Title
Mechanistic studies of CRISPR proteins in living cells

Author
Knight, Spencer Charles

Publication Date
2017

Peer reviewed|Thesis/dissertation

Mechanistic studies of CRISPR proteins in living cells

By

Spencer Charles Knight

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jennifer A. Doudna, Co-Chair
Professor Robert Tjian, Co-Chair
Professor Ming C. Hammond
Professor Roberto Zoncu

Spring 2017

Abstract

Mechanistic studies of CRISPR proteins in living cells

by

Spencer Charles Knight

Doctor of Philosophy in Chemistry

University of California, Berkeley

Professor Jennifer A. Doudna, Co-Chair

Professor Robert Tjian, Co-Chair

The discovery of the CRISPR-Cas9 protein has enabled facile genome editing in living cells and organisms. Structural and biochemical studies of Cas9 endonucleases have provided critical information about the core molecular requirements of the RNA-guided DNA cleavage reaction, but questions remain about how this bacterial protein navigates chromatin to identify DNA targets within living eukaryotic cells. In particular, the in vivo kinetics of on- versus off-target binding and Cas9 dependence on chromatin environment remain largely unknown. Here we present a single-molecule analysis of Cas9 searching in living mammalian cells. We provide evidence for a diffusion-dominated search mechanism and show that off-target binding at PAMs and short seed sequences is, on average, short-lived (milleseconds to seconds) by both single-molecule and bulk imaging techniques. The differential behavior of Cas9 in closed-off (heterochromatic) versus open (euchromatic) regions of the genome is also explored via heterochromatin protein 1 (HP1) staining and nuclear masking of single-particle trajectories. Comparative analysis of trajectories from these two regions suggests that Cas9 undersamples heterochromatin and moves more compactly within these regions. These data provide the first direct visualization of Cas9 searching in living cells and offer quantitative insights into how Cas9 navigates hierarchical organization of DNA within a eukaryotic nucleus. We additionally present mechanistic investigations of unrelated CRISPR-C2c2 proteins and identify two distinct enzymatic activities– pre-CRISPR RNA (pre-crRNA) processing and crRNA-stimulated general RNase activity. Collectively, this work expands our mechanistic understanding of CRISPR biology and highlights the utility of these enzymatically diverse proteins to be harnessed for biotechnological applications.

Table of Contents

List of Figures and Tables

Acknowledgments

This thesis is the culmination of five years of work, and none of it would have been possible without the unwavering support of friends, family, classmates, coworkers, and advisors. First and foremost– thank you Mom and Dad! You two have been my anchors during this volatile journey. While we are physically separated by nearly 2,000 miles, you have been there with me at every step of the way.

I am extremely grateful to my two PhD advisors– Robert Tjian and Jennifer Doudna. As a troubled second year seeking new research directions, you took a big chance on me and gave me extraordinary freedom to work on a risky and difficult project that paid off tremendously. Moreover, I learned so much about science, leadership, management, and innovation from the two of you. I feel incredibly lucky to have had both of you as mentors during my time here.

A big shout-out goes to a very special friend and roommate– Scott Evan Miller. While you were of little help with respect to analyzing imaging data or interpreting RNA gels, you were nonetheless a constant source of emotional support and entertainment in my life. Also, my plants would have died years ago were it not for your green thumb, so thanks for that!

I would like to thank some of my classmates, especially Zachary Hallberg, Omer Ad, Allegra Aron, and Thomas Brewer. Zach– you are an incredible scientist and an even better friend, and I look forward to seeing the (big) places you will go in the near future! And always remember: “Tough never tired!” Omer– you have kept me organized and focused during this chaotic journey. Allegra and Thom– you have been stellar friends, and– for a brief period of time– coworkers. I wish you the best of luck in your future endeavors!

I am grateful to labmates from both the Tjian and Doudna labs, especially Sharon Torigoe, Frank (Liangqi) Xie, Lana Bosanac, James (Zhe) Liu, Wulan Deng, Hervé Marie-Nelly, Lea Witkowsky-Gainous, Benjamin Guglielmi, Alexandra East-Seletsky, and Mitch O’Connell. Sharon– you have been my emotional rock in lab, and I will sincerely miss our daily lunches and life chats. I look forward to seeing your career develop at Lewis & Clark! Frank, Lana, James, and Wulan– you provided invaluable mentorship during my transition from synthetic chemistry to biophysics, and the Science paper would not have happened without you! Hervé– thank you for working with me on the translational readthrough project. I learned a tremendous amount about machine learning and data science from you, and on a more personal note I thoroughly enjoyed our conversations about philosophy and the meaning of life. Lea and Benjamin– I would like to express my gratitude for technical assistance and critical discussions relating to the Science paper. Alex and Mitch– thank you for working with me on the C2c2 project. I enjoyed my brief foray into RNA biochemistry and developed a nuanced appreciation for non-Cas9 CRISPR systems.

I would like to thank several professors at Berkeley who have provided invaluable mentorship and feedback along this journey, especially Xavier Darzacq, Carolyn Bertozzi (you will always be a Cal Bear in my mind!), Ming Chen Hammond, Roberto Zoncu, and Christopher Chang. Xavier– it has been a real treat having you in Li Ka Shing! You are a fountain of imaging knowledge, and you have been like a third advisor to me. Carolyn– you are an inspiration to young LGBT scientists, and I could not have

asked for a better chairperson for my qualifying exam committee. Ming and Roberto– I appreciate the two of you serving on my qualifying committee and thank you for critically reading this manuscript. Chris– I am grateful to you for gracefully handling my transition into the Tjian and Doudna labs. I have wandered very far from “chemistry” since starting graduate school, but the Chang lab will always hold a special place in my heart.

Last, but certainly not least, I would like to thank my two best friends back in Minnesota– Alison Key and Cameron Thorne. You two goons have been there for every victory and every setback since we first met at the tender age of 19. We have come a long way since our days of illegal trespassing and cocktail parties with Cynthia Kraznepick, but one thing that has been constant has been your unrelenting support and companionship. It is impossible for me to imagine life without the two of you.

Cheers!

Chapter 1: Introduction

Recent advances in molecular biology and bioinformatics have allowed for rapid and cost-effective perturbation of a gene of interest. The emergence of new technologies in these fields has ushered in a renaissance in molecular genetics and precision medicine. This chapter provides an overview of these technologies– especially CRISPR-based genome editing. We additionally discuss recent advances in single-molecule microscopy and outline an imaging-based strategy to study the search mechanism of CRISPR proteins within living cells.

The Central Dogma: Regulation and Perturbation

The expression of genes and proteins that govern an organism's response to its environment must be spatially and temporally controlled, a process known as gene regulation (Rockman and Kruglyak, 2006). Of particular importance is controlling the synthesis of proteins, the building blocks of life, from DNA, the universal genetic material (the Central Dogma) (Crick, 1970). The first step of the Central Dogma involves transcription, the process whereby a template DNA strand is used to synthesize a complementary messenger RNA (mRNA) (Browning and Busby, 2004; Lee et al., 2012; Levine and Tjian, 2003; Levine et al., 2014). The sequence of nucleotides from the resulting mRNA is then read out by ribosomes to produce proteins in a process called translation (Hinnebusch, 2014; Jackson et al., 2010; Kapp and Lorsch, 2004; Myasnikov et al., 2009).

Dynamic gatekeeping mechanisms associated with transcription and translation serve as powerful checkpoints for regulating gene expression. While the human genome consists of 3 billion (3,000,000,000) DNA base pairs, only a fraction of it is expressed at any given time. Transcription is tightly controlled by DNA-binding proteins (transcription factors), mediator proteins, proximal and distal sequence motifs (e.g. promoters and enhancers), non-coding RNAs, and the epigenetic state of local chromatin (Cech and Steitz, 2014; Conaway and Conaway, 1993; Dynan and Tjian, 1983; Kadonaga, 1998; Levine et al., 2014; Voss and Hager, 2013). Translation is gated by initiaton factors, sequence and structural motifs within the mRNA (e.g. Kozak, polyA), the mTOR pathway, non-coding RNAs, and ribosome levels, among other factors (Hall et al., 1982; Hinnebusch, 2014; Jackson et al., 2010; Kozak, 2005; Sachs et al., 1997; Sonenberg and Hinnebusch, 2009). All of these parameters can be rapidly and combinatorially adjusted, providing a vast gene expression landscape in response to different environmental stimuli.

Given the profound effect that gene expression has on phenotype, there has been great interest in perturbing biological systems at the molecular level to gain new insights (Schwanhäusser et al., 2011). Systematic studies of the components of gene expression have allowed biologists, engineers, and physicians to tackle salient problems in engineering and medicine. Examples include: genetic studies of transcription factors leading to the discovery of stem cell reprogramming (Takahashi and Yamanaka, 2006); and studies of histone chemical modifications leading to new cancer drugs (Chi et al., 2010; Falkenberg and Johnstone, 2014).

Paramount to understanding the causal relationship between a gene and a particular phenotype is the ability to singularly perturb that gene. To this end, a number of technologies have been developed over the years that have allowed for targeted modulation of a gene of interest. Early efforts in this area focused on the development of plasmid-based methods for overexpressing a protein of interest (Prelich, 2012). While this approach has successfully predicted the function of many genes, it is prone to artifacts including: (a) Proteins often behave atypically at unnaturally high concentrations; (b) Excessive taxation of the cellular translational machinery to overexpress one gene may reduce the cell's ability to appropriately regulate other genes.

An orthogonal approach to this has been to selectively upregulate or downregulate a protein of interest using chemical approaches. Efforts towards this end

began with the use of naturally occurring ligands that specifically modulate the activity of a target protein (Braun and Schulman, 1995; Newton, 1995). This later expanded to library and directed evolution approaches to identify non-native effector ligands to selectively probe gene function (Bishop et al., 2001; Chockalingam et al., 2005; Cravatt et al., 2008). Examples include: (a) synthetic chemical inhibitors of transcription factors to study their role in oncogenesis (Yeh et al., 2013); (b) checkpoint kinase inhibitors from combinatorial libraries to study the mechanism of cell cycling (Cohen et al., 1998); and (c) the development of blood vessel receptor antagonists via phage display to map the human vasculature (Arap et al., 2002).

At the level of RNA, a major breakthrough occurred in the late 1990s with the discovery of RNA interference (RNAi), a eukaryotic RNA-based immune system (Fire et al., 1998; Wilson and Doudna, 2013; Zamore et al., 2000). In RNAi, a small RNA molecule– either a microRNA (miRNA) or small interfering RNA (siRNA)– binds to a complementary mRNA or DNA molecule to initialize a cascade of events that ultimately results in either translational or transcriptional silencing, respectfully. In the case of miRNA, a single-stranded hairpin primary transcript is synthesized in the nucleus, further processed by Drosha and Dicer proteins to form a short double-stranded RNA (termed shRNA), and then directed by the RISC complex to target and degrade a complementary mRNA. In siRNA systems, double-stranded RNA is cleaved by Dicer to generate short siRNA fragments that are directed by the RITS complex to silence transcription of a complementary RNA (Wilson and Doudna, 2013).

The realization that siRNA could be repurposed as a tool to selectively silence a gene of interest ushered in a renaissance in genetics and medicine. Short interfering RNAs could be readily synthesized in high throughput at a low cost, which allowed for rapid and systematic interrogation of individual gene function (Kamath et al., 2003). Within the span of a decade, RNAi enhanced our understanding of genetic circuits within eukaryotes (Dietzl et al., 2007; Moffat et al., 2006); improved crops by reducing toxin levels and increasing resistance to pathogens (Perrimon et al., 2010); and achieved limited success in treating a range of diseases via targeted viral delivery of siRNAs (gene therapy) (Yin et al., 2014).

Despite its advantages, RNAi suffers from a number of drawbacks. First, gene silencing by RNAi is only effective for the lifetime of the siRNA. Thus, a recurring supply of the appropriate siRNA(s) is required to achieve continuous silencing. Second, not all siRNA sequences are equally effective at silencing. Knockdown levels can range from 0-100%, and variation can be extremely high across multiple replicates with significant off-target effects (Jackson et al., 2003). Third, because RNAi technology hijacks the natural machinery of the host cell, there may be unintended artifacts associated with disturbing cellular homeostasis that are not directly related to the gene of interest. Most importantly, it has been shown in many cases that a knockdown is not genetically equivalent to a full knockout, making RNAi experiments difficult to draw meaningful causal conclusions from (Shalem et al., 2015; Wang et al., 2014).

A more direct way to study gene function is to mutate or excise the DNA associated with its expression (genome editing). This has the advantage of permanently silencing or altering the gene while minimizing indirect disruption to other cellular processes (Carroll, 2016). While a number of technologies have been developed general pipeline for genome editing is as follows: (a) A DNA-cleaving enzyme (a

nuclease) generates a double-stranded break (DSB) at the site of interest; (b) The DSB triggers recruitment of DNA repair enzymes and/or a short donor DNA sequence to be integrated at the break site; (c) The break is repaired, either with or without the donor DNA (Urnov et al., 2010) (Figure 1.1).

The vast majority of DNA repair occurs through one of two pathways: non-homologous end joining (NHEJ) or homology directed recombination (HDR) (Burma et al., 2006; Haber, 2000; Iyama and Wilson, 2013; Liang et al., 1998; Mehta and Haber, 2014). In mammalian NHEJ, the MRN complex– consisting of Mre11, Rad50, and Nbs1– is recruited to the break site to bridge both ends of the DNA (van den Bosch et al., 2003). Binding of the MRN complex results in recruitment of the ATM kinase complex, which locally phosphorylates several proteins– including CHK2, NBS1, and the histone variant H2AX– and effects recruitment of the X family DNA polymerases λ and μ (Pols λ and μ, respectively) (Iyama and Wilson, 2013). Pols λ and μ fill in 5'- and 3'-gaps at the break sites, allowing for the DNA ligase IV complex to ligate the two ends to fully repair the break (Grawunder et al., 1998).

Mammalian HDR similarly begins with recruitment of the MRN complex, which is followed by end resection by EXO1 and BLM nucleases to generate 3'-single strands (Mehta and Haber, 2014). The single-stranded ends are bound by Rad51 proteins, which aid in invasion of another DNA strand that shares a high degree of sequence similarity to the single-stranded ends (Mehta and Haber, 2014). The invaded DNA strand– either a chromosome or an exogenously introduced oligonucleotide– serves as a template for DNA polymerase to fill in gaps via a Holliday junction, which is resolved via nuclease and ligase activity to generate repaired, recombinant DNA strands.

NHEJ, HDR, and other DNA repair pathways compete with each other in the context of a living cell (Iyama and Wilson, 2013). NHEJ is thought to occur more frequently in mammals but is prone to high error rates. HDR occurs less frequently but with higher fidelity. From a genome editing perspective, HDR offers the additional advantage of allowing for incorporation of non-native DNA sequence information at a locus of interest via exogenous DNA donors that bear homology arms matching the break site.

Historically, targeted DNA editing has been extremely challenging given the vast size and redundancy of eukaryotic genomes (International Human Genome Sequencing Consortium, 2001; Celera Genomics, 2001). A major breakthrough occurred in 1991 with the publication of the first crystal structure of a zinc finger DNA-binding domain (ZF) (Pavletich and Pabo, 1991). This seminal structure demonstrated a generalizable mode of 3 base pair (bp) recognition by individual fingers and outlined a path forward for engineering ZFs to bind specifically to arbitrary DNA sequences of interest. Less than one decade later, Bibikova, Carroll, *et al.* demonstrated targeted genome editing by fusing the cleavage domain of FokI nucleases to a sequence-specific ZF (ZFN) (Bibikova et al., 2001; 2002; Smith et al., 2000). When two ZFN proteins bound to DNA within proximity to each other, the FokI domains dimerized resulting in site-specific cleavage.

Despite their promise, adoption of ZFNs for genome editing in research labs was relatively limited primarily because of the intractability of engineering new ZFNs for every sequence of interest. Not all fingers targeted their corresponding nucleic acid triplets with equal levels of specificity, and in some cases the amino acid sequence of a

preceding finger would unexpectedly alter the specificity and/or nucleotide preference of adjacent fingers (Gabriel et al., 2011; Pattanayak et al., 2011). Moreover, cloning and testing ZFNs for every new DNA sequence of interest proved extremely cumbersome.

The mid 2000s ushered in renewed interest in genome editing with the discovery of transcription activator-like effector proteins (TALEs) in *Xanthomonas* bacteria (Boch et al., 2009; Moscou and Bogdanove, 2009). TALE proteins consist of variable ~34 amino acid motifs chained linearly in sequence space. Each of these motifs preferentially binds to one nucleotide, resulting in a larger sequence preference across the entire TALE. Analogous to zinc finger domains, fusion of TALEs to FokI nuclease domains (TALENs) allowed for site-specific genome editing (Christian et al., 2010; Miller et al., 2010).

TALENs proved far easier to engineer than ZFNs, and they were rapidly adopted by the agricultural community to improve crops (Li et al., 2012; Wood et al., 2011). Despite intense industry interest, academics were sluggish to pick up the technology, primarily because of the difficulty of cloning extremely large plasmid vectors for every new target genomic sequence. The large amount of DNA real estate required for each TALE also rendered them impractical for gene therapy applications, since AAV viral vectors have a cargo limit of ~4.5 kilobases (kb) (Kotterman and Schaffer, 2014).

Arguably the most important advance in genome editing to date was the discovery and biochemical characterization of the Cas9 protein from *Streptococcus pyogenes* in 2012 (Gasiunas et al., 2012; Jinek et al., 2012). Derived from the bacterial CRISPR immune system, Cas9 is a site-specific endonuclease that targets DNA based on complementarity to a guide RNA (sgRNA) that it carries. Rather than designing a new protein for every new target, site-specific DNA cleavage is achieved by tuning the sequence of sgRNA. The low cost and ease of short RNA synthesis made CRISPR-Cas9 the first genome editing tool to be amenable to high-throughput knockout experiments within the confines of an academic research lab, and as such it has spawned a revolution in genome engineering that is unrivaled by any other genome editing technology (Lin et al., 2014; Richardson et al., 2016).

A CRISPR Perspective on Genome Engineering

The story of CRISPR begins with fundamental investigations of genomic sequences in bacteria. In the 1980s, Japanese researchers reported a series of nearly identical inverted repeat sequences within the *E. coli* genome interspaced between variable ~30 nt sequences, termed spacers (Chen et al., 2013; 2016a). Over the next two decades, a handful of researchers identified hundreds of similar genomic motifs– termed CRISPRs (Clustered Regularly Interspaced Short Palindromic Repeats)– across a plethora of unrelated species of bacteria and archaea (Chen et al., 2013; 2016b; Qin et al., 2017; Shao et al., 2016). Further genomic analysis of spacer sequences led to the startling realization that they mapped to foreign viral and plasmid sequences, which in turn led to the hypothesis in 2005 that CRISPRs serve as an immune system to protect bacteria from foreign nucleic acids (Chen et al., 2013). Only two years later, scientists definitively demonstrated via phage challenge experiments that transcription of CRISPR loci spacers conferred immunity against viruses via RNA-guided interference (Chen et al., 2016b).

