding or R-Loop formation, which in turn guides cleavage action as per our unpublished study. To probe its effect in cleavage action of Cpf1, the radio-labeled denaturing gel-electrophoresis experiments (Figure 4.15b) were performed to measure lifetime of cleavage with and without the nick. Values of which are shown in (b). Cleavage is 2-3-fold slower in presence of nick. Here, the values without nick are same as the ones shown in Figure 4.15c. (c) Schematic of single-molecule cleavage product release assay as described in Figure 4.15d. These experiments were done in reducing and non-reducing conditions i.e. with and without DTT. Lifetime of cleavage product

release with and without DTT is summarized in (d). Here, the values without nick are the same as the ones shown in Figure 4.15f. FnCpf1 and AsCpf1 activity were similar, as we have shown previously that their activity is not much affected by reducing conditions. But LbCpf1 which we have shown to be strongly dependent on reducing conditions had the most effect in absence of reducing conditions as shown.



**Figure 4.22 | Effect of an extra guanine (G) base in guide-RNA for LbCpf1-RNA activity.**

(a) Schematic of the single-molecule cleavage product release assay as described in Figure 4.15d and Figure 4.18. (b) Two different guide-RNA sequences tested for LbCpf1-RNA. Former is the canonical guide-RNA for LbCpf1 as has been described<sup>11</sup>. The latter is the guide-RNA we *in-vitro* transcribed for testing with LbCpf1, which has an extra guanine (G) at the 5' end. (c)  $t_{\text{cleavage}}$  by LbCpf1 with guide-RNA with the extra G at the 5' end. (d) Extra G is quite deleterious for LbCpf1-RNA activity as shown by many-fold increase in the lifetime of cleavage-product release.



**Figure 4.23 | FnCpf1 activity is not affected by the imaging buffer components.**

DNA cleavage & binding by FnCpf1 analyzed by 4% native gel electrophoresis and SYBR Gold II staining of nucleic acids. Lanes colored with black legends used the reaction buffer (50 mM Tris-HCl (pH 8.0), 100 mM NaCl, 10 mM MgCl<sub>2</sub>, 5 mM Dithiothreitol (DTT)) while the lane colored with red legend used reaction buffer containing Trolox and BSA i.e. imaging buffer components (50 mM Tris-HCl (pH 8.0), 100 mM NaCl, 10 mM MgCl<sub>2</sub>, 5 mM Dithiothreitol (DTT), 0.2 mg/ml BSA and saturated Trolox (>5 mM)). Sequences of DNA targets and guide-RNA used for these experiments is in Table 4.3 and Table 4.5.

## 4.5 MATERIALS AND METHODS

### 4.5.1 Preparation of DNA target and guide-RNA

Single-stranded DNA (ssDNA) oligonucleotides were purchased from Integrated DNA Technologies. ssDNA target and non-target (labeled with Cy3) strands and a biotinylated adaptor strand were mixed. Excess target strand was used to ensure near complete hybridization of non-target strand with the target strand. Upon surface immobilization of the assembled DNA target, any free target strand can be washed away because it does not contain biotin. The non-target strand was created by ligating two component

strands, one with Cy3 and the other containing the protospacer region to avoid having to synthesize modified oligos for each mismatch construct. For schematics, see Figure 4.1a.

Fully duplexed DNA targets but with a nick were also used. The oligonucleotide containing Cy3 is referred to as “Cy3 oligo” and is in part, complementary to a “biotin oligo”. Hybridization of the two oligos results in a biotin-Cy3 adaptor, which has a 14 nt overhang complementary to the “target oligo” that contains the protospacer region. Finally, the non-target “oligo” complementary to the target oligo was used to complete the duplexed DNA target (Figure 4.1a). DNA targets were prepared by mixing all of the four component oligos in the buffer containing 50 mM NaCl, 20 mM Tris-HCl (pH 8.0), which was then heated to 90° C followed by slow-cooling to room temperature over 3 hr. The mixing ratio of component oligos was 1:1:2:3 for Cy3 oligo: biotin oligo: target oligo: non-target oligo. An excess of target and non-target oligo was used to ensure that any Cy3 oligo detected on the surface is in complex with three other oligos. The Cy3 fluorophore is located 4 bp upstream of the protospacer adjacent motif (PAM: 5'-YTTN-3') and was conjugated via Cy3 N-hydroxysuccinimido (Cy3-NHS; GE Healthcare) to the Cy3 oligo at amino-group attached to a modified thymine through a C6 linker (amino-dT) using NHS ester linkage. smFRET experiments were done with both sets of DNA targets (with or without a nick) and no significant differences were found between them. Table 4.1 shows all DNA targets used. For single-molecule cleavage product release experiments, a non-target strand with the Cy3 relocated in a different position was used. Cy3 label was conjugated onto the amine modification (amino-dT) using Cy3-NHS, as described above. Schematic of these DNA targets is in the Figure 4.18 and their sequences in Table 4.5.

DNA targets for gel electrophoresis experiments were prepared and hybridized as described above. For radio-labeled gel electrophoresis experiments, the target strand was 5' radiolabeled with T4 polynucleotide kinase (New England BioLabs) and  $\gamma$ -<sup>32</sup>P ATP (Perkin Elmer). The target and non-target strands were annealed with the non-target strands in excess. For single molecule experiments, guide-RNA was purchased from IDT with modifications for Cy5 labeling as described in Table 4.5. Cy5 was conjugated via Cy3 N-hydroxysuccinimido (Cy5-NHS; GE Healthcare) to the RNA as described

previously<sup>83,125</sup>. For all other experiments, unmodified guide-RNA was used and they were either *in vitro* transcribed or purchased from IDT. Guide-RNA sequences used in this study is available in Table 4.5

#### **4.5.2 Preparation of Cpf1-RNA.**

The Cpf1-RNA was freshly prepared prior to each experiment by mixing the guide-RNA (50 nM) and Cpf1 in 1:3.5 ratio in the following reaction buffers and incubated for at least 10 min at room temperature. 50 mM Tris-HCl (pH 8.0) 100 mM NaCl, 10 mM MgCl<sub>2</sub>, (FnCpf1 and LbCpf1) and 50 mM HEPES (pH 7.0) 100 mM NaCl, 10 mM MgCl<sub>2</sub>, (AsCpf1). 5mM DTT was only used in the buffer when specified. 0.2 mg/ml Bovine serum albumin (BSA), 1 mg/ml glucose oxidase, 0.04 mg/ml catalase, 0.8% dextrose and saturated Trolox (>5 mM)) were additional contents of the reaction buffers for single-molecule imaging experiments. Excess Cpf1 was used to achieve highest extent of complexation of all the available guide-RNA and the concentration of guide-RNA was used as the concentration of Cpf1-RNA. Cpf1 activity using the similar guide-RNA and on DNA targets with same protospacer and PAM have been characterized previously<sup>11</sup>. Fluorophore labeling of either DNA targets or guide-RNA did not impair Cpf1 activity (Figure 4.3).

#### **4.5.3 Single-molecule fluorescence imaging and data analysis.**

DNA targets were immobilized on the polyethylene glycol-passivated flow chamber surface (purchased from Johns Hopkins University Microscope Supplies Core or prepared following protocols reported previously<sup>89</sup> using neutrAvidin-biotin interaction and imaged in the presence of Cas9-RNA at the stated concentration using two-color total internal reflection fluorescence microscopy. For DNA unwinding by surface-tethered Cas9-RNA smFRET assay, 20 nM of biotin-labeled Cas9-RNA was immobilized on surface before adding FRET pair labeled DNA targets. All the imaging experiments were done at room temperature in a Cas9-RNA activity buffer with oxygen scavenging for extending photostability (20 mM Tris-HCl, 100 mM KCl, 5 mM MgCl<sub>2</sub>, 5% (v/v) glycerol, 0.2 mg ml<sup>-1</sup> BSA, 1 mg ml<sup>-1</sup> glucose oxidase, 0.04 mg ml<sup>-1</sup> catalase, 0.8% dextrose and saturated Trolox (>5 mM)<sup>101</sup>. Time resolution was 100 ms

unless stated otherwise. Detailed methods of single-molecule FRET (smFRET) data acquisition and analysis have been described previously<sup>89</sup>. Video recordings obtained using EMCCD camera (Andor) were processed to extract single molecule fluorescence intensities at each frame and custom written scripts were used to calculate FRET efficiencies. Data acquisition and analysis software can be downloaded from <https://cplc.illinois.edu/software/>. FRET efficiency ( $E$ ) of the detected spot was approximated as  $E = I_A/(I_D+I_A)$ , where  $I_D$  and  $I_A$  are background and leakage corrected emission intensities of the donor and acceptor, respectively.

#### **4.5.4 Expression and purification of Cpf1.**

The methods of Cpf1 protein expression and purification were adapted from a protocol described previously<sup>9</sup>. Codon optimized Cpf1 gene sequence cloned into a bacterial expression vector (6-His-MBP-TEV-FnCpf1, a pET based vector) was cloned in house or purchased from GenScript. The vector was transformed into Rosetta (DE3) pLyseS (EMD Millipore) cells and cells were plated onto LB-Kanamycin agar plates and grown at 37 °C overnight. Single colony from the agar plates was then cultured overnight in 10 ml of SOC medium (Thermo Fisher Scientific). The overnight miniculture of Rosetta (DE3) pLyseS cells containing the Cpf1 expression construct were inoculated (1:500 dilution) into 4 liters of Terrific Broth (Sigma Aldrich) growth media containing 50 µg/ml Kanamycin. Growth media with the inoculant was grown at 37 °C in a shaker at 100 revolution per minute (rpm) until the cell density reached 0.2 OD600, at which point the temperature was lowered to 21 °C. Growth was continued and 6-His-MBP-TEV-Cpf1 protein expression was induced when cells reached 0.6 OD600 by addition of IPTG (Sigma) to 0.5 mM final concentration in the growth media. The induced culture was kept for 14–18 hr at 21 °C after which the cells were harvested by centrifugation at 5000 rpm for 30 min at 4 °C. The harvested cells were quickly stored at –80 °C until further purification. The harvested cells were then suspended in 200 ml of lysis buffer (50 mM HEPES [pH 7], 2M NaCl, 5 mM MgCl<sub>2</sub>, 20 mM imidazole) supplemented with protease inhibitors (Roche complete, EDTA-free) from Roche and lysozyme (Sigma Aldrich) and incubated at 4 °C for 30–45 minutes. After homogenization, cells were further lysed by sonication (Fisher

Model 500 Sonic Dismembrator; Thermo Fisher Scientific) at 30% amplitude in 3 cycles of 2 s sonicate-2 s relax mode, each cycle lasting 1 min. Following lysis, cell solution was centrifuged at 15,000 ×g for 30-45 minutes, the cellular debris was discarded and the supernatant of lysate was collected. The clear lysate was then incubated at 4 °C with Ni-NTA slurry (Qiagen) for 45 min in a shaker at 30 rpm. The lysate with the Ni-NTA slurry was then applied to a column and multiple cycles of lysis buffer were used to wash the Ni-NTA slurry through the column. The 6-His-MBP-TEV-Cpf1 was eluted in a single step with 300 mM imidazole buffer (50 mM HEPES [pH 7], 2 M NaCl, 5 mM MgCl<sub>2</sub>, 300 mM imidazole). TEV protease (Sigma Aldrich) was then added, and the sample was dialyzed using Slide-A-Lyzer™ dialysis cassettes (Thermo Fisher Scientific) overnight into the buffer suitable for TEV protease activity (500 mM NaCl, 50 mM HEPES [pH 7], 5 mM MgCl<sub>2</sub>, 2 mM DTT). TEV protease activity resulted in the deconstitution of 6-His-MBP-TEV-Cpf1 into 6-His-MBP and Cpf1, which was confirmed by SDS-PAGE. The free 6-His-MBP was removed by another round of Ni-NTA chromatography resulting in the solution containing only Cpf1. Sample was then injected on to a HiLoad 26/600 S200 size exclusion column equilibrated with gel filtration buffer (50 mM Tris-HCl pH 8.0, 100 mM NaCl, 10 mM MgCl<sub>2</sub>, 5mM DTT). Fractions containing Cpf1 were pooled, concentrated, and then flash frozen in liquid nitrogen. Final sample was stored at -80 °C until used in experiments.

#### **4.5.5 FRET histograms, Cas9-RNA bound DNA fraction and unwound fraction**

Neutravidin-biotin interaction was used to immobilize the biotinylated Cy3-labeled DNA targets on the polyethylene glycol (PEG) passivated flow chamber surface prepared following protocols reported previously<sup>89</sup> or purchased from Johns Hopkins Microscope Supplies Core. Cy5-labeled Cpf1-RNA or unlabeled Cpf1-RNA (both referred to as Cpf1-RNA for brevity) was added to the flow chamber. The flow chamber was then illuminated with green laser and imaged with two color total internal reflection fluorescence microscopy. A buffer suitable for single-molecule imaging and Cpf1 activity was used and is referred to as the imaging-reaction buffer (50 mM Tris-HCl (pH 8.0) 100 mM NaCl, 10 mM MgCl<sub>2</sub>, 0.2 mg/ml Bovine serum albumin (BSA), 1 mg/ml glucose oxidase, 0.04 mg/ml catalase, 0.8% dextrose and

saturated Trolox (>5 mM)) (Figure 4.23). 50 mM HEPES (pH 7.0) was used in place of 50 mM Tris-HCl (pH 8.0) for AsCpf1 experiments only, unless stated otherwise. 5 mM DTT was only added to these buffers for experiments done and stated to be in reducing conditions (chiefly LbCpf1 only). Technical details of single-molecule imaging, data acquisition and analysis have been described previously<sup>89</sup>. Video recordings obtained using EMCCD camera (Andor) were processed to extract single molecule fluorescence intensities at each frame and custom written scripts were used to calculate FRET efficiencies. Data acquisition and analysis software can be downloaded from <https://cplc.illinois.edu/software/>. FRET efficiency of the detected spot was approximated as  $FRET = I_A/(I_D+I_A)$ , where  $I_D$  and  $I_A$  are background and leakage corrected emission intensities of the donor and acceptor respectively. For single-molecule cleavage experiments, series of snapshots of different imaging areas were taken at different time points, under same green laser illumination via total internal reflection. The snapshots were then analyzed to estimate the changing number of Cy3-labeled DNA targets on the surface. Time resolution for all the experiments was 100 ms unless stated otherwise.

#### **4.5.6 FRET efficiency histograms and Cpf1-RNA bound DNA fraction**

A smFRET time-trajectory is a series of  $E$  values every 100 ms. First five  $E$  values of each single-molecule trace were pooled together to build single molecule  $E$  histograms. Cpf1-RNA bound DNA fraction ( $f_{\text{bound}}$ ) was calculated as a ratio between the number of molecules with  $E > 0.2$  and the total number of molecules in the  $E$  histograms.  $E$  histograms shown in Figure 4.6 were constructed by combining data from two independent experiments (except for AsCpf1; PAM-less DNA).

#### **4.5.7 Determination of binding kinetics**

For DNA targets that showed real-time reversible binding/dissociation of Cpf1-RNA, idealization of smFRET traces via Hidden Markov Model<sup>90</sup> analysis yielded two pre-dominant FRET states, of zero ( $E < 0.2$ ) and bound state ( $E > 0.2$ ). Lifetime of the unbound state,  $\tau_{\text{unbound}}$ , was calculated by fitting survival probability of dwell times of unwound state ( $E < 0.2$ ) vs time to a single-exponential decay ( $\exp[-$



$t/\tau_{\text{unbound}}]$ ). The survival probability of the bound state required a double-exponential decay for adequate fitting ( $A \cdot \exp[-t/\tau_1] + [1-A] \cdot \exp[-t/\tau_2]$ ), and the average lifetime was calculated as  $\tau_{\text{avg}} = A \tau_1 + (1-A) \tau_2$ .

The bimolecular association rate constant  $k_{\text{on}}$ , binding rate  $k_{\text{binding}}$  and dissociation rate  $k_{\text{off}}$  were calculated as follows.

$$k_{\text{binding}} = \tau_{\text{unbound}}^{-1}$$

$$k_{\text{off}} = \tau_{\text{bound}}^{-1}$$

$$k_{\text{on}} = k_{\text{binding}} / [\text{Cpf1-RNA}]$$

Due to under-sampled binding events,  $\tau_{\text{avg}}$  of FnCpf1 for PAM-less DNA and DNA with 2  $n_{\text{pp}}$  were calculated as the algebraic average of  $E > 0.2$  dwell-times.

Cy5 labeling efficiency of guide-RNA was  $\sim 90\%$  and thus  $f_{\text{bound}}$  and  $\tau_{\text{unbound}}$  were appropriately corrected.

Due to high noise, the smFRET traces from experiments involving AsCpf1 could not be idealized with high accuracy thus preventing their  $k_{\text{off}}$  and  $k_{\text{on}}$  analysis.

#### 4.5.8 Estimation of dissociation constant ( $K_d$ )

To estimate  $K_d$ , Cpf1-RNA bound DNA fraction ( $f_{\text{bound}}$ ) vs Cpf1-RNA concentration ( $c$ ) was fit using  $f_{\text{bound}} = M \times c / (K_d + c)$  where  $M$  is the maximum observable  $f_{\text{bound}}$ .  $M$  is typically less than 1 because inactive or missing acceptors or because not all of the DNA on the surface are capable of binding Cpf1-RNA.

#### 4.5.9 Overall lifetime of release of cleavage products

Single-molecule experiments were used to estimate the lifetime of the release of cleavage products by fitting the decreasing number of Cy3 spots (loss of spots due to Cpf1-RNA induced cleavage and release) to a single-exponential decay. The time of binding ( $k_{\text{on}} \times 50 \text{ nM}$ ) and time of cleavage ( $\tau_{\text{cleavage}}$ ) were subtracted from the obtained lifetime to get the true lifetime of the release ( $\tau_{\text{release}}$ ) of cleavage products.

But since  $\tau_{\text{cleavage}}$  was not measured for LbCpf1, its reported  $\tau_{\text{release}}$  is without the  $\tau_{\text{cleavage}}$  and time of binding subtraction.

#### **4.5.10 Gel electrophoresis experiments**

All the biochemical experiments were done in the following reaction buffers: 50 mM Tris-HCl (pH 8.0) 100 mM NaCl, 10 mM MgCl<sub>2</sub>, 5 mM DTT (FnCpf1 and LbCpf1) and 50 mM HEPES (pH 7.0) 100 mM NaCl, 10 mM MgCl<sub>2</sub>, 5 mM DTT (AsCpf1).

#### Gel electrophoresis experiments involving visualization of nucleic acid bands via SYBR Gold II staining.

All experiments were conducted by mixing DNA targets and Cpf1-RNA in 1:5 ratio in the reaction buffer. The reaction was incubated for 4.5-5 hr (unless stated otherwise) before being resolved by 4% native/denaturing agarose gel electrophoresis and SYBR Gold II staining of nucleic acids using the precast gels containing SYBR Gold II, purchased from Thermo Fisher Scientific. For native gel electrophoresis, the reaction aliquots were directly loaded onto the gels. All the reactions were incubated at the room temperature, 37 °C or 4 °C and indicated in the presentation of their results. The gel electrophoresis was run at room temperature for experiments incubated at room temperature/37 °C and at 4 °C for experiments incubated at 4 °C. The cleaved-uncleaved DNA target with/without the bound Cpf1-RNA along with other nucleic acids were stained by SYBR Gold II and imaged by blue laser illumination (480 nm; GE Amersham Molecular Dynamics Typhoon 9410 Molecular Imager and 488 nm; Amersham Imager 600). For all of these experiments, the concentration of the DNA targets ranged from 20 nM to 60 nM and consequently the effective concentration of Cpf1-RNA ranged from 100 nM to 300 nM respectively. Volume of aliquots used for gel loading ranged from 10 to 20  $\mu\text{L}$  per lane. For the time-lapse denaturing gel electrophoresis experiments, the acquired gel-images were quantified using ImageJ<sup>126</sup>. Entire panel of DNA targets used in these gel-electrophoresis experiments is available in

Table 4.3 and Table 4.4. Tris-HCl at pH 8.0 was used in the reaction buffers for all the experiments except for the ones reported in Figure 4.3 and Figure 4.12 where Tris-HCl at pH 8.5 was used.

Gel electrophoresis experiments and autoradiography. Experiments containing radiolabeled DNA substrates were performed as above. However, samples were quenched, in buffer containing 95% formamide, 0.01% SDS 0.01% bromophenol blue, 0.01% xylene cyanol, and 1 mM EDTA and incubated at 95 °C for 5min then on ice for 2min. Volume ratio of quenching buffer to reaction was 5:1. Samples were loaded on to denaturing polyacrylamide gels (10% acrylamide, 50%(w/v) urea) and allowed to separate. Amount of sample loaded on to gel was normalized to 10,000 counts per sample. Gels were imaged via phosphor screens. Entire panel of DNA targets used in these gel-electrophoresis experiments is available in Table 4.1.