While CRISPR systems vary widely in terms of repeat sequence as well as CRISPR-associated (Cas) genes, the CRISPR immune response can broadly be divided into three stages (Deng et al., 2015; Guan et al., 2017; Ma et al., 2015; 2016; Qin et al., 2017; Shao et al., 2016; Wang et al., 2016). Upon initial viral infection, small fragments (~20-50 nucleotides) from the viral genome must be recognized, excised, and integrated into the host genome to provide a molecular memory of infection (adaptation/acquisition). If re-infected, the entire CRISPR repeat-array is then transcribed from an AT-rich promoter and processed by RNA nucleases to generate mature crRNAs (crRNA biogenesis/processing). These mature crRNAs serve as templates to guide Cas proteins to nucleolytically degrade complementary, foreign nucleic acids (interference).

Of the three stages of CRISPR immunity, acquisition is the most conserved and generally requires the signature genes Cas1 and Cas2 (Konermann et al., 2015; Qin et al., 2017; Shao et al., 2016; Wang et al., 2016). Type II systems additionally require the interference-related protein Cas9, which assists Cas1 and Cas2 in protospacer recognition via binding to the protospacer adjacent motif (PAM) (Qin et al., 2017). Cas1 and Cas2 form a multimeric complex that recognizes and cleaves short fragments of foreign nucleic acid proximal to a short, 2-5 nucleotide PAM (Chen et al., 2016c; Esvelt et al., 2013; Ma et al., 2015; 2016). The processed protospacer is then integrated at the leader-adjacent repeat of the host genome via nucleophilic attack by the 3'-OH of the protospacer.

Biogenesis of crRNAs is evolutionally more diverse across different CRISPR systems (Charpentier et al., 2015; Hochstrasser and Doudna, 2015; Li, 2015). In most Type I and Type III systems, a dedicated endonuclease (Cas6) binds to pre-crRNA inverted repeats with femtomolar affinity, which then triggers cleavage ~5-8 nucleotides (nt) upstream of an individual spacer element via a molecular ruler mechanism (Carte et al., 2008b; Haurwitz et al., 2010; 2012). In Type I-C systems, which lack Cas6, processing is performed by Cas5 endonucleases in a mechanism that is less well understood due to lack of a target-bound crystal structure (Garside et al., 2012; Nam et al., 2012). Type II systems do not require a dedicated processing endonuclease and instead couple a general host endonuclease (e.g. RNase III) to a partially complementary trans-encoded RNA (tracrRNA) to generate mature crRNAs (Deltcheva et al., 2012). In other systems, crRNA biogenesis and RNA-guided interference activities are compacted into two separate catalytic domains within the same effector enzyme (e.g. CpfI in Type V systems) (Fonfara et al., 2016). The incredible diversity of processing mechanisms across bacterial phyla is a striking example of convergent evolution in biology.

Similarly to crRNA biogenesis, RNA-guided interference in CRISPR systems varies widely both in terms of Cas genes as well as substrate preference (van der Oost et al., 2014; Wright et al., 2016). Type I systems preferentially target dsDNA via a multi-subunit, 300+ kD Cascade complex, which consists of several different Cas genes arranged in a seahorse-like structure (Brouns et al., 2008; Jackson et al., 2014; Jore et al., 2011; Mulepati et al., 2014). The molecular composition of Cascade itself varies across different species, although the signature protein Cas3 is common to all of them (Makarova et al., 2015). Structural and biochemical studies have demonstrated that Cas3 possesses both helicase and nuclease activity, while the other Cascade proteins

serve as a scaffold for appropriately positioning the crRNA and target DNA for target degradation via PAM recognition and R-loop formation (Hochstrasser et al., 2014; Huo et al., 2014; Jackson et al., 2014; Mulepati et al., 2014; Sinkunas et al., 2011). Analogous to Type I systems, Type III systems require multisubunit Csm or Cmr complexes for interference but preferentially target RNA as well as actively transcribing DNA (Hale et al., 2009; Liu et al., 2017b; Rouillon et al., 2013; Samai et al., 2015; Staals et al., 2013; Taylor et al., 2015).

Several CRISPR systems (e.g. Type II, V, and VI) have distilled the process of nucleic acid interference to a single polypeptide, the most prominent of which is Cas9 (Abudayyeh et al., 2016; East-Seletsky et al., 2016; Gasiunas et al., 2012; Jinek et al., 2012; Shmakov et al., 2015; Zetsche et al., 2015). In 2012, several research laboratories independently identified Cas9 as sufficient for RNA-guided dsDNA interference in Type II CRISPR systems (Gasiunas et al., 2012; Jinek et al., 2012). The protein required both tracrRNA and crRNA, which could be concatenated into a single guide RNA (sgRNA) via a short 4-nt linker loop (Jinek et al., 2012). These researchers further showed that Cas9 could be programmed to specifically target arbitrary DNA sequences by varying a short 20-nt region within the sgRNA.

The potential of this technology for genome editing was immediately realized. Within less than 6 months, three research labs independently demonstrated that Cas9 could be utilized to make targeted DSBs within the genomes of mammalian cells (Cong et al., 2013; Jinek et al., 2013; Mali et al., 2013). By co-transfecting cells with short, homologous DNA oligonucleotides, researchers could further incorporate new sequence information at the break site via homology-directed repair.

These seminal papers catalyzed a renaissance in genome engineering. Within the span of just one year, CRISPR technology was repurposed for genome editing in a plethora of living organisms, including bacteria, yeast, flies, rodents, zebrafish, primates, pigs, and plants, among others (DiCarlo et al., 2013; Gratz et al., 2013; Hai et al., 2014; Hwang et al., 2013; Niu et al., 2014; Shan et al., 2013; Wang et al., 2013). The technology demonstrated a remarkable robustness to different cell types enabling rapid and high-throughput generation of genetic knockout models for research (Shalem et al., 2015; Wang et al., 2014). Moreover, the low cost and ease of synthesizing sgRNAs arguably made it the first genome editing technology amenable to use by a wide academic research audience (Doudna and Charpentier, 2014).

Outside of the research laboratory, CRISPR has generated great commercial and medical interest, and a number of applications are actively being explored (Doudna and Charpentier, 2014; Hsu et al., 2014a; Terns and Terns, 2014; Wright et al., 2016). Because of its efficacy across many different cell types, CRISPR-Cas9 is being considered for gene therapy-based disease treatment (Xiao-Jie et al., 2015). It has shown particular promise for blood conditions that are genetically well characterized (e.g. sickle cell anemia), given the ease of intravenous delivery to hematopoietic cells (Traxler et al., 2016; Ye et al., 2016). Cas9 has already been adopted for re-engineering livestock and plant genomes to improve yields and pathogen resistance, and genome engineering in pigs is additionally being considered as a means of streamlining the process of organ donation (Shan et al., 2013; Yang et al., 2015). By humanizing antigens and receptors via editing, porcine organs could potentially be made compatible for transplantation into humans. Yet another real-world application is the use of Cas9 for

genetic control of wildlife populations (a gene drive), whereby an artificial gene of interest is propagated over many generations to dominate over a less desirable wild-type gene (Gantz et al., 2015; Hammond et al., 2015). The quintessential example is the introduction of long-term malarial resistance into mosquito populations. Briefly, a DNA sequence encoding the malarial resistance gene, Cas9, and sgRNA is incorporated into the genome of a parent mosquito. The Cas9-sgRNA complex then targets DNA that does not contain the resistance gene but leaves DNA that does intact. After many generations and subsequent DNA repair cycles by homologous recombination, the resistance gene becomes dominant within the population.

Beyond any other application of CRISPR, editing of human embryos has received the greatest amount of attention and discussion (Bosley et al., 2015; Evitt et al., 2015; Lanphier et al., 2015; 2015). Can CRISPR be used to correct mutations or enhance desirable traits within the germline? If so, how would appropriate boundaries be drawn to avoid dystopian, unethical outcomes? If editing effects negative outcomes in a human being, can it be reversed past the embryo stage? There are no obvious answers to these questions. Written perspectives by the scientific community, ethics panels, and international summits have provided some clarity on the matter, but these conversations must continue to happen as societal perspectives evolve and as our collective understanding of the genetic landscape grows (Bosley et al., 2015; Evitt et al., 2015; Baltimore et al., 2015).

Dissecting the Dissector

Concomitant with the CRISPR genome engineering revolution has been a series of careful mechanistic studies to better understand the structure and function of Cas9 itself (Nishimasu and Nureki, 2017). How much RNA-DNA complementarity is required for Cas9 to make a double-stranded break? What is the 3-dimensional structure of Cas9? Which amino acid residues are catalytic? How often does Cas9 cut at a mismatched site? Answers to these questions as well as others have over the years provided researchers with a better understanding of the limitations of CRISPR-Cas9 and have helped to inform engineering efforts to make CRISPR a more efficient and reliable tool (Kleinstiver et al., 2016; Slaymaker et al., 2015).

Early biochemical experiments identified the core molecular components required for the *S. pyogenes* Cas9 cleavage reaction (Jinek et al., 2012). Critically, Cas9 requires a 3-nt PAM (NGG) on the non-complementary DNA strand proximal to the cleavage site. This PAM serves as an anchor for Cas9 to initialize contact with the target nucleic acid and allows it to discriminate between self versus non-self DNA. Following PAM contact and DNA melting, the complementary strand base pairs with the crRNA or sgRNA, resulting in cleavage of both strands. Mutational analysis showed that a catalytic HNH nuclease histidine cleaves the complementary strand, while a catalytic RuvC nuclease aspartate cleaves the non-complementary strand (H840A and D10A in *S. pyogenes*, respectively) (Jinek et al., 2012). In vitro single-molecule and gel-shift experiments revealed that Cas9 binds to its target DNA with extremely high affinity (K_d < 1 nM) and dissociates on the order of hours after cleavage in vitro (Sternberg et al., 2014).

Crystal and electron micrograph structures illustrate that Cas9 adopts a bi-lobed architecture, with a positively charged cleft situated between alpha-helical recognition

(REC) and nuclease (NUC) lobes for accommodating nucleic acids (Anders et al., 2015; Jiang et al., 2016; 2015; Jinek et al., 2014; Nishimasu et al., 2014). These structures showcase the dramatic ~100° relative rotation between the two lobes that occurs upon RNA and substrate binding to form an R-loop competent for cleavage (Jiang et al., 2016; Jinek et al., 2014; Szczelkun et al., 2014). A series of positively charged amino acid residues engage in ionic contacts with the PAM, the sgRNA, and the backbone of the 18-20 base pairs of the DNA duplex that map to the sgRNA (Anders et al., 2015; Jiang et al., 2015; 2016). Critically, these residues highlight the importance of the seed region, the ~8-12 base pairs of the DNA duplex immediately proximal to the PAM (Jinek et al., 2012; Kuscu et al., 2014; Sternberg et al., 2014). Cas9 target recognition occurs through PAM recognition followed by ATP-independent 3'-5' melting of the DNA double helix via Brownian ratcheting (Sternberg et al., 2014); thus, the protein is especially sensitive to mutations within this seed region. As a corollary, numerous studies have concluded that off-target cleavage events correlate with homology in this region.

Cas9 maintains its specific DNA binding properties in the absence of its nuclease activity (Jinek et al., 2012; Sternberg et al., 2014). As such, the catalytically dead version of the protein, termed dCas9, has been extensively explored for non-editing applications (Doudna and Charpentier, 2014; Hsu et al., 2014a). Early efforts focused on attenuating transcription at a specific locus based on dCas9 occupancy at core promoters (CRISPRi) (Gilbert et al., 2013; Larson et al., 2013). Upon recognizing that both the C-terminus of Cas9 and the linker region of the sgRNA were amenable to modification, this grew into fusing dCas9 to: (a) transcriptional activator and repressor domains to dynamically alter expression levels in a high-throughput fashion (Gilbert et al., 2014; Liu et al., 2017a; Zalatan et al., 2015); (b) fluorescent proteins to image genomic loci in vivo (Chen et al., 2013; 2016c; Deng et al., 2015; Qin et al., 2017); and (c) DNA and histone modifying enzymes, such as histone acetylases and DNA methylases, to selectively modulate the epigenetic landscape at a locus of interest (Hilton et al., 2015; Liu et al., 2016).

How does the bacterial Cas9 work so effectively as a genome editing tool in higher organisms? DNA within the context of a eukaryotic cell poses a number of unique challenges to a bacterial protein, including compartmentalization of DNA within a nucleus, histones and other DNA binding proteins, and a genome space that is redundant and orders of magnitude larger than that of *S. pyogenes* (Bartholomew, 2014; International Human Genome Sequencing Consortium, 2001; Celera Genomics, 2001; Kadonaga, 1998). A number of genomics studies have attempted to more thoroughly investigate the relationship between Cas9 and chromatin. Genome occupancy studies of dCas9 confirmed the mechanistic importance of the seed region in vivo and suggest a genome sampling bias in favor of open chromatin (Kuscu et al., 2014; O'Geen et al., 2015; Wu et al., 2014). Other deep-sequencing approaches have better contextualized the frequency and nature of off-target cleavage within eukaryotic cells, resulting in more robust rules for sgRNA design (Frock et al., 2014; Koo et al., 2015; Naito et al., 2015; Singh et al., 2015; Tsai et al., 2014). Other work has explored the influence of cell cycling by synchronizing cells and precisely timing delivery of Cas9 ribonucleotide protein (RNP) complex (Gutschner et al., 2016; Lin et al., 2014). Despite mechanistic gains from these studies, none of them provide a real-time, dynamic picture of the Cas9 search process in vivo. Questions about in vivo binding times, the ability of

Cas9 to navigate heterochromatin, the target search time, and the relative amount of sliding versus 3D diffusion are beyond the scope of biochemistry and genomics.

Seeing is Believing

The emergence of imaging technologies has in recent years vastly enhanced our understanding of biological systems (Buxbaum et al., 2014; Diezmann et al., 2017; Lippincott-Schwartz et al., 2001; Liu et al., 2015). Imaging offers several advantages over biochemical and genomics approaches including: (a) spatiotemporal resolution of molecular events; and (b) visualization of biological processes in the context of a living cell. At the core of the vast majority of imaging techniques is the principle of fluorescence, whereby an electron from a small-molecule (fluorophore) is promoted to a higher energy state by a specific wavelength of excitation light (Lichtman and Conchello, 2005). As this electron relaxes to its ground state configuration, it emits light on the timescale of nanoseconds that is red-shifted relative to the excitation wavelength (Stokes shift). It is this characteristic emission that enables real-time visualization of that molecule under a microscope.

Historically, the utility of imaging to answer complex biological questions has been by limited by resolution constraints. As light passes through a lens via a medium (e.g. air, water, or oil), its wavelength is shifted in a process called diffraction. Diffraction imposes a limit on the precise localization of a light-emitting molecule by increasing the apparent size of its emission profile. This limit is inversely proportional to the numerical aperture of the lens and is on the order of a few hundred nanometers for most visible-light fluorophores (Lichtman and Conchello, 2005).

Several technological developments over the last few decades have enabled physicists and biologists to go beyond the diffraction limit. The invention of confocal microscopy in the 1950s increased resolution via a pinhole that restricts light passing through the confocal plane (Webb, 1996). Later on, the discovery that sample illumination could be reduced to 200 nm planes using total internal reflection (TIRF) further pushed the resolution limit of biological imaging to ~50-200 nm, although this was limited to thin samples near the surface of the objective (Axelrod, 2001).

Concomitant with mechanical improvements to microscopes has been the development of better fluorophores. Since the discovery of GFP in the jellyfish *Aequorea victoria*, fluorescent proteins have been engineered to be brighter, photoactivatable, and to absorb and emit light at a range of wavelengths across the visible spectrum (Campbell et al., 2002; Frommer et al., 2009; Gurskaya et al., 2006; Nagai et al., 2001; Ormo et al., 1996; Patterson and Lippincott-Schwartz, 2002; Tsien, 1998). Additionally, chemical genetics has enabled biomolecule tagging with synthetic fluorophores via covalent and irreversible attachment to a *C*- or *N*-terminal enzymatic domain. Prominent examples include SNAP and HaloTag, which were engineered from naturally occurring DNA methyltransferase and dehalogenase domains to bind to synthetic fluorophores (Keppler et al., 2002; Los et al., 2008). These non-traditional tags have become increasingly popular in the wake of rapid improvements to the photophysical properties of synthetic dyes (Grimm et al., 2015; 2016; Lavis and Raines, 2008).

A major breakthrough in imaging occurred when scientists realized that individual fluorophores could be robustly resolved by exploiting time as a parameter for the

deconvolution of signal (Diezmann et al., 2017). While an individual fluorophore emits light in a point spread function (PSF) that is on the order of hundreds of nanometers in width, the center of this PSF is readily identifiable within ~1-2 nm of precision (Fig. 1.2a). Two molecules within 10 nm of each other will be impossible to resolve if they emit light simultaneously, but what if their emission is separated in time? Physicists began to explore this possibility starting in the late 1990s (Betzig, 1996; Hell and Wichmann, 1994); ten years later, several time-resolved super-resolution methods were independently reported. These techniques, which include STED, PALM, STORM and others, enabled for the first time ~10 nm resolution within fixed cells (Betzig et al., 2006; Dyba et al., 2003; Rust et al., 2006). The applicability of super-resolution microscopy to studying biological structures was immediately realized, and within less than ten years, STED and PALM were the subject of the 2014 Nobel Prize in chemistry.

The premise of a super-resolution experiment is conceptually simple. A photo-excited electron can relax via multiple mechanisms, one of which is fluorescence (Fig. 1.2b) (Lichtman and Conchello, 2005). Other modes include non-fluorescent vibrational relaxation as well as intersystem crossing to a parity-forbidden triplet state. From the triplet state, an electron may: (a) react with triplet oxygen (photobleach); (b) relax non-radiatively to its ground state; (c) relax radiatively to its ground state (phosphorescence); or (d) be promoted to its singlet excited state, where it may fluorescently relax to the ground state. Re-excitation to the singlet state is parity-forbidden and thus occurs slowly and stochastically. If all of the molecules are synchronously converted to the dark triplet state, they can be asynchronously converted to a singlet state at the appropriate excitation wavelength, leading to fluorescence. Thus, molecules that are spatially very close can be resolved in time at high resolution (Betzig, 1996; Betzig et al., 2006; Rust et al., 2006).

The core principle of super-resolution microscopy– namely, precise center-of-mass localization of PSFs– was later adapted to study the dynamics of single molecules within living cells. In a landmark set of experiments, the laboratory of Sunny Xie reported the first in vivo visualization of single transcription factors in *E. coli* (Elf et al., 2007). By expressing yellow fluorescent protein (YFP)-tagged lacI at low levels, Elf *et al.* were able to precisely locate and track TF dynamics, measure non-specific versus specific binding times, estimate diffusion coefficients, and quantify 1D-sliding along the DNA.

This seminal paper set the stage for a number of improvements to computational methods, instrumentation, and fluorophores that enabled single-molecule, single-cell studies in eukaryotic cells. Examples include: (a) monitoring the dynamics of the ubiquitous tumor suppressor protein p53 (Gebhardt et al., 2013); (b) unraveling the hierarchical assembly of transcription factors within embryonic stem cells (Chen et al., 2014); (c) elucidating differences in protein search modalities (Izeddin et al., 2014); (d) structural characterization of Huntington's aggregates within living cells (Li et al., 2016); and (e) discovery of a generalizable clustering behavior of transcription factors (Liu et al., 2014). Taken together, these studies highlight the utility of imaging to answer exciting questions about the structurally dynamic nature of biological systems.

CRISPR In Focus

In the context of eukaryotic genome editing, Cas9 must navigate chromatin and other structural elements in order to efficiently identify and cleave a DNA target. Our molecular understanding of this process has come largely from biochemical, structural, and genomics data. While these studies have provided mechanistic insights that have informed engineering efforts, they offer only a static picture of the CRISPR-Cas9 target search process in mammalian cells.

Imaging is uniquely poised to answer questions about dynamic molecular processes in vivo. In Chapter 2, we discuss the development and application of single-molecule imaging methods to track fluorescently labeled dCas9-HaloTag proteins in live mouse cells (Knight et al., 2015). To our knowledge, these experiments offer the first direct visualization of Cas9 searching in vivo. Our results provide evidence for a general, 3D-diffusion-dominated search mechanism in living cells and illuminate differences in how Cas9 explores open versus closed regions of the genome (euchromatin and heterochromatin, respectively).

Single-subunit interference proteins from other CRISPR systems have recently attracted interest as orthogonal genome editing tools. The Type V Cpf1 protein has expanded the CRISPR genome editing toolkit since its alternative PAM preference (5'-TTN-3') allows for targeting sequences that are not compatible with Cas9 (Fonfara et al., 2016; Yamano et al., 2016; Zetsche et al., 2015). In Type VI systems, C2c2 has been suggested as a potential tool for RNA-guided RNA targeting since its active domains (HEPN) preferentially cleave RNA over DNA (Abudayyeh et al., 2016; East-Seletsky et al., 2016; Shmakov et al., 2015). While Cas9 has been repurposed to target RNA, a CRISPR protein that natively targets RNA could be exploited to image RNA or to knock down mRNA in an RNAi-like fashion (Nelles et al., 2016; O'Connell et al., 2015).

In Chapter 3, we discuss mechanistic studies of C2c2 proteins from a variety of organisms (East-Seletsky et al., 2016). We outline two distinct enzymatic activities of C2c2– crRNA processing and RNA-triggered general RNase activity– mediated by separate domains within the protein. Collectively, this work enhances both our understanding of CRISPR biology as well as our ability to repurpose CRISPR systems for relevant, real-world applications.

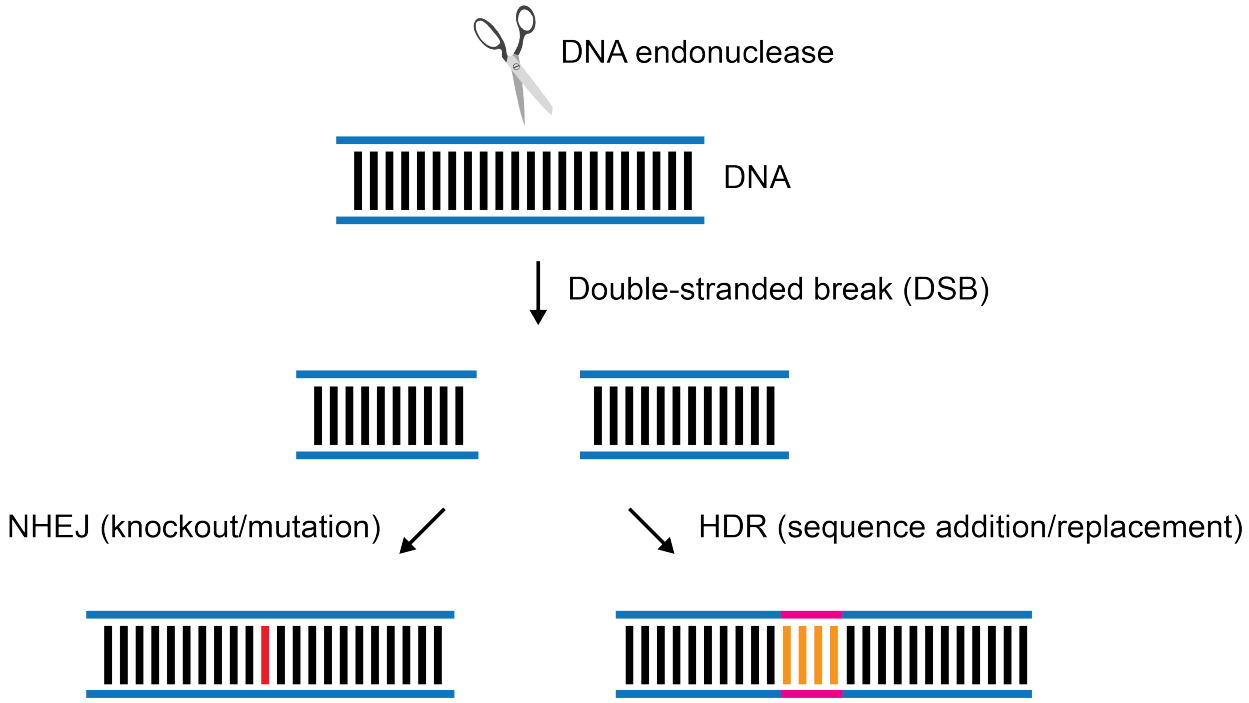
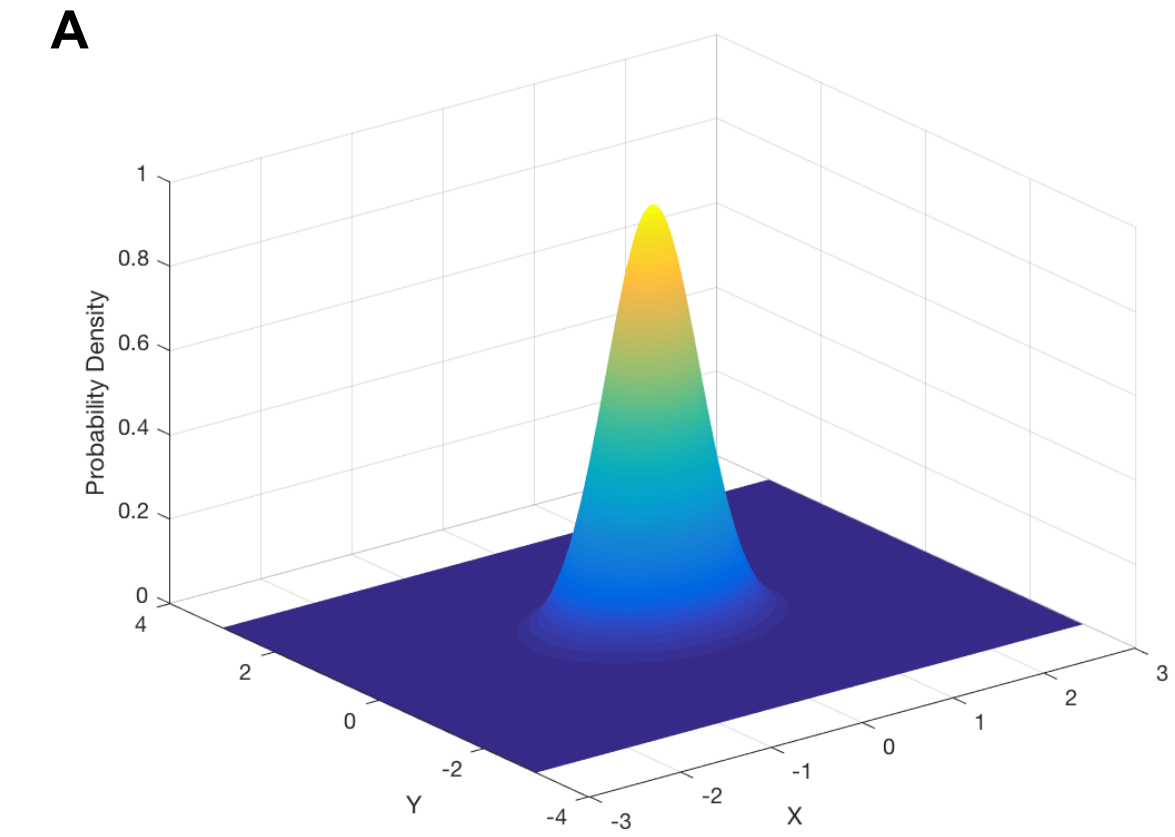


Figure 1.1. Common pipeline for a genome editing experiment. In the first step, a DNA endonuclease (e.g. ZFN, TALEN, or CRISPR-Cas9) makes a targeted double-stranded break (DSB) at a genomic locus of interest. The DSB triggers recruitment of cellular machinery to repair the break via one of several pathways, which can result in incorporation of new DNA information (pink/orange, HDR pathway) at the break site.


A
Probability Density
1
0.8
0.6
0.4
0.2
0
4
2
0
-2
-4
Y
3
2
1
0
-1
-2
-3
X

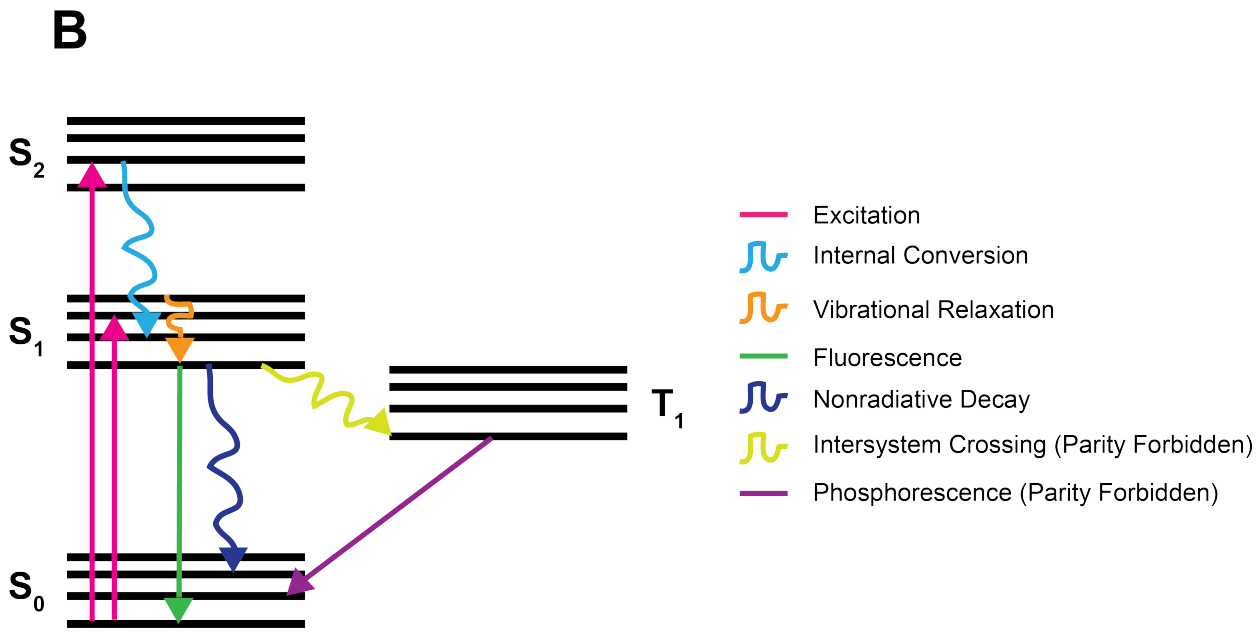

B
S2
S1
S0
T1
Excitation
Internal Conversion
Vibrational Relaxation
Fluorescence
Nonradiative Decay
Intersystem Crossing (Parity Forbidden)
Phosphorescence (Parity Forbidden)

Figure 1.2. Photophysical principles of fluorescence. (A) Mock illustration of a point spread function (PSF) for a light-emitting fluorophore. The center of the PSF can be localized with nanometer-level precision in the absence of convolution from proximal fluorophores. **(B)** Jablonski diagram illustrating a subset of transitions that an electron may undergo upon being excited to a higher energy state. In a superresolution experiment, molecules are synchronously switched to a dark triplet state (T_1) and then stochastically re-excited to a singlet state (S_1) in a parity-forbidden transition. Relaxation of a single molecule from the singlet state results in a characteristic PSF as shown in Panel A.

Chapter 2: Dynamics of CRISPR-Cas9 genome interrogation in living cells

The RNA-guided CRISPR-associated protein Cas9 is used for genome editing, transcriptional modulation and live-cell imaging. Cas9-guide RNA complexes recognize and cleave double-stranded DNA sequences based on 20-nucleotide RNA-DNA complementarity, but the mechanism of target searching in mammalian cells is unknown. Here we use single particle tracking to visualize diffusion and chromatin binding of Cas9 in living cells. We show that three-dimensional diffusion dominates Cas9 searching in vivo, and off target binding events at PAMs and short seed sequences are on average short-lived (milleseconds to seconds). Searching is dependent on local chromatin environment, with less sampling and slower movement within heterochromatin. These results reveal how the prokaryotic Cas9 protein interrogates mammalian genomes and navigates eukaryotic chromatin structure.

This work was done collaboratively and was originally published in *Science* Magazine:

Knight, S.C., Xie, L., Deng, W. Guglielmi, B., Witkowsky, L.B., Bosanac, L., Zhang, E.T., El Beheiry, M., Masson, J.-B., Dahan, M., Liu, Z., Doudna, J.A., Tjian, R. Dynamics of CRISPR-Cas9 genome interrogation in living cells. Science *350*, 823-826 (2015).

Co-authors have consented to reprinting the original publication for this thesis. Reprinted with permission from AAAS.

Introduction

The RNA-guided endonuclease Cas9 uses RNA-DNA complementarity to target and cleave double-stranded DNA upstream of a protospacer adjacent motif (PAM) (Gasiunas et al., 2012; Jinek et al., 2012). Cas9 can be programmed with a single guide RNA (sgRNA) to cleave specific DNA sequences within eukaryotic cells, facilitating its use as a tool for genome engineering (Doudna and Charpentier, 2014; Hsu et al., 2014b; Terns and Terns, 2014). Biochemical and genome occupancy studies have established the PAM and adjacent ~5-8 base pairs (the "seed" region) of the DNA target site as the basis for Cas9 DNA interrogation and off-target activity (Anders et al., 2015; Hsu et al., 2013; Jinek et al., 2014; 2012; Kuscu et al., 2014; Nishimasu et al., 2014; Pattanayak et al., 2013; Sternberg et al., 2014; Wu et al., 2014). Nonetheless, how Cas9 explores large eukaryotic genomes and identifies targets within the context of chromatin remains largely unknown. In particular, the in vivo kinetics of on- versus off-target binding and Cas9 dependence on chromatin environment have not yet been examined in living eukaryotic cells.

Results and Discussion

To investigate the live-cell dynamics of Cas9 target searching, we tracked single, fluorescently-labeled, catalytically-inactive *Streptococcus pyogenes* Cas9 (dCas9) molecules to determine their diffusion and chromatin binding properties in live mouse cell nuclei (Chen et al., 2013). dCas9 was fused at its *C*-terminus with a HaloTag domain and stably integrated into the genome of NIH 3T3 cells under a doxycycline-inducible TRE-tight promoter (Fig. 2.1A, figs. A1.1-A1.3) (Los et al., 2008). Guide RNAs were transiently expressed from a BFP reporter plasmid. Covalent linkage of a cell-permeable, fluorescent HaloTag ligand (JF549) allowed for visualization of single Cas9-HaloTag molecules under leaky expression (Figs. 2.1A & B, fig. A1.4) (Grimm et al., 2015).

To study dCas9-HaloTag binding dynamics at endogenous genomic loci, we transfected cells with a guide RNA targeted to B2 SINEs (short interspersed nuclear elements). The B2 elements are repeated ~350,000 times throughout the mouse genome, often in intragenic regions, with a single element per insertion site (Espinoza et al., 2007; Jurka et al., 2005). We reasoned that the abundance of these loci would shift the global equilibrium of Cas9-HaloTag binding, allowing us to observe otherwise rare target binding events. Two-photon fluorescence correlation spectroscopy (FCS) experiments revealed a significant reduction in global dCas9-HaloTag mobility for B2 sgRNA-transfected cells relative to apo (no sgRNA) protein (Fig. 2.1C). Both apo and B2-loaded dCas9-HaloTag displayed biphasic kinetic behavior in our FCS measurements, reflecting slowly and rapidly moving populations. The magnitude of the diffusion coefficient for the slow population was ~45-fold lower for B2-loaded Cas9 compared to the apo protein ($0.006 \pm 0.003\ \mu m^2\ s^{-1}$ vs. $0.26 \pm 0.13\ \mu m^2\ s^{-1}$, fig. A1.5).

We conducted 2D tracking experiments at short (10 *ms*) exposure times in cells transfected with a plasmid encoding either B2 or phage-derived "nonsense" guide bearing minimal homology to the 3T3 genome (figs. A1.6-A1.8). A nonsense sgRNA has the potential to direct Cas9 off-target interactions through millions of PAMs and short seed sequences within the genome and thus serves as a proxy for a Cas9 protein in the process of searching (Jinek et al., 2012; Sternberg et al., 2014). The log diffusion

coefficient histograms for these two sgRNAs showed a significant fraction of highly immobile (D < 0.1 $\mu m^2\ s^{-1}$) Cas9 molecules for B2 sgRNA relative to nonsense sgRNA or no-guide controls, consistent with more chromatin binding for the B2-loaded Cas9 (Fig. 2.2A; fig. A1.9). In similar experiments, a B2 guide with mismatches proximal to the target PAM gave rise to Cas9 diffusion histograms similar to the nonsense guide; in contrast, a B2 guide with homology mismatches distal to the target PAM gave rise to a distribution more similar to the cognate B2 guide (B2_0M and B2_13M, respectively, Fig. 2.2A, fig. A1.6). These observations are consistent with the role of the seed region in driving Cas9's RNA-guided interaction with DNA (Kuscu et al., 2014; Sternberg et al., 2014; Wu et al., 2014).