**Table 4.1 | List of DNA targets used in smFRET experiments to study DNA interrogation by Cpf1-RNA.**

Description	DNA Sequences
0/0	5' - CAGTCCTGCTGGTCGT-TCGGTACCCGGGA <span style="background-color: yellow;">CC</span> <span style="background-color: yellow;">TTT</span> <span style="background-color: yellow;">AGAGAAGTCATTTAATAAGGCCACT</span> GTTAAAAAGCTTGGCGTAATCA- 3' 3' - <span style="color: red;">-GTCAGGACGACCAGCA</span> AGCCATGGGCCCTAGGAAAT <span style="color: red;">CTCTTCAGTAAATTAATTCGGTGACA</span> CAATTTTTTCGAACCGCATTAGT- 5' ● 16 nucleotide biotinylated adaptor for surface immobilization
2	5' - CAGTCCTGCTGGTCGT-TCGGTACCCGGGA <span style="background-color: yellow;">CC</span> <span style="background-color: yellow;">TTT</span> <span style="background-color: yellow;">AGAGAAGTCATTTAATAAGGCCAGAG</span> TAAAAAGCTTGGCGTAATCA- 3' 3' - <span style="color: red;">-GTCAGGACGACCAGCA</span> AGCCATGGGCCCTAGGAAAT <span style="color: red;">CTCTTCAGTAAATTAATTCGGTCTCA</span> ATTTTTTCGAACCGCATTAGT- 5'
4	5' - CAGTCCTGCTGGTCGT-TCGGTACCCGGGA <span style="background-color: yellow;">CC</span> <span style="background-color: yellow;">TTT</span> <span style="background-color: yellow;">AGAGAAGTCATTTAATAAGGCCG</span> TAGTTAAAAAGCTTGGCGTAATCA- 3' 3' - <span style="color: red;">-GTCAGGACGACCAGCA</span> AGCCATGGGCCCTAGGAAAT <span style="color: red;">CTCTTCAGTAAATTAATTCGGCACTCA</span> ATTTTTTCGAACCGCATTAGT- 5'
6	5' - CAGTCCTGCTGGTCGT-TCGGTACCCGGGA <span style="background-color: yellow;">CC</span> <span style="background-color: yellow;">TTT</span> <span style="background-color: yellow;">AGAGAAGTCATTTAATAAGCCG</span> TAGTTAAAAAGCTTGGCGTAATCA- 3' 3' - <span style="color: red;">-GTCAGGACGACCAGCA</span> AGCCATGGGCCCTAGGAAAT <span style="color: red;">CTCTTCAGTAAATTAATTCGCCACTCA</span> ATTTTTTCGAACCGCATTAGT- 5'
7	5' - CAGTCCTGCTGGTCGT-TCGGTACCCGGGA <span style="background-color: yellow;">CC</span> <span style="background-color: yellow;">TTT</span> <span style="background-color: yellow;">AGAGAAGTCATTTAATAAGCCG</span> TAGTTAAAAAGCTTGGCGTAATCA- 3' 3' - <span style="color: red;">-GTCAGGACGACCAGCA</span> AGCCATGGGCCCTAGGAAAT <span style="color: red;">CTCTTCAGTAAATTAATTCGCCACTCA</span> ATTTTTTCGAACCGCATTAGT- 5'
8	5' - CAGTCCTGCTGGTCGT-TCGGTACCCGGGA <span style="background-color: yellow;">CC</span> <span style="background-color: yellow;">TTT</span> <span style="background-color: yellow;">AGAGAAGTCATTTAATAAGCCG</span> TAGTTAAAAAGCTTGGCGTAATCA- 3' 3' - <span style="color: red;">-GTCAGGACGACCAGCA</span> AGCCATGGGCCCTAGGAAAT <span style="color: red;">CTCTTCAGTAAATTAATTCGCCACTCA</span> ATTTTTTCGAACCGCATTAGT- 5'
9	5' - CAGTCCTGCTGGTCGT-TCGGTACCCGGGA <span style="background-color: yellow;">CC</span> <span style="background-color: yellow;">TTT</span> <span style="background-color: yellow;">AGAGAAGTCATTTAATAAGCCG</span> TAGTTAAAAAGCTTGGCGTAATCA- 3' 3' - <span style="color: red;">-GTCAGGACGACCAGCA</span> AGCCATGGGCCCTAGGAAAT <span style="color: red;">CTCTTCAGTAAATTAATTCGCCACTCA</span> ATTTTTTCGAACCGCATTAGT- 5'
16	5' - CAGTCCTGCTGGTCGT-TCGGTACCCGGGA <span style="background-color: yellow;">CC</span> <span style="background-color: yellow;">TTT</span> <span style="background-color: yellow;">AGAGAAGTCATTTAATAAGCCG</span> TAGTTAAAAAGCTTGGCGTAATCA- 3' 3' - <span style="color: red;">-GTCAGGACGACCAGCA</span> AGCCATGGGCCCTAGGAAAT <span style="color: red;">CTCTTCAGTAAATTAATAAGGCCACTCA</span> ATTTTTTCGAACCGCATTAGT- 5'
17	5' - CAGTCCTGCTGGTCGT-TCGGTACCCGGGA <span style="background-color: yellow;">CC</span> <span style="background-color: yellow;">TTT</span> <span style="background-color: yellow;">AGAGAAGTGTAAATTAATTCGGT</span> GAGTTAAAAAGCTTGGCGTAATCA- 3' 3' - <span style="color: red;">-GTCAGGACGACCAGCA</span> AGCCATGGGCCCTAGGAAAT <span style="color: red;">CTCTTCAGTAAATTAATAAGGCCACTCA</span> ATTTTTTCGAACCGCATTAGT- 5'
18	5' - CAGTCCTGCTGGTCGT-TCGGTACCCGGGA <span style="background-color: yellow;">CC</span> <span style="background-color: yellow;">TTT</span> <span style="background-color: yellow;">AGAGAAGTGTAAATTAATTCGGT</span> GAGTTAAAAAGCTTGGCGTAATCA- 3' 3' - <span style="color: red;">-GTCAGGACGACCAGCA</span> AGCCATGGGCCCTAGGAAAT <span style="color: red;">CTCTTCAGTAAATTAATAAGGCCACTCA</span> ATTTTTTCGAACCGCATTAGT- 5'
24/24	5' - CAGTCCTGCTGGTCGT-TCGGTACCCGGGA <span style="background-color: yellow;">CC</span> <span style="background-color: yellow;">TTT</span> <span style="background-color: yellow;">ACTCTTCAGTAAATTAATTCGGT</span> GAGTTAAAAAGCTTGGCGTAATCA- 3' 3' - <span style="color: red;">-GTCAGGACGACCAGCA</span> AGCCATGGGCCCTAGGAAATGAGAAGTCATTTAATAAGGCCACTCAATTTTTTCGAACCGCATTAGT- 5'
2	5' - CAGTCCTGCTGGTCGT-TCGGTACCCGGGA <span style="background-color: yellow;">CC</span> <span style="background-color: yellow;">TTT</span> <span style="background-color: yellow;">ACTGAAGTCATTTAATAAGGCCACT</span> GTTAAAAAGCTTGGCGTAATCA- 3' 3' - <span style="color: red;">-GTCAGGACGACCAGCA</span> AGCCATGGGCCCTAGGAAATGAC <span style="color: red;">TTCAGTAAATTAATTCGGTGACA</span> CAATTTTTTCGAACCGCATTAGT- 5'
4	5' - CAGTCCTGCTGGTCGT-TCGGTACCCGGGA <span style="background-color: yellow;">CC</span> <span style="background-color: yellow;">TTT</span> <span style="background-color: yellow;">ACTCTAGTCATTTAATAAGGCCACT</span> GTTAAAAAGCTTGGCGTAATCA- 3' 3' - <span style="color: red;">-GTCAGGACGACCAGCA</span> AGCCATGGGCCCTAGGAAATGAGAT <span style="color: red;">TCAGTAAATTAATTCGGTGACA</span> CAATTTTTTCGAACCGCATTAGT- 5'
0/0 NO PAM	5' - CAGTCCTGCTGGTCGT-TCGGTACCCGGGA <span style="background-color: yellow;">CC</span> <span style="background-color: yellow;">CAGGAGAGAAGTCATTTAATAAGGCCACT</span> GTTAAAAAGCTTGGCGTAATCA- 3' 3' - <span style="color: red;">-GTCAGGACGACCAGCA</span> AGCCATGGGCCCTAGGTGCC <span style="color: red;">CTCTTCAGTAAATTAATTCGGTGACA</span> CAATTTTTTCGAACCGCATTAGT- 5'
24/24 NO PAM	5' - CAGTCCTGCTGGTCGT-TCGGTACCCGGGA <span style="background-color: yellow;">CC</span> <span style="background-color: yellow;">CAGGCTCTTCAGTAAATTAATTCGGT</span> GAGTTAAAAAGCTTGGCGTAATCA- 3' 3' - <span style="color: red;">-GTCAGGACGACCAGCA</span> AGCCATGGGCCCTAGTCCGAGAGAAGTCATTTAATAAGGCCACTCAATTTTTTCGAACCGCATTAGT- 5'

● Biotin         Protospacer Adjacent Motif (PAM)         Thymine modification for Cy3. IDT code: *HAmmc67*

DNA sequences complementary to guide-RNA are shown in red (Cognate).

Originally DNA targets were created with a 'nick' at     -cca near PAM.

The nick did not affect binding and 2 replicates of smFRET experiments to study DNA interrogation by Cpf1-RNA was done without nick and one with nick respectively.

**Table 4.2 | List of DNA targets used in all radio-labeled gel electrophoresis experiments.**

Description	DNA Sequences
0/0	<sup>32</sup> P -CAGTCCTGCTGGTCGTTCCGGTACCCGGGATCC TTTA GAGAAGTCATTTAATAAGGCCACT GTTAAAAAGCTTGGCGTAATCA- 3' 3'- AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTTATCCGGTGA CAATTTTTTCGAACCGCATTAGT- 5'
2	<sup>32</sup> P -CAGTCCTGCTGGTCGTTCCGGTACCCGGGATCC TTTA GAGAAGTCATTTAATAAGGCCA GAGTAAAAAGCTTGGCGTAATCA- 3' 3'- AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTTATCCGGTCTCAATTTTTTCGAACCGCATTAGT- 5'
4	<sup>32</sup> P -CAGTCCTGCTGGTCGTTCCGGTACCCGGGATCC TTTA GAGAAGTCATTTAATAAGCGTGAGTAAAAAGCTTGGCGTAATCA- 3' 3'- AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTTATCCGCACTCAATTTTTTCGAACCGCATTAGT- 5'
6	<sup>32</sup> P -CAGTCCTGCTGGTCGTTCCGGTACCCGGGATCC TTTA GAGAAGTCATTTAATAAGCGTGAGTAAAAAGCTTGGCGTAATCA- 3' 3'- AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTTATCCGCACTCAATTTTTTCGAACCGCATTAGT- 5'
7	<sup>32</sup> P -CAGTCCTGCTGGTCGTTCCGGTACCCGGGATCC TTTA GAGAAGTCATTTAATAA CCGTGAGTAAAAAGCTTGGCGTAATCA- 3' 3'- AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTTATGGCCACTCAATTTTTTCGAACCGCATTAGT- 5'
8	<sup>32</sup> P -CAGTCCTGCTGGTCGTTCCGGTACCCGGGATCC TTTA GAGAAGTCATTTAATA TCCGTGAGTAAAAAGCTTGGCGTAATCA- 3' 3'- AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTTATAGGCCACTCAATTTTTTCGAACCGCATTAGT- 5'
9	<sup>32</sup> P -CAGTCCTGCTGGTCGTTCCGGTACCCGGGATCC TTTA GAGAAGTCATTTAAT TCCGTGAGTAAAAAGCTTGGCGTAATCA- 3' 3'- AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTTAAGGCCACTCAATTTTTTCGAACCGCATTAGT- 5'
16	<sup>32</sup> P -CAGTCCTGCTGGTCGTTCCGGTACCCGGGATCC TTTA GAGAAGTC TAAATTTCCGTGAGTAAAAAGCTTGGCGTAATCA- 3' 3'- AGCCATGGGCCCTAGGAAATCTCTTCAGATTTAATAAGGCCACTCAATTTTTTCGAACCGCATTAGT- 5'
17	<sup>32</sup> P -CAGTCCTGCTGGTCGTTCCGGTACCCGGGATCC TTTA GAGAAGTG TAAATTTCCGTGAGTAAAAAGCTTGGCGTAATCA- 3' 3'- AGCCATGGGCCCTAGGAAATCTCTTCACATTTAATAAGGCCACTCAATTTTTTCGAACCGCATTAGT- 5'
18	<sup>32</sup> P -CAGTCCTGCTGGTCGTTCCGGTACCCGGGATCC TTTA GAGAAG TAAATTTCCGTGAGTAAAAAGCTTGGCGTAATCA- 3' 3'- AGCCATGGGCCCTAGGAAATCTCTTC TCAATTTAATAAGGCCACTCAATTTTTTCGAACCGCATTAGT- 5'
24/24	<sup>32</sup> P -CAGTCCTGCTGGTCGTTCCGGTACCCGGGATCC TTTA CTCTTCAGTAAATTTATCCGTGAGTAAAAAGCTTGGCGTAATCA- 3' 3'- AGCCATGGGCCCTAGGAAATGAGAAGTCATTTAATAAGGCCACTCAATTTTTTCGAACCGCATTAGT- 5'
2	<sup>32</sup> P -CAGTCCTGCTGGTCGTTCCGGTACCCGGGATCC TTTA CTGAAGTCATTTAATAAGGCCACT GTTAAAAAGCTTGGCGTAATCA- 3' 3'- AGCCATGGGCCCTAGGAAATGACTTCAGTAAATTTATCCGGTGA CAATTTTTTCGAACCGCATTAGT- 5'
4	<sup>32</sup> P -CAGTCCTGCTGGTCGTTCCGGTACCCGGGATCC TTTA CTCTAGTCATTTAATAAGGCCACT GTTAAAAAGCTTGGCGTAATCA- 3' 3'- AGCCATGGGCCCTAGGAAATGAGATCAGTAAATTTATCCGGTGA CAATTTTTTCGAACCGCATTAGT- 5'
0/0 NO PAM	<sup>32</sup> P -CAGTCCTGCTGGTCGTTCCGGTACCCGGGATCCACGGGAGAAGTCATTTAATAAGGCCACT GTTAAAAAGCTTGGCGTAATCA- 3' 3'- AGCCATGGGCCCTAGGTGCCCTCTCTTCAGTAAATTTATCCGGTGA CAATTTTTTCGAACCGCATTAGT- 5'
24/24 NO PAM	<sup>32</sup> P -CAGTCCTGCTGGTCGTTCCGGTACCCGGGATCCAGGCCCTCTTCAGTAAATTTATCCGTGAGTAAAAAGCTTGGCGTAATCA- 3' 3'- AGCCATGGGCCCTAGGTCCGGAGAAGTCATTTAATAAGGCCACTCAATTTTTTCGAACCGCATTAGT- 5'

■ Protospacer Adjacent Motif (PAM) <sup>32</sup>P Radio-label DNA sequences complementary to guide-RNA are shown in red (Cognate).

**Table 4.3 | List of DNA targets used in gel electrophoresis experiments involving use of SYBR Gold II for staining and visualization of nucleic acid bands.**

Description	DNA Sequences
0/0	<p>5' - CCTTTAGAGAAGTCATTTAATAAGGCCACTGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATATTCCGGTGACAATTTTTTCGAACCGCATTAGT - 5'</p>
2	<p>5' - CCTTTAGAGAAGTCATTTAATAAGGCCAGAGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATATTCCGGTCTCAATTTTTTCGAACCGCATTAGT - 5'</p>
4	<p>5' - CCTTTAGAGAAGTCATTTAATAAGCCGTGAGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATATTCCGCACTCAATTTTTTCGAACCGCATTAGT - 5'</p>
6	<p>5' - CCTTTAGAGAAGTCATTTAATAAGCCGTGAGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATATTCCGCCACTCAATTTTTTCGAACCGCATTAGT - 5'</p>
7	<p>5' - CCTTTAGAGAAGTCATTTAATAACCGGTGAGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATATTGGCCACTCAATTTTTTCGAACCGCATTAGT - 5'</p>
8	<p>5' - CCTTTAGAGAAGTCATTTAATAACCGGTGAGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATATTAGGCCACTCAATTTTTTCGAACCGCATTAGT - 5'</p>
9	<p>5' - CCTTTAGAGAAGTCATTTAATTTCCGGTGAGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTAAGGCCACTCAATTTTTTCGAACCGCATTAGT - 5'</p>
16	<p>5' - CCTTTAGAGAAGTCATAATTTCCGGTGAGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCAGATTTAATAAGGCCACTCAATTTTTTCGAACCGCATTAGT - 5'</p>
17	<p>5' - CCTTTAGAGAAGTGTAATTTCCGGTGAGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCACATTTAATAAGGCCACTCAATTTTTTCGAACCGCATTAGT - 5'</p>
18	<p>5' - CCTTTAGAGAAGAGTAAATTTCCGGTGAGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCATTTAATAAGGCCACTCAATTTTTTCGAACCGCATTAGT - 5'</p>
24/24	<p>5' - CCTTTACTCTTCAGTAAATTTCCGGTGAGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATGAGAAGTCATTTAATAAGGCCACTCAATTTTTTCGAACCGCATTAGT - 5'</p>
2	<p>5' - CCTTTACTGAAGTCATTTAATAAGGCCACTGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATGACTTCAGTAAATATTCCGGTGACAATTTTTTCGAACCGCATTAGT - 5'</p>
4	<p>5' - CCTTTACTCTAGTCATTTAATAAGGCCACTGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATGAGATCAGTAAATATTCCGGTGACAATTTTTTCGAACCGCATTAGT - 5'</p>
0/0 NO PAM	<p>5' - CCACGGGAGAAGTCATTTAATAAGGCCACTGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGTGCCCTCTTCAGTAAATATTCCGGTGACAATTTTTTCGAACCGCATTAGT - 5'</p>
24/24 NO PAM	<p>5' - CCAGGCCCTTCAGTAAATTTCCGGTGAGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGTCCGGAGAAGTCATTTAATAAGGCCACTCAATTTTTTCGAACCGCATTAGT - 5'</p>

■ Protospacer Adjacent Motif (PAM)

DNA sequences complementary to guide-RNA are shown in red (Cognate).

**Table 4.4 | Pre-unwound DNA targets used in gel electrophoresis experiments involving use of SYBR Gold II for staining and visualization of nucleic acid bands.**

Description	DNA Sequences
1-2 <sub>mm</sub> 1-2 <sub>uw</sub>	<p>5' - CCTTTA GAGAAGTCATTTAATAAGGCCACTGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATGACTTCAGTAAATTATTCGGTGACAATTTTCGAACCGCATTAGT - 5'</p>
1-2 <sub>uw</sub>	<p>5' - CCTTTA CTGAGAAGTCATTTAATAAGGCCACTGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTATTCGGTGACAATTTTCGAACCGCATTAGT - 5'</p>
1-4 <sub>mm</sub> 1-4 <sub>uw</sub>	<p>5' - CCTTTA GAGAAGTCATTTAATAAGGCCACTGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATGAGATCAGTAAATTATTCGGTGACAATTTTCGAACCGCATTAGT - 5'</p>
1-4 <sub>mm</sub>	<p>5' - CCTTTA CTCTAGTCATTTAATAAGGCCACTGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTATTCGGTGACAATTTTCGAACCGCATTAGT - 5'</p>
16-24 <sub>uw</sub>	<p>5' - CCTTTA GAGAAGTCATTTAATTTCCGGTGAGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTATTCGGTGACAATTTTCGAACCGCATTAGT - 5'</p>
16-24 <sub>mm</sub> 16-24 <sub>uw</sub>	<p>5' - CCTTTA GAGAAGTCATTTAATTTCCGGTGAGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTATTCGGTGACAATTTTCGAACCGCATTAGT - 5'</p>
9-24 <sub>uw</sub>	<p>5' - CCTTTA GAGAAGTC TAAATTATTCGGTGAGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTATTCGGTGACAATTTTCGAACCGCATTAGT - 5'</p>
9-24 <sub>mm</sub> 9-24 <sub>uw</sub>	<p>5' - CCTTTA GAGAAGTCATTTAATAAGGCCACTGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCAGATTTAATAAGGCCACTCAATTTTCGAACCGCATTAGT - 5'</p>
7-24 <sub>uw</sub>	<p>5' - CCTTTA GAGAAGAGTAAATTATTCGGTGAGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTATTCGGTGACAATTTTCGAACCGCATTAGT - 5'</p>
7-24 <sub>mm</sub> 7-24 <sub>uw</sub>	<p>5' - CCTTTA GAGAAGTCATTTAATAAGGCCACTGTTAAAAAGCTTGGCGTAATCA - 3'</p> <p>3' - AGCCATGGGCCCTAGGAAATCTCTTCATTTAATAAGGCCACTCAATTTTCGAACCGCATTAGT - 5'</p>

■ Protospacer Adjacent Motif (PAM)

DNA sequences complementary to guide-RNA are shown in red (Cognate).

**Table 4.5 | List of guide-RNA used in all experiments and DNA targets used in single-molecule cleavage product release assay.**

Description	RNA Sequences
guide-RNA (AsCpf1 & FnCpf1)	5' -AAUUUCUACUCUUGUAGAU <b>GAGAAGUCAUUUAAUAAGCCACU-3'</b>
Modified guide-RNA for Cy5 labeling (AsCpf1 &FnCpf1)	5' -AAUUUCUACUCU <b>UGUAGAU</b> <b>GAGAAGUCAUUUAAUAAGCCACU-3'</b>
Modified guide-RNA for Cy5 labeling (LbCpf1)	5' -UAAUUUCUACUAAG*UGUAGAU <b>GAGAAGUCAUUUAAUAAGCCACU-3'</b>
guide-RNA (LbCpf1)	5' -GUAAUUUCUACUAAGUGUAGAU <b>GAGAAGUCAUUUAAUAAGCCACU-3'</b>

■ Thymine modification for Cy5 labeling. IDT code: **/iAmMC6T/**

\*amino-modifier phosphoramidite for smFRET binding experiments. Functionally interchangeable with the above mentioned thymine modification. IDT code: **/iUniAmM/**

RNA sequences with thymine modification or amino-modifier phosphoramidite were used for or smFRET experiments to study DNA interrogation by Cpf1-RNA

RNA sequences complementary to the protospacer in a cognate DNA target are shown in red (Cognate).

Unmodified guide-RNA for LbCpf1 has an extra G at the 5' end, compared to the canonical guide-RNA of LbCpf1. This is due to the T7 Transcription limitation which is one of the most widely used methods for both *in vivo* and *in vitro* transcription.

Description	DNA Sequences
0/0	<p>5' - CAGTCTGCTGGTCGT-TCGGTACCCGGGATCC <b>TTTA</b> <b>GAGAAGTCATTTAATAAGGCCACTGT</b> <b>AAAAAGCTTGGCGTAATCA</b>- 3'</p> <p>3' <b>-*GTCAGGACGACCAGCA</b> <b>AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTTATCCGGTGACAATTTTTCGAACCGCATTAGT</b>- 5'</p> <p>● 16 nucleotide biotinylated adaptor for surface immobilization</p>
6	<p>5' - CAGTCTGCTGGTCGT-TCGGTACCCGGGATCC <b>TTTA</b> <b>GAGAAGTCATTTAATAAGCGGTGAGT</b> <b>AAAAAGCTTGGCGTAATCA</b>- 3'</p> <p>3' <b>-*GTCAGGACGACCAGCA</b> <b>AGCCATGGGCCCTAGGAAATCTCTTCAGTAAATTTATCCGCCACTCAATTTTTCGAACCGCATTAGT</b>- 5'</p>

● Biotin      ■ Protospacer Adjacent Motif (PAM)      ■ Thymine modification for Cy3. IDT code: **/iAmMC6T/**

DNA sequences complementary to guide-RNA are shown in red (Cognate).



#### **4.6 AUTHOR CONTRIBUTIONS.**

Digvijay Singh and Taekjip Ha designed the experiments. Digvijay Singh performed single molecule experiments experiments. John Mallon performed radio-labeled gel electrophoresis experiments. Digvijay Singh performed gel electrophoresis experiments involving SYBR staining of nucleic acids. Digvijay Singh performed smFRET DNA unwinding experiments. Digvijay Singh, John Mallon, Ramreddy Tipanna expressed and purified Cpf1s. Digvijay Singh wrote the MATLAB package for data analysis and performed it with help from Anustup Poddar. Anustup Poddar assisted with the PEG passivation of some slides. Olivia Yang and Yanbo Wang assisted with some experiments.

## CONCLUSION AND OUTLOOK

There are many outstanding questions about the CRISPR system which can be addressed using sm approaches. For example, it should be possible to employ 3-4 color smFRET<sup>127,128</sup> for simultaneous observation of the DNA unwinding and nuclease and proofreading domain dynamics, helping us understand the molecular details of allosteric communication between RNA-DNA heteroduplex, DNA unwinding and cleavage action. While the majority of single molecule investigations have focused on CRISPR's DNA binding and its conformational changes upon DNA binding, a largely unexplored territory has been the assembly of CRISPR-RNA complex. smFRET is a powerful technique to probe the folding pathways of RNA<sup>129</sup> and thus could be used to probe structural rearrangements within guide-RNA and also within CRISPR enzymes that ensures efficient complexation of guide-RNA with CRISPR enzyme. The information from such experiments will aid in the rational design of efficient guide-RNAs and explain the variations in activity of different guide-RNA sequences<sup>31</sup>.

Another unexplored territory is the use of high resolution force (optical traps<sup>130</sup> & nanopore sequencer<sup>131</sup>) or correlative force-fluorescence spectroscopy. A suitable geometry of force application, with and without fluorescence visualization, can be used to extract additional information about transitions states between different dominant intermediates and also probe mechanical stability of CRISPR-RNA-DNA complex<sup>132</sup>. CRISPR-RNA-DNA is an ultrastable complex that persists even post DNA cleavage, thus masking the cleaved DNA sites from genome editing machinery. The chromatin *in vivo* is likely to be under tension and torsional stress which would be altered during the process of transcription and replication<sup>133,134</sup>. DNA stretching and twisting may help in quick dissociation of CRISPR-RNA-DNA complex and help expose cleaved DNA sites for genome editing machinery. Therefore, probing mechanical stability of CRISPR-RNA-DNA complex may be useful in this regard.

Dead CRISPR enzymes fused with transcriptional effectors are being increasingly used to achieve *in vivo* site-specific transcriptional control as the fused transcriptional effectors recruit additional effectors at CRISPR targeted sites<sup>97</sup>, but the molecular details of their recruitment are unknown. This recruitment assembly can be studied using another sm technique called Co-localization Single Molecule Spectroscopy (COSMOS)<sup>135</sup>, wherein entry and exit of fluorescently labeled biomolecules of interest (additional effectors in this case) can be studied at a particular site or a molecule (CRISPR targeted site in a DNA molecule).

Identification of critical steps in various stages of DNA targeting by CRISPR will not only help in designing strategies to improve its efficiency and reduce off-target effects but will be useful for evaluating the rational design of new CRISPR enzyme and their chemical and enzymatic regulators, which are being increasingly pursued for additional control over CRISPR enzymes<sup>136-138</sup>. For example, a chemical regulator may be rationally designed to target an important intermediate of CRISPR enzyme. sm approaches allow for precise characterization of differences between different CRISPR enzymes. This characterization of differences can be used to explain the variation in activity of different CRISPR enzymes in different conditions and organisms. CRISPR enzymes with different biophysical and biochemical parameters may be employed for different applications, expanding the functionalities of CRISPR toolbox.

# Chapter 5. Protocols

## **5.1 INTRODUCTION**

The goal of this chapter is to provide a detailed protocol for the key fundamental experiments that underlie this dissertation. While the previous chapters provide the methods for all the reported experiments, it is important to have protocols with detailed step by step procedures. The writing of the protocols in this chapter has been deliberately kept conversational to aid the reader in visualizing these experiments. Particular emphasis has been given to common mistakes and scenarios that result in the failure of these experiments. These protocols are based on previously described methods<sup>4,83,89,125,139,140</sup>.

## 5.2 EXPRESSION and PURIFICATION of CRISPR proteins

This protocol will describe the steps for the over-expression of CRISPR proteins and their purification. It will also describe the vector (plasmids) that will be used for these over-expression. Specific warnings have been added at many steps to emphasize the important points of that particular step. While this protocol is specific for Cas9 protein, it is broadly applicable to other Cas9 and Cpf1 variants and orthologs with necessary changes to some important buffers etc.

### 5.2.1 Materials

- A vector for over-expression of Cas9 or Cas9 mutant protein in *E. coli* cells.
- One Shot® BL21(DE3) Chemically Competent *E. coli*.
- Ampicillin for Anti-biotic resistance.
- LB-Agar Plates with Ampicillin as Anti-biotic resistance.
- 37 °C incubator and shaker.
- Lysogeny broth (LB) powder or LB media.
- Terrific Broth (TB) powder or media.
- Nalgene flasks for harvesting the cells.
- Ultra-centrifuge.
- Nalgene tubes, capable of handling ultra-centrifugation steps for collecting the cell-lysate.
- Ni-NTA resin (or IMAC resin). The protein being over-expressed has a histidine tag so Ni-NTA resin will be used to purify the protein.
- Basic salts needed for making various buffers. Double check that you have all of them.
- A column for Ni-NTA resin assisted protein purification.
- A gentle shaker.
- Cold room.

- Dialysis cassettes (for e.g. Slide-A-Lyzer™ Dialysis Cassettes) for the dialysis of the final retrieved protein.
- All the components involved in the expression, purification and final preparation of the protein must be certified RNase/DNase free components.
- Lysis Buffer – 500 mL:
  - 25 mL of 1 M Tris-HCl, pH 7.5 (50 mM Tris-HCl, pH 7.5 final concentration)
  - 14.66 g of NaCl (500mM NaCl)
  - 1 mL of 0.5M TCEP solution (1mM TCEP). Add it right before the lysis, do not keep it added in the lysis solution
  - 25 mL of 100% Glycerol (5% glycerol)
  - 500 µL of 200mM of Phenylmethanesulfonyl fluoride (PMSF) in its isopropyl solution (0.5mM PMSF). Add it right before the lysis, do not keep it added in the lysis buffer solution
  - Protease inhibitors (Roche). 2 tablets per 100mL of the lysis solution. Generally, 100 mL of lysis solution is used for 1-2 liters of bacterial solution used for over-expressing the protein. Add it right before the lysis, do not keep it added in the lysis buffer solution
- Wash Buffer – 2000mL:
  - 100mL of 1M Tris-HCl, pH 7.5 (50mM Tris-HCl, pH 7.5)
  - 58.44g of NaCl (500mM NaCl)
  - 4mL of 0.5M TCEP solution (1mM TCEP). Add it right before the wash, do not keep it added in the wash buffer solution
  - 1.36g of Imidazole powder (10mM imidazole)
  - 100mL of 100% Glycerol (5% glycerol)
- Elution Buffer – 20mL:
  - 1mL of 1M Tris-HCl, pH 7.5 (50mM Tris-HCl, pH 7.5)
  - 2.5mL of 4M NaCl (500mM NaCl)

- 0.4mL of 0.5 M TCEP (1mM TCEP)
- 2mL of 3M imidazole (300mM imidazole)
- 1mL of 100% glycerol (5% glycerol)
- Check the pH of this solution. With the presence of Tris-HCl, the pH should be regulated but high concentration of imidazole will cause pH increase. The pH should not be higher than 8.0.
- TEV buffer-1.5 liters. This is the buffer in which the Ni-NTA eluted protein will be dialyzed to, with TEV protease added to it which will cleave the TEV site tag.
  - 30mL of 1M Tris-HCl, pH 7.5 (20mM Tris-HCl, pH 7.5)
  - 13.97 g of KCl powder (125mM KCl)
  - 3mL of 0.5 M TCEP (1mM TCEP)
  - 75mL of 100% glycerol (5% glycerol)
- All the components involved in the expression, purification and final preparation of the protein must be certified RNase/DNase free components.

### **5.2.2 Basic introduction about the pET based vector for Cas9 protein production**

pET (Plasmid for Expression by T7 RNA polymerase) 28a vector generally has the gene of interest inserted under the T7 promoter as shown below. The T7 RNA polymerase can only transcribe the gene of interest since our gene of interest is present under the T7 promoter. The cell line (BL21 cell lines) which we are using does not have any endogenous levels of T7 RNA polymerase i.e. does not produce T7 RNA polymerase in normal conditions, but will produce them when we switch the production on. The cell line (BL21 cell lines) has been artificially engineered and has the T7 RNA polymerase gene inserted within the bacterial genome under the lac operator. Please note that this T7 RNA polymerase gene and the lac operator is in the bacterial chromosome and is most likely not present in your plasmid of interest (which has your gene of interest). But if they are present in the plasmid, they have the similar job i.e. the Lac operator would be present 5' to your gene of interest. When Lac repressor is bound to the Lac operator, it



will prevent any transcription from that region i.e. Lac operator genes and the gene of interest present downstream of it, hence no protein production. In a normal condition, the lac repressor protein (present endogenously) is bound to the lac operator and thus preventing the transcription of T7 RNA polymerase gene in bacteria. This prevents the production of any T7 Polymerase in the cell which obviously prevents the transcription of our genes of interest (in the plasmid we transformed) in the pET 28a vector. When the IPTG is added, it arrests all the Lac repressor from its bound site at the lac operator. The T7 RNA polymerase is translated in bulk from the bacterial genome which then starts binds the T7 promoter in the pET28 vector and starts transcribing the gene of interest. Our gene of interest is the one that codes for Cas9 or mutants. It has a maltose binding sequence (to improve the solubility of the protein) conjugated right next to a TEV protease site. The TEV protease will be added to cleave out the maltose binding site from the protein of interest after majority of purification steps.

### **5.2.3 Transformation of the cells with the vector containing your gene of interest**

- 1) Prepare Ice.
- 2) Get the competent cells ([One Shot® BL21\(DE3\) Chemically Competent \*E. coli\*](#)) from -80 °C freezer.
- 3) Never vortex the competent cells. These are the cells we discussed briefly above that have the artificially engineered gene in them that will produce the T7 RNA polymerase which will then transcribe the gene of interest in our vector.
- 4) Thaw the competent cells on ice, store the competent cells on ice at all times while aliquoting.
- 5) Bring a clean Eppendorf tube or something similar and keep it in the ice for pre-chilling. It is essential that tube is placed on ice before the competent cells are aliquoted directly into each pre-chilled tube.

- 6) Gently mix the competent cells. Aliquot 50 $\mu$ L of the competent cells into the appropriate number of tubes, generally only 1 will be required.
- 7) Dilute XL 10-Gold Beta mercaptoethanol mix (BME) provided with the appropriate kit 1:10 with nuclease free water. The cells suffer from a tremendous oxidative stress and hence the BME is added which is a strong reducing agent for relieving the oxidative stress.
- 8) Each aliquot of cells (50 $\mu$ L) requires 1 $\mu$ L of diluted Beta mercaptoethanol. Add 1 $\mu$ L of BME into each of the aliquots.
- 9) Add a very small amount of the plasmid from the plasmid stock to the cells, so that the its final amount in the solution ranges from 1-50 ng.
- 10) Incubate the reactions on ice for 30 minutes.
- 11) Heat pulse each transformation reaction in a bath preset at 42 °C for 20-30 seconds. The duration of the heat pulse is critical. The heat pulses will porate the cell membrane and few of the plasmid molecules will be able to get inside the cells.
- 12) Put the tube back onto the ice and let it incubate for 5 minutes.
- 13) Add 950 $\mu$ L of a freshly prepared SOC (Super Optimal broth with Catabolite repression) media to each transformation and incubate the reactions at 37 °C for 1 hour in the shaker shaking at 225-250 revolutions per minute (rpm). Please make sure that the 1000 $\mu$ L sample is properly aerated and that it is in a big enough flask. Please do not use small tubes for growing this culture. Cells

growing in a small tube tend to have problems wherein many cells are cramped into the bottom of the solution where they are forced to grow without proper oxygen levels. And these cells are extremely bad for protein production. The amount of bacterial culture in a flask with respect to its maximum volume should be a very low fraction. Please also make sure that the cap to the flask is very lightly placed, so that it does not fall off during the shaking but should be able provide enough aeration. I personally putting aluminum foil to cover the flask. The aluminum foils do not fall off and yet provide good aeration, if they are lightly placed.

14) Meanwhile, take out two LB-Agar plates with appropriate anti-biotic resistance and put them in the 37 °C incubator for ~5-10 minutes.

15) After 1 hr of bacterial growth in the shaker. Create the following dilutions of the culture you grew:

16) 5 times dilution: 20 $\mu$ L culture + 80-100 $\mu$ L of SOC media.

17) 2 times dilution: 75 $\mu$ L culture + 75 $\mu$ L of SOC media.

18) Take our the LB-Agar plates (from 37 °C incubator) and pour the above two diluted culture solutions into the two plates, right at the center.

19) Meanwhile, light up the Bunsen burner and use a glass spreader or something similar (kept in ethanol) and burn it for few seconds under a propane flame. Then let it cool down. Now the spreader has been sterilized and you can use the spreader to spread the suspended culture onto the agar plates. Make sure you spread it evenly. Please make sure the spreader has been cooled down to room temperature before spreading the culture solution.

- 20) Keep the agar plates upside down in incubator at 37 °C. Please make sure to keep the plates away from each other and away from other plates. You do not want any contamination. Keep changing gloves while you are doing all of this to avoid any possible chance of cross-contamination.
  
- 21) Transformants will appear as colonies following overnight incubation at 37 °C. All the cells that took up the pET28 vector (that produced the antibiotic resistance) will be the only ones surviving. So this is a very easy selection for cells that took up the plasmid.

#### **5.2.4 Preparing the media for large-scale culture**

- 1) Take 2 four liter flasks and add ~36 grams of Terrific Broth to each one of them and add 750mL of extremely clean nuclease free water to it. Please remember the point, I wrote earlier about giving the cells enough room for good aeration. Hence do not grow the 1-liter or similar culture in a 2-liter flask. That is too small a flask for a 2-liter culture. Please also make sure that the cap to the flask is very lightly placed, so that it does not fall off during the shaking but should provide enough aeration. I personally putting aluminum foil to cover the flask. The aluminum foil does not fall off easily and yet provide good aeration, if they are lightly placed.
  
- 2) Add 6mL of 100% glycerol to each of the flasks and mix well to dissolve everything.
  
- 3) Using commercial Terrific Broth powder is highly recommended. Or you can make your own Terrific Broth powder.
  
- 4) It is also recommended to use [granulated terrific broth powder](#) .These are dust-free and dissolve much easily. Some of the commercial available ones are packed in gelatin capsules. I would avoid using those. There is no point adding gelatin in your culture media. Hence granulated

gelatin free Terrific Broth powder are the best.

- 5) Put the aluminum foil on the flask and then the autoclave tapes on them and autoclave the flasks containing the media.
- 6) Take them out of the autoclave machine and let them cool down to the room temperature, at which point add 75mg of Ampicillin to them. This is for the selection via anti-biotic resistance. The desired concentration of the Ampicillin the media is 100 $\mu$ g/mL. Hence 75mg to each 0.75 liter of the media achieves that. Mix everything well.
- 7) Please do not add Ampicillin powder while the media is still hot from the auto-clave step, high temperature may degrade the Ampicillin compound.
- 8) Store a small amount from this formulation, to be used a base-line correcting solution(media) when measuring the optical density at 600 nanometers wavelength (OD<sub>600</sub>) for the culture growth. Please note that OD<sub>600</sub> is a decent estimate for the density of cells in the solution, but far from accurate. The best way would be to calibrate your spectrophotometer so that you can derive an empirical relationship between the OD<sub>600</sub> obtained from your spectrophotometer and the density of the cells (for e.g. cells/ml).

### **5.2.5 Preparing the inoculant for large-scale culture**

- 1) Take 4mL of the LB medium in a 50mL Falcon tube. Please remember the point, I wrote earlier about giving the cells enough room for good aeration. Hence do not grow the 4mL culture in a 10mL tube. That is too small a tube for a 4mL culture. Please also make sure that the cap to the flask is very lightly placed, so that it does not fall off during the shaking but should provide enough aeration. I personally putting aluminum foil to cover the flask. The aluminum foil does

not fall off and yet provide good aeration, if they are lightly placed.

- 2) Take a single colony from the LB agar plates from the step number 28. Choose a colony which is well separated from other colonies, yet be in the middle of the plate. The middle of the plates tends to have more nutrition and grow better and hence such colonies are better than the colonies at the edges. But please make sure that you pick a single colony. All the cells in your culture that you will be growing need to be from a single progeny. And a single colony always starts off with a single cell hence ensuring that all the bacterial cells would be coming from a single parent.
- 3) Take the single colony using a 10-200 $\mu$ L pipette tip. This tip works the best to pick the colony. Put the tip on a pipette and aspirate the colony onto the 4mL small inoculant culture. Mix everything well and discard the pipette tip into bacterial cell culture trash.
- 4) Put the 4mL culture into a shaker at 37 °C, rotating at 250 rpm. Make sure that the cap/or foil for the flask is just lightly placed for the proper aeration.
- 5) The steps of preparing the inoculant for large-scale culture and preparing the media for large-scale culture can be done largely side by side.

#### **5.2.6 Transferring the inoculant from 4mL culture to the large-scale culture**

- 1) First of all, put the 4-liter flask, each containing 0.75 liter of the media supplemented with 75mg of Ampicillin into the shaker at 37 °C. The large media has to be brought to the same temperature as the small 4mL inoculant culture. The transfer must take place at the same temperature.

- 2) Let the inoculate grow to a point where the expected  $OD_{600}$  is  $\sim 1.0$ . i.e. the solution will turn milky. Do not go beyond that. The cells may just enter a toxic phase of growth if the cells in the inoculant culture are grown too much without transferring them to the large-scale culture. From the addition of a small single colony to the desired inoculant culture, the inoculant culture is grown for  $\sim 4-6$  hrs before transferring them to the big culture. In 4 hrs, the solution will turn quite milky.
- 3) In my experience, please do not transfer a prematurely grown inoculant into the large-scale culture. Bacterial cells employ quorum sensing. And my armchair hypothesis is that a dilute inoculant will dilute further in the large scale culture. And it will take even longer for the cells to grow to expected ODs. Let the inoculant turn milky before transferring it to the large scale culture.
- 4) Once the temperature of the 4 liter flasks media has reached  $37\text{ }^{\circ}\text{C}$ . Equally distribute the 4mL culture between the two flasks. The 'equally' here is very important because then the amount of cells ( $OD_{600}$ ) in both the 1-liter culture will always be the same and their temperature changes/induction/harvesting etc. can all be carried out concurrently without any delay between them. The cells divide and multiply and hence even a small change in the amount of inoculant that is distributed between them will lead to a big subsequent changes in their  $OD_{600}$ .
- 5) After the transfer is over, discard the inoculant flask in bacterial trash. And let the large scale (0.75 liter) culture (s) grow at  $37\text{ }^{\circ}\text{C}$  till the temperature changes/induction are required.

### **5.2.7 Setting up the $OD_{600}$ measurement**

- 1) A small amount from the large scale 0.75 liter (after adding Ampicillin) media was stored for using as a blank for the OD measurements. It is best to use a dedicated spectrophotometer for the

OD<sub>600</sub> measurements i.e. the labs usually have a small machine set for measuring OD at precisely 600 nm wavelength point.

- 2) Baseline with the blank solution that had been stored from the large-scale media. Now the machine is ready for measuring the OD of actual large scale culture.

### **5.2.8 OD<sub>600</sub> check**

- 1) Keep checking the OD<sub>600</sub> of the large-scale culture.
- 2) Check the OD<sub>600</sub> of all the flasks separately, there may be significant changes of OD<sub>600</sub> between and thus you do not want to rely on the OD<sub>600</sub> from one flask and assume it for the other flasks as well.
- 3) As mentioned before, do not set up a culture of more than 0.75- 1L in a 4L flask. The bacteria do not grow well and goes into an anaerobic condition, if they are grown in a very large culture i.e. 3 liters in a 4-liter flask, it prevents proper aeration of the cells. A good check of whether cells are growing well or not is to check the doubling time, at 37 °C the doubling time should be about 20 minutes and at 21 °C, it should be about 40 minutes.

### **5.2.9 Induction with (Isopropyl β-D-1-thiogalactopyranoside) IPTG**

- 1) Keep checking the OD<sub>600</sub> of the large-scale culture. The temperature should now be 37 °C.
- 2) When the OD<sub>600</sub> reaches 0.5, induce the production of the protein by supplementing both the culture with 0.5mM IPTG. i.e. to each 1-liter culture, add 1mL of 1000X IPTG (0.5M solution).
- 3) Now lower the temperature to 18 °C.



- 4) It is important to understand why the temperature was decreased to 18 °C, sometimes some big proteins cannot take the excess temperature well. They will fold and be expressed properly when they are at low enough temperature. Same goes for this protein, it is important for the protein to be actively formed that the temperature is not too high. 18 °C is the most optimum temperature for this protein to be stably over expressed.
- 5) Let the cells/culture grow at 18 °C for 16-18 hrs and then harvest the cells. I have noticed that the culture left for longer than 16-18 hrs, tended to have lesser over-expression. Perhaps the protein of interest was degrading. Anyways 16-18 hrs is sufficient for massive expression of your for Cas9 proteins as I have observed.

#### **5.2.10 Harvesting the cells**

- 1) After 16-18 hours, depending on the type of centrifuge, pour the bacterial culture into appropriate Nalgene flasks and centrifuge the culture at 5000-7000 rpm for 20-30 minutes at 4 degrees.
- 2) The cells would have collected at the bottom of each flasks as bacterial pellets. Discard the supernatant media. Quickly store the pellets at -80 °C for later purification. Or proceed directly for the purification.
- 3) Keep a very small amount of the pellet for the SDS-Page to confirm the protein overexpression. The protocol for the SDS-Page with the pellets is attached at the end of this protocol.
- 4) You now need 100mL of lysis solution per 1 liter of culture from which the cells were harvested. You would have only the ions/salts in the lysis solution. Please add the following to 100mL of

lysis solution in the given order strictly.

- Two protease inhibitor tablets (Ones from Roche work well). These are chemicals that inhibit the protease from acting.
  - Add 100 $\mu$ L 200mM of isopropyl solution of PMSF (0.5mM PMSF final concentration). PMSF is extremely toxic. So please be very careful.
  - 200  $\mu$ L 0.5M TCEP solution (1mM TCEP final concentration in lysis buffer solution).
  - Add appropriate amount of lysozyme (Sigma Aldrich) so that its final concentration is 1mg/mL. Add the lysozyme powder and not lysozyme chloride or any other salt of it.
- 5) Mix the lysis solution well to make sure that everything dissolves. You cannot vortex it or give it harsh mechanical shakeup. Mix and dissolve all things gently. Let it take time. Do not rush.
- 6) Re-suspend the harvested cells in the above lysis buffer solution and mix them well but gently. Do not give mechanical shakeups to the solution at any subsequent solution at any stage because that will denature your protein and lysozyme etc.
- 7) Mix the cells thoroughly and gently in the lysis buffer solution. You would have to break up the chunks of the cells in the DNA. Doing this will significantly increase the yield of your protein. Take a spatula and gently break up the cellular chunks manually one by one. This can be a boring process but it is important and well worth it.

### **5.2.11 Lysing the cells**

- 1) Put the cells suspended into the lysis buffer solution into ice and take it to the sonicator.

- 2) Set the amplitude of the sonicator to 30%. Please follow the settings appropriate for your sonicator.
- 3) Break the cells by sonication in cycles of 2s sonicate-2s rest. Each cycle is to last 1 minute and you have to do 3 cycles. The sonication heats up the sample that is why the cells have to be surrounded by ice throughout this process.
- 4) Put the lysate now into Polyallomer tubes. These tubes are very strong and can withstand the rigors of the ultra-centrifugation which would be the next step. Do not use tubes which are not certified for use in ultra-centrifugation, as they will break during the process.
- 5) Centrifuge the lysate now at 15,000g -17,000g for 30 minutes to remove the cell lysate away from the other hard cellular junks which will settle down. And the desired cell lysate, containing protein, will be in the supernatant.

### **5.2.12 Ni-NTA Chromatography for protein purification**

A small note about Ni-NTA Chromatography. Since your protein of interest has a polyhistidine-tag. Ni-NTA can be used to selectively retrieve your protein of interest leaving behind the other proteins present in the cell-lysate. It uses the ability of Histidine to bind nickel. A polyhistidine-tag is an amino acid motif in proteins that consists of at least five histidine (His) residues, often at the N- or C-terminus of the protein. It is also known as polyhistidine-tag, and by the trademarked name His-tag. Nickel is bound to an agarose bead by chelation using Nitroloacetic acid (NTA) beads. Several companies produce these beads as His-tagged proteins are some of the most commonly used. The general method is to batch absorb the His-tagged protein onto the column packed with the Ni-NTA agarose beads where solutions of low concentrations of phosphate and imidazole are used to remove low affinity bound proteins (i.e. the ones which do not have the His-tag) at a time when the His-tagged protein is strongly bound to the Ni-NTA

beads. And finally a high concentration of the Imidazole is passed through the column to elute out the protein of interest which is as we discussed remains firmly bound to the beads at low concentrations of imidazole but disassociates and gives away in the higher concentration of Imidazole.