Compared to a binding dominant protein (e.g., H2B) or to a protein that demonstrates a mixture of binding and diffusion (e.g. Sox2), both the nonsense-loaded and apo Cas9 showed significantly more apparent 3D diffusion in cell nuclei (Fig. 2.2B; fig. A1.9). In addition, 3D multifocus tracking experiments with the nonsense guide showed that Cas9-guide-RNA complexes employ diffusion-dominated target searching throughout the entirety of the cell nucleus (Fig. 2.2) (Abrahamsson et al., 2012). These results underscore the dominance of 3D diffusion over binding during DNA interrogation by Cas9, demonstrating an in vivo target search mechanism similar to what has been observed in vitro (*6*).

To determine the relative kinetics of on- vs. off-target binding, we measured in vivo residence times of dCas9-HaloTag molecules bound to chromatin. We performed time-lapse experiments at a constant exposure time (20 *ms*) while varying the lapse time between successive frames (Fig. 2.3A). From these movies, we plotted the probability that a dCas9-HaloTag molecule would remain stationary as a function of time (survival probability, Fig. 2.3B). Re-scaling and concatenation of these plots allowed us to extract an average off-target residence time of 0.75 ± 0.1 *s* for Cas9 containing nonsense guide (τ_{ns}, fig. A1.10) (Gebhardt et al., 2013; Normanno et al., 1AD). We note that a small fraction of the binding events in our concatenated plot were longer than 10 *s,* which might be attributed to rare genomic sequences with higher homology to the nonsense guide (fig. A1.10) (Sternberg et al., 2014; Szczelkun et al., 2014). We also measured the binding of nonsense-loaded protein in dCas9-eGFP stable cell lines using fluorescence recovery after photobleaching (FRAP), a bulk technique for assessing protein mobility based on exchange between bleached and unbleached molecules within a region of interest. We observed nearly full recovery within 10 *s*, indicating mostly transient (milliseconds to seconds) chromatin interactions intermixed with diffusion (Figs. 2.3C & D) (Sprague et al., 2004).

Although nonsense guide-loaded dCas9-eGFP recovered rapidly after photobleaching in our FRAP curves, the B2 guide-loaded protein resulted in a large immobile fraction even when measured out to 5 min (Figs. 2.3C & D). Similarly, survival probability plots of B2 guide-loaded Cas9 showed substantially longer residence times compared to those with nonsense guide (Fig. 2.3B). These data suggest that Cas9 binding at bona fide targets (τ_s) could be significantly longer (e.g., minutes or more) in vivo relative to short-lived (milliseconds to seconds) binding typical of PAMs and very short seed sequences (τ_{ns}) (Sternberg et al., 2014). We refrain from more precisely estimating τ_s here due to: (1) a likely mixture of off-target and on-target binding in the immobile fraction, (2) imaging limitations due to photobleaching in our single-molecule

measurements (curved tails, Fig. 2.3B) and (3) known complications with extracting residence times from FRAP data (Mueller et al., 2010).

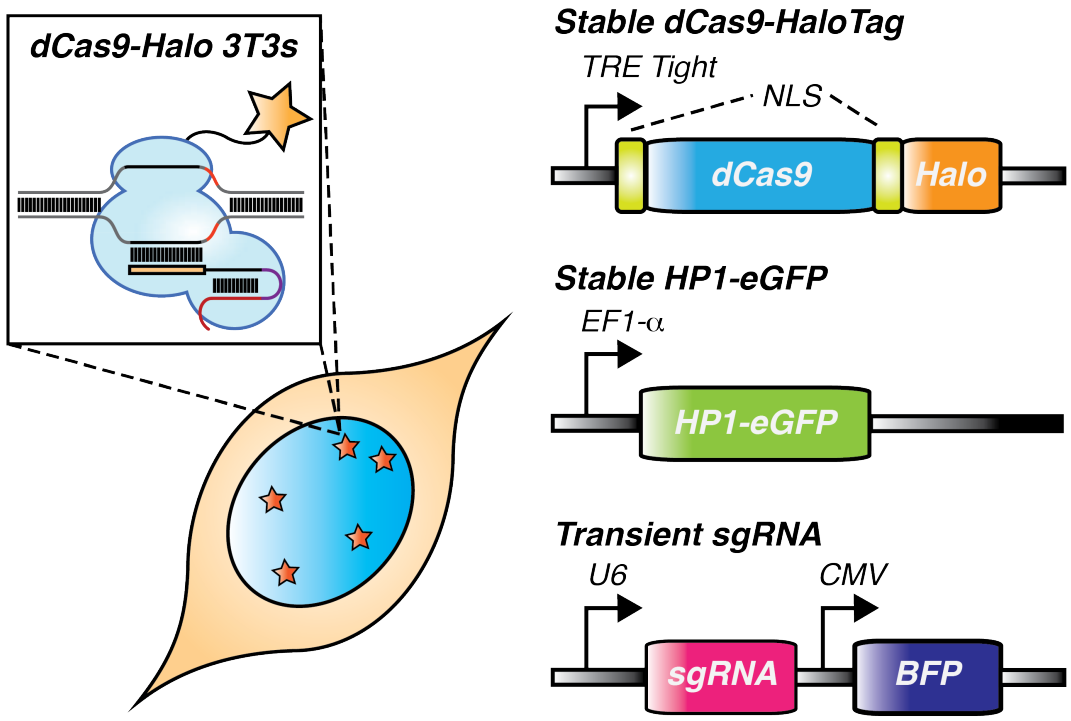
The ability of Cas9 to target heterochromatic regions (HRs) is important for its application to genome editing. To study Cas9 behavior in HRs, we performed tracking experiments in cells with eGFP-labeled heterochromatin protein 1 (HP1, fig. A1.11, Methods) (Eissenberg and Elgin, 2015; Liu et al., 2014; Manley et al., 2008). dCas9-HaloTag molecules with nonsense sgRNA were stochastically excited and tracked in live-cell nuclei, and the trajectories were overlaid onto HP1-labeled nuclear images to visualize searching with respect to heterochromatin. The resulting composite image shows significant depletion of tracks within HRs (30 ± 9% track density reduction, fig. A1.11). Diffusion analysis of tracks within HRs revealed that dCas9 diffusion is moderately slower in these regions (Fig. 2.4A; fig. A1.12) (Beheiry et al., 2015). We also performed jumping angle analysis on three-point sliding windows of our Cas9 trajectories to monitor the anisotropy of searching in HRs (Methods) (Izeddin et al., 2014). The resulting angle distributions revealed a slight bias towards reverse (180°) angles, suggesting more compact exploration and a tendency of Cas9 to return to its starting point while interrogating heterochromatin (Fig. 2.4B, fig. A1.13). Together, these results show that Cas9 search efficiency is reduced, but not eliminated, in HRs.

To test whether dCas9 can bind to target sites in heterochromatin, we transfected cells expressing dCas9-HaloTag with a plasmid encoding a sgRNA targeted to pericentromeric DNA sequences within heterochromatin. We observed distinct puncta within HRs of fixed cells colocalized with dense Hoechst staining, consistent with successful dCas9 targeting to pericentromeres (Fig. 2.4C). This result strongly suggests that Cas9 is able to bypass chromatin obstacles and faithfully engage with HR target sites despite reduced sampling efficiency within these regions.

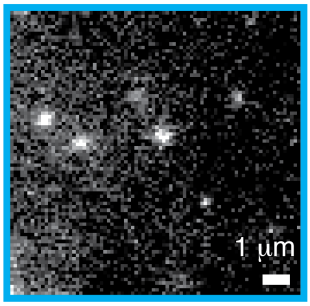
Conclusion

Our data provide a direct visualization of DNA interrogation by Cas9 in mammalian cells. The target search mechanism involves rapid three-dimensional diffusion of Cas9 around the nucleus, with occasional forays into heterochromatic regions. Our imaging approach complements chromatin immunoprecipitation (ChIP) experiments by capturing many of the more transient interactions with DNA that predominate as Cas9 scans vast mammalian genomes in search of its target site. Overall, our results provide a quantitative understanding of Cas9 dynamics in living cells and offer insight into how Cas9 navigates hierarchical organization of DNA within a eukaryotic nucleus.

A


dCas9-Halo 3T3s
Stable dCas9-HaloTag
TRE Tight
NLS
dCas9
Halo
Stable HP1-eGFP
EF1-α
HP1-eGFP
Transient sgRNA
U6
CMV
sgRNA
BFP

B


1 μm

C

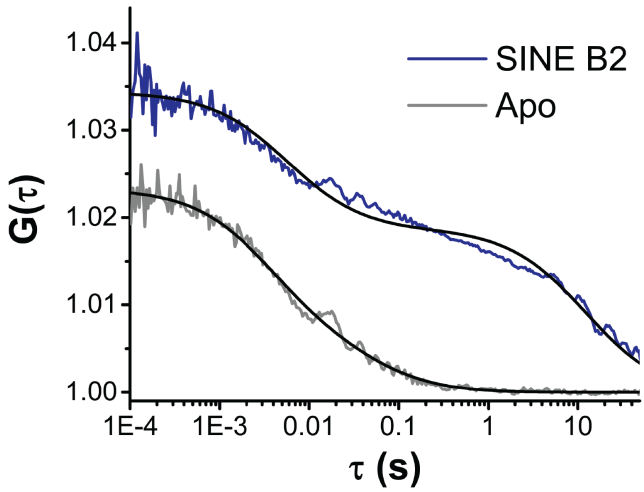

SINE B2
Apo
G(τ)
1.04
1.03
1.02
1.01
1.00
1E-4
1E-3
0.01
0.1
1
10
τ (s)

Figure 2.1. Visualization of single dCas9 molecules in living cells. (A) Overview of the imaging system for tracking single dCas9-HaloTag molecules in living cells. dCas9-HaloTag and eGFP-tagged Heterochromatin Protein 1 (HP1) were stably integrated into 3T3 cells, and sgRNAs were transiently transfected. **(B)** 2D single-molecule visualization of dCas9-HaloTag molecules within live 3T3 nuclei at a 10 *ms* exposure time. **(C)** Two-photon FCS correlation curves and mathematical fits for dCas9-HaloTag in the absence of sgRNA (apo, grey) or loaded with cognate SINE B2 sgRNA (blue). Fluorescence correlation was measured within diffraction-limited volumes over time at random locations within cell nuclei (N = 11 cells for each condition).

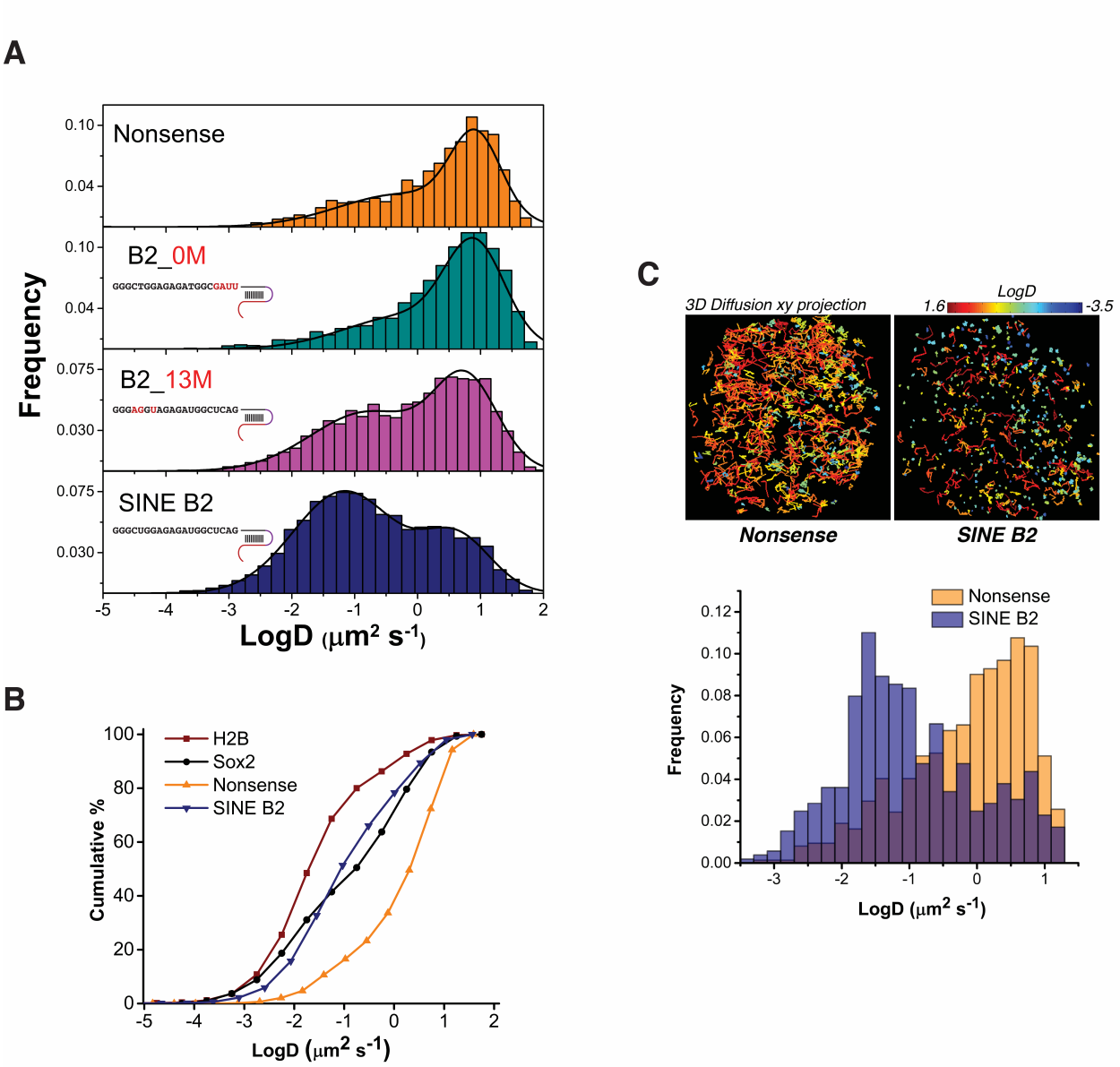
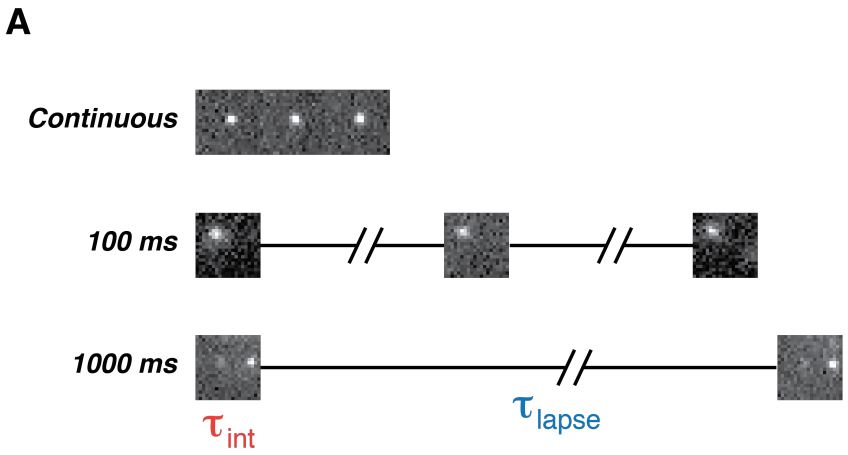
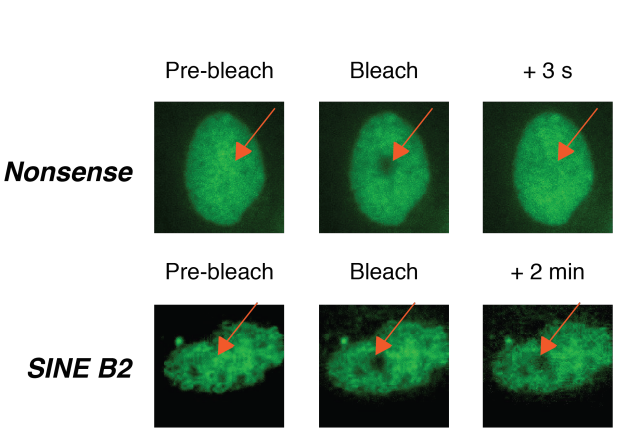
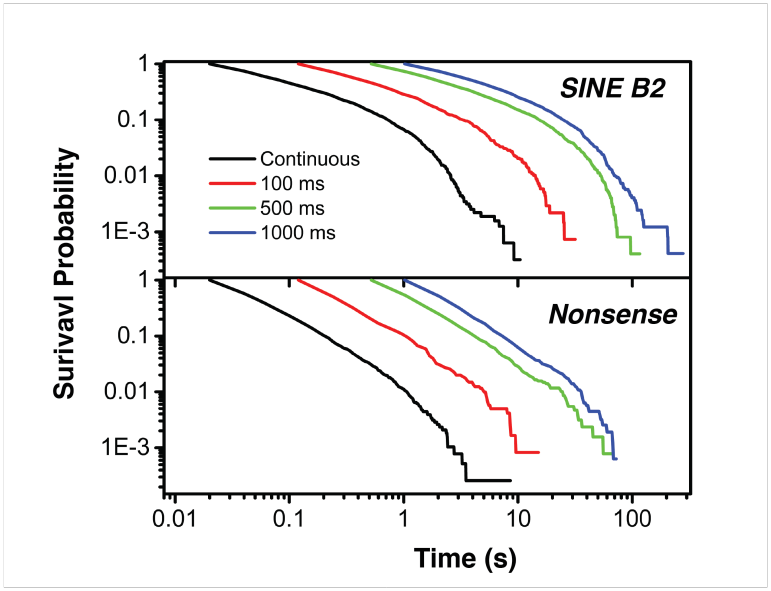

A
Nonsense
B2_0M
GGGCTGGAGAGATGGCGAUU
B2_13M
GGGAGGUAGAGAUGGCUCAG
SINE B2
GGGCUGGAGAGAUGGCUCAG
Frequency
LogD (μm² s⁻¹)
B
H2B
Sox2
Nonsense
SINE B2
Cumulative %
LogD (μm² s⁻¹)
C
3D Diffusion xy projection
LogD
1.6
-3.5
Nonsense
SINE B2
Nonsense
SINE B2
Frequency
LogD (μm² s⁻¹)

Figure 2.2. Cas9 exploration is dominated by 3D diffusion while searching for target sites in vivo. (A) Normalized histograms and two-component Gaussian fits illustrating the log diffusion coefficient distributions for dCas9-HaloTag with different sgRNAs ($N \geq 12$ cells for each condition). For reference, chromatin-bound H2B molecules can be fitted with a single Gaussian with $\log D \approx 0.1\ \mu m^2\ s^{-1}$. **(B)** Cumulative distribution plots quantifying the log diffusion coefficient for SINE B2 or nonsense-loaded dCas9-HaloTag relative to histone H2B or Sox2. **(C)** (*Top*) 2D projections of single particle trajectories obtained from 3D imaging using a multifocus microscope and (*Bottom*) histograms showing the log*D* distribution of trajectories. The trajectories are color coded according to diffusion coefficient. 3D movies were collected at a 30 *ms* exposure time, and diffusion coefficients were extracted directly from the 3D trajectories using MSD analyzer (Methods, $N = 2$ cells for each condition).