- 1) To the clean cell-lysate, add 3-5mL of [Ni-NTA resin](#). Shake the resin well before adding it to the cell-lysate. 3-5mL of this resin is appropriate for 100mL of the cell-lysate. Incubate in the cold room (in the gentle-shaker) for 30-45 minutes. Put it for a slow and gentle shake.
- 2) Do not add copious amount of this Ni-NTA resin, otherwise it will be very slow for the column. Like mentioned, 3-5mL is just fine.
- 3) The solution can be directly added to a column and the resin can be washed with the wash buffer in the column, but letting the cell-lysate drip through the column and then letting the wash buffer drip through the column is a painfully slow process largely because the cell-lysate solutions are quite viscous. It is at this step that the user can lose patience and try to hurry things, which may mean less wash of the Ni-NTA resin/slurry and less wash could mean that the protein prep could end up with some non-specific protein components. And it would be extremely deleterious, if those non-specific proteins are nucleases.

Here's describing a much more effective way of washing the Ni-NTA resin.

- 4) After adding the 3-5mL of Ni-NTA resin/slurry to the cell -lysate and 30-45 minutes incubation. All the purification steps are to be done at 4 °C, so cold room is the ideal place to carry out the purification and required incubations.
- 5) After the incubation, transfer the cell-lysate solution to falcon tubes in equal amounts.

- 6) Centrifuge at 2400 rpm (*no higher!*) for 3 minutes, and gently pour off supernatant into the sink immediately after spin is over. Please do not remove all the cell-lysate solution. You let a full portion of Ni-NTA immersed in the solution. The Ni-NTA resin/slurry must never get dry and be exposed to air.
- 7) These pellets are quite fragile, and will begin to dissipate in the supernatant rather quickly so remove the supernatant quickly. Avoid accidentally pouring off your Ni-NTA resin.
- 8) Now add wash buffer solution to these falcon tubes (fill to ~50 mL mark) containing the Ni-NTA resin, which would be at the bottom of the tube. Put them in the gentle shaker for 10-15 minutes. This will wash the Ni-NTA resin more. Centrifuge again at 2400 rpm for 3 minutes, and gently pour off some supernatant into the sink immediately after spin is over, leaving behind substantial wash buffer solution so that the Ni-NTA resin is completely submerged. Repeat this step 3 more times for complete wash of the Ni-NTA resin.
- 9) Now pour the Ni-NTA resin along with the wash buffer solution (do not remove all the supernatant as discussed before) into a column. Please make sure that the prior to the use, column has been properly cleaned.
- 10) Now pass about 50mL- 100mL of the wash buffer through the column and collect the dripping solution in the falcon tubes. The protein of interest is still bound to the Ni-NTA resin/slurry.
- 11) Do this till the most of the wash buffer solution has almost run through the column. Please do not let the Ni-NTA resin/slurry *ever* run dry. You let the entire portion of Ni-NTA resin immersed in the wash solution. The Ni-NTA resin/slurry must never get dry and be exposed to air.

12) All the above processes with the wash buffer solution is to get rid of all the proteins that bound Ni-NTA resin/slurry non-specifically. This wash step is extremely critical, because if the non-specific proteins are not washed out from the Ni-NTA resin, they may elute with your protein of interest and create all sorts of problems. Chiefly, these non-specific proteins maybe nucleases and will remain as a nuclease contaminant in your CRISPR protein samples. Hence, it is very critical that the resin/slurry be washed extensively.

13) The protein of interest is still bound to the Ni-NTA resin/slurry.

### **5.2.13 Elution of the protein of interest**

When the wash buffer level just reaches the very top of the Ni-NTA resin/slurry level i.e. when the wash buffer is about to get over. That is the time to elute. Adding elution buffer while lot of wash buffer is still in the column makes no sense, because the wash buffer present in the column will not only significantly dilute the high concentration imidazole necessary to elute the protein of interest but it will also carry some unwanted protein which was non-specifically stuck in the Ni-NTA resin and is now part of the wash buffer. So just when the column is about to run dry, that is when you add the elution buffer. Typically, do not add more than 10-15mL of elution. 10-15mL is enough to retrieve all the protein of interested bound to the Ni-NTA resin. 10-15mL is enough for the pellets collected from 0.75-liter culture, for more amount of pellets, you may need more elution buffer. Never let the Ni-NTA resin/slurry run dry.

- 1) Blank the spectrophotometer using the elution buffer.
- 2) Elute your protein of interest in a single step by adding 10-15mL of elution buffer containing 300mM Imidazole Buffer Solution (pH 8.0) into the column.

- 3) After adding 10-15mL elute buffer, close the cap of the column, this will stop the column flow and give enough time to the elute buffer to retrieve all possible Hist-tagged protein of interest from the Ni-NTA resin. Collect all the elutant in a brand new and ultra-clean nuclease free tube.
- 4) Measure the absorbance at 280 nm and record this, along with the volume, to give us the overall yield. Also, print out the absorbance spectrum for your notebook, since this has information on the purity of protein (or, if it is co-purifying with any nucleic acid).
- 5) Absorbance at 280 nm is important to measure the purity of the protein (against nucleic acids) as well as to figure out the amount of TEV protease that will be needed.

#### **5.2.14 Dialysis and obtaining Apo-Cas9 by utilizing TEV protease to remove Hist-Tag and MBP-Tag**

As we discussed above that the protein was designed along with a maltose binding site (MBP Tag) which increases the solubility of the protein so that lot of it can be retrieved from the cell pellet and be brought into the cell-lysate solution. After the purification of the protein, we must get rid of this maltose binding site in the protein. The way we had designed the original protein construct was that it had His-Tag-MBP-Tag-TEV-Cas9. His-Tag for purification using Ni-NTA chromatography as performed above. MBP-Tag is for higher protein solubility. TEV is the protease site for the TEV protease, which will make the cut at the TEV protease site thus removing both the Hist-Tag and MBP-Tag from the rest of the protein i.e. Cas9 sequence.

- 1) Add the appropriate amount of TEV protease directly to the elutant containing protein of interest. Figure out the appropriate amount by using the absorbance 280 nm and thus the amount of the protein of interest.

- 2) Prepare the TEV buffer as described in the materials section above. This is the buffer to which the elutant/elute obtained will be dialyzed to.
- 3) After adding the TEV protease to the elutant, now transfer it to a dialysis cassette that can hold atleast 10mL of the solution. Be very careful in adding the solution/sample to the cassette, the cassette membranes are very gentle. Any damage/leak you introduce will lead to complete loss of your protein of interest. After adding the solution/sample to the cassette, double-triple check to make sure that there is no leak at all before adding it to the vast dialysis buffer.
- 4) Make sure to suck out the air from the cassette, too much air reduces the surface area and will slow the dialysis process.
- 5) Now drop the cassette into the TEV buffer. Make sure to use a floater so that cassette can be floated and it does not drop down to the bottom of the giant dialysis solution. Then, it will be difficult to retrieve the cassette afterwards.
- 6) A reminder again that all of these steps must be taking place in the cold room at 4 °C.
- 7) Do the dialysis of the protein sample in TEV buffer for at least 6 hours in cold room at 4°C to remove excess NaCl, Imidazole and everything other than protein of protein of interest i.e. Cas9. During the dialysis, the TEV protease will cut out the MBP site from your protein of interest leaving Apo-Cas9.
- 8) It is critical that this dialysis not proceed longer, as Cas9 is not very soluble at low salt and can precipitate with prolonged incubation. The disadvantage of raising the salt concentration is that



Cas9 will no longer bind the SP column efficiently (i.e. FPLC steps).

- 9) Retrieve the dialysis cassette and slowly aspirate out the sample/solution inside the cassette. This solution now has the following cleaved Apo-Cas9 without His-Tag and MBP tag, fusion of His-Tag-MBP-Tag, and TEV protease. Commonly used TEV protease have a His-tag in it as well. So, all the components (except our desired Apo-Cas9) has the Hist-Tag. Therefore, this can be used to separate Apo-Cas9 from other unwanted components. Add a small amount of Ni-NTA resin again to this sample and run it through another fresh and ultra-clean column. This time, collect the wash/flow through, because the wash/flow through contains desired Apo-Cas9. All other unwanted components will be stuck in the column in the Ni-NTA resin.
  
- 10) The retrieved flow through is now ready for further FPLC purification steps. Please carry them out as per the specifications described in the methods section of the chapter 2 and chapter 3 of this dissertation.

## 5.3 IN VITRO TRANSCRIPTION AND PURIFICATION OF RNA

### 5.3.1 Introduction

Chemically synthesized RNAs are extremely expensive to buy. If you are not looking for any special modification, you should almost always in vitro transcribe the RNA instead. You can make any amount of RNA from a much cheaper DNA template and you can keep making more should you need more.

### 5.3.2 Materials

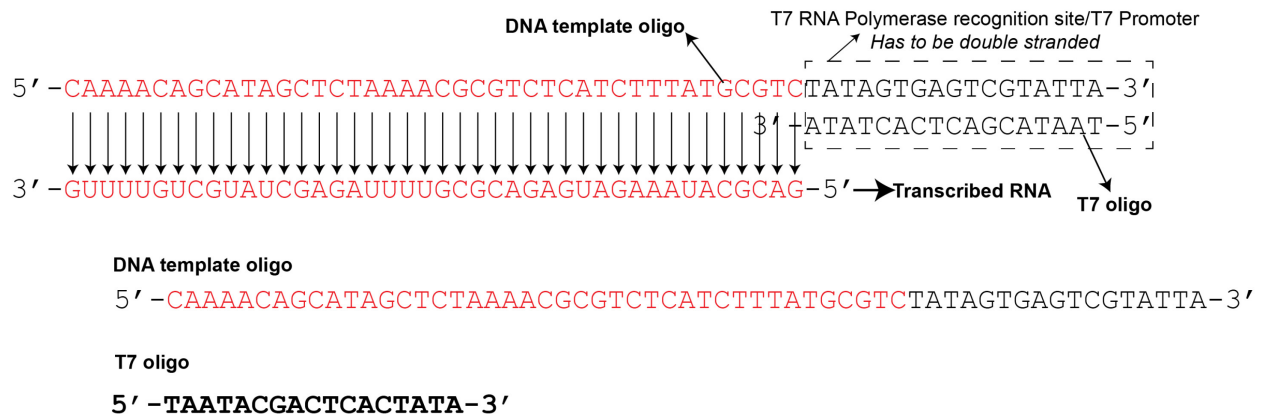
- DNA template oligo which depends on the RNA sequence you want to transcribe.
- T7 oligo which is the same for all RNA that you want to transcribe.

T7 oligo =5'-TAATACGACTCACTATA-3'. Hybridization of DNA template oligo and T7 oligo produces the DNA template.

- HiScribe™ T7 Quick High Yield RNA Synthesis Kit
- Zymos RNA Clean & Concentrator™-25
- T50 buffer (20mM Tris- HCl, pH 8.0 and 50mM NaCl).
- Ultra-Centrifuge.
- DNase I (RNase-free) from New England Biolabs (NEB).
- E-Gel™ EX Agarose Gels, 4%
- E-Gel machine
- Do not use any component which is not certified nuclease free.

### 5.3.3 Procedure

Schematic of the DNA template and the RNA it transcribes. Please take a look at this carefully. We already have the indicated T7 oligo, you would just need to order a DNA template oligo.



**Figure 5.1 | Schematic of DNA template for invitro transcription**

The presence of C right after the T7 promoter region is another design consideration to be included which is explained below.

### Importance of C/CC nucleotide(s) for high efficiency transcription

- Template for transcribing RNA 1 with a single C:

5' -CAAAACAGCATAGCTCTAAAACGCGTCTCATCTTTATACATC[TATAGTGAGTCGTATTA] 3'  
 3' -[ATATCACTCAGCATAAT] +5'

Transcribed RNA 1:

3' -GUUUUGUCGUAUCGAGAUUUUGCGCAGAGUAGAAUAUGUAG-5'

- Template for transcribing RNA 2 with two CC:

5' -CAAAACAGCATAGCTCTAAAACGCGTCTCATCTTTATACA[CC]TATAGTGAGTCGTATTA] 3'  
 3' -[ATATCACTCAGCATAAT] +5'

Transcribed RNA 2:

3' -GUUUUGUCGUAUCGAGAUUUUGCGCAGAGUAGAAUAUGUGG-5'

- Sequence for transcribing RNA 3 without C:

5' -CAAAACAGCATAGCTCTAAAACGCGTCTCATCTTTATACAT[T]TATAGTGAGTCGTATTA] 3'  
 3' -[ATATCACTCAGCATAAT] +5'

Transcribed RNA 3:

3' -GUUUUGUCGUAUCGAGAUUUUGCGCAGAGUAGAAUAUGUAA-5'

- The dsDNA region in the box indicate the minimal T7 promoter region.
- The letters highlighted in the cyan indicate a C/CC required for high efficiency of T7 induced transcription.
- CC will do better than C. But a single C is enough. You need not shoot for CC.
- If you RNA design prevents you from using these C/CC. Then you can still transcribe without them. The efficiency will be low. But if you scale up your reaction, you can still get enough amount of your transcribed RNA.
- The base highlighted in the green indicate a T replacing C.

**Figure 5.2 | Importance of C/CC at +1 and +2 sites for high efficiency invitro transcription**

#### 5.3.4 Preparation of partial dsDNA template for invitro transcription

- 1) Mix the DNA template oligo and T7 oligo in 1:1 ratio at 10µM concentration in a T50 buffer (20mM Tris- HCl, pH 8 and 50mM NaCl).
- 2) Heat the mixture at 95 °C for 2 minutes and let it cool down slowly to the room temperature over the next 1-2 hours.

- 3) 10 uM of the partial dsDNA template is now ready for use in the invitro transcription reaction.

### 5.3.5 Setting up the invitro transcription reaction

- 1) Add the required components in exactly the following order:

Nuclease-free water	x $\mu$ l
rNTP Buffer Mix	10 $\mu$ l
Template DNA	y $\mu$ l (1 $\mu$ g)
T7 RNA Polymerase Mix	2 $\mu$ l

Use only the components/samples from the same HiScribe Kit. Do not use rNTP from other kit.

Different manufactures keep their components in different buffer and different concentration.

You never want to mix stuff from different kits for any experiments.

- 2) Choose x such that the total reaction volume is 30 $\mu$ L if the RNA transcript < 0.3kb. If the RNA transcript >0.3kb, then the total reaction volume needs to be 20 $\mu$ L.
- 3) Mix thoroughly and Incubate at 37 °C. Incubate at 37°C for 4 hrs if the length of the RNA transcript < 0.3kb. Incubate at 37 °C for 2 hrs if the length of the RNA transcript > 0.3kb. The timing is especially critical if the RNA > 0.3 kB.
- 4) After the incubation, the RNA is ready. It now needs to be purified. We also need to use DNase to chop off the DNA template. Please note that the concentration/amount of DNA template << RNA transcribed, so for many applications, it may not be overtly important to get rid of the DNA template. But I suggest that you do it anyways.

### 5.3.6 DNase treatment to degrade the partial dsDNA template

The reaction is quite viscous from all the RNA that has been transcribed. We need to dilute it first by adding 30  $\mu\text{L}$  nuclease-free water to the 20  $\mu\text{L}$  reaction. Then add 2  $\mu\text{L}$  of DNase I (RNase-free, supplied by NEB), mix and incubate for 15 minutes at 37°C. Use only the DNase I from NEB if you are following this protocol.

### 5.3.7 Purification of the RNA

- 1) The total volume of the solution containing the transcribed RNA along with tons of rNTP is 50  $\mu\text{L}$ .
- 2) Now we will proceed to use Zymos RNA Clean & Concentrator<sup>TM</sup>-25 for purifying the RNA.
- 3) Add 2 volumes ( $2 \times 50\mu\text{L}$  in this case) RNA Binding Buffer to the sample and mix. Total volume is now 150 $\mu\text{L}$ .
- 4) Add an equal volume, 150  $\mu\text{L}$  in this case, of pure ethanol and mix.
- 5) Transfer the sample to the Zymo-Spin<sup>TM</sup> IIC Column in a Collection Tube and centrifuge for 30 seconds. Discard the flow-through.
- 6) Add 400 $\mu\text{L}$  RNA Prep Buffer to the column and centrifuge for 30 seconds. Discard the flow-through.
- 7) Add 700 $\mu\text{L}$  RNA Wash Buffer to the column and centrifuge for 30 seconds. Discard the flow-through.

- 8) Add 400 $\mu$ L RNA Wash Buffer to the column and centrifuge for 2 minutes to ensure complete removal of the wash buffer.
- 9) Transfer the column carefully into an RNase free tube.
- 10) Elution. Add 50  $\mu$ L T50 buffer directly to the center of the column matrix and centrifuge for 30 seconds. The desired transcribed RNA will be eluted in this T50 buffer. The efficiency of this column in removing small molecules like rNTP is pretty high. 99% of the rNTP must have been removed and you would have a very pure RNA in the T50 buffer which can be stored in -80C for long term storage.
- 11) To further remove small amounts of possible residual rNTP and obtain even ultra-pure RNA. I generally do the same Zymos column purification again using new columns and tubes. So repeat the section 6.6.7 again and that should give you an ultra-pure RNA sample stock.

Measure the concentration of the RNA in the Nanodrop using the absorbance at 260nm. You can measure the extinction coefficient of the RNA using any of the online tools available like IDT OligoAnalyzer.

Do not add Magnesium ions to any RNA stock for long term storage. Nuclease contaminants need Magnesium ions to function and by denying them that you are protecting your RNA from possible nuclease contamination activity. But the only source of nuclease contamination is you and your mistakes. Do not use any component which is not certified nuclease free.

You can also use PAGE to purify the RNA. PAGE involves working with more buffers and other components and involves heating. It could be deleterious to subject RNA to such conditions. I

personally prefer using the Zymos RNA concentrator kit for RNA purification. Please note that RNA degradation can not only be caused by the nuclease contaminants but also chemical contaminants.

### **5.3.8 Checking the integrity of the RNA**

- 1) The easiest way to check if you have the right size and fully integral RNA, just use a pre-cast 4% Agarose gel (E-Gel™ EX) (or 2% depending on the length of the RNA).
- 2) Load 4 $\mu$ L of the purified RNA into one of the lanes. Put RNA ladder into the other lanes or any other reference RNA you might have.
- 3) Run the Gel in the E-Gel™ machine. The Gel is pre-mixed with the SYBR Gold for the staining of the nucleic acids, so you will be able to visualize the RNA.
- 4) Visualize the RNA bands post electrophoresis. How does it compare to the ladder? Most RNA will atleast show two bands because this is the native gel and RNA generally exists in a folding equilibrium between two states. Do not be worried about 2 states, they are likely the two states of the same RNA. You can run a denaturing gel to confirm this. But the main point of the native gel was to make sure that your RNA is there and has not been degraded. If the RNA is degraded, then you will likely not see any band or see a band with a big smear.



## 5.4 NHS ESTER LABELING OF NUCLEIC ACIDS WITH FLUORESCENT LABELS

N-Hydroxysuccinimide (NHS) is an organic compound with the formula  $C_4H_5NO_3$ . NHS reacts with carboxylic acids to form NHS-Ester linkage which provide for an excellent and easy chemical linkage platform that are widely used in Biochemistry and Molecular biology for labeling purposes. Please see the reaction-schematic diagram below.

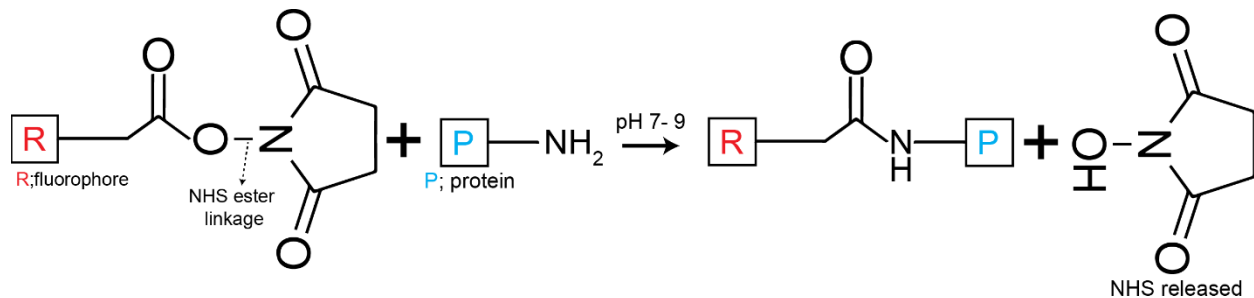
### 5.4.1 Materials

- Anhydrous Dimethyl sulfoxide (DMSO)
- DNA (or RNA) sample with reactive amine group at the desired labeling location.
- Dyes (monofunctional NHS ester form i.e. Dyes with NHS ester linkage as described in the diagram below. R in the diagram below is the Dye.
  - Alexa and other dyes (Invitrogen)
  - Cy3 (GE Healthcare PA13101)
  - Cy5 (GE Healthcare PA15101)
  - Cy5.5 (GE Healthcare PA15601)
  - or can be anything that you want conjugate onto the RNA/DNA.
- Ethanol (cold).
- Labeling Buffer (pH 8.5). Dissolve 384mg of sodium tetraborate (Borax) in 10mL of nuclease free  $dH_2O$  (for a final concentration of 0.1M). Add 65 $\mu$ L of 12.1M HCl or equivalent to bring the pH to 8.5. Prepare fresh before use. Please see below about the exact formula and specs of the chemical required. pH is extremely important. So please double check the pH. We will see below why it is so important.
- 3M Sodium Chloride Solution.

- Nuclease free dH<sub>2</sub>O. Nuclease acids are very sensitive to any nuclease contamination. Please be sure to use ultra-pure chemicals (NaCl, DMSO, dyes) for use. The dH<sub>2</sub>O being used should also be a certified nuclease free dH<sub>2</sub>O.
- Equipments:
  - Centrifuge
  - Eppendorf (or any other) tubes
  - Spectrophotometer

### 5.4.2 Procedure

Reaction-schematic. R here is any dye that you want to conjugate on your DNA/RNA. P is any DNA/RNA that you want to label with the dye of choice (R). P i.e. DNA/RNA needs to have a reactive amine group at the desired labeling location. pH is very important for the efficiency of this reaction, which we will look at more details later.

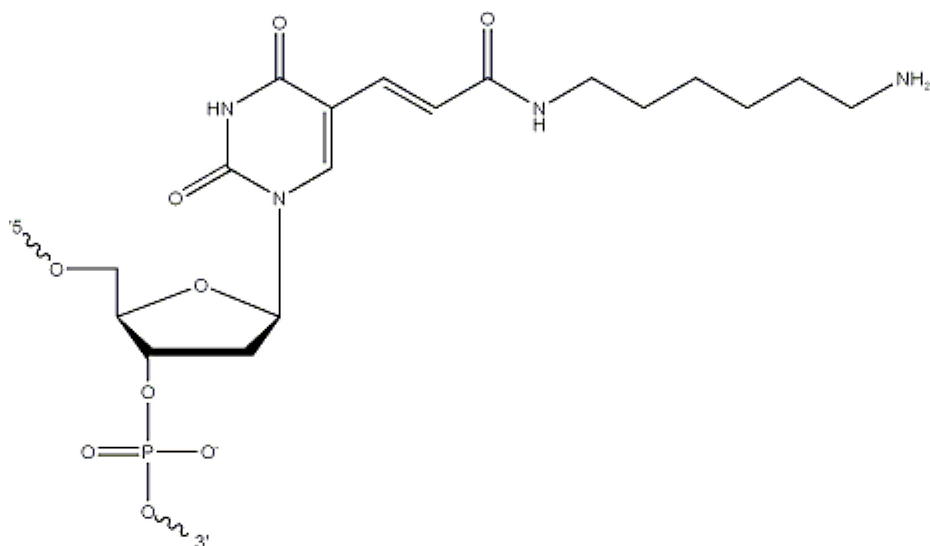


**Figure 5.3 | Schematic describing the reaction between NHS ester reagent (conjugated to fluorophore R) with the -NH<sub>2</sub> group in the nucleic acid or other target of interest.**

For nucleic acids, we typically use an internally modified thymine for labeling. As seen below, you will see that the thymine base is modified to have a reaction amine (NH<sub>2</sub>) group. So the molecule P-NH<sub>2</sub> in the reaction schematic above is the nucleic acid molecule below. More details and references at:

<http://www.valuegene.com/site/Catalog/Modifications/Product/1388>

For Integrated DNA technology (IDT) orders: For 3' and 5' ends of the nucleic acids, a different thymine modification code is used.



**Figure 5.4 | -NH<sub>2</sub> (amine) group present in the modified thymine.**

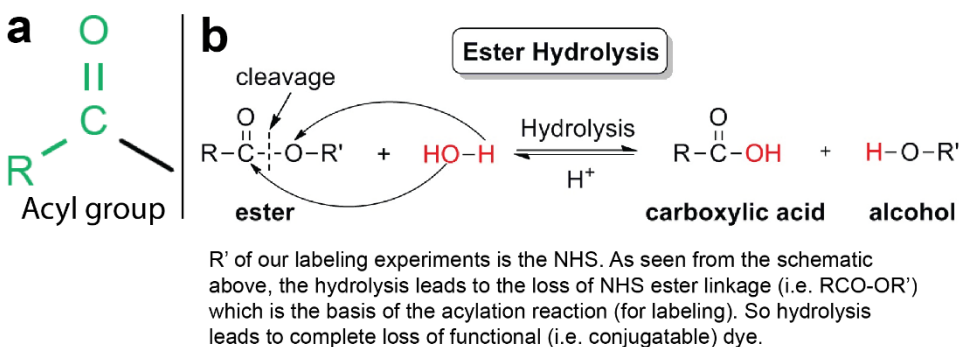
Reference for the figure: <https://www.idtdna.com/site/Catalog/Modifications/Product/1388>

The commercial dyes typically come in tubes containing 1mg of dye in solid form. Add 56 $\mu$ L of DMSO (~20 mM) to the entire tube containing 1mg of dye. You would not need the whole 1mg dye (now in DMSO solution for a single experiment). But it is extremely difficult to take out <1mg of dye out from this tube and weigh it accurately. Hence the best solution is to add 56 $\mu$ L of DMSO to 1mg of dye and then use small aliquots of DMSO dissolved dye for your labeling reaction. As soon as you add DMSO to the dye, take out the required amount for your labeling reaction and store the rest of DMSO dissolved dye solution at -80 °C, you can make aliquots if you like. The dye is not super stable in the solution at the room temperature, so you should quickly put it in -80°C and never keep the dyes dissolved in the DMSO solution at room temperature for too long. Some people take the dye aliquots (dye dissolved in DMSO) and dry them for long term storage. If you dry it out for too long at room temperature, in my opinion this could do more harm than good. The reason is that over-exposure to the atmospheric humidity will lead to the hydrolysis of the ester (dye is conjugated to an ester) and reduce the quality. So you have to take the right call. Dye aliquots in DMSO in liquid form can be stored in -80 °C for long time.

1) In an Eppendorf tube, add 5 $\mu$ L of dye in DMSO.

2) To the above tube, add 27.5 $\mu$ L of freshly prepared labeling buffer (pH 8.5).

pH of this labeling buffer is extremely important, hence double check the pH before adding this to the tube containing dye dissolved in DMSO. A high pH enhances the acylation rate; however, it also increases the hydrolysis of the esters. The addition of acyl group (shown below in green) is called acylation. Our labeling reaction is an acylation reaction i.e. the addition of the acyl group where the R is the dye or any molecule of interest. So high pH enhances the rate of this reaction but it comes at a pretty bad cost i.e. the ester themselves get hydrolyzed quicker at higher pH conditions. NHS esters half-life is few minutes at pH>8.6. See below. So a right balance of pH is an absolute must for the success of this experiment and pH 8.5 is the most optimum pH for the high efficiency of the labeling.



**Figure 5.5 | Hydrolysis of the ester.**

3) Add 2.5/2 $\mu$ L of 1mM DNA or equivalent (~2.5/2nmol total).

4) This composition results in about 40/50 dye molecules for each reactive amine group respectively. If there is not enough DNA available, linearly rescale the amount of the dye and the labeling buffer.

- 5) Some protocols suggest to use 20:1 ratio for the dye to DNA/RNA. I have found that, for such compositions, the labeling efficiency is ~80%. So if you want a near 100% labeling efficiency, I would use the 40:1 or 50 :1 dye to DNA ratio. For single molecule experiments, the amount of DNA/RNA that we require is fairly little. So the better way to achieve a higher Dye to DNA ratio is to reduce the amount of DNA/RNA in the labeling reaction instead of using more dyes. Dyes are pretty expensive and one successful labeling reaction should be enough for many replicates of single molecule experiments.
- 6) Cover the tube with aluminum foil to restrict any stray light exposure for the dye molecules.
- 7) Incubate the mixture for 6 hours at room temperature with gentle mixing. Some protocols would also suggest overnight at 4°C with gentle mixing in the dark. Room temperature is the best, 4°C incubation may result in lower labeling efficiency, even if done for longer hours.

#### **5.4.3 Purifying the labeled nucleic acids from the labeling reaction**

- 1) Add 87.5µL of ethanol and 3.5µL of 3 M NaCl to the mixture and keep it at –80 °C for 30 min. Salt is added to help with the ethanol precipitation of the nucleic acids. Read more here at : [http://physiology.med.cornell.edu/faculty/mason/lab/zumbo/files/ETHANOL\\_PRECIPITATION.pdf](http://physiology.med.cornell.edu/faculty/mason/lab/zumbo/files/ETHANOL_PRECIPITATION.pdf). Some people put the labeling reaction (after addition of ethanol and salt) at 4 degrees for slow precipitation. In my opinion, this could be a bad step and may influence the quality of the dyes for experiments. 30 minutes at -80C is more than enough. Infact, you will see the nucleic acid being precipitated out as soon as you add ethanol, when you spin the reaction in a centrifuge. But you must still subject it to –80 °C treatment for 30 minutes.

- 2) Centrifuge at 14,000g for 30 min at 4 °C. The DNA or RNA will have pelleted at the bottom of the tube. The pellet will be colored.
  
- 3) Remove the supernatant carefully and rinse the pellet with cold ethanol several times very gently. Make sure to not disturb the pellet. Multiple such steps will lead to the wash of all the free dyes from the solution. Some people prefer to use spin columns like P<sub>6</sub> or P<sub>30</sub> to remove the free dyes. In my opinion, those are not needed. The disadvantage of using spin columns to remove free dyes is that they will also lead to a considerable loss of your labeled nucleic acids as well. The procedure described above will guarantee a very high efficiency of the removal of free dyes.
  
- 4) Dry the labeled nucleic acid pellet to evaporate the ethanol.
  
- 5) Dissolve the pellet in an appropriate buffer solution. 20mM Tris HCl (pH 8.0) and 50mM NaCl buffer is perfect for this. Be sure to use nuclease free buffers for all the purposes.
  
- 6) Check the labeling efficiency by comparing the absorption spectra of the nucleic acid (260 nm) and the conjugated dye. Typically, it is close to 100%. If not, run an additional purification such as denaturing polyacrylamide gel electrophoresis to separate the labeled and unlabeled DNA.



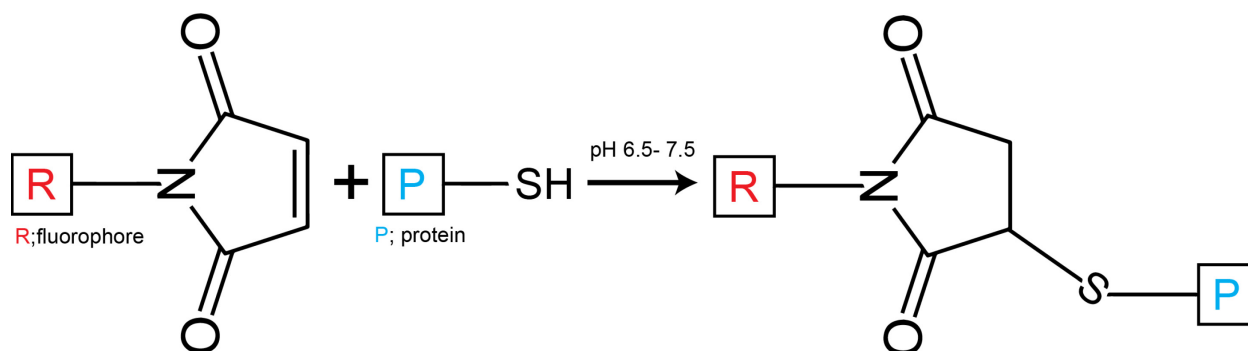
## 5.5 CYSTEINE MALEIMIDE LABELING OF PROTEINS WITH FLUORESCENT LABELS

This protocol is valid for any protein with cysteine in it that needs to be labeled with a molecule of interest containing the Maleimide linkage. The thiols or sulfhydryl present in the Cysteine amino acid residues of the proteins can be used for site-specific labeling. The labeling reaction schematic is described below.

### 5.5.1 Materials

- Protein of interest (at  $< 10\mu\text{M}$  concentration) to be labeled in buffer containing (20mM Tris-Cl pH 7.5, 200mM KCl, 5% glycerol, 1mM TCEP). This is a fairly regular buffer, most proteins that we use for biophysical experiments should be stable and well behaved in this condition. When you are HPLC/FPLC purifying the protein, it would be best to purify it out in this buffer so that protein can be directly used for labeling. Any component like DTT or BME should be totally avoided in the protein buffer. Because they contain -SH group which will compete away all the available Maleimide reagent. TCEP does not have any -thiol (-SH) group so it is fine.
- Anhydrous DMSO.
- Molecule of interest with Maleimide functional group that needs to be conjugated to the protein.

R in the schematic below.



**Figure 5.6 | Schematic describing the reaction between Maleimide reagent (conjugated to fluorophore R) with the -SH group of cysteine in the protein of interest which is to be labeled.**



### 5.5.2 Procedure

- 1) The molecule of interest with Maleimide functional group (R /the dye) would mostly likely be supplied as a dried anhydrous powder. Please keep it away from moisture and humidity as much as possible, as the humidity can hydrolyze the Maleimide functional group.
- 2) Mix the molecule of interest with Maleimide functional group in Anhydrous DMSO so that its concentration is 10mM.
- 3) Typically, the amount of supplied R will be very less i.e. 1 -2mg. It is hard to scoop out and accurately measure small amount of R from such a stock. The best way is to add n  $\mu$ L of anhydrous DMSO to it so that its final concentration = 10mM. And store this stock in -80 °C for long term storage.
- 4) After adding the n  $\mu$ L of the anhydrous DMSO, mix very well. DMSO is an excellent dissolving agent and will dissolve all the R. Be sure that all the R has completely dissolved.
- 5) Now you can use x  $\mu$ L of the R dissolved in anhydrous DMSO for any number of future labeling reactions.
- 6) In an Eppendorf tube (1.5mL), take 24.5  $\mu$ L of the protein of interest (<10 $\mu$ M; preferably 4  $\mu$ M concentration) in the buffer with 20mM Tris-Cl pH 7.5, 200mM KCl, 5% glycerol, 1mM TCEP.
- 7) Add 0.5 $\mu$ L of the R dissolved in anhydrous DMSO (10mM stock).

- 8) This will produce a labeling reaction of the following conditions:
- [Protein of interest to be labeled] = 4 $\mu$ M
  - [R]= 200 $\mu$ M.
  - Buffer conditions = 20mM Tris-Cl pH 7.5, 200mM KCl, 5% glycerol, 1mM TCEP
  - % Anhydrous DMSO in labeling reaction solution = 2%

Anhydrous DMSO in labeling reaction solution should not exceed 5% at any cost, for it can affect the protein. Moreover, it can be deleterious for the downstream applications because if you directly use the protein from this labeling reaction, your downstream experiments will end up having a small amount of DMSO in the reaction solutions.

The Maleimide reaction is most effective at the pH 6.5-7.5. This is on the lower side; most proteins may not be at their 'happiest' in this pH range. The best solution is to do the reaction at the pH 7.5. That is why it was recommended to keep the protein in the buffer which has 20mM Tris-Cl **pH 7.5**, 200mM KCl, 5% glycerol, 1mMTCEP.

- 9) Incubate the reaction in the dark for 2 hours at the room temperature (~23-25 °C). Following this, incubate the reaction at 4 °C (like cold room) in dark for another 8-10 hrs.

- 10) Optional. Reactions can be quenched adding x  $\mu$ L of 0.5M DTT so that its final concentration in the reaction solution = 10mM. DTT also has -SH group so such a high concentration of DTT will compete away all the Maleimide reagent away from the -SH site in the protein of interest.

I would not add DTT for quenching. Infact, I don't think quenching is particularly required. You can let the reaction solution as it is and directly use for your experiments. DTT causes severe problems for single

molecule fluorescence experiments, hence I will totally avoid it.

Any component like DTT or BME should be totally avoided in the reaction mixture. Because they contain -SH group and will compete away all the available Maleimide reagent. TCEP does not have any -thiol (-SH) group so it is fine.

### **5.5.3 Purifying the labeled protein from free R or dyes**

The purification depends on the experiments you want to do. In majority of experiments, the protein will be immobilized on the surface and the free protein (along with excess dyes) will be washed out. So I do not need to worry about free dyes interfering as a background etc. But if you must purify out the labeled protein from the free R (or dyes). You can do HPLC/FPLC and even easier solution would be to use a dialysis membrane. The protein of interest will remain inside the membrane while the free dyes/R will diffuse out the membrane into the large reservoir. This can also be a good way, should you need to exchange your protein into a different buffer. But all these purification procedures require that you have atleast 1mL of the reaction solution for small amounts are hard to run through these procedures. So you can prepare labeling reaction at a much higher volume or dilute the sample for purification.

## 5.6 PEGYLATION PROTOCOL

### 5.6.1 Introduction

For TIR or any microscopy, clean and a fully passivated surface i.e. passive to any non-specific binding so that only specifically bound molecules (labeled with fluorophores) can be visualized. Polyethylene Glycol (PEG) is a polymer that has been well studied and established to be a powerful repellent for any non-specific binding (especially of proteins) and nucleic acids to some extent. The two most commonly attributed reasons are:

Physical: PEG polymer create a globular mesh around itself thus masking the surface it is attached to it and preventing any non-specific binding. It itself is a hydrophobic polymer and the biomolecules in the aqueous solution largely have hydrophilic surfaces exposed (for better interaction with the water). This hydrophilic-hydrophobic anti-match is one of the reasons why proteins do not interact with PEG directly that well. Chemical, which have been previously described in details<sup>141-144</sup>.

But achieving a high quality and clean PEG surface is not trivial. In this protocol, I will go over the detailed protocol and go over some key points which are critical for achieving a high quality PEG and clean PEG surface. The process of chemically attaching PEG polymer onto a surface is called PEGylation.

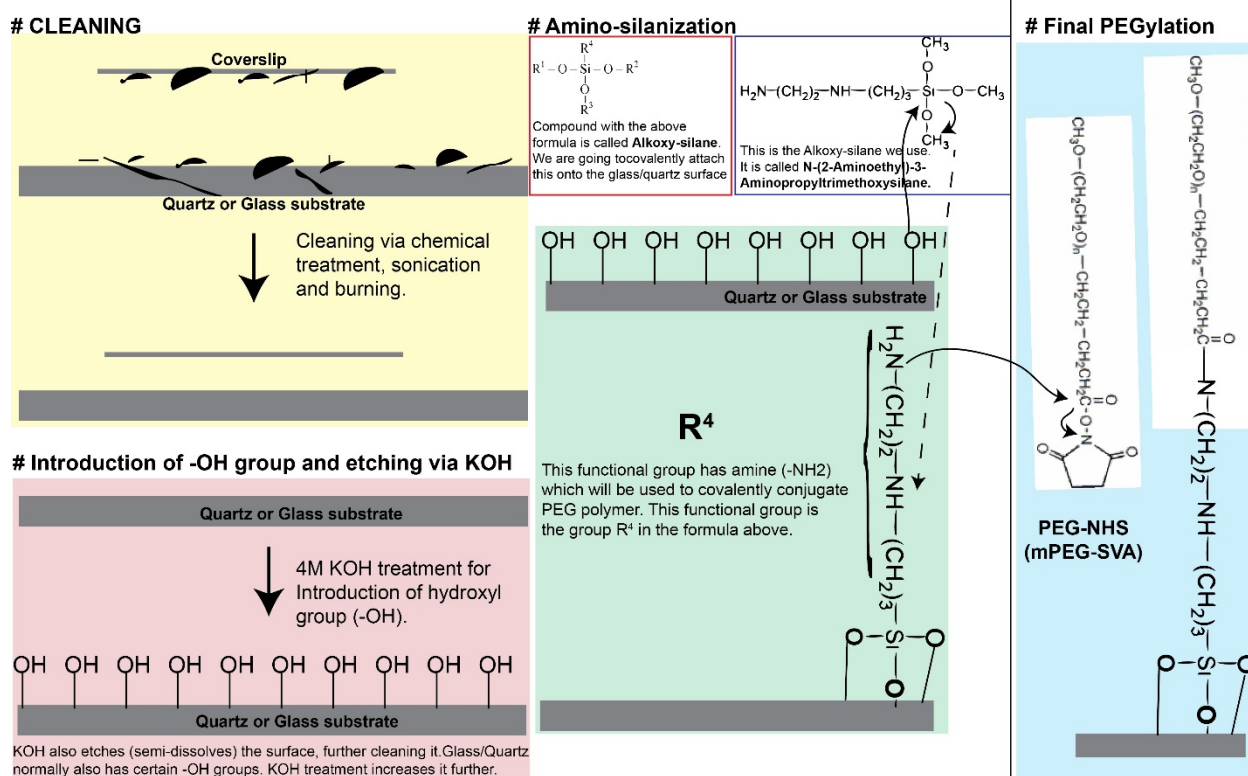
### 5.6.2 Materials

- mPEG SVA : This is the PEG polymer that will be grafted onto the slides and coverslips.
- Biotin PEG SVA : The PEG polymer above will be doped with this PEG which has a Biotin on its end. The Biotin-Neutravidin interaction is used for the specific immobilization of desired molecules on the surface.

- Coverslips: Preferably 24\*40 mm size. 24\*50mm will be good as well. Longer ones may have trouble fitting well into the holder in the microscope. Coverslips cannot be reused. Always take out new ones for new batch of PEGylation.
- Quartz slides: Quartz slides are better than glass. Because they have less 'impurities' and thus cause slightly lower background levels in images. But Quartz is more expensive than glass. 22\$ per piece. You should reuse the Quartz slides.
- Acetone. Purer the better.
- Methanol. Purer the better.
- Sodium Bicarbonate.
- dH2O.
- Amino silane.
- 4M KOH solution. Purer the better.
- Propane flame for burning the slides and the coverslips.

### 5.6.3 Procedure

The schematic shown below is the full chemical schematic of the PEG protocol. It is broadly classified into 4 steps which are Cleaning, Introduction of -OH group and etching via KOH, Amino-silanization, and Final PEGylation



**Figure 5.7 | Complete chemical schematic of the PEGylation protocol**

### 5.6.3.1 Cleaning

Please do not make more than 5 slides at once. We are tempted to make >5 slides in one go, thinking that they will be used for many long term experiments. But managing >5 slides can be difficult and takes time and bring the quality of slides down. You must be wearing and regularly changing the gloves all the time during this entire process.

- 1) Cleaning with Acetone dipped lens tissue paper. Take the fresh lens paper wipes and put acetone in it. Clean each of the slides with this acetone dipped lens tissue wipes thoroughly by rubbing the slides with these wipes. Use the razor to scrap out any unwanted dirt or epoxy or any other in the slide. Following which, clean them with acetone dipped lens tissue paper once again. You can use

KimWipes instead of lens tissue paper, but they will introduce scratches in the slide.

- 2) Place the slides in a glass staining jar.
- 3) Now take out 6 cover slips. Preferred size 24 x 40 mm or 24 x 50 mm. If you use longer coverslips, it may prevent the slide-coverslip chimera to fit into the microscope holders. So double check. Put these cover slips into the glass staining jar.
- 4) Cleaning with dH<sub>2</sub>O: Rinse the slides and coverslips with MilliQ H<sub>2</sub>O. Repeat it 3 times.
- 5) Now sonicate the slides and coverslips with MilliQ H<sub>2</sub>O for 20 min to remove the remaining dirt. Dispose the water and rinse the slides and coverslips 3 times again with MilliQ H<sub>2</sub>O.
- 6) Optional. At this stage, sometime I prefer to let the slides and coverslips soak in water overnight and begin the subsequent steps the next day.
- 7) Cleaning with acetone. Replace the MilliQ dH<sub>2</sub>O with acetone in the jar containing slides and coverslips. Sonicate them with acetone for 20-30 min.
- 8) Cleaning with methanol. Replace the acetone with methanol in the jar containing slides and coverslips. Sonicate them with methanol for 20-30 min.
- 9) Rinsing with dH<sub>2</sub>O. Discard the methanol and rinse the slides with MilliQ dH<sub>2</sub>O. Repeat it for 2 times in order to remove any methanol residue. The acetone and methanol must not be discarded into the sink. In every lab, there must be separate waste bottles for the acetone and methanol waste. Use that.

Why does sonication help in cleaning? Sonication causes lot of agitation. This agitation forces the contaminants to break free from the surface it is attached to. This agitation also forces the water molecules through narrow openings and holes thus efficiently cleaning them.

Why methanol and acetone are used? These solvents help in dissolving out the contaminants and impurities from the slides and coverslips which might not have dissolved well with the water.

- 10) Burning. Start the propane flame and use tweezers etc. to burn both sides of the slides. Burn each surface completely for about 1-2 minutes. Burn well near the holes in the slides. The burning will further get rid of contaminants.

Burning coverslips will be far more difficult, they are extremely thin and brittle. Over burning will quickly break them. So be very careful in limiting the duration of the flame over the coverslips. Please do not exceed 1-2 seconds and burn the coverslips in 3 repeats, each lasting 1-2 seconds. Burn both the sides of the coverslips.

### 5.6.3.2 Introduction of –OH group and etching via KOH

- 1) Put the slides and the coverslips in a glass staining jar and add 4M KOH to it.
- 2) Sonicate for 20-30 minutes. KOH etches the surface of the slides and coverslips. This etching process produces large number of hydroxyl (-OH) group on the slides and coverslips as shown in Figure 5.7. It is extremely important that the slides and coverslips be not subjected to any more burning or cleaning process after KOH etching. The –OH groups can be destroyed by such processes.
- 3) Rinse the slides and coverslips with dH<sub>2</sub>O for 2 times to remove traces of KOH. Any trace of KOH will interfere with downstream PEGylation steps, for e.g. Aminosilanization.



- 4) Rinse the slides and coverslips with methanol for 2 times to remove traces of dH<sub>2</sub>O. Any trace of dH<sub>2</sub>O will interfere with downstream PEGylation steps, for e.g. Aminosilanization.
- 5) Let the slides and coverslips be immersed in methanol.

### 5.6.3.3 Aminosilanization

The process of covalent addition of a silane with an amine (-NH<sub>2</sub>) is called Aminosilanization. This reaction for slides and coverslips is shown in Figure 5.7.

- 1) Take out an ultra-clean and dry 500 mL flask. It must be devoid of any traces of dH<sub>2</sub>O.
- 2) Add methanol solution to it and just sonicate the flask with methanol in it for 20 minutes. Discard the methanol
- 3) Add 200 mL of 100% anhydrous methanol to it. 200 mL is the right amount for 5 slides. For more or less number of slides, you have to scale accordingly.
- 4) Add 10 ml of Glacial Acetic acid. Please do not use plastic pipette tips for transferring Glacial Acetic acid as it reacts with plastic creating unwanted products. Use a glass pipette (for e.g. Pasteur pipettes) for this.
- 5) Add 6 mL of commercial Aminosilane solution.