A
Continuous
100 ms
1000 ms
τint
τlapse


C
Pre-bleach
Bleach
+ 3 s
Nonsense
Pre-bleach
Bleach
+ 2 min
SINE B2


B
SINE B2
Continuous
100 ms
500 ms
1000 ms
Nonsense
Survival Probability
Time (s)

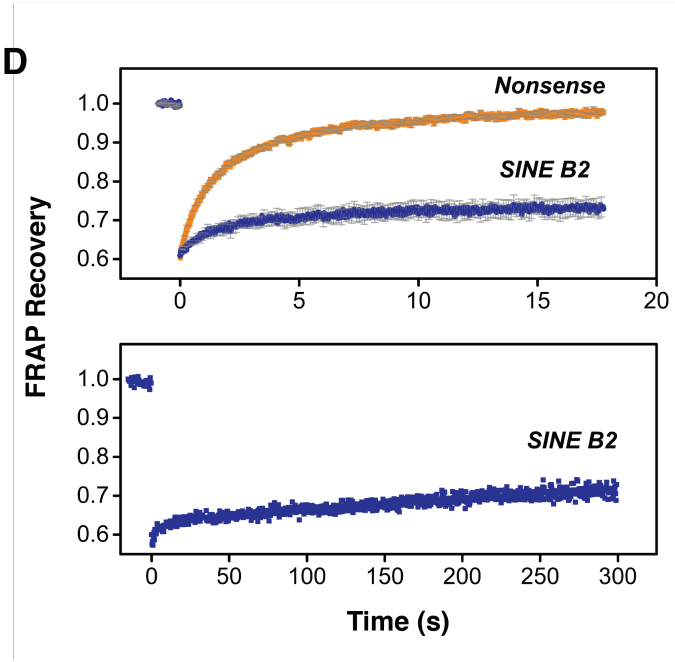

D
Nonsense
SINE B2
SINE B2
FRAP Recovery
Time (s)

Figure 2.3. Binding at on- and off-target sites by dCas9 (A) Time-lapse imaging of dCas9-HaloTag at constant exposure time (20 *ms*, τ_{int}) with varying lapse times (0, 100, 500, or 1000 *ms*, τ_{lapse}). **(B)** Quantification of survival probability for stationary molecules at different lapse times with different sgRNAs. Data were re-scaled and linearly fit to extract the average off-target residence time for the nonsense sgRNA (Methods and fig. A1.10). **(C)** FRAP images of dCas9-eGFP in live mouse cells with either nonsense (top) or SINE B2 (bottom) sgRNA. **(D)** Quantification of FRAP images for dCas9-eGFP using different sgRNAs (*N* = 17 cells for each condition).

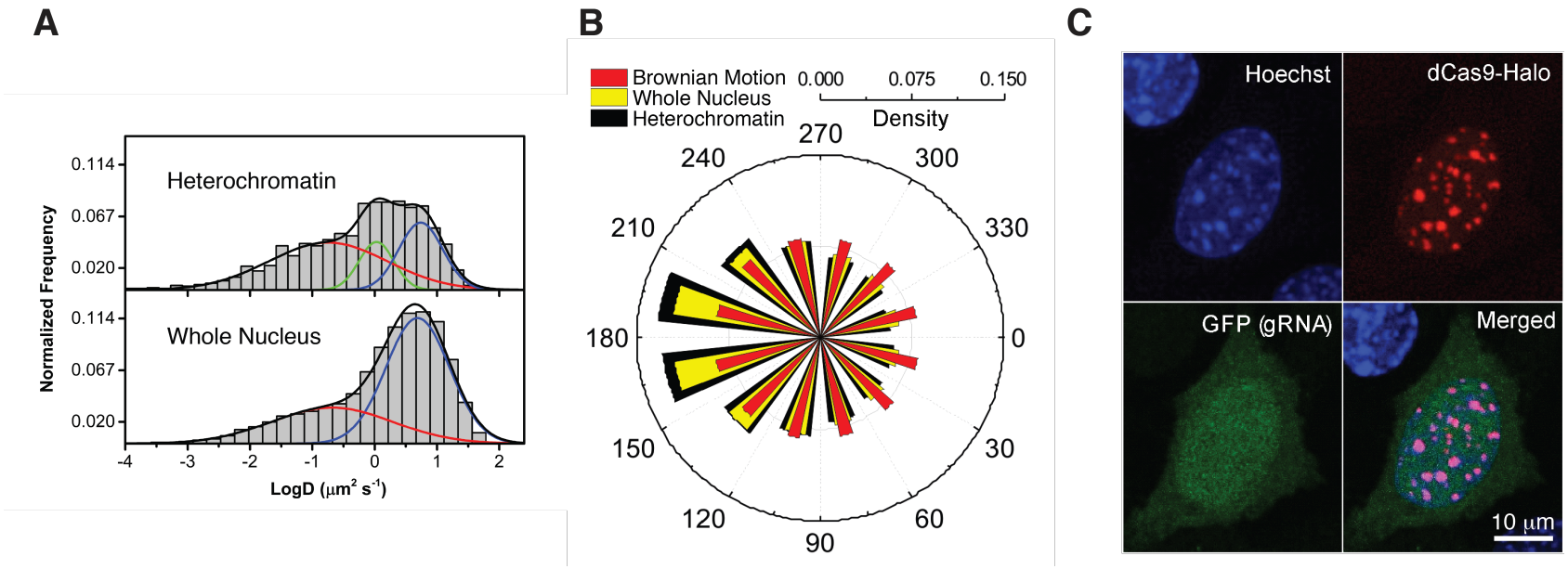

A
Heterochromatin
Whole Nucleus
Normalized Frequency
0.114
0.067
0.020
-4
-3
-2
-1
0
1
2
LogD (μm² s⁻¹)
B
Brownian Motion
Whole Nucleus
Heterochromatin
0.000
0.075
0.150
Density
270
240
300
210
330
180
0
150
30
120
60
90
C
Hoechst
dCas9-Halo
GFP (gRNA)
Merged
10 μm

Figure 2.4. Cas9 search efficiency is reduced, but not eliminated, in heterochromatic regions. (A) Log diffusion coefficient histograms and Gaussian fits for dCas9-HaloTag in HRs versus the entire cell nucleus (*N* = 11 cells). **(B)** Jumping angle analysis of diffusion anisotropy within HRs relative to the entire cell nucleus (*N* = 5 cells). **(C)** Epi-fluorescence image illustrating puncta formation in cells transfected with pericentromere-targeted sgRNA. Cells were fixed and co-stained with Hoechst 33258 for orthogonal labeling of pericentromeres.

Materials and Methods

Cell culture

Mouse 3T3 and mouse embryonic fibroblast cells (MEF) were maintained on round borosilicate plates (25 mm, Warner Instruments). Cells were grown in FluoroBrite™ DMEM (Life Technologies) supplemented with 15% FBS, 1 mM glutamax, 1 mM sodium pyruvate, and 1X antibiotic-antimycotic (Life Technologies). Imaging experiments were performed in the same media, and a Tokai-hit PI live-cell chamber and GM-8000 digital gas mixer were used to maintain culturing conditions (37 °C and 5% CO_2).

Plasmid construction and cell line generation

Codon optimized *S. pyogenes* Cas9 was amplified by PCR from a previously reported plasmid (laboratory of Jennifer Doudna) and inserted into a custom Piggybac (PB) transposon vector harboring a TRE-Tight Dox inducible promoter (Clonetech), an rtTA-V16 gene (CBh promoter) and a NeoR gene (Lin et al., 2014). The Cas9 sequence was flanked by *N*- and *C*-terminal nuclear localization signals (NLS), a *C*-terminal V5 epitope tag, and a *C*-terminal HaloTag® domain (map available upon request). The catalytically inactive (dCas9) version of the same construct was generated by introducing the D10A and H840A point mutations as described previously (Jinek et al., 2012). All sgRNA constructs were cloned from a custom vector featuring a U6 promoter, a *lacZ* sequence, and a downstream mTagBFP (SPT) or mCherry (FRAP) reporter gene under the control of a CMV promoter (maps available upon request). Excision of the *lacZ* sequence using type II BsmBI restriction enzyme allowed for ligation and incorporation of arbitrary sgRNA spacer sequences into our vector. An extended hairpin tracrRNA sequence was used to prevent premature termination of sgRNA transcription (Chen et al., 2013). HP1-eGFP was inserted into a custom PB plasmid harboring an EF1-α promoter and PuroR gene as described previously (Liu et al., 2014).

Stable cell lines were generated by co-transfecting the PB-Cas9/dCas9-HaloTag and PB-HP1-eGFP constructs with helper plasmid overexpressing Super PB Transposase (System Biosciences, 1:1 molar ratio). Cells were transfected for 24 h using Lipofectamine® 3000 and then subjected to selection under G418 (500 μg/mL) and Puromycin (2 μg/mL) for 10 d. Cells were subsequently maintained in culture medium supplemented with G418 (500 μg/mL) and grown to 90% confluency prior to freezing down in media supplemented with 10% DMSO.

Transfection of sgRNA and preparation for imaging

3T3 cells were grown to 70% confluency and then transfected with sgRNA-BFP plasmid for 24 h using Lipofectamine® 3000 (2:1 mass ratio). Transfected cells were stained with JF549 HaloTag ligand (100 nM final concentration) for 30 s and then washed 3x with culture media. After 15 min, cells were washed an additional time with culture media, and the cover glass was transferred to a live-cell metal holder for imaging. Samples were mounted individually on the microscope and maintained under culturing conditions (37 °C and 5% CO_2) throughout the course of imaging experiments.

2D single-molecule imaging with epi-illumination

2D single-molecule experiments were conducted on a Nikon Eclipse Ti2000 microscope equipped with a 100X Oil-immersion Objective lens (Nikon, N.A. = 1.41), a Lumencor light source, two filter wheels (Lambda 10-3, Sutter Instrument), perfect focusing systems, and EMCCD (iXon3, Andor). Proper emission filters (Semrock) were switched in front of the cameras for BFP, GFP, or JF549 emission, and a band mirror (405/488/561/633 BrightLine quad-band bandpass filter, Semrock) was used to reflect the laser into the objective. For 2D tracking experiments, JF549 dye was excited using a 561 nm laser (MPB Lasertech) at an intensity of ~800 W cm^{-2} and imaged at a 10 *ms* exposure time. The microscope and laser output were controlled using NIS Elements (Nikon).

2D single-molecule localization, tracking and diffusion analysis

For 2D single-molecule tracking, the spot localization (*x, y*) was obtained through 2D Gaussian fitting based on MTT algorithms using a home-built Matlab program (Sergé et al., 2008). The localization and tracking parameters in SPT experiments are listed in Table 2.1. For time-lapse measurements of residence times at different temporal length scales (continuous, 100 *ms*, 500 *ms*, 1 *s*), 0.05 $\mu m^2/s$ was set as maximum expected diffusion coefficient (D_{max}) for tracking. The D_{max} works as a limit constraining the maximum distance (r_{max}) between two frames for a particle translocation. Only molecules localized within r_{max} for at least two consecutive frames were considered as bound molecules; events that appeared in single frames were discarded. The duration of individual tracks (dwell time) was directly calculated based on the track length. This was used to calculate survival probabilities to extract the average off-target residence time(Normanno et al., 1AD). The MTT algorithm was used for fast tracking of dCas9-HaloTag (10 *ms*) and diffusion analysis measurements. Diffusion coefficients were calculated from tracks with at least 5 consecutive frames by the MSDanalyzer with a minimal fitting R^2 of 0.8 (Tarantino et al., 2014).

FRAP experiments with dCas9-eGFP cell lines

Fluorescence recovery after photobleaching (FRAP) experiments were performed on a Nikon Eclipse Ti2000 microscope equipped with a Lumencor light source, two filter wheels (Lambda 10-3, Sutter Instrument), perfect focusing systems and EMCCD (iXon3, Andor). Proper emission filters (Semrock) were switched in front of the cameras for eGFP and mCherry emission. Local photobleaching of dCas9-eGFP was achieved using an OBIS 488 nm laser (Coherent, Inc.) passed through a 100X oil immersion objective (100X). The microscope and FRAP laser were controlled using NIS Elements (Nikon).

Photobleaching was generated within a ~1 μm circular region of cell nuclei using ~28 mW of laser power and 250 *ms* total irradiation time. Images were collected at 18 *ms* exposure times with 50 frames acquired before bleaching for normalization purposes. For longer timescale FRAP experiments with SINE B2 sgRNA, a 300 *ms* dark lapse time between frames was applied to minimize photobleaching. To quantify FRAP curves, we measured the evolution of the radial intensity profile *versus* time ($I_r(t)$). A custom homebuilt MatLab program was used to subtract the background and correct for

eGFP photobleaching. After normalization, $I_r(t)$ was fitted to a constant function with a Gaussian flank according to Equation 2.1 (Sprague et al., 2004):

$$I_r(t) = \begin{cases} A & , \text{for } r \le r_c \\ 1-(1-A)\exp\left(-\dfrac{\left(r-r_c\right)^2}{2\sigma^2}\right) & , \text{for } r > r_c \end{cases} \qquad \textbf{Eq. 2.1}$$

Here, r_c is the radius of constant value, and σ is the width of the Gaussian fit of the bleach profile. We recorded intensity profiles for multiple cells, normalized the intensities between 0.6 and 1.0, and then averaged across all cells to obtain the intensity profiles shown in Fig. 2.3D.

MFM 3D single-molecule imaging

3D single-molecule tracking experiments were performed using the same microscope as for 2D experiments. The multifocus optical elements were appended after the primary image plane. The details for the multifocus microscopy instrumentation are described in previous publications (Abrahamsson et al., 2012; Chen et al., 2014; Liu et al., 2014). Briefly, the diffractive multifocus grating was placed in the Fourier plane in the emission pathway to form the multifocus image. It was followed by the chromatic correction grating (custom made by Tessera) and prism (custom made by Rocky Mountain Instruments), which allowed images from different focal planes to be refocused at different positions of the camera chip. The images were then reconstructed into 3D using 200 nm fluorescence beads, which provide a transformation matrix enabling the focal planes to be superimposed with an accuracy of ~10 nm. For experiments with JF549 dyes in 3D, we used a 561-nm laser (MPB Lasertech) of excitation intensity ~1KW cm^{-3} and the acquisition time was 30 *ms*.

3D PSF model, 3D single-molecule localization and tracking

3D localization *(x,y,z)* was conducted using FISH-QUANT software (Mueller et al., 2013). The PSF Model can be described by the following equation:

$$I(x,y,z) = \left(A_0 e^{-\frac{(x-x_0)^2}{2\sigma_{xy}^2}} e^{-\frac{(y-y_0)}{2\sigma_{xy}^2}} e^{-\frac{(z-z_0)^2}{2\sigma_z^2}} \right)_{PSF} + B \qquad \textbf{Eq. 2.2}$$

Here, A_0 is the signal amplitude; σ is the Standard Deviation (S.D.) of the Gaussian fit in the standard deviations in the *x* and *y* directions were the same. A U-track algorithm was used for 3D single particle tracking (Jaqaman et al., 2008). Diffusion coefficients were calculated from tracks with at least 5 consecutive frames by the MSDanalyzer with a minimal fitting R^2 of 0.8 (Tarantino et al., 2014).

Heterochromatin mask definition, TF diffusion analysis and Localization Density Calculation

We took HP1-eGFP images before and after SPT experiments to ensure that cell nuclei and HRs did not move during the ~4 min acquisition. After background subtraction, the intensity map for heterochromatic regions in single cells was directly calculated by normalizing pixel intensity in the HP1-eGFP channel with the highest pixel intensity in the image. A binary mask for HRs was calculated by applying a threshold cutoff of 0.2 to the intensity map. Single-molecule localization events and 2D single-molecule tracks were divided into localizations / track segments inside and outside of the mask. For reconstructing diffusion co-efficient histograms, track segments from each group were pooled, and diffusion coefficients were calculated from tracks with at least 5 consecutive frames by the MSDanalyzer with a minimal fitting R^2 of 0.8 (Tarantino et al., 2014). To calculate localization density, the number of localizations in each group was divided by the total area (in the unit of the number of pixels) of that group. We further normalized the In-mask density with the out-mask density.

Brownian motion simulation in the nucleus

The mean square displacement of Brownian motion is described as:

$$\left\langle r^2 \right\rangle = 2dDt \qquad \textbf{Eq. 2.3}$$

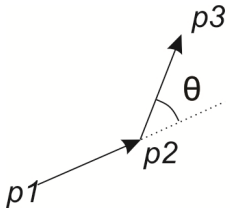
d, dimensionality (*d* = 2 in our case)
D, diffusion coefficient

To computationally simulate Brownian motion in the Cartesian coordinate system, we uncoupled each jump to *x, y* one-dimensional steps defined by the equation below:

$$\begin{pmatrix} x(t+\delta t) \\ y(t+\delta t) \end{pmatrix} = \begin{pmatrix} x(t) \\ y(t) \end{pmatrix} + \sqrt{2D\delta t} \begin{pmatrix} N_1 \\ N_2 \end{pmatrix} \qquad \textbf{Eq. 2.4}$$

Here, N_i are independent random numbers obeying a Gaussian distribution with a zero mean and a variance of 1, and *dt* is the sampling interval. Simulation of diffusion was performed with MathWorks Matlab 2013a. Specifically, we limited the 2D diffusion of the protein to a nucleus with a radius of 5 μm and diffusion coefficient of 2 $\mu m^2\ s^{-1}$.

Diffusion jumping-angle analysis



Tracks from different categories (whole nucleus, heterochromatin, or simulation) were pooled, and a sliding window of 3 points was applied to each track. The angle between the vectors of the first two and the last two points was calculated by the acos() function in the Matlab 2013a. The program iteratively processed all tracks in each category and individual angles were pooled and binned accordingly for the angular Rose histogram.