- 6) Mix the above solutions well. Avoid any water in the solution. The Aminosilanization reaction solution is now ready.
- 7) Remove methanol from the glass jar containing the slides and coverslips.
- 8) Now add Aminosilanization reaction solution into the glass staining jar containing slides and the coverslips.
- 9) Incubate the jar with Aminosilanization reaction solution for 30-35 minutes. No sonication or ultra-mixing is recommended at this stage.
- 1) Wash the slides and coverslips with copious amounts of MilliQ dH<sub>2</sub>O. A common mistake people at this stage is that they do not clean the slides/coverslips well enough at this step. Any amount of residual Aminosilane left in the solution or on the slides/coverslips will be a chemical target for the PEG polymer you are trying to conjugate. But you want them conjugated only on the silane groups on the slides/coverslips and not on the free flowing ones. So please clean the slides/coverslips with copious amounts of water.

#### **5.6.3.4 Final PEGylation**

- 1) Take 84 mg of Sodium Bicarbonate and mix it in 10 mL MilliQ dH<sub>2</sub>O to get 0.1 M Sodium Bicarbonate solution.
- 2) Take 5 (slides) × 70 μL = 350 μL of the above solution for making the final PEG reaction solution.
- 3) Add ~160-200 mg of mPEG-SVA to 350 μL of the solution above. Please note that different versions of mPEG are commercially available and do not use the incorrect mPEG and use mPEG-

SVA. The incorrect mPEG may not have the functional group necessary for the required reaction so everything will be a waste. Please double check to make sure it is mPEG-SVA.

- 4) Dope the above solution with ~2-5 mg of Biotin-PEG-SVA. Please double check to make sure that this is the right Biotin-PEG polymer.
- 5) Mix the above very well and centrifuge them in 4,000g for 2 minutes.
- 6) Put the slides in chamber with tons of water beneath/around it to make sure that things stay hydrated. PEGylation efficiency is poor in dry and non-humid environment. An easy way of making a hydrated chamber is to use the old pipette tip boxes. Fill the tip box with MilliQ dH<sub>2</sub>O and add tips to it. And position the slides onto the tips.
- 7) Now add 70  $\mu$ L of the PEG reaction solution to each quartz slide and put a single coverslip on each one of them.

Close the lid of chamber and place them in the dark for incubation. Please check from time to time to make sure that the slides and coverslips are perfectly horizontal. Even if they are slightly titled, the PEG reaction solution will slowly drip out, rendering the entire region dry without any PEG for PEGylation of the surface. This is a very common mistake, please avoid it. The PEG slides may also get stuck in the top/roof of the tip box so please keep a check to make sure that the slides and coverslips are oriented in the right way in the tip boxes for the PEGylation reaction.

- 8) The PEG reaction is almost over in 4 hrs. So please do incubate the slides for more than 4-4.5 hours. Beyond this time, the PEG coating on your surface is largely degrading. In some ways,

over-incubation is equivalent to putting PEGylated slides on the room temperature for n hrs, n being the number of hours after 4-4.5 hours incubation.

- 9) Clean the slides with copious amount of water. Under cleaning the slides is another common mistake. Please think of it this way, this is the last chance you will get to clean the slides before your experiments so clean the slides with copious amount of MilliQ dH<sub>2</sub>O, otherwise your surface may substantial auto fluorescent junk which will interfere with your experiments.
  
- 10) Dry the slides and coverslips with Nitrogen and Mark the Un-PEGylated sides of the coverslips and slides with 'N'.
  
- 11) Put each pair of coverslip and slide into a falcon tube with their PEGylated sides facing outward.
  
- 12) Put each falcon tube, with partially closed caps, into a food saver bag and vacuum seal them.
  
- 13) Put them into -80 °C or -20 °C for long term storage.

## **5.7 A VERSATILE PROTOCOL FOR SINGLE MOLECULE FRET EXPERIMENT USING CRISPR ENZYMES**

### **5.7.1 Introduction**

The protocol presented below has been written for smFRET experiments to investigate DNA interrogation by Cas9-RNA. But the protocol is broadly applicable for other CRISPR enzymes, DNA targets and experiments that investigate DNA unwinding etc. The goal of the protocol is to inform critical aspects of performing single molecule fluorescent experiments to investigate CRISPR toolbox. We use an in-house built software called smCamera for movie acquisition and data analysis. The software is available for download at <https://cplc.illinois.edu/software/>

### **5.7.2 Materials**

- PEGylated quartz slides and coverslips. Please make sure they have been made to the highest possible quality.
- All the components involved in these experiments must be certified RNase/DNase free components.
- Basic salts needed for making various buffers. Double check that you have all of them. Prior to starting any CRISPR experiments, remake all your buffers using certified RNase/DNase free components. Store them in an ultra-clean space to avoid contamination. These experiments use RNA, which are very sensitive to degradation, therefore experiments can suffer a lot if you do not keep everything super clean and free of nuclease contamination.
- DNA or CRISPR enzyme or RNA substrates labeled with fluorophores for smfluorescence or smFRET and biotin for surface immobilization.
- CRISPR protein. These must be the highest quality and should be in the aliquots stored at -80 °C.

- Guide-RNA. This will be a chimera of tracrRNA and crRNA (Cy5; acceptor labeled) pre-mixed in aliquots stored at -80 °C.
- Cas9-RNA imaging buffer recipe. Please note that this buffer is without gloxy, we will add gloxy right before the imaging.
  - 2 × 20μL of 1M Tris-HCl, pH 8.0 (20mM Tris-HCl, final concentration)
  - 2 × 33.3μL of 3M KCl (100mM KCl)
  - 2 × 10μL of 0.5M MgCl<sub>2</sub> (5mM MgCl<sub>2</sub>)
  - 2 × 50μL 100% Glycerol (5% v/v glycerol)
  - 2 × 10μL NEB Bovine Serum Albumin (BSA) stock (0.2mg/mL BSA)
  - 2 × 875μL of Saturated Trolox+ glucose solution (recipe below)
  - Total = 2 tubes of 1000μL each. 1-2 tubes will be required per set of experiment.

I like to use a single buffer for all of the steps of the experiments including Neutravidin immobilization, DNA immobilization, adding Cas9-RNA etc. I do all of this in a single buffer i.e. this Cas9-RNA imaging buffer. Hence we have made 2mL of it, which should be enough for a single set of experiments (i.e. 5-6 channels). I do not like to juggle around between different kinds of buffers while doing experiments. Moreover, it is best to treat everything i.e. the surface, DNA only etc. with a single buffer to avoid any artifact of other buffers.

- Saturated Trolox+ glucose solution recipe. To 10mL of dH<sub>2</sub>O add the following:
  - 15mg of Trolox powder.
  - 80mg of Glucose powder
  - 10μL of 5M NaOH
  - Mix at room temperature in a gentle shaker. DO NOT cover it in aluminum foil, this is a common mistake people make. The Trolox requires initial light activation when it is being mixed in the water. After ~8-10hrs of mixing, then be sure to cover it in the aluminum foil

and keep it in 4 °C or -20 °C. Lasts about 7 days in 4 °C and much longer in -20 °C.

### 5.7.3 Procedure

Glove Policy. It is also important to keep changing gloves. While working, gloves will accrue all sorts of bad contaminants which can then end up in your samples. So keep changing glove every few minutes while working. Extra gloves are worth nothing as compared to the quality of good experiments, your time and efforts, so please do not mind using extra gloves.

### 5.7.4 Preparing the chambers for smFRET experiments

- 1) Assemble the PEG slides as per the conventional protocol described previously<sup>89</sup>. No touching the slides with bare hands i.e. without gloves. Hands have tons of nucleases (commonly referred to as fingerases) and they can degrade your sample so please avoid that. None of what you do should be without gloves. Please check out the glove policy.
- 2) Do 5-6 chamber worth of experiments in a single microscope use. 5-6 is the most optimum number, trying to do more is bad because the latter experiments can lose quality. Doing only 1-2 chamber is not worth the effort of setting up the whole experiment. Moreover, I do not want to refreeze the slides after taking them out to room temperature. A single slide has 5-6 channels/chamber so the idea is to use a single slide for 5-6 set of experiments (1 experiment per channel).
- 3) After the slide has been assembled, add 30µL of Cas9-RNA imaging buffer to each channel. The BSA in this buffer will further passivate the surface against any residual non-specific binding possibilities. Incubate for ~3-5 minutes. This is an extremely important step. BSA must be present in all your buffers at the specified concentration (see recipes in the materials section).

- 4) Prepare a Neutravidin solution by mixing 1  $\mu$ L of Neutravidin stock (200 nM stock will suffice) with 100  $\mu$ L of Cas9-RNA imaging buffer. Keep the Neutravidin back to the public stock box.
- 5) Now add 20  $\mu$ L of this diluted Neutravidin solution to each channel/chamber and incubate for 2 minutes.
- 6) Wash out the unbound Neutravidin with 40  $\mu$ L of Cas9-RNA imaging buffer for each channel.
- 7) The slides are now ready for immobilization of biomolecules of interest.
- 8) Put the slide in the microscope and arrange all the required optical components to achieve TIR condition.

### **5.7.5 TIR Condition and TIR Spot**

Applicable if you are using PRISM based TIR microscope.

- 1) Adjust the TIR (Total internal reflection) spot. The TIR spot must be a clean, regular, elliptical shaped blob in the middle. The TIR spot must be just the right size i.e. not too small and not too big. Too small would mean that the illuminating area will not fill the imaging area. Too big would mean that the intensity of the single molecule spots will be very low and would require high laser power and your TIR illumination will not be sharp. The TIR spots must be right size and right shape. Take your time in setting up the TIR spot, please do not rush.
- 2) Once the right TIR spot has been achieved, move the slide around to make sure that the entire assembly is working fine and that slide can be moved upwards and sideways without losing the TIR spot.



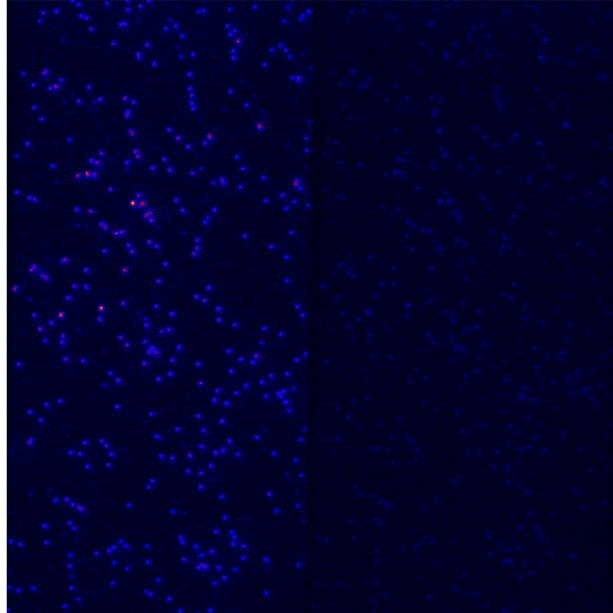
### 5.7.6 Immobilizing DNA target molecules on the surface

Before putting the DNA molecules on the surface, decide on the identity of DNA target that you will be testing. As a good practice, it is important to put some DNA targets that we know will show high extent of Cas9-RNA binding (i.e. Cognate DNA target) and no stable binding (ones without PAM) The success of experiment, seen via stable high FRET for cognate DNA and no stable FRET for PAMless DNA, will be a good confirmation that the Cas9-RNA complex you have prepared is a high quality and active complex and that smFRET signal is due to specific binding. So for 5-6 channels, I would use the following set of DNA targets. Please refer to the chapter 2 for the naming of the DNA targets as shown below.

- Channel 1: Cognate DNA target (0-0<sub>mm</sub>)
- Channel 2: 13-20<sub>mm</sub> DNA target.
- Channel 3: 8-20<sub>mm</sub> DNA target.
- Channel 4: 5-20<sub>mm</sub> DNA target.
- Channel 5: 1-20<sub>mm</sub> DNA target.
- Channel 6: 1-20<sub>mm</sub>NOPAM DNA target.

- 1) To a 100 $\mu$ L Cas9-RNA imaging buffer solution. Add 1  $\mu$ L of the Gloxy and then add 0.2 $\mu$ L of the DNA from the 5nM stock. Add 20 $\mu$ L of this on a particular channel and immediately visualize the density on the surface. The density has to be just right, too less is bad and too much is even worse as it can result in multiple molecules for many fluorescent spots.
- 2) The number of spots will continue to increase for up to 10 minutes, so if you have achieved optimal density per imaging area kind of density then, immediately flow out the remaining unbound DNA from the channel with 30 $\mu$ L of plain Cas9-RNA imaging buffer.

- 3) The above two steps are extremely critical, so please do them very well. Here's how you single molecule spots for a given imaging area should look like in our setup. The optimal density will be different for different setups, please determine this accordingly.



**Figure 5.8| Representative image of Cognate DNA target in Cas9-RNA imaging buffer.**

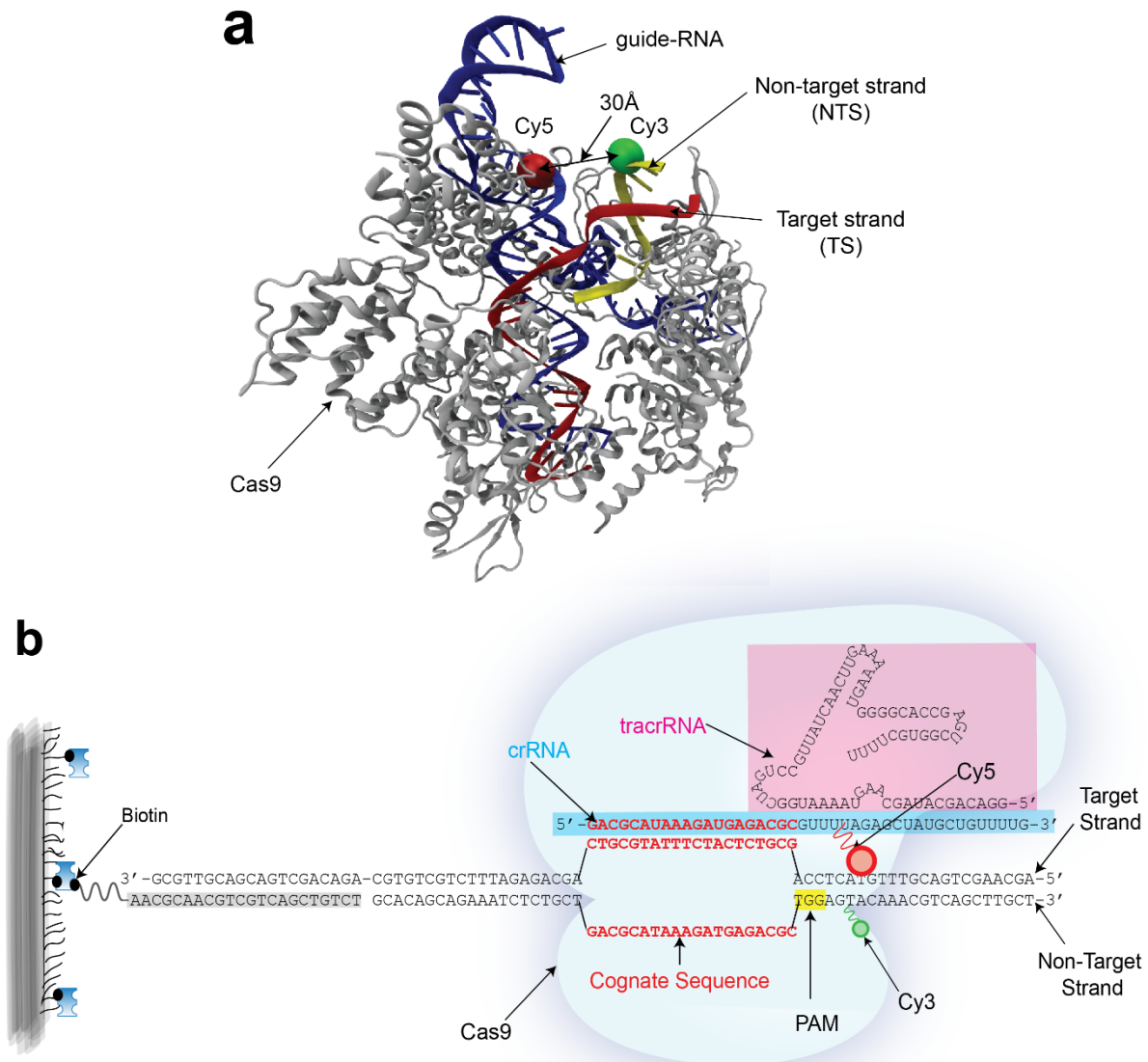
The spots on the donor channel (left channel) indicate Cy3 labeled DNA target molecules. Since there is no acceptor Cas9-RNA in solution, there are no 'FRETing' molecules on the acceptor channel (on the right).

- 4) You can now immobilize DNA targets of different sequence on all 1 or 2 more channels. In my opinion, it is best to get ready with DNA targets in 2-3 channels before adding Cas9-RNA and measuring binding. It also helps with the fact that you can now put all the DNA stocks back in the  $-80^{\circ}\text{C}$  and clean up your workspace/bench and focus totally on the next part. The idea is to do things step by step. i.e. whenever you are immobilizing DNA targets, do not even bring out Cas9 or RNA. Do not even think about it. Just focus on getting the right immobilization for all the DNA molecules.

- 5) Now when all the DNA targets are immobilized properly on the channels. Take a very small break, drink some water, catch some air and regain your focus. Now is the time to dwell into the main part.

### **5.7.7 Preparation of Guide-RNA**

The guide-RNA can be stored as aliquots of hybridized-chimera consisting of crRNA and TracrRNA in T50 buffer (50 mM NaCl, 10mM Tris-HCl pH 8). Please do not store any RNA in a buffer that has Magnesium ions. Magnesium ions are required for nuclease activity, keeping the RNA storage buffer devoid of magnesium provides an extra insurance against the possible nuclease contamination activity. The crRNA is labeled with Cy5 for smFRET imaging. TracrRNA is unlabeled. The aliquot may contain 2.5 $\mu$ L of 4 $\mu$ M guide RNA (i.e. 4 $\mu$ M of Cy5 labeled crRNA+ 5 $\mu$ M TracrRNA) in T50 buffer (10mM Tris-HCl, pH 8.0 and 50mM NaCl). The guide-RNA is already annealed, so you can use it directly. This chimera of TracrRNA and crRNA is known as guide-RNA. RNA in Cas9-RNA refers to this guide-RNA. The crRNA;TracrRNA hybridized chimera is shown below. Also shown is the base pairing region between guide-RNA (crRNA;TracrRNA is referred to as guide-RNA, which then is referred to as RNA when saying Cas9-RNA for brevity).



**Figure 5.9 | FRET probe labeling locations in the Cas9-RNA-DNA complex.**

**(a)** Cy3 and Cy5 labeling locations shown in the crystal structure of Cas9-RNA bound to a cognate DNA target (PDB ID: 4UN3)<sup>65</sup>. The strand hybridized with the guide RNA to form the RNA-DNA heteroduplex is referred to as the target strand while the other strand, containing the PAM (5'-NGG-3'), is the non-target strand. **(b)** Schematic of a bound Cas9-RNA-DNA complex showing the base pairing between different components. The sequences shown in red denote the cognate sequence of the DNA target and the complementary guide sequence of the crRNA. The DNA sequence highlighted in light gray

is a separate 22 nucleotide-long biotinylated adaptor used for surface immobilization of DNA target molecules.

### 5.7.8 Preparation of Cas9-RNA

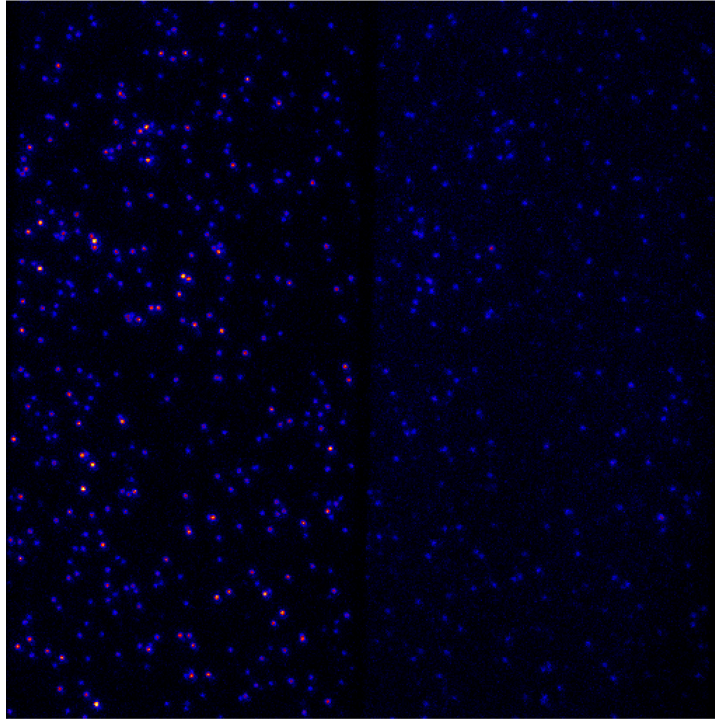
- 2) The Cas9 protein must be stored in aliquots at  $-80^{\circ}\text{C}$ . Preferably 2-3  $\mu\text{L}$  of 4  $\mu\text{M}$  concentration in Cas9 storage buffer (20mM Tris-HCl, 100mM KCl, 5mM  $\text{MgCl}_2$ , 5 % v/v glycerol).
- 3) Mix the 1 $\mu\text{L}$  of the annealed guide-RNA (4  $\mu\text{M}$  stock) from with 2-3 $\mu\text{L}$  of Cas9 (4  $\mu\text{M}$  stock)  
Total volume 4 $\mu\text{L}$ .
- 4) Incubate them on ice for atleast 10-15 minutes. Cas9-RNA complex is made with 2:1 or 3:1 ratio between Cas9 and guide-RNA.  $K_D$  of Cas9 and guide-RNA is extremely low at about 10  $\text{pM}$ <sup>145</sup>. But excess of Cas9 is added to ensure near 100% complexation of acceptor labeled guide-RNA with Cas9.
- 5) After incubation, Cas9-RNA complex labeled with Cy5 is now ready for use in the smFRET experiment.
- 6) As reminder note of caution, I must warn that all the tubes, pipette tips, pipette and almost anything you are using to do this experiment must be certified RNase/DNase free. I thoroughly clean my pipettes with 70% Ethanol from time to time to make sure that they are not a breeding ground for contamination.
- 7) The total volume of Cas9-RNA complex at this point is 4 $\mu\text{L}$ . Add 196 $\mu\text{L}$  of Cas9-RNA imaging buffer to this Cas9-RNA complex, this will bring the total volume of Cas9-RNA complex to

be 100 $\mu$ L with the concentration of Cas9-RNA at 20nM which is what we will use in the experiment. This Cas9-RNA is now ready to be added to the channels with DNA targets immobilized on them, but do not yet add Gloxy on them. Gloxy leads to drop in the drop in the pH problem, so Gloxy is to be added right before the actual imaging.

- 8) Take 25 $\mu$ L of 20nM Cas9-RNA complex (from the 200 $\mu$ L stock prepared above) and add 0.3 $\mu$ L of relatively fresh stock of Gloxy. This is now ready to be added to a given channel. The remaining Cas9-RNA of the initial 200 $\mu$ L stock should still be sitting in the ice.
- 9) Add the above ~25 $\mu$ L of 50nM of Cas9-RNA (labeled with Cy5) to a particular channel. Wait for ~5-10 minutes for the reaction to reach equilibrium. Now you can begin single molecule imaging. Each channel must be imaged separately i.e. you add 25 $\mu$ L of 20nM of Cas9-RNA (supplemented with Gloxy, right before imaging) to a channel. Wait 10 minutes and then acquire all the movies for that channel and then proceed to the next channel, which till now should be sitting with just the DNA without any Cas9-RNA.
- 10) As a reminder, Gloxy should be added to the mini 25 $\mu$ L sample as it is just about to be added to the channel for imaging. Do not add any Gloxy to original 200 $\mu$ L 50nM Cas9-RNA stock which should now be sitting in the ice.

### **5.7.9 Single Molecule Imaging/Data Acquisition**

- 1) After waiting for about 10 minutes as mentioned before, begin the image acquisition. This is what your donor (left) and acceptor (right) channels should now look like:



**Figure 5.10 | Representative image of Cognate DNA target (Cy3) incubated with 20nM Cas9-RNA complex (Cy5). The Spots on the right indicate FRETing molecules which indicate the interaction between DNA (with Cy3) and Cas9-RNA (with Cy5) bound to the DNA target.**

2) Please check the following:

- Density of the spots.
- Intensity of the spots. Intensity is about 10% less than what I think would be ideal, so you might want to use slightly higher laser power. But these parameters are highly dependent on your instrument, the gain (I always work at near maximum gain) of your camera. So you should use laser power that gives good signal to noise ratio and enables fairly long duration of detection prior to photobleaching.
- Constant and uniform illumination of the entire imaging area, if the TIRF spot is not ideal, you will not get this and such experiment will be a waste. So please make sure that you have the right TIRF spot.

- Check the Gain value. I like to max it so that I get the highest possible signal for a given laser power. Allows you to get good intensity while still not having to use very high laser power.
- 3) Make 15-20 movies each imaging channel (i.e. each DNA target) each lasting about 20-30 frames. This is for making the FRET histograms i.e. a distribution of FRET value of the first few frames of each of the detected single molecule.
  - 4) Get an absolute tight focus for each movie, people tend to make multiple movies by just moving different areas without re-adjusting the focus. But whenever you move, you lose the focus a slightest bit and the data will then be bad with single molecule spots being slightly de-focused. So for each of the 15-20 movies, move to a new imaging area re-adjust the focus till you are convinced that is the tightest possible focus and then image. It may look cumbersome but you have done so much effort to get to this point, it would be a terrible waste of efforts, if you acquired sub-standard images, just because you do not want to re-focus.
  - 5) Now, another important part is the long duration movies. You have to make about 3 movies each lasting atleast 3000-4000 frames (4000 frames, 100 ms/frame == 400 seconds). Longer than 4000 frames would be even better, but it is quite likely that the fluorophores would photobleach before it.
  - 6) Before you start acquiring the long-duration movies. Readjust the focus, get the tightest possible focus and now start the recording. Do not move the optical table or anything in the room while you are recording the long -duration movies. Now when you are making movies that long, the biggest problem you will face is the de-focusing problem. Over the course of the movie acquisition for the long -duration movie, the focus will shift and your spots will not be sharp anymore. If you keep moving things around the room or in the optical table, your movies will

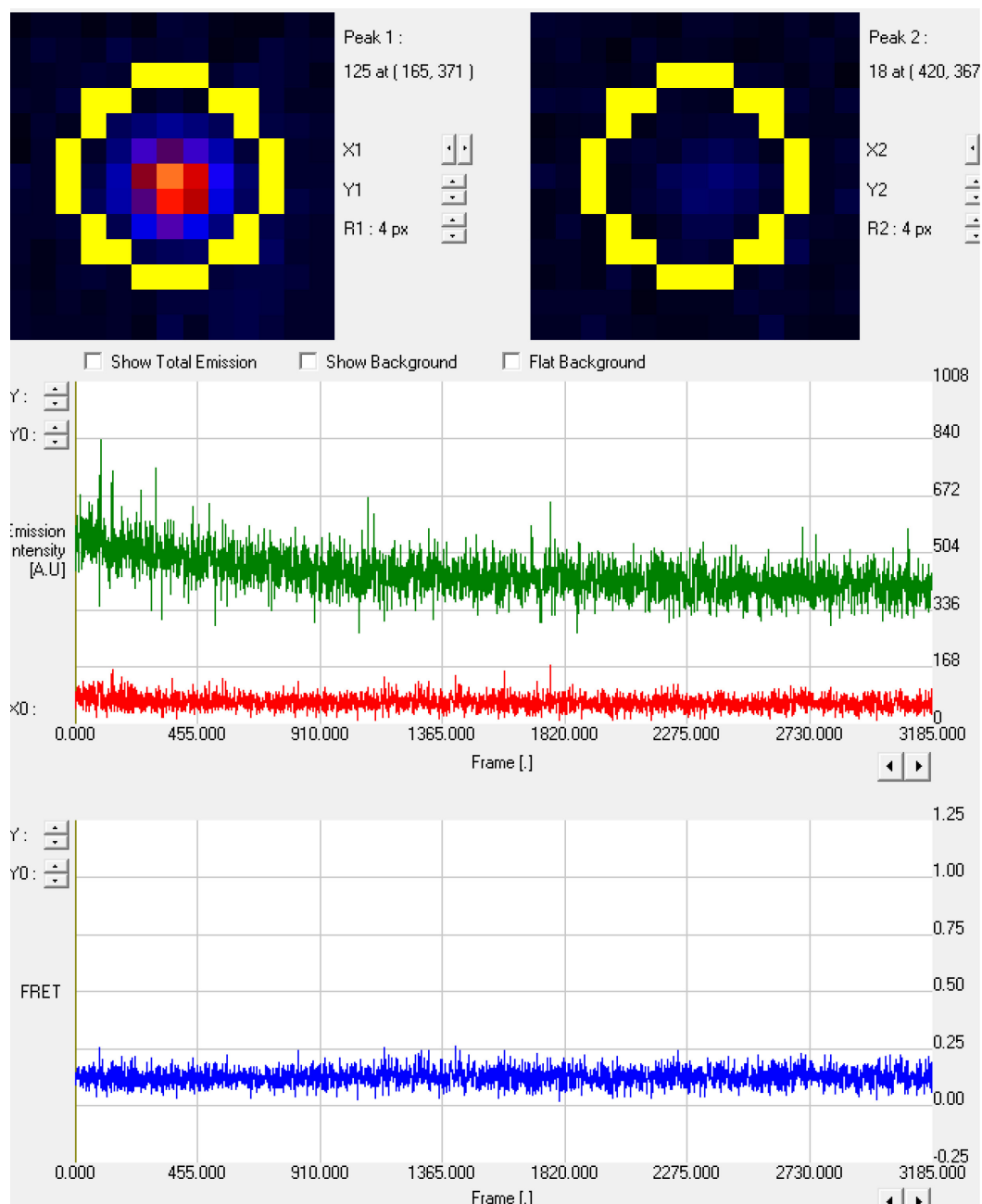


lose the focus and that movie will be totally useless.

- 7) It is important to understand why even slight defocusing in the movies is totally useless. The biggest power of single molecule experiment is to detect transient and minute things. If you are interested in broader interaction, you might as well run a gel. To capture the minutest/transient interaction/event, your movies have to be extremely sharp, otherwise the real minute/transient event will be buried deep within the noise of the single molecule trajectories. So be very careful while making long duration movies.
  
- 8) The more obvious reason why you need good quality long -duration movies is that these movies will be used for different kinds of kinetic analysis for Cas9-RNA binding/disassociation etc. So those movies are really the backbone/end-goal of these experiments.

#### **5.7.10 Double checking the consistency of the focus in the long-duration movies**

I have emphasized the importance of the long duration movies a lot. While you are doing the experiment, it is important to know whether a long movie you are acquiring is good or not. A simple check is the check of the eye, you will see that the spots will lose their sharpness and shape if the focus is being gradually lost. But this is not always obvious, infact it is quite difficult to see just by the naked eye. To resolve this problem, just quickly analyze the movie using smCamera software (in-house software, described in the introduction) and check the smFRET time trajectory in the smCamera software itself. If the traces look like this i.e. slight gradual delay of the signal, then it is a movie with defocusing problem and you have to remake the same movie again in a different imaging area ofcourse.



**Figure 5.11 | A representative smFRET time-trajectory from a long duration movie**

**(Top Panel)** Representative single molecule fluorescent spot in donor (left) and acceptor (right) channel.

**(Middle Panel)** The green and red curve as a function of time denotes the donor and acceptor channel intensity of the spot respectively. **(Bottom Panel)** The blue curve as a function of time is the FRET

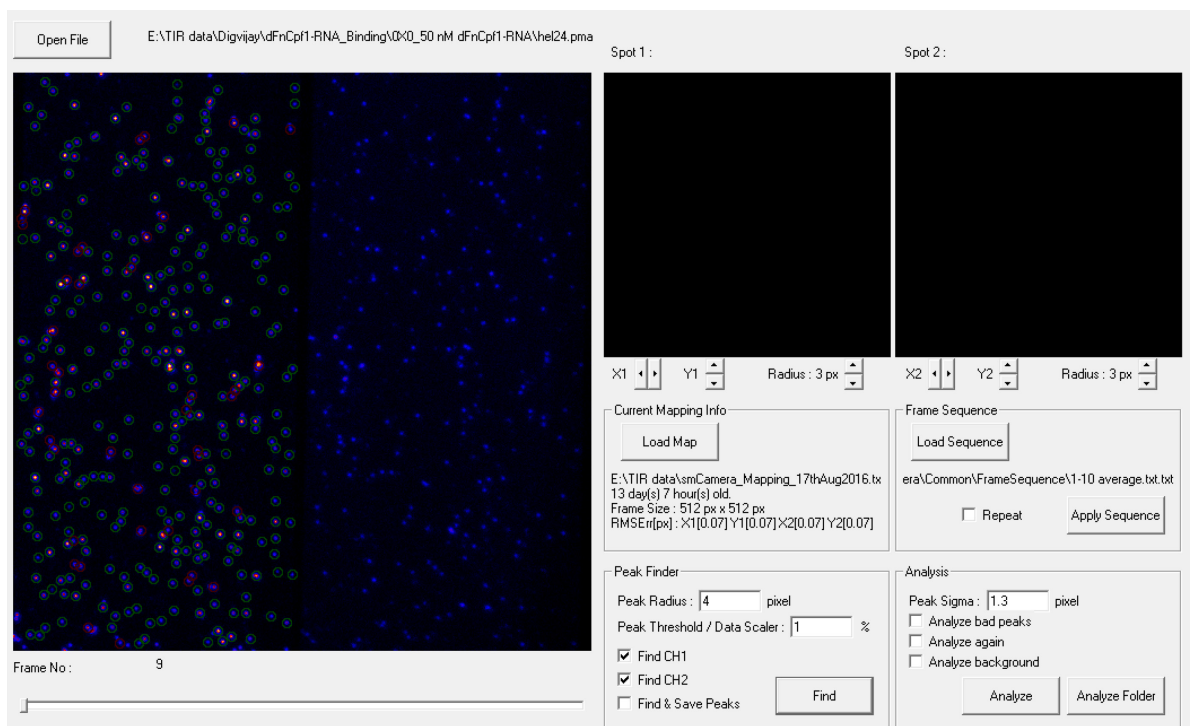
efficiency curve for the single molecule spot. The gradual decay of the signal indicates that the focus was

being lost gradually as the movie was acquired. So this movie cannot be used for any kinetic or any lifetime analysis and you would have to acquire another long duration movie.

15-20 short movies for FRET histograms and 3 long -duration movies with no defocusing issues is enough for a single DNA target.

### 5.7.11 Analysis of the single molecule movies

- 1) After you are done taking movies for all the DNA targets. You can analyze the movie using the smCamera software.
- 2) Here's a snapshot of the Analysis parameters that you need to use while analysing these movies. Please check the following important parameters.



**Figure 5.12 | A representative image showing the various parameters for the movie analysis in the smCamera Software.**

- 3) You should do the mapping for each set of experiments and use it for the analysis.

Things move in any microscope, for smFRET experiments analysis, where there has to be a one to one correlation between the spots in two different channels. A bad mapping is the worst possible 'sin' for these experiments. You have done all the hard work, your reagents and substrates worked, you saw good smFRET signal, but if you had the bad mapping file, all that information is essentially lost. Think about it.

- 4) Mapping file, the one relating the coordinate in the donor channel to the acceptor channel, being used for the analysis should not be more than 1-2 days old.

## BIBLIOGRAPHY

- 1 Marraffini, L. A. & Sontheimer, E. J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature reviews. Genetics* **11**, 181-190, doi:10.1038/nrg2749 (2010).
- 2 Barrangou, R. *et al.* CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712, doi:10.1126/science.1138140 (2007).
- 3 Brouns, S. J. *et al.* Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960-964, doi:10.1126/science.1159689 (2008).
- 4 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821, doi:10.1126/science.1225829 (2012).
- 5 Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E2579-2586, doi:10.1073/pnas.1208507109 (2012).
- 6 Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62-67, doi:10.1038/nature13011 (2014).
- 7 Wang, H., La Russa, M. & Qi, L. S. CRISPR/Cas9 in Genome Editing and Beyond. *Annual review of biochemistry* **85**, 227-264, doi:10.1146/annurev-biochem-060815-014607 (2016).
- 8 Barrangou, R. & Doudna, J. A. Applications of CRISPR technologies in research and beyond. *Nature biotechnology* **34**, 933-941, doi:10.1038/nbt.3659 (2016).
- 9 Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nature reviews. Microbiology* **13**, 722-736, doi:10.1038/nrmicro3569 (2015).
- 10 Ran, F. A. *et al.* In vivo genome editing using *Staphylococcus aureus* Cas9. *Nature* **520**, 186-191, doi:10.1038/nature14299 (2015).
- 11 Zetsche, B. *et al.* Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759-771, doi:10.1016/j.cell.2015.09.038 (2015).

- 12 Jore, M. M. *et al.* Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nature structural & molecular biology* **18**, 529-536, doi:10.1038/nsmb.2019 (2011).
- 13 Mulepati, S., Heroux, A. & Bailey, S. Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science* **345**, 1479-1484, doi:10.1126/science.1256996 (2014).
- 14 Wright, A. V., Nunez, J. K. & Doudna, J. A. Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. *Cell* **164**, 29-44, doi:10.1016/j.cell.2015.12.035 (2016).
- 15 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821, doi:10.1126/science.1225829 (2012).
- 16 Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E2579-E2586, doi:10.1073/pnas.1208507109 (2012).
- 17 Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62-67, doi:10.1038/nature13011 (2014).
- 18 Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096, doi:10.1126/science.1258096 (2014).
- 19 Wu, X., Kriz, A. J. & Sharp, P. A. Target specificity of the CRISPR-Cas9 system. *Quantitative biology* **2**, 59-70, doi:10.1007/s40484-014-0030-x (2014).
- 20 O'Geen, H., Yu, A. S. & Segal, D. J. How specific is CRISPR/Cas9 really? *Curr Opin Chem Biol* **29**, 72-78, doi:10.1016/j.cbpa.2015.10.001 (2015).
- 21 Sapranaukas, R. *et al.* The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res* **39**, 9275-9282, doi:10.1093/nar/gkr606 (2011).

- 22 Semenova, E. *et al.* Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proc Natl Acad Sci U S A* **108**, 10098-10103, doi:10.1073/pnas.1104144108 (2011).
- 23 Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819-823, doi:10.1126/science.1231143 (2013).
- 24 Fu, Y. *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol* **31**, 822-826, doi:10.1038/nbt.2623 (2013).
- 25 Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol.* **31**, doi:10.1038/nbt.2647 (2013).
- 26 Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat Biotechnol* **31**, 233-239, doi:10.1038/nbt.2508 (2013).
- 27 Jinek, M. *et al.* RNA-programmed genome editing in human cells. *Elife* **2**, e00471, doi:10.7554/eLife.00471 (2013).
- 28 Mali, P. *et al.* CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat Biotechnol* **31**, 833-838, doi:10.1038/nbt.2675 (2013).
- 29 Pattanayak, V. *et al.* High-throughput profiling of off-target DNA cleavage reveals RNA-programmed Cas9 nuclease specificity. *Nat Biotechnol* **31**, 839-843, doi:10.1038/nbt.2673 (2013).
- 30 Cho, S. W. *et al.* Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res* **24**, 132-141, doi:10.1101/gr.162339.113 (2014).
- 31 Doench, J. G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nature biotechnology* **32**, 1262-1267, doi:10.1038/nbt.3026 (2014).
- 32 Jinek, M. *et al.* Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* **343**, 1247997, doi:10.1126/science.1247997 (2014).

- 33 Gagnon, J. A. *et al.* Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs. *PLoS One* **9**, e98186, doi:10.1371/journal.pone.0098186 (2014).
- 34 Smith, C. *et al.* Whole-genome sequencing analysis reveals high specificity of CRISPR/Cas9 and TALEN-based genome editing in human iPSCs. *Cell Stem Cell* **15**, 12-13, doi:10.1016/j.stem.2014.06.011 (2014).
- 35 Shen, B. *et al.* Efficient genome modification by CRISPR-Cas9 nickase with minimal off-target effects. *Nat Methods* **11**, 399-402, doi:10.1038/nmeth.2857 (2014).
- 36 Zhang, Y. *et al.* Comparison of non-canonical PAMs for CRISPR/Cas9-mediated DNA cleavage in human cells. *Sci Rep* **4**, 5405, doi:10.1038/srep05405 (2014).
- 37 Frock, R. L. *et al.* Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat Biotechnol* **33**, 179-186, doi:10.1038/nbt.3101 (2015).
- 38 Iyer, V. *et al.* Off-target mutations are rare in Cas9-modified mice. *Nat Methods* **12**, 479, doi:10.1038/nmeth.3408 (2015).
- 39 Kim, D. *et al.* Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat Methods*. **12**, doi:10.1038/nmeth.3284 (2015).
- 40 Paulis, M. *et al.* A pre-screening FISH-based method to detect CRISPR/Cas9 off-targets in mouse embryonic stem cells. *Sci Rep* **5**, 12327, doi:10.1038/srep12327 (2015).
- 41 Singh, R., Kuscu, C., Quinlan, A., Qi, Y. & Adli, M. Cas9-chromatin binding information enables more accurate CRISPR off-target prediction. *Nucleic Acids Res* **43**, e118, doi:10.1093/nar/gkv575 (2015).
- 42 Tan, E.-P., Li, Y., Velasco-Herrera, M. D. C., Yusa, K. & Bradley, A. Off-target assessment of CRISPR-Cas9 guiding RNAs in human iPS and mouse ES cells. *Genesis* **53**, 225-236, doi:10.1002/dvg.22835 (2015).
- 43 Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol*. **33**, doi:10.1038/nbt.3117 (2015).



- 44 Wang, X. *et al.* Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. *Nat Biotechnol* **33**, 175-178, doi:10.1038/nbt.3127 (2015).
- 45 Xu, H. *et al.* Sequence determinants of improved CRISPR sgRNA design. *Genome Res* **25**, 1147-1157, doi:10.1101/gr.191452.115 (2015).
- 46 Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* **34**, 184-191, doi:10.1038/nbt.3437 (2016).
- 47 Cencic, R. *et al.* Protospacer adjacent motif (PAM)-distal sequences engage CRISPR Cas9 DNA target cleavage. *PLoS One* **9**, e109213, doi:10.1371/journal.pone.0109213 (2014).
- 48 Duan, J. *et al.* Genome-wide identification of CRISPR/Cas9 off-targets in human genome. *Cell Res* **24**, 1009-1012, doi:10.1038/cr.2014.87 (2014).
- 49 Kuscu, C., Arslan, S., Singh, R., Thorpe, J. & Adli, M. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat Biotechnol* **32**, 677-683, doi:10.1038/nbt.2916 (2014).
- 50 Wu, X. *et al.* Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat Biotechnol* **32**, 670-676, doi:10.1038/nbt.2889 (2014).
- 51 O'Geen, H., Henry, I. M., Bhakta, M. S., Meckler, J. F. & Segal, D. J. A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. *Nucleic Acids Res* **43**, 3389-3404, doi:10.1093/nar/gkv137 (2015).
- 52 Polstein, L. R. *et al.* Genome-wide specificity of DNA binding, gene regulation, and chromatin remodeling by TALE- and CRISPR/Cas9-based transcriptional activators. *Genome Res* **25**, 1158-1169, doi:10.1101/gr.179044.114 (2015).
- 53 Rutkauskas, M. *et al.* Directional R-Loop Formation by the CRISPR-Cas Surveillance Complex Cascade Provides Efficient Off-Target Site Rejection. *Cell reports*, doi:10.1016/j.celrep.2015.01.067 (2015).

- 54 Josephs, E. A. *et al.* Structure and specificity of the RNA-guided endonuclease Cas9 during DNA interrogation, target binding and cleavage. *Nucleic Acids Res* **43**, 8924-8941, doi:10.1093/nar/gkv892 (2015).
- 55 Blosser, T. R. *et al.* Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. *Molecular Cell* **58**, 60-70, doi:10.1016/j.molcel.2015.01.028 (2015).
- 56 Knight, S. C. *et al.* Dynamics of CRISPR-Cas9 genome interrogation in living cells. *Science* **350**, 823-826, doi:10.1126/science.aac6572 (2015).
- 57 Joo, C., Balci, H., Ishitsuka, Y., Buranachai, C. & Ha, T. Advances in single-molecule fluorescence methods for molecular biology. *Annual review of biochemistry* **77**, 51-76, doi:10.1146/annurev.biochem.77.070606.101543 (2008).
- 58 Ragunathan, K., Joo, C. & Ha, T. Real-time observation of strand exchange reaction with high spatiotemporal resolution. *Structure* **19**, 1064-1073, doi:10.1016/j.str.2011.06.009 (2011).
- 59 Lee, J. Y. *et al.* DNA RECOMBINATION. Base triplet stepping by the Rad51/RecA family of recombinases. *Science* **349**, 977-981, doi:10.1126/science.aab2666 (2015).
- 60 Ragunathan, K., Liu, C. & Ha, T. RecA filament sliding on DNA facilitates homology search. *Elife* **1**, e00067, doi:10.7554/eLife.00067 (2012).
- 61 Szczelkun, M. D. *et al.* Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 9798-9803, doi:10.1073/pnas.1402597111 (2014).
- 62 Redding, S. *et al.* Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. *Cell* **163**, 854-865, doi:10.1016/j.cell.2015.10.003 (2015).
- 63 Ha, T. *et al.* Probing the interaction between two single molecules: fluorescence resonance energy transfer between a single donor and a single acceptor. *Proc Natl Acad Sci U S A* **93**, 6264-6268 (1996).
- 64 Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nature Methods* **5**, 507-516, doi:10.1038/nmeth.1208 (2008).

- 65 Anders, C., Niewoehner, O., Duerst, A. & Jinek, M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**, 569-573, doi:10.1038/nature13579 (2014).
- 66 McKinney, S. A., Joo, C. & Ha, T. Analysis of Single-Molecule FRET Trajectories Using Hidden Markov Modeling. *Biophysical Journal* **91**, 1941-1951, doi:10.1529/biophysj.106.082487 (2006).
- 67 Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84-88, doi:10.1126/science.aad5227 (2016).
- 68 Kleinstiver, B. P. *et al.* High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490-495, doi:10.1038/nature16526 (2016).
- 69 Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935-949, doi:10.1016/j.cell.2014.02.001 (2014).
- 70 Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31**, 827-832, doi:10.1038/nbt.2647 (2013).
- 71 Bae, S., Park, J. & Kim, J.-S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473-1475, doi:10.1093/bioinformatics/btu048 (2014).
- 72 Heigwer, F., Kerr, G. & Boutros, M. E-CRISP: fast CRISPR target site identification. *Nat Methods*. **11**, doi:10.1038/nmeth.2812 (2014).
- 73 Ren, X. *et al.* Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized sgRNA parameters in Drosophila. *Cell Rep* **9**, 1151-1162, doi:10.1016/j.celrep.2014.09.044 (2014).
- 74 Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84, doi:10.1126/science.1246981 (2014).
- 75 Chari, R., Mali, P., Moosburner, M. & Church, G. M. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat Methods* **12**, 823-826, doi:10.1038/nmeth.3473 (2015).