Pericentromere staining of dCas9-HaloTag cells

Mouse embryonic fibroblast (MEF) cells were transduced with constructs containing dCas9-HaloTag and sgRNA against major satellite (sgMajSat) by lipofectamine 3000 (Life Technologies). 48 hours post transfection, cells were incubated with medium containing 100nM TMR conjugated Halo ligand for 15 min, followed by brief washes with PBS and incubation in fresh medium for 30 min. Cells were fixed with 4% formaldehyde for 10 min and stained by Hoechst 33258 before imaging. The transcribed sequence of the guide is below, with the 20-nucleotide spacer sequence highlighted in red:

5'-GG**CCAUAUUCCACGUCCUACAG**GUUUAAGAGCUAUGCUGGAAACAGCA
UAGCAAGUUUAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGU
CGGUGC-***3'***

2-photon FCS measurements

In order to calculate diffusion coefficients of apo dCas9-HaloTag and sgRNA B2 loaded dCas9-HaloTag, we performed two-photon Fluorescence correlation spectroscopy (FCS) measurements. An 80-MHz Ti: Sapphire laser (Chameleon Ultra II, Coherent) at wavelength of 1020 nm was used to excite JF549 through an inverted Olympus IX81 microscope with a 60X water immersion objective (N.A. = 1.2) (UplanSApo 60XW, Olympus, Japan). Fluorescence emission was collected after passing through a short-pass filter (FF01-720SP, Semrock) and a band pass filter (FF02-617/73, Semrock) and directed with a fiber-coupled (100 µm-core, multi-mode fiber, AFS105/125Y, Thorlabs) avalanche photo-diode (SPCM-AQRH-14-FC, Perkin-Elmer, Canada). Autocorrelation was calculated through an external autocorrelator (Flex03LQ-01, correlator.com). The data analysis was performed using a custom software package (provided by V.Iyer, Janelia Research Campus). The autocorrelation curves were fitted with two-component 3D diffusion model according to Equation 2.5 (Schwille et al., 1997):

$$G(\tau) = \frac{1}{N}\left((1-f)(1+\frac{\tau}{\tau_{D1}})^{-1}(1+\frac{\tau}{\kappa^2\tau_{D1}})^{-1/2} + f(1+\frac{\tau}{\tau_{D2}})^{-1}(1+\frac{\tau}{\kappa^2\tau_{D2}})^{-1/2} \right) \qquad \textbf{Eq. 2.5}$$

Here, τ_{D1} and τ_{D2} are lag times for the two populations, *f* is the fraction corresponding to τ_{D2}, and *N* is the number of molecules in the confocal volume.

Cloning of His_6-dCas9-HaloTag and His_6-Cas9-HaloTag for in vitro assays

Bacterial expression dCas9(D10A/H840A) and wide type spCas9 were PCR amplified from plasmids pMJ841 and pMJ915, respectively, and cloned into the EcoRI/XhoI site of the Champion pET302/NT-His vector (Invitrogen) by introducing the C-terminal SpeI restriction site. The HaloTag was PCR amplified with a C-terminal NLS signal and inserted into the SpeI/XhoI site of the pET302/NT-His-dmCas9 construct. Constructs were verified by sequencing.

In vitro purification of Cas9-HaloTag proteins

His_6-dCas9-HaloTag and His_6-Cas9-HaloTag were expressed in *E. coli* BL21-star (DE3), and BL21(DE3)pLysS-Rosetta respectfully. Cells were grown in Terrific Broth medium at 18°C for 16 hr after induction with 0.5 mM IPTG (Invitrogen) for His_6-dCas9-HaloTag and 0.2mM IPTG for His_6-Cas9-HaloTag. Cells from 1 L of liquid culture were lysed by sonication in lysis buffer consisting of 500 mM NaCl, 50 mM HEPES pH 7.6, 5% glycerol, 1 mM TCEP (for the His_6-dCas9-HaloTag, none for His_6-Cas9-HaloTag), 1% Triton X-100, 1 mM benzamidine, 1:1000 aprotinin, 0.5 mM PMSF, and 10 mM imidazole, and EDTA-free protease inhibitor cocktail (Roche). Lysate was clarified by centrifugation and incubated with 1 mL packed volume of Ni-NTA agarose resin (Qiagen) for 1-2 hrs. The resin was washed extensively with lysis buffer supplemented with either 10 mM or 25 mM imidazole, and the protein was eluted with lysis buffer containing 250 mM imidazole. The His_6-dCas9-HaloTag protein was then dialyzed for 1 hour into 200 mM NaCl, 50 mM HEPES pH 7.6, 5% glycerol, 1 mM TCEP, and 1 mM DTT and then for 2 hours into 125 mM NaCl, 50 mM HEPES pH 7.6, 5% glycerol, 1 mM TCEP, and 1 mM DTT. The His_6-dCas9-HaloTag protein was further purified on a custom ~8 mL SP-Sepharose FF column (GE Life Sciences) over a gradient from 125 mM to 2 M NaCl. Fractions containing dCas9 were pooled and further purified on a Superdex200 column (GE Life Sciences) (fig. A1.1). After the nickel column, the His_6-Cas9-HaloTag protein was diluted in ion exchange buffer to bring the salt concentration to 200 mM NaCl. The protein was then flowed over tandem Q and SP HiTrap 5mL columns (GE Life sciences) with Q first in line. The tandem columns were washed with 200 mM NaCl (50 mM Hepes pH 7.6, 5% glycerol, 1 mM DTT, 0.5 mM PMSF) until the A280 absorbance returned to baseline. The Q column was then removed, and the SP was eluted with a linear gradient from 200 mM NaCl (50 mM Hepes pH 7.6, 5% glycerol, 1 mM DTT, 0.5 mM PMSF) to 1 M NaCl (50 mM Hepes pH 7.6, 5% glycerol, 1 mM DTT, 0.5 mM PMSF) over 10 column volumes. Pooled fractions were then dialyzed into storage buffer (200 mM NaCl, 50 mM Hepes pH 7.6, 5% glycerol, 1 mM DTT).

In vitro sgRNA preparation and electrophoretic mobility shift assay (EMSA)

For the synthesis of sgRNA, two DNA oligonucleotides were mixed, annealed and filled in through one PCR cycle reaction to produce the template for in vitro T7 transcription. DNA was ethanol precipitated and used as a template for sgRNA synthesis with the T7 Quick High Yield RNA Synthesis Kit (NEB) according to vendor instructions. The sgRNA was purified via polyacrylamide gel, with the following modifications: the gel slice containing the RNA was crushed and incubated overnight in five volumes of sodium acetate (0.3 M) plus five volumes of phenol chloroform (Nilsen,

2013). The aqueous phase was ethanol precipitated and resuspended in RNase free water. The quality of the RNA was verified on a 6% acrylamide 3 M urea gel.

Oligonucleotides:

5'-TTAATACGACTCACTATAGACATGTTGATTTCCTGAAAGTTTTAGAGCTAG AAATAGC-***3'***

5'-AAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTT ATTTTAACTTGCTATTTCTAGCTCTAAAAC-***3'***

The sense DNA probe used for the gel shift assay was derived from the lambda genome and obtained from annealing the following oligonucleotides:

5'-CGGAACTGGAAAACCGACATGTTGATTTCCTGAAACGGGATATCATCAAA-***3'***

5'-TTTGATGATATCCCGTTTCAGGAAATCAACATGTCGGTTTTCCAGTTCCG-***3'***

Cleavage assay

Purified His_6-Cas9-HaloTag protein was pre-incubated with sgRNA in reaction buffer (20 mM Tris-HCl pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 5% glycerol, and 1 mM DTT) for 10 min at 37 °C prior to adding target-containing plasmid DNA (p1eBlueScript SK+, fig. S3) to a final concentration of 500 nM sgRNA, 100 nM Cas9, and 20 nM DNA. The reaction was mixed and then incubated at 37 °C for 1 h prior to running on a 0.5% agarose gel (fig. A1.3). Cleavage of the plasmid DNA with XbaI (NEB) served as a control for the linearized plasmid band.

Bayesian analysis of heterochromatin data

Spatial dependence of Cas9 diffusion was analyzed using a Bayesian inference mapping algorithm(Beheiry et al., 2015). Trajectories of Cas9 inside the nucleus were spatially partitioned using a hierarchical (quad-tree) mesh (fig. A1.12). Dimensions of the zones in this type of mesh were adapted to the characteristic size of the trajectory steps within them, hence accounting for spatially dependent heterogeneities in diffusive behavior. For each zone, the diffusion was presumed to be constant and was calculated by considering all trajectory steps within it (the total length of the trajectory is not consequential). Trajectories were modeled by an overdamped Langevin equation, allowing for physical parameters governing single-molecule movement (e.g. diffusion and interaction energies) to be distinguished. The diffusion coefficient within each zone was calculated as the result of a maximum *a posteriori* estimate from a Bayesian inference calculation. Generated parameter maps showed that Cas9 diffusion is markedly less inside heterochromatic regions relative to other regions inside the nucleus.

Table 2.1. Parameters used for 2D single-molecule tracking and analysis.

Parameter	Diffusion Analysis (Fig. 2.2)	Time-Lapse Experiments (Fig. 2.3)
exposure time (ms)	10	20
laser power (mW cm^{-2})	~800	~800
λ_{ex} (nm)	561	561
λ_{em} (nm)	590	590
pixel size (nm)	160	160
numerical aperture	1.41	1.41
Expected D_{max} (μm^2 s^{-1})	5	0.05

Chapter 3: RNA targeting by C2c2 proteins from Type VI CRISPR systems

Bacterial adaptive immune systems employ CRISPRs (clustered regularly interspaced short palindromic repeats) and CRISPR-associated (Cas) proteins for RNA-guided nucleic acid cleavage (van der Oost et al., 2014; Wright et al., 2016). Although most prokaryotic adaptive immune systems generally target DNA substrates (Brouns et al., 2008; Garneau et al., 2010; Marraffini and Sontheimer, 2008), the Type III and Type VI CRISPR systems direct interference complexes against single-stranded RNA (ssRNA) substrates (Abudayyeh et al., 2016; Hale et al., 2009; Samai et al., 2015; Staals et al., 2013). In Type VI systems, the single-subunit C2c2 protein functions as an RNA-guided RNA endonuclease (Abudayyeh et al., 2016; Shmakov et al., 2015). How this enzyme acquires mature CRISPR RNAs (crRNAs) that are essential for immune surveillance and how it carries out crRNA-mediated RNA cleavage remain unclear. Here we show that the bacterial C2c2 possesses a unique ribonuclease activity responsible for CRISPR RNA maturation that is distinct from its RNA-activated ssRNA-degradation activity. These dual ribonuclease functions are chemically and mechanistically different from each other and from the crRNA-processing behavior of the evolutionarily unrelated CRISPR enzyme Cpf1 (Fonfara et al., 2016). We show that the two ribonuclease activities of C2c2 enable multiplexed processing and loading of guide RNAs that in turn allow for sensitive cellular transcript detection.

This work was done collaboratively and was originally published in *Nature*:

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

*Indicates co-first author

All co-authors have consented to reprinting the original publication for this thesis. Reprinted with permission from Nature Publishing Group. Alexandra East-Seletsky, Spencer C. Knight, and Mitchell O'Connell conceived the study and designed experiments with guidance from Jamie H.D. Cate, Robert Tjian, and Jennifer A. Doudna. David Burstein performed bioinformatic analyses. Alexandra East-Seletsky and Mitchell O'Connell performed primary experiments for the Nature publication with technical assistance from Spencer C. Knight. All authors wrote and discussed the manuscript.

Introduction

The first step of CRISPR immune surveillance requires processing of precursor crRNA transcripts (pre-crRNAs), consisting of repeat sequences flanking viral spacer sequences, into individual mature crRNAs that each contain a single spacer (Charpentier et al., 2015; Hochstrasser and Doudna, 2015; Li, 2015). CRISPR systems use three known mechanisms to produce mature crRNAs: a dedicated endonuclease (for example, Cas6 or Cas5d in Type I and III systems) (Carte et al., 2008a; Haurwitz et al., 2010; Nam et al., 2012), coupling of a host endonuclease (for example, RNase III with a trans-activating crRNA (tracrRNA) in Type II systems) (Deltcheva et al., 2012), or a ribonuclease activity intrinsic to the effector enzyme itself (for example, Cpf1, Type V systems) (Fonfara et al., 2016).

Results and Discussion

Since Type VI CRISPR loci lack an obvious Cas6 or Cas5d-like endonuclease or tracrRNA (Shmakov et al., 2015), we wondered whether C2c2 itself might possess pre-crRNA processing activity, and if so, whether the mechanism would be distinct from Cpf1, an unrelated class 2 CRISPR effector recently demonstrated to process pre-crRNAs (Fonfara et al., 2016). Using purified recombinant C2c2 protein homologs from three distinct branches of the C2c2 protein family (Fig. 3.1, figs. A2.1-A2.3), we found that all three C2c2 enzymes cleave 5′-end radiolabeled pre-crRNA substrates consisting of a full-length consensus repeat sequence and a 20 nucleotide (nt) spacer sequence (Fig. 3.1C). We mapped the cleavage site for each pre-crRNA:C2c2 homolog pair, revealing that processing occurs at a position either two or five nucleotides upstream of the predicted repeat-sequence hairpin structure, depending on the C2c2 homolog (Fig. 3.1C and fig. A3.4A). Surprisingly, our biochemically mapped 5'-cleavage sites do not agree with previously reported cleavage sites for *Leptotrichia shahii* (LshC2c2) or *Listeria seeligeri* (LseC2c2) pre-crRNAs (Shmakov et al., 2015). Our own analysis of Shmakov *et al.*'s RNA sequencing data set indicates agreement of the in vivo cleavage site with the in vitro site reported here (fig. A2.4B-I). Furthermore, cleavage assays using C2c2 from *Leptotricia buccalis* (LbuC2c2) and a larger pre-crRNA comprising a tandem hairpin-repeat array resulted in two products resulting from two separate cleavage events (fig. A2.5A), consistent with a role for C2c2 in processing precursor crRNA transcripts generated from Type VI CRISPR loci.

To understand the substrate requirements and mechanism of C2c2 guide RNA processing, we generated pre-crRNAs harboring mutations in either the stem loop or the single-stranded flanking regions of the consensus repeat sequence and tested their ability to be processed by LbuC2c2 (Fig. 3.2). We found that C2c2-catalyzed cleavage was attenuated upon altering the length of the stem in the repeat region (fig. 3.2A). Inversion of the stem loop or reduction of the loop length also reduced C2c2's processing activity, while contiguous 4-nt mutations including or near the scissile bond completely abolished it (fig. A2.5B). A more extensive mutational analysis of the full crRNA repeat sequence revealed two distinct regions on either side of the hairpin with marked sensitivity to base changes (Fig. 3.2B). By contrast, there was no dependence on the spacer sequence for kinetics of processing (fig. A2.5B). This sensitivity to both flanking regions of the hairpin is reminiscent of the sequence and structural motifs required by many Cas6 and Cas5d enzymes (Charpentier et al., 2015; Li, 2015). In

contrast, Cpf1 does not have any dependence on the 3′ hairpin flanking region, as the variable spacer region abuts the hairpin stem (Fonfara et al., 2016).

The processing activity of LbuC2c2 was unaffected by the presence of divalent metal ion chelators EDTA or EGTA (Fig. 3.2C), indicative of a metal ion-independent RNA hydrolytic mechanism. Metal ion-independent RNA hydrolysis is typified by the formation of a 2′, 3′-cyclic phosphate and 5′-hydroxide on the 5′ and 3′ halves of the crRNA cleavage products, respectively (2011a). To determine the end-group chemical identity of C2c2-processed substrates, we further incubated the 5′ flanking products with T4 polynucleotide kinase, which removes 2′,3′-cyclic phosphates to leave a 3′-hydroxyl. We observed altered denaturing-gel migration of the 5′ flanking product after kinase treatment, consistent with the removal of a 3′ phosphate group (fig. A2.5d). The divalent metal ion independence of C2c2's pre-crRNA processing activity is in stark contrast with the divalent metal ion dependency of Cpf1, the only other single-protein CRISPR effector shown to perform guide processing (Fonfara et al., 2016). Collectively, these data indicate that C2c2-catalyzed pre-crRNA cleavage is a divalent metal ion-independent process that likely uses a general acid-base catalysis mechanism (2011b).

After maturation, crRNAs typically bind with high affinity to Cas effector protein(s) to create RNA-guided surveillance complexes capable of sequence-specific nucleic acid recognition (Jinek et al., 2012; van der Oost et al., 2014; Wright et al., 2016). In agreement with previous work using LshC2c2 (Abudayyeh et al., 2016), LbuC2c2 catalyzed efficient target RNA cleavage only when such substrates could base pair with a complementary sequence in the crRNA (figs. A2.6-A2.8). Given the promiscuous pattern of cleavage observed for C2c2 (fig. A2.7), we tested the ability of LbuC2c2 to act as a crRNA-activated non-specific RNA endonuclease in *trans* (fig. A2.6B). In striking contrast to non-target cleavage experiments performed in *cis* and consistent with observations for LshC2c2 (Abudayyeh et al., 2016), we observed rapid degradation of non-target RNA in *trans* (fig. A2.6B). This result shows that target recognition activates C2c2 for general non-specific degradation of RNA. Importantly, the similar RNA cleavage rates and near-identical cleavage products observed for both *cis* on-target cleavage and *trans* non-target cleavage of the same RNA substrate implicate the same nuclease center in both activities (fig. A2.6B).

crRNA-mediated cleavage of target ssRNA occurs ~80-fold faster than pre-crRNA processing (Fig. 3.3A), and in contrast to pre-crRNA processing, RNA-guided target cleavage is abolished in the presence of EDTA, indicating that this activity is divalent metal ion-dependent (Fig. 3.3A, fig. A2.6c). Given these clear differences, we reasoned that C2c2 might possess two orthogonal RNA cleavage activities: one for crRNA maturation, and the other for crRNA-directed, non-specific RNA degradation. To test this hypothesis, we systematically mutated several residues within the conserved HEPN motifs of LbuC2c2 (Abudayyeh et al., 2016; Anantharaman et al., 2013; Niewoehner and Jinek, 2016; Sheppard et al., 2016), and assessed pre-crRNA processing and RNA-guided RNase activity of the mutants (Fig. 3.3 and fig. A2.8D). Double and quadruple mutants of conserved HEPN residues (R472A, R477A, R1048A and R1053) retained robust pre-crRNA cleavage activity (Fig. 3.3C). By contrast, all HEPN mutations abolished RNA-guided cleavage activity while not affecting crRNA or ssRNA-binding ability (fig. A2.8D) (Abudayyeh et al., 2016).

We sought mutations that would abrogate pre-crRNA processing activity without disrupting target RNA cleavage. Given that we were unable to predict any other potential RNase motifs beyond the HEPN motifs, and that C2c2 proteins bear no homology to Cpf1, we opted to systematically mutate the charged residues throughout LbuC2c2. We identified an arginine residue (R1079A) that upon mutation resulted in severely attenuated pre-crRNA processing activity (Fig. 3.3C). This C2c2 mutant enzyme retained crRNA-binding ability as well as RNA target cleavage activity (fig. A2.9D). Taken together, our results show that distinct active sites within the C2c2 protein catalyze pre-crRNA processing and RNA-directed RNA cleavage.