- 76 Farboud, B. & Meyer, B. J. Dramatic enhancement of genome editing by CRISPR/Cas9 through improved guide RNA design. *Genetics* **199**, 959-971, doi:10.1534/genetics.115.175166 (2015).
- 77 Moreno-Mateos, M. A. *et al.* CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat Methods* **12**, 982-988, doi:10.1038/nmeth.3543 (2015).
- 78 Stemmer, M., Thumberger, T., Del Sol Keyer, M., Wittbrodt, J. & Mateo, J. L. CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. *PLoS One* **10**, e0124633, doi:10.1371/journal.pone.0124633 (2015).
- 79 Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol* **33**, 187-197, doi:10.1038/nbt.3117 (2015).
- 80 Wong, N., Liu, W. & Wang, X. WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biol* **16**, 218, doi:10.1186/s13059-015-0784-0 (2015).
- 81 Labun, K., Montague, T. G., Gagnon, J. A., Thyme, S. B. & Valen, E. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res*, doi:10.1093/nar/gkw398 (2016).
- 82 Wu, X., Kriz, A. J. & Sharp, P. A. Target specificity of the CRISPR-Cas9 system. *Quantitative biology* **2**, 59-70, doi:10.1007/s40484-014-0030-x (2014).
- 83 Singh, D., Sternberg, S. H., Fei, J., Doudna, J. A. & Ha, T. Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. *Nature communications* **7**, 12778, doi:10.1038/ncomms12778 (2016).
- 84 Dagdas, Y. S., Chen, J. S., Sternberg, S. H., Doudna, J. A. & Yildiz, A. A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9. *Science Advances* **3**, doi:10.1126/sciadv.aao0027 (2017).
- 85 Chen, J. S. *et al.* Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature*, doi:10.1038/nature24268 (2017).

- 86 Szczelkun, M. D. *et al.* Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc Natl Acad Sci U S A* **111**, 9798-9803, doi:10.1073/pnas.1402597111 (2014).
- 87 Blosser, T. R. Two distinct DNA binding modes guide dual roles Of a CRISPR-Cas protein complex. **58**, 60-70, doi:10.1016/j.molcel.2015.01.028 (2015).
- 88 Lim, Y. *et al.* Structural roles of guide RNAs in the nuclease activity of Cas9 endonuclease. *Nature communications* **7**, 13350, doi:10.1038/ncomms13350 (2016).
- 89 Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nature methods* **5**, 507-516, doi:10.1038/nmeth.1208 (2008).
- 90 McKinney, S. A., Joo, C. & Ha, T. Analysis of single-molecule FRET trajectories using hidden Markov modeling. *Biophys J* **91**, 1941-1951, doi:10.1529/biophysj.106.082487 (2006).
- 91 Boyle, E. A. *et al.* High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 5461-5466, doi:10.1073/pnas.1700557114 (2017).
- 92 Anders, C. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. **513**, 569-573, doi:10.1038/nature13579 (2014).
- 93 Jiang, F. *et al.* Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* **351**, 867-871, doi:10.1126/science.aad8282 (2016).
- 94 Richardson, C. D., Ray, G. J., DeWitt, M. A., Curie, G. L. & Corn, J. E. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nat Biotechnol* **34**, 339-344, doi:10.1038/nbt.3481 (2016).
- 95 Zuo, Z. & Liu, J. Cas9-catalyzed DNA Cleavage Generates Staggered Ends: Evidence from Molecular Dynamics Simulations. *Scientific reports* **5**, 37584, doi:10.1038/srep37584 (2016).
- 96 Sternberg, S. H., LaFrance, B., Kaplan, M. & Doudna, J. A. Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature* **527**, 110-113, doi:10.1038/nature15544 (2015).

- 97 Gilbert, L. A. *et al.* CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* **154**, 442-451, doi:10.1016/j.cell.2013.06.044 (2013).
- 98 Chen, B. *et al.* Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell* **155**, 1479-1491, doi:10.1016/j.cell.2013.12.001 (2013).
- 99 Palermo, G., Miao, Y., Walker, R. C., Jinek, M. & McCammon, J. A. CRISPR-Cas9 conformational activation as elucidated from enhanced molecular simulations. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 7260-7265, doi:10.1073/pnas.1707645114 (2017).
- 100 Fu, Y., Sander, J. D., Reyon, D., Cascio, V. M. & Joung, J. K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nature biotechnology* **32**, 279-284, doi:10.1038/nbt.2808 (2014).
- 101 Rasnik, I., McKinney, S. A. & Ha, T. Nonblinking and long-lasting single-molecule fluorescence imaging. *Nature methods* **3**, 891-893, doi:10.1038/nmeth934 (2006).
- 102 Marraffini, L. A. & Sontheimer, E. J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature reviews. Genetics* **11**, 181-190, doi:10.1038/nrg2749 (2010).
- 103 Gasiunas, G., Barrangou, R., Horvath, P. & Siksnys, V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc Natl Acad Sci U S A* **109**, E2579-2586, doi:10.1073/pnas.1208507109 (2012).
- 104 Wright, A. V., Nuñez, J. K. & Doudna, J. A. Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. *Cell* **164**, 29-44, doi:10.1016/j.cell.2015.12.035 (2016).
- 105 Shmakov, S. *et al.* Diversity and evolution of class 2 CRISPR-Cas systems. *Nature reviews. Microbiology* **15**, 169-182, doi:10.1038/nrmicro.2016.184 (2017).
- 106 Burstein, D. *et al.* New CRISPR-Cas systems from uncultivated microbes. *Nature* **542**, 237-241, doi:10.1038/nature21059 (2017).

- 107 Kleinstiver, B. P. *et al.* Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nat Biotechnol* **34**, 869-874, doi:10.1038/nbt.3620 (2016).
- 108 Kim, D. *et al.* Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nat Biotechnol* **34**, 863-868, doi:10.1038/nbt.3609 (2016).
- 109 Fonfara, I., Richter, H., Bratovič, M., Le Rhun, A. & Charpentier, E. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **532**, 517-521, doi:10.1038/nature17945 (2016).
- 110 Josephs, E. A. *et al.* Structure and specificity of the RNA-guided endonuclease Cas9 during DNA interrogation, target binding and cleavage. *Nucleic acids research* **44**, 2474, doi:10.1093/nar/gkv1293 (2016).
- 111 Blosser, T. R. *et al.* Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. *Mol Cell* **58**, 60-70, doi:10.1016/j.molcel.2015.01.028 (2015).
- 112 Dagdas, Y. S., Chen, J. S., Sternberg, S. H., Doudna, J. A. & Yildiz, A. A Conformational Checkpoint Between DNA Binding And Cleavage By CRISPR-Cas9. *bioRxiv* (2017).
- 113 Chen, J. S. *et al.* Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *bioRxiv*, doi:10.1101/160036 (2017).
- 114 Yamano, T. *et al.* Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. *Cell* **165**, 949-962, doi:10.1016/j.cell.2016.04.003 (2016).
- 115 Bates, M., Blosser, T. R. & Zhuang, X. Short-range spectroscopic ruler based on a single-molecule optical switch. *Physical review letters* **94**, 108101, doi:10.1103/PhysRevLett.94.108101 (2005).
- 116 Dong, D. *et al.* The crystal structure of Cpf1 in complex with CRISPR RNA. *Nature* **532**, 522-526, doi:10.1038/nature17944 (2016).
- 117 Stella, S., Alcon, P. & Montoya, G. Structure of the Cpf1 endonuclease R-loop complex after target DNA cleavage. *Nature* **546**, 559-563, doi:10.1038/nature22398 (2017).

- 118 Lee, W., von Hippel, P. H. & Marcus, A. H. Internally labeled Cy3/Cy5 DNA constructs show greatly enhanced photo-stability in single-molecule FRET experiments. *Nucleic acids research* **42**, 5967-5977, doi:10.1093/nar/gku199 (2014).
- 119 Singh, D. *et al.* Mechanisms of improved specificity of engineered Cas9s revealed by single molecule analysis. *bioRxiv* (2017).
- 120 Wang, S., Su, J. H., Zhang, F. & Zhuang, X. An RNA-aptamer-based two-color CRISPR labeling system. *Scientific reports* **6**, 26857, doi:10.1038/srep26857 (2016).
- 121 Oakley, J. L. & Coleman, J. E. Structure of a promoter for T7 RNA polymerase. *Proc Natl Acad Sci U S A* **74**, 4266-4270 (1977).
- 122 Guschin, D. Y. *et al.* A rapid and general assay for monitoring endogenous gene modification. *Methods in molecular biology (Clifton, N.J.)* **649**, 247-256, doi:10.1007/978-1-60761-753-2\_15 (2010).
- 123 Kan, Y., Ruis, B., Lin, S. & Hendrickson, E. A. The mechanism of gene targeting in human somatic cells. *PLoS genetics* **10**, e1004251, doi:10.1371/journal.pgen.1004251 (2014).
- 124 Revyakin, A., Liu, C., Ebright, R. H. & Strick, T. R. Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching. *Science* **314**, 1139-1143, doi:10.1126/science.1131398 (2006).
- 125 Joo, C. & Ha, T. Labeling DNA (or RNA) for single-molecule FRET. *Cold Spring Harbor protocols* **2012**, 1005-1008, doi:10.1101/pdb.prot071027 (2012).
- 126 Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* **9**, 671-675 (2012).
- 127 Hohng, S., Joo, C. & Ha, T. Single-Molecule Three-Color FRET. *Biophysical Journal* **87**, 1328-1337, doi:10.1529/biophysj.104.043935 (2004).
- 128 Lee, J. *et al.* Single-molecule Four-color FRET. *Angewandte Chemie (International ed. in English)* **49**, 9922-9925, doi:10.1002/anie.201005402 (2010).



- 129 Suddala, K. C. & Walter, N. G. Riboswitch structure and dynamics by smFRET microscopy. *Methods in enzymology* **549**, 343-373, doi:10.1016/b978-0-12-801122-5.00015-5 (2014).
- 130 Perkins, T. T. Angstrom-precision optical traps and applications. *Annual review of biophysics* **43**, 279-302, doi:10.1146/annurev-biophys-042910-155223 (2014).
- 131 Derrington, I. M. *et al.* Sub-angstrom single-molecule measurements of motor proteins using a nanopore. *Nature biotechnology* **33**, 1073-1075, doi:10.1038/nbt.3357 (2015).
- 132 Hohng, S. *et al.* Fluorescence-force spectroscopy maps two-dimensional reaction landscape of the Holliday junction. *Science* **318**, 279-283, doi:10.1126/science.1146113 (2007).
- 133 Bermúdez, I., García-Martínez, J., Pérez-Ortín, J. E. & Roca, J. A method for genome-wide analysis of DNA helical tension by means of psoralen–DNA photobinding. *Nucleic acids research* **38**, e182-e182, doi:10.1093/nar/gkq687 (2010).
- 134 Ljungman, M. & Hanawalt, P. C. Localized torsional tension in the DNA of human cells. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 6055-6059 (1992).
- 135 Hoskins, A. A. *et al.* Ordered and Dynamic Assembly of Single Spliceosomes. *Science* **331**, 1289-1295, doi:10.1126/science.1198830 (2011).
- 136 Maji, B. *et al.* Multidimensional chemical control of CRISPR-Cas9. *Nature chemical biology* **13**, 9-11, doi:10.1038/nchembio.2224 (2017).
- 137 Bondy-Denomy, J. *et al.* Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. *Nature* **526**, 136-139, doi:10.1038/nature15254 (2015).
- 138 Shin, J. *et al.* Disabling Cas9 by an anti-CRISPR DNA mimic. *Science advances* **3**, e1701620, doi:10.1126/sciadv.1701620 (2017).
- 139 Chandradoss, S. D. *et al.* Surface passivation for single-molecule protein studies. *Journal of visualized experiments : JoVE*, doi:10.3791/50549 (2014).
- 140 Joo, C. & Ha, T. Single-molecule FRET with total internal reflection microscopy. *Cold Spring Harbor protocols* **2012**, doi:10.1101/pdb.top072058 (2012).

- 141 Besseling, N. A. M. Theory of Hydration Forces between Surfaces. *Langmuir* **13**, 2113-2122, doi:10.1021/la960672w (1997).
- 142 Dhruv, H. *Controlling Nonspecific Adsorption of Proteins at Bio-Interfaces for Biosensor and Biomedical Applications*. (2017).
- 143 Besseling, N. A. M. & Scheutjens, J. M. H. M. Statistical Thermodynamics of Molecules with Orientation-Dependent Interactions in Homogeneous and Inhomogeneous Systems. *The Journal of Physical Chemistry* **98**, 11597-11609, doi:10.1021/j100095a048 (1994).
- 144 Besseling, N. A. M. & Lyklema, J. Equilibrium Properties of Water and Its Liquid-Vapor Interface. *The Journal of Physical Chemistry* **98**, 11610-11622, doi:10.1021/j100095a049 (1994).
- 145 Wright, A. V. *et al.* Rational design of a split-Cas9 enzyme complex. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 2984-2989, doi:10.1073/pnas.1501698112 (2015).

## **CURRICULUM VITAE**

### **Digvijay Singh**

#### **Birth**

June 8, 1989                      Raipur, Chhattisgarh, India

#### **Education:**

2012-Present                      Ph.D. candidate, Johns Hopkins University School of Medicine, Baltimore, MD. Program in Molecular Biophysics. Anticipated completion in the Spring of 2018.

2007-2012                      Integrated B. S + M. S. in Chemistry, Indian Institute of Technology, Kharagpur, West Bengal, India.

#### **Research**

2013 – present                      Graduate student, Department of Biophysics, Johns Hopkins University School of Medicine, USA *transferred from* Center for Biophysics & Quantitative Biology, University of Illinois at Urbana-Champaign, USA.

Principal investigator : Prof. Taekjip Ha

- Single molecule imaging and biochemical assays to characterize molecular mechanisms and specificity of DNA targeting by CRISPR-Cas9 and CRISPR-Cpf1 family of RNA-guided endonucleases which are being widely used for various genome engineering applications. Collaborated with Prof. Jennifer Doudna ( UC Berkeley) and Prof. Scott Bailey (Johns Hopkins

School of Public Health).

- Collaboration with Prof. Venigala B. Rao (CUA) to characterize mechanism of coordination between different subunits of bacteriophage T4 DNA packaging motor using single molecule imaging.
- Design and implementation of data-analysis packages for single molecule studies.

6/2012 – 8/2012      Visiting research assistant; Principal investigator: Prof. Robert Best (now at NIH), University of Cambridge, UK.

- *Theoretical Biophysics* - Construction of multi-dimensional free energy surfaces of protein folding using certain select coordinates.

8/2011 – 3/2012      Master thesis student; Principal investigator: Prof. Swagata Dasgupta, Indian Institute of Technology, Kharagpur.

- *Computational Biophysics* – Modeling of amyloid beta multimers via protein structure prediction (*Rosetta*).

5/2011 – 7/2011      Visiting research assistant; Principal investigator: Prof. Collin M. Stultz, Massachusetts Institute of Technology, USA

- *Computational Biophysics* – Construction of structural library of intrinsically disordered amyloid beta protein, using molecular dynamics simulations, for creation of its conformational ensembles.

- 11/2010 – 12/2010 Visiting research assistant; Principal investigator: Prof. Anne S. Ulrich, Karlsruhe Institute of Technology, Germany
- *Solid State NMR* - Synthesis of membrane active peptides and evaluation of its alignment in lipid bilayer from NMR (Nuclear Magnetic Resonance) signals of <sup>19</sup>F labels on the peptides.
- 11/2010 – 12/2010 Visiting research assistant; Principal investigator: Prof. Anne S. Ulrich, Karlsruhe Institute of Technology, Germany
- *Solid State NMR* - Synthesis of membrane active peptides and evaluation of its alignment in lipid bilayer from NMR (Nuclear Magnetic Resonance) signals of <sup>19</sup>F labels on the peptides.
- 5/2010 – 7/2010 Visiting research assistant; Principal investigator: Prof. Martin Gruebele, University of Illinois at Urbana-Champaign, USA
- *Protein Folding* - Expression and purification of FRET probe labeled protein construct for its use in fast relaxation imaging following temperature shocks.
- 11/2009 – 12/2009 Research intern; Unilever, Bangalore, India
- Investigation of binding affinity of tea polyphenols with milk caseins.
- 4/2009 – 7/2009 Research intern; General Electric, Bangalore, India
- Synthesis of radio labeled indoles & amides with high binding affinity to certain specific receptors found in nervous system for

its use in PET (Positron Emission Tomography) imaging.

**Peer-Reviewed Publications:**

**Digvijay Singh**, Samuel H. Sternberg, Jingyi Fei, Jennifer A. Doudna, Taekjip Ha. “Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9”. *Nature Communications* (2016).

Jingyi Fei, **Digvijay Singh**, Qiucen Zhang, Seongjin Park, Divya Balasubramanian, Ido Golding, Carin K. Vanderpool, Taekjip Ha. “Determination of in vivo target search kinetics of regulatory non-coding RNA”. *Science* (2015).

Boyang Hua, Kyu Young Han, Ruobo Zhou, Hajin Kim, Xinghua Shi, Sanjaya C. Abeysirigunawardena, Ankur Jain, **Digvijay Singh**, Vasudha Aggarwal, Sarah A. Woodson, Taekjip Ha. “An improved surface passivation method for single-molecule studies”. *Nature Methods* (2014).

A. Dhar, K. Girdhar, **D. Singh**, S. Ebbinghaus and M. Gruebele, “Different protein stability and folding kinetics in the nucleus, endoplasmic reticulum, and cytoplasm of living cells,” *Biophys. J.* 101, 421-430 (2011).

**Pre-print and under-review Publications:**

**Digvijay Singh**, Yanbo Wang, John Mallon, Olivia Yang, Jingyi Fei, Anustup

Poddar, Damon Ceylan, Scott Bailey, Taekjip Ha. "Mechanism of improved specificity of engineered Cas9s revealed by single molecule analysis". *BioRxiv* (2017).

**Digvijay Singh**, John Mallon, Anustup Poddar, Yanbo Wang, Ramreddy Tipanna, Olivia Yang, Scott Bailey, Taekjip Ha. "Real-time observation of DNA target interrogation and product release by RNA-guided endonuclease CRISPR-Cpf1". *BioRxiv* (2017).

**Digvijay Singh**, Taekjip Ha. "Understanding the molecular mechanism of CRISPR toolbox using single-molecule approaches". *Submitted*.

Boyang Hua, Yanbo Wang, Kyu Young Han, Seongjin Park, **Digvijay Singh**, Jin H. Kim, Wei Cheng, Taekjip Ha. "Single-molecule centroid localization algorithm improves the accuracy of fluorescence binding assays". *Submitted*.

Li Dai, **Digvijay Singh**, Reza Vafabakhsh, Marthandan Mahalingam, Vishal Kottadiel, Yann Chemla, Taekjip Ha, Venigalla B Rao. "Mechanism of coordination of the bacteriophage T4 DNA packaging motor analyzed by real-time single molecule fluorescence assay". *In Preparation*.

### **Presentations:**

Talk. "Role of co-factors in Rhodopsin: An action potential story of vision". *Wednesday Morning Seminar Series, Biophysics, JHMI* (2017). Literature Survey.

Talk. "DNA targeting by CRISPR-Cas at the single molecule level". *Student Evening Seminar Series, Biophysics, JHMI* (2017).

Talk. "Investigation of DNA binding, nucleolysis and product release specificity of RNA guided endonuclease CRISPR-Cpf1 family reveals important differences from Cas9-RNA". *Biophysical Society Meeting* (2017).

Talk. "Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9". *Physics of Living Systems Conference, Arlington, VA* (2015).

Talk. "Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9". *Biophysical Society Meeting* (2015).

Talk. "Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9". *Center for Physics of Living Cells Symposium, University of Illinois* (2015).

Poster. **Digvijay Singh**, Yanbo Wang, John Mallon, Samuel H. Sternberg, Olivia Yang, Jingyi Fei, Anustup Poddar, Damon Ceylan, Jennifer A. Doudna, Scott Bailey, Taekjip Ha. "Mechanistic basis of the specificity improving mutations of Engineered Cas9s". *NCI RNA Biology* (2017).

Poster. **Digvijay Singh**, Yanbo Wang, John Mallon, Samuel H. Sternberg, Olivia Yang, Jingyi Fei, Anustup Poddar, Damon Ceylan, Jennifer A. Doudna, Scott Bailey, Taekjip



Ha. "Mechanistic basis of the specificity improving mutations of Engineered Cas9s". *Chesapeake Bay Area Single Molecule Meeting* (2017).

Poster. **Digvijay Singh**, Samuel H. Sternberg, Jingyi Fei, Jennifer A. Doudna, Taekjip Ha. "Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9". *Bayview Research Symposium* (2017).

Poster. Li Dai, **Digvijay Singh**, Reza Vafabakhsh, Marthandan Mahalingam, Vishal Kottadiel, Yann Chemla, Taekjip Ha, Venigalla B Rao. "Mechanism of coordination of the bacteriophage T4 DNA packaging motor analyzed by real-time single molecule fluorescence assay". *Biophysical Society Meeting* (2016).

Poster. **Digvijay Singh**, Samuel H. Sternberg, Jingyi Fei, Jennifer A. Doudna, Taekjip Ha. "Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9". *Biophysical Society Meeting* (2016).

Poster. Boyang Hua, Ruobo Zhou, Hajin Kim, Xinghua Shi, Ankur Jain, **Digvijay Singh**, Vasudha Aggarwal, Taekjip Ha. "An improved surface passivation method for single-molecule studies". *Biophysical Society Meeting* (2014).

Poster. Seongjin Park, Sultan Doğanay, Jingyi Fei, **Digvijay Singh**, Taekjip Ha. "Super-resolution imaging of immune response proteins against viral infection". *Biophysical Society Meeting* (2014).

Poster. Jingyi Fei, **Digvijay Singh**, Qiucen Zhang, Seongjin Park, Ido Golding, Carin K Vanderpool, Taekjip Ha. "Watching gene regulation by small RNA in bacteria with super-resolution Imaging". *Biophysical Society Meeting* (2013).

Poster. Jingyi Fei, Seongjin Park, **Digvijay Singh**, Yanxin Liu, John E. Stone, Klaus Schulten, Kannanganattu V.Prasanth and Taekjip Ha. "Organization of long non-coding RNAs in nuclear bodies revealed by super resolution imaging". *EMBO | EMBL Symposia: The Complex Life of mRNA* (2012).

### **Courses and Workshops:**

2017                      Summer Student; Optical Microscopy & Imaging in the Biomedical sciences-2017, Marine Biological Laboratory, University of Chicago

### **Teaching experience:**

Fall 2014-present      Mentor; Directly mentored one undergraduate and two graduate students.

2013-2015              Instructor; Center for the Physics of Living Cells (CPLC) summer schools, University of Illinois at Urbana-Champaign, USA.

2014-2015              Teaching assistant; Advanced Biophysics course (smFRET module), Department of Physics, University of Illinois at Urbana-Champaign, USA.