We recognized that the robust RNA-stimulated cleavage of substrates might be employed as a means of detecting specific RNAs within a pool of transcripts. While many polymerase-based methods have been developed for RNA amplification and subsequent detection, few approaches are able to directly detect the target RNA without significant engineering or stringent design constraints for each new RNA target (Cordray and Richards-Kortum, 2012; Rohrman et al., 2012). As a readily-programmable alternative, we tested whether C2c2's RNA-guided *trans* endonuclease activity could be harnessed to cleave a fluorophore-quencher-labeled reporter RNA substrate, thereby resulting in increased fluorescence upon target RNA-triggered RNase activation (Fig. 3.4A). LbuC2c2 was loaded with bacteriophage λ-targeting crRNAs and tested for its ability to detect the corresponding λ ssRNA targets spiked into HeLa cell total RNA. We found that upon addition of as little as 1-10 pM complementary λ target-RNA, a substantial crRNA-specific increase in fluorescence occurred within 30 min (Fig. 3.4B and fig. A2.10A). Control experiments with either C2c2:crRNA complex alone or in the presence of crRNA and a non-complementary target RNA resulted in negligible increases in fluorescence relative to an RNase A positive control (Fig. 3.4B and fig. A2.10A). We note that at the 10 pM concentration of a λ target RNA, only ~0.02% of the C2c2:crRNA complex is predicted to be in the active state, yet the observed fluorescent signal reflected ~25-50% cleavage of the reporter RNA substrate, depending on the RNA target. Fragment size resolution of the background RNA in these reactions revealed significant degradation, even on highly structured tRNAs (fig. A2.10B). Since reporter RNA cleavage occurs in the presence of a vast excess of unlabeled RNA, we conclude that LbuC2c2 is a robust multiple-turnover enzyme capable of at least 10^4 turnovers per target RNA recognized. Thus, in contrast to previous observations(Abudayyeh et al., 2016), crRNA-directed *trans* cleavage is potent and detectable even at extremely low levels of activated protein.

To extend this LbuC2c2 RNA detection system, we designed a crRNA to target endogenous beta-actin mRNA. We observed a measurable increase in fluorescence in the presence of human total RNA relative to *E. coli* total RNA, demonstrating the specificity of this method (Fig. 3.4C). Furthermore, given that C2c2 processes its own guide, we combined pre-crRNA processing and RNA detection in a single reaction by designing tandem crRNA-repeat containing spacers complementary to target RNAs A and λ2. LbuC2c2 incubated with this unprocessed tandem guide RNA in the detection assay generated a significant increase in fluorescence similar in magnitude and sensitivity to experiments using mature crRNAs (Fig. 3.4B, D). Taken together, these data highlight the exciting opportunity to take advantage of C2c2's two distinct RNase activities for a range of biotechnological applications (Fig. 3.4E).

In bacteria, C2c2 likely operates as a sentinel for viral RNAs (Abudayyeh et al., 2016). We propose that when invasive transcripts are detected within the host cell via base pairing with crRNAs, C2c2 is activated for promiscuous cleavage of RNA in *trans* (Fig. 3.4E). As a defense mechanism, this bears striking similarity to RNase L and caspase systems in eukaryotes, whereby a cellular signal triggers promiscuous ribonucleolytic or proteolytic degradation within the host cell, respectively, leading to apoptosis (Choi et al., 2015; McIlwain et al., 2013). While the RNA targeting mechanisms of Type III CRISPR systems generally result in RNA cleavage within the protospacer-guide duplex (Samai et al., 2015), recent examples of associated nucleases Csx1 (Sheppard et al., 2016) and Csm6 (Niewoehner and Jinek, 2016) provide compelling parallels between the Type VI systems and the multi-component Type III inference complexes.

Conclusion

Our data show that CRISPR-C2c2 proteins represent a new class of enzyme capable of two separate RNA recognition and cleavage activities. Efficient pre-crRNA processing requires sequence and structural motifs within the CRISPR repeat which prevent non-endogenous crRNA loading and helps to reduce the potential toxicity of this potent RNase. The entirely different pre-crRNA processing mechanisms of C2c2 and the Type V CRISPR effector protein Cpf1 indicate that each protein family has converged upon independent activities encompassing both the processing and interference functions of their respective CRISPR pathways. Furthermore, the two distinct catalytic capabilities of C2c2 can be harnessed in concert for RNA detection, as the activation of C2c2 to cleave thousands of *trans*-RNAs for every target RNA detected enables potent signal amplification. The capacity of C2c2 to process its own guide RNAs from arrays could also allow the use of tissue-specific Pol II promoters for guide expression, in addition to target multiplexing for a wide range of applications. The C2c2 enzyme is unique within bacterial adaptive immunity for its dual RNase activities, and highlights the utility of harnessing CRISPR proteins for precise nucleic acid manipulation in cells and cell-free systems.

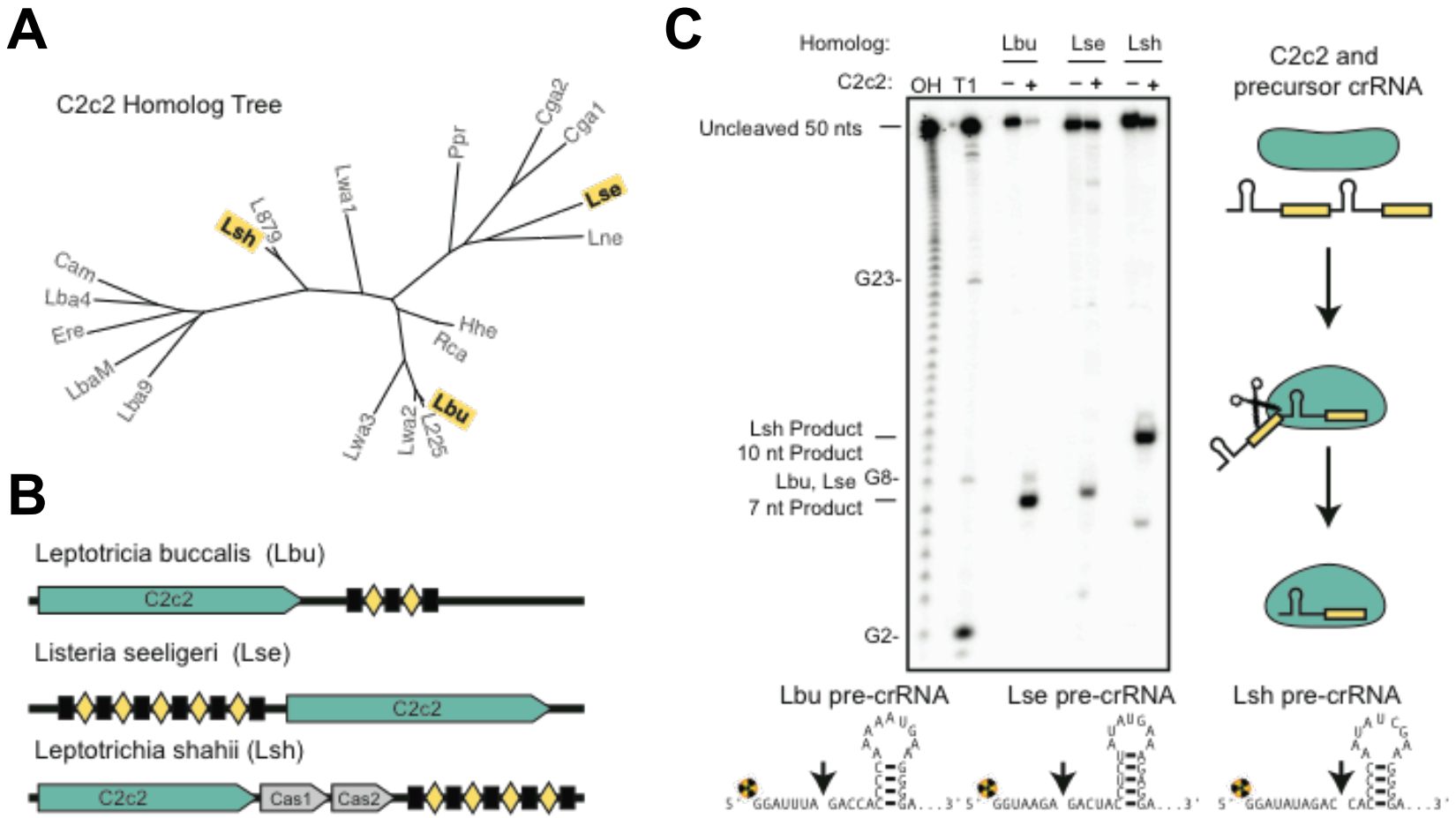

A
C2c2 Homolog Tree
Cga2
Cga1
Ppr
Lwa1
L879
Lsh
Lse
Lne
Cam
Lba4
Ere
LbaM
Lba9
Hhe
Rca
Lwa3
Lwa2
L225
Lbu
B
Leptotricia buccalis (Lbu)
C2c2
Listeria seeligeri (Lse)
C2c2
Leptotrichia shahii (Lsh)
C2c2
Cas1
Cas2
C
Homolog:
Lbu
Lse
Lsh
C2c2:
OH
T1
– +
– +
– +
Uncleaved 50 nts
G23-
Lsh Product
10 nt Product
Lbu, Lse
7 nt Product
G8-
G2-
C2c2 and
precursor crRNA
Lbu pre-crRNA
Lse pre-crRNA
Lsh pre-crRNA
5' GGAUUUA GACCAC GA...3'
5' GGUAAGA GACUAC GA...3'
5' GGAUAUAGAC CAC GA...3'

Figure 3.1. C2c2 proteins process precursor crRNA transcripts to generate mature crRNAs. **(A)** Maximum-likelihood phylogenetic tree of C2c2 proteins. Homologues used in this study are highlighted in yellow. **(B)** Diagram of the type VI CRISPR loci used in this study. Black rectangles denote repeat elements, yellow diamonds denote spacer sequences. Cas1 and Cas2 are only found in the genomic vicinity of LshC2c2. **(C)** C2c2-mediated cleavage of pre-crRNA derived from the LbuC2c2, LseC2c2 and LshC2c2 CRISPR repeat loci. OH, alkaline hydrolysis ladder; T1, RNase T1 hydrolysis ladder. Processing cleavage reactions were performed with 100 nM C2c2 and <1 nM pre-crRNA. Schematic of cleavage is depicted on the right and predicted pre-crRNA secondary structures are shown below, with arrows indicating the mapped C2c2 cleavage sites (nt, nucleotides).

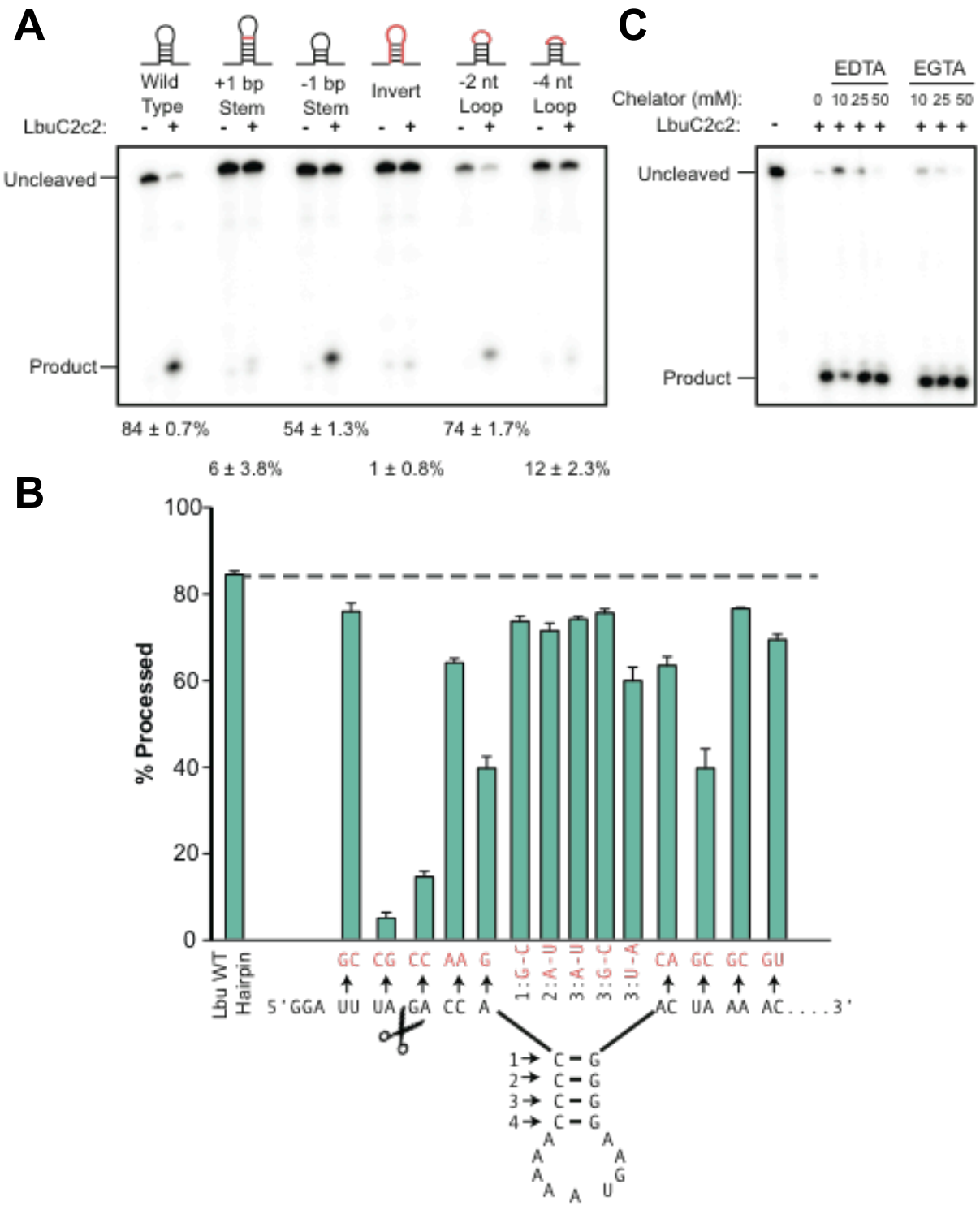

A
Wild Type
+1 bp Stem
-1 bp Stem
Invert
-2 nt Loop
-4 nt Loop
LbuC2c2:
Uncleaved
Product
84 ± 0.7%
54 ± 1.3%
74 ± 1.7%
6 ± 3.8%
1 ± 0.8%
12 ± 2.3%
C
EDTA
EGTA
Chelator (mM):
0 10 25 50 10 25 50
LbuC2c2:
Uncleaved
Product
B
% Processed
100
80
60
40
20
0
Lbu WT Hairpin
5'GGA UU UA GA CC A GC CG CC AA G 1:G-C 2:A-U 3:A-U 3:G-C 3:U-A CA GC GC GU AC UA AA AC....3'
1→ C-G
2→ C-G
3→ C-G
4→ C-G

Figure 3.2. LbuC2c2-mediated crRNA biogenesis depends on both structure and sequence of CRISPR repeats. (A) Representative cleavage assay by LbuC2c2 on pre-crRNAs containing structural mutations within the stem and loop regions of hairpin. Processed percentages listed below are quantified at 1 h (mean ± s.d., $n = 3$). **(B)** Bar graph showing the dependence of pre-crRNA processing on the CRISPR repeat sequence. The wild-type (WT) repeat sequence is shown below with individual bars representing tandem nucleotide mutations as noted in red. The cleavage site is indicated by cartoon scissors. Percentage processed was measured after 1 h (mean ± s.d., $n = 3$). **(C)** Divalent metal ion dependence of the crRNA processing reaction was tested by the addition of 10–50 mM EDTA and EGTA to standard reaction conditions.

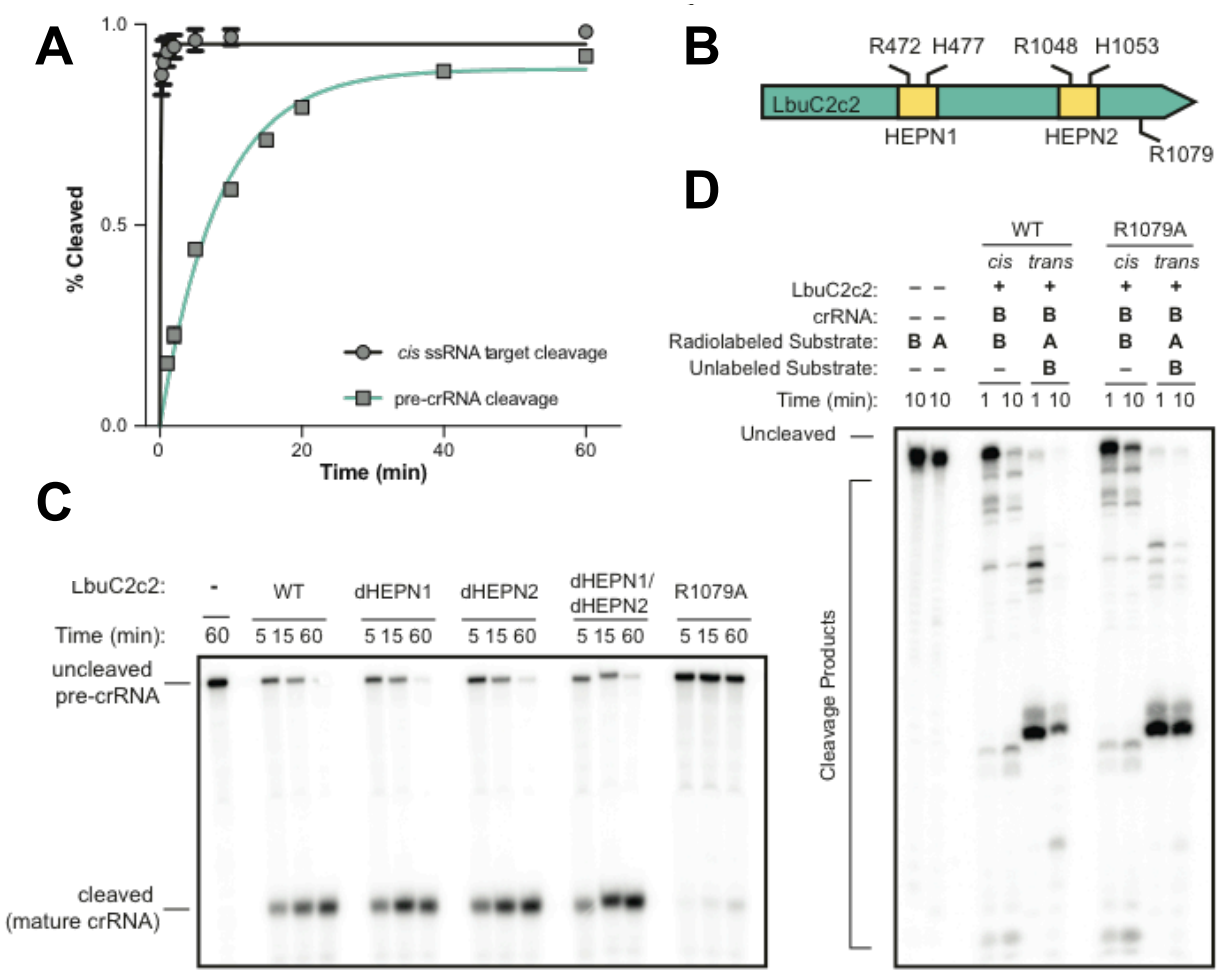


Figure 3.3. LbuC2c2 contains two distinct RNase activities. (A) Quantified time-course data of *cis* ssRNA target (black) and pre-crRNA (teal) cleavage by LbuC2c2. Exponential fits are shown as solid lines ($n = 3$), and the calculated pseudo-first-order rate constants (k_{obs}) (mean ± s.d.) are $9.74 \pm 1.15\ min^{-1}$ and $0.12 \pm 0.02\ min^{-1}$for *cis* ssRNA target and pre-crRNA cleavage, respectively. **(B)** LbuC2c2 architecture depicting the location of HEPN motifs and processing-deficient point mutant **(C,D)** Representative ribonuclease activity of LbuC2c2 mutants for pre-crRNA processing in **(C)** and ssRNA targeting in **(D).**

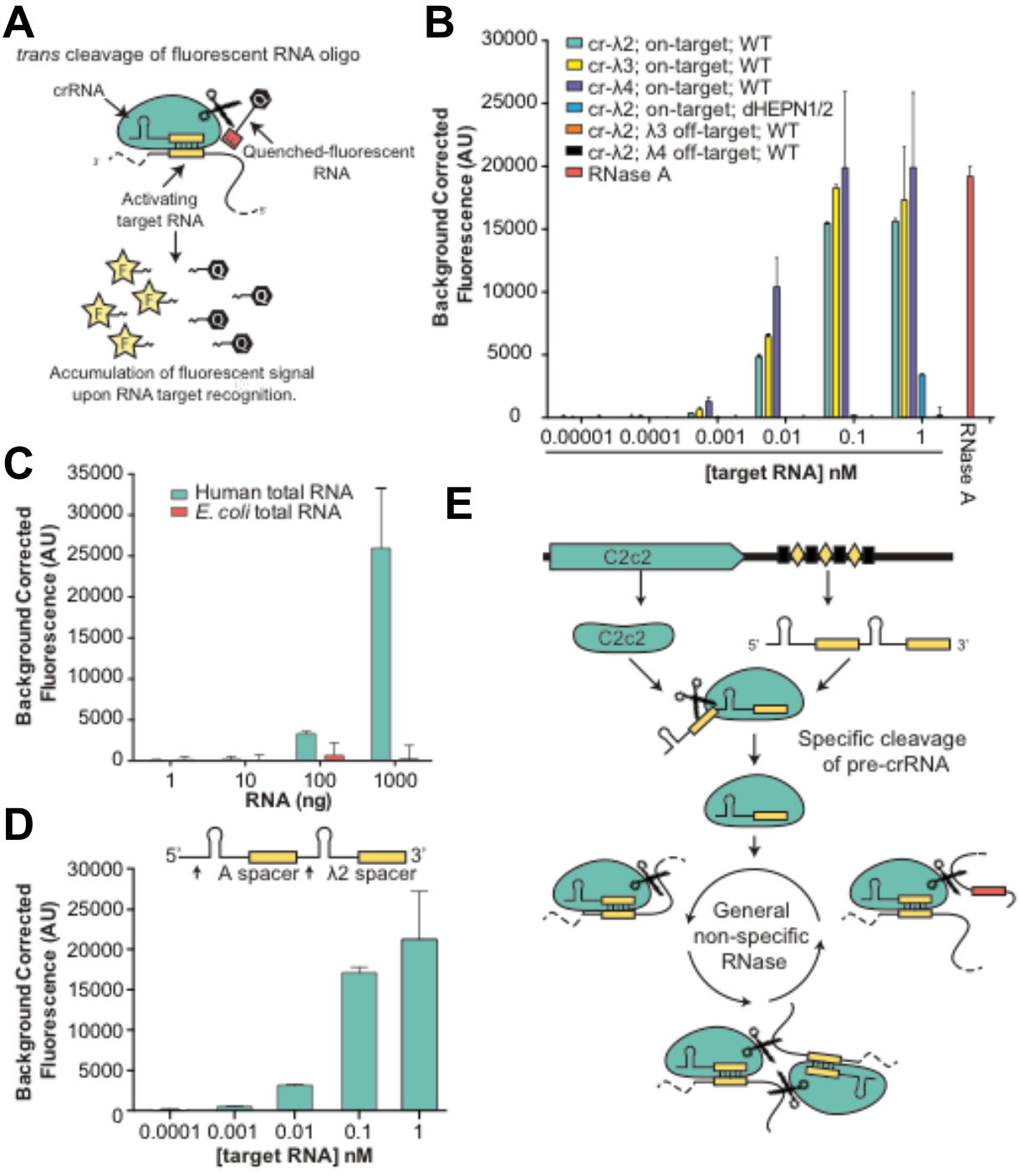

A
trans cleavage of fluorescent RNA oligo
crRNA
Quenched-fluorescent RNA
Activating target RNA
Accumulation of fluorescent signal upon RNA target recognition.
B
cr-λ2; on-target; WT
cr-λ3; on-target; WT
cr-λ4; on-target; WT
cr-λ2; on-target; dHEPN1/2
cr-λ2; λ3 off-target; WT
cr-λ2; λ4 off-target; WT
RNase A
Background Corrected Fluorescence (AU)
30000
25000
20000
15000
10000
5000
0
0.00001 0.0001 0.001 0.01 0.1 1
RNase A
[target RNA] nM
C
Human total RNA
E. coli total RNA
Background Corrected Fluorescence (AU)
35000
30000
25000
20000
15000
10000
5000
0
1 10 100 1000
RNA (ng)
D
5'
A spacer
λ2 spacer
3'
Background Corrected Fluorescence (AU)
30000
25000
20000
15000
10000
5000
0
0.0001 0.001 0.01 0.1 1
[target RNA] nM
E
C2c2
C2c2
5'
3'
Specific cleavage of pre-crRNA
General non-specific RNase

Figure 3.4. C2c2 provides sensitive detection of transcripts in complex mixtures. **(A)** Illustration of LbuC2c2 RNA detection approach using a quenched fluorescent RNA reporter. **(B)** Quantification of fluorescence signal generated after 30 min by wild-type or catalytically dead (dHEPN1/2) LbuC2c2 loaded with either a λ2-, λ3- or λ4-targeting crRNA (cr-; as indicated) in the presence of varying concentrations of λ2−λ4 target ssRNA and human total RNA. RNase A shown as positive RNA degradation control (mean ± s.d., $n = 3$). AU, arbitrary units. **(C)** Quantification of fluorescence signal generated by LbuC2c2 loaded with a β-actin targeting crRNA after 3 h for varying amounts of human total RNA or bacterial total RNA (as a β-actin-null negative control) (mean ± s.d., $n = 3$). **(D)** Tandem pre-crRNA processing also enables RNA detection (mean ± s.d., $n = 3$). **(E)** Model of the type VI CRISPR pathway highlighting both of the C2c2 RNase activities.

Materials and Methods

C2c2 phylogenic and candidate selection

C2c2 maximum-likelihood phylogenies were computed using RAxML (Stamatakis, 2014) with the PROTGAMMALG evolutionary model and 100 bootstrap samplings. Sequences were aligned by MAFFT with the 'einsi' method (Katoh and Standley, 2013).

C2c2 protein production and purification

Expression vectors for protein purification were assembled using synthenic gBlocks ordered from Integrated DNA Technologies. The codon-optimized C2c2 genomic sequence was *N*-terminally tagged with a His_6-MBP-TEV cleavage site, with expression driven by a T7 promoter. Mutant proteins were cloned via site-directed mutagenesis of wild-type C2c2 constructs. Expression vectors were transformed into Rosetta2 *E. coli* cells grown in 2xYT broth at 37 °C. *E. coli* cells were induced during log phase with 0.5 M ITPG, and the temperature was reduced to 16 °C for overnight expression of His-MBP-C2c2. Cells were subsequently harvested, resuspended in lysis buffer (50 mM Tris-HCl pH 7.0, 500 mM NaCl, 5% glycerol, 1 mM TCEP, 0.5mM PMSF, and EDTA-free protease inhibitor (Roche)) and lysed by sonication, and the lysates were clarified by centrifugation. Soluble His-MBP-C2c2 was isolated over metal ion affinity chromatography, and protein-containing eluate was incubated with TEV protease at 4 °C overnight while dialyzing into ion exchange buffer (50 mM Tris-HCl pH 7.0, 250 mM KCl, 5% glycerol, 1 mM TCEP) in order to cleave off the His_6-MBP tag. Cleaved protein was loaded onto a HiTrap SP column and eluted over a linear KCl (0.25-1.5M) gradient. Cation exchange chromatography fractions were pooled and concentrated with 30 kD cutoff concentrators (Thermo Fisher). The C2c2 protein was further purified via size-exclusion chromatography on an S200 column and stored in gel filtration buffer (20 mM Tris-HCl pH 7.0, 200 mM KCl, 5% glycerol, 1 mM TCEP) for subsequent enzymatic assays. Expression plasmids are deposited with Addgene. Representative PAGE analysis of purification fractions and size-exclusion chromatography trace is shown in Fig A2.3.

Generation of RNA

All RNAs used in this study were transcribed in vitro except for crRNA AES461 which was ordered synthetically (Integrated DNA Technologies) [see Table 3.1]. In vitro transcription reactions were performed as previously described with the following modifications: the T7 polymerase concentration was reduced to 10 μg/mL, and the UTP concentration was reduced to 2.5 mM (Sternberg et al., 2012). Transcriptions were incubated at 37°C for 1-2 hr to reduce non-template addition of nucleotides and quenched via treatment with DNase I at 37°C for 0.5-1 hr. Transcription reactions were purified by 15% denaturing polyacrylamide gel electrophoresis (PAGE), and all RNAs were resuspended in cleavage buffer (20 mM HEPES pH 6.8, 50 mM KCl, 5 mM $MgCl_2$, and 5% glycerol). For radioactive experiments, 5' triphosphates were removed by calf intestinal phosphate (New England Biolabs) prior to radiolabeling and ssRNA substrates were then 5'-end labeled using T4 polynucleotide kinase (New England Biolabs) and [γ-^{32}P]-ATP (Perkin Elmer) as described previously (Sternberg et al., 2012).

Pre-crRNA processing assays

Pre-crRNA cleavage assays were performed at 37 °C in RNA processing buffer (20 mM HEPES pH 6.8, 50 mM KCl, 5 mM $MgCl_2$, 10 µg/mL BSA, 100 µg/mL tRNA, 0.01% Igepal CA-630 and 5% glycerol) with a 100-fold molar excess of C2c2 relative to 5'-labeled pre-crRNA (final concentrations of 100 nM and <1 nM, respectively). Unless otherwise indicated, reaction was quenched after 60 min with 1.5X RNA loading dye (100% formamide, 0.025 w/v% bromophenol blue, and 200 µg mL^{-1} heparin). After quenching, reactions were denatured at 95 °C for 5 min prior to resolving by 12% or 15% denaturing PAGE (0.5X TBE buffer). Metal dependence of the reaction was tested by addition of EDTA or EGTA to reaction buffer at concentrations varying from 10-100 mM. Bands were visualized by phosphorimaging and quantified with ImageQuant (GE Healthcare). The percent cleavage was determined as the ratio of the product band intensity to the total intensity of both the product and uncleaved pre-crRNA bands and normalized for background within each measured substrate using ImageQuant TL Software (GE Healthcare) and fit to a one phase exponential association using Prism (GraphPad).

Product Size Mapping and 3' end moiety identification

Cleavage product length was determined biochemically by comparing gel migration of product bands to alkaline hydrolysis and RNase T1 digestion ladders using the RNase T1 Kit from Ambion. For hydrolysis ladder, 15 nM full-length RNA substrates were incubated at 95°C in 1X alkaline hydrolysis buffer (Ambion) for 5 min. Reactions were quenched with 1.5X RNA loading buffer, and cooled to -20°C to immediately stop hydrolysis. For RNase T1 ladder, 15 nM full length RNA substrates were unfolded in 1X RNA sequencing buffer (Ambion) at 65°C. Reactions were cooled to ambient temperature, and then 1 U of RNase T1 (Ambion) was added to reaction. After 15 min, reactions were stopped by phenol-chlorofrom extraction and 1.5X RNA loading buffer was added for storage. Hydrolysis bands were resolved in parallel to cleavage samples on 15% denaturing PAGE and visualized by phosphorimaging. For 3' end moiety identification, products from the processing reaction were incubated with 10 U of T4 polynucleotide kinase (New England Biolabs) for 1 hr at 37°C in processing buffer. Reactions were quenched with 1.5X RNA loading buffer, resolved on 20% denaturing PAGE and visualized by phosphorimaging.

Small RNA sequencing analysis

RNA reads from Shmakov *et al.* (2015) were downloaded from SRA runs SRR3713697, SRR3713948, and SRR3713950. The paired-end reads were locally mapped to the reference sequences using Bowtie2 (2012) with the following options: "--reorder --very-fast-local --local". The mapping was then filtered to retain only alignments that contained no mismatch using mapped.py (https://github.com/christophertbrown/mapped) with the "-m 0 -p both" options. BAM file of the resulting mapping are in the supplementary files for this manuscript. Read coverage was visualized using Geneious and plotted using Prism (GraphPad).

Target cleavage assays

Target cleavages assays were performed at 25 °C or 37 °C in cleavage buffer (20 mM HEPES pH 6.8, 50 mM KCl, 5 mM $MgCl_2$, and 5% glycerol). crRNA guides were pre-folded by heating to 65 °C for 5 min and then slowly cooling to ambient temperature in cleavage buffer. C2c2:crRNA complex formation was performed in cleavage buffer, generally at a molar ratio of 2:1 protein to crRNA at 37 °C for 10 min, prior to adding 5'-end labeled target and/or other non-radiolabeled RNA target substrates. Unless otherwise indicated, final concentrations of protein, guide, and targets were 100 nM, 50 nM, and <1 nM respectively for all reactions. Reactions were quenched with 1.5X RNA loading dye and resolved by 15% denaturing PAGE (0.5X TBE buffer). Bands were visualized by phosphorimaging and quantified with ImageQuant (GE Healthcare). The percent cleavage was determined as the ratio of total banding intensity for all shorter products relative to the uncleaved band and normalized for background within each measured substrate using ImageQuant TL Software (GE Healthcare) and fit to a one phase exponential association using Prism (GraphPad).

crRNA filter-binding assays

Filter binding assays was carried out in RNA processing buffer (20 mM HEPES pH 6.8, 50 mM KCl, 5 mM $MgCl_2$, 10 µg/mL BSA, 100 µg/mL yeast tRNA, 0.01% Igepal CA-630 and 5% glycerol). LbuC2c2 was incubated with radiolabeled crRNA (<0.1 nM) for 1hr at 37°C. Tufryn, Protran and Hybond-N+ were assembled onto a dot-blot apparatus in the order listed above. The membranes were washed twice with 50µL Equilibration Buffer (20 mM HEPES pH 6.8, 50 mM KCl, 5 mM $MgCl_2$ and 5% glycerol) before the sample was applied to the membranes. Membranes were again washed with 50 µL Equilibration Buffer, dried and visualized by phosphorimaging. Data were quantified with ImageQuant TL Software (GE Healthcare) and fit to a binding isotherm using Prism (GraphPad Software). All experiments were carried out in triplicate. Dissociation constants and associated errors are reported in the figure legends.

Electrophoretic mobility-shift assays

In order to avoid the dissociation of the LbuC2c2-dHEPN1/dHEPN2: crRNA complex at low concentrations during ssRNA-binding experiments, binding reactions contained a constant excess of LbuC22c2-dHEPN1/dHEPN2 (200 nM), and increasing concentrations of crRNA-A and < 0.1 nM target ssRNA. Assays were carried out in C2c2 EMSA buffer (20 mM HEPES pH 6.8, 50 mM KCl, 10 µg/mL BSA, 100 µg/mL yeast tRNA, 0.01% Igepal CA-630 and 5% glycerol). LbuC2c2-crRNA-A complexes were pre-formed as described above for 10 min at 37°C before the addition of 5'-radiolabelled ssRNA substrate and a further incubation for 45 min at 37°C. Samples were then resolved by 8% native PAGE at 4°C (0.5X TBE buffer). Gels were imaged by phosphorimaging, quantified using ImageQuant TL Software (GE Healthcare) and fit to a binding isotherm using Prism (GraphPad Software). All experiments were carried out in triplicate. Dissociation constants and associated errors are reported in the figure legends.

Fluorescent RNA detection assay

LbuC2c2:crRNA complexes were preassembled by incubating 1μM of Lbu-C2c2:C2c2 with 500 nM of crRNA for 10 min at 37°C . These complexes were then diluted to 100nM LbuC2c2: 50 nM crRNA-λ2 in RNA processing buffer (20 mM HEPES pH 6.8, 50 mM KCl, 5 mM $MgCl_2$, 10 μg/mL BSA, 10 μg/mL yeast tRNA, 0.01% Igepal CA-630 and 5% glycerol) in the presence of 185 nM of RNAase-Alert substrate (Thermo-Fisher), 100 ng of HeLa total RNA and increasing amounts of target 60 nt ssRNA (0-1 nM). These reactions were incubated in a fluorescence plate reader for up to 120 min at 37°C with fluorescence measurements taken every 5 min (λ_{ex}: 485 nm; λ_{em}: 535 nm). Background-corrected fluorescence values were obtained by subtracting fluorescence values obtained from reactions carried out in the absence of target ssRNA. Maximal fluorescence was measured by incubating 50 nM RNaseA with 185 nM of RNAase-Alert substrate. For measurement of crRNA-ACTB mediated LbuC2c2 activation by *beta-actin* mRNA in human total RNA, LbuCas9:crRNA complexes were preassembled by incubating 1μM of LbuC2c2 with 500 nM of crRNA-ACTB for 10 min at 37°C and reactions were carried out in the conditions above in the presence of increasing amounts (0-1 μg) of either HeLa cell total RNA or E. Coli total RNA (as a negative control). These reactions were incubated in a fluorescence plate reader for up to 180 min at 37°C with fluorescence measurements taken every 5 min (λ_{ex}: 485 nm; λ_{em}: 535 nm). Background-corrected fluorescence values were obtained by subtracting fluorescence values obtained from reactions carried out in the absence of target ssRNA. For coupled pre-crRNA processing and RNA detection assays, LbuCas9-crRNA complexes were preassembled by incubating 1μM of LbuC2c2 with 500 nM of pre-crRNA-A-λ2 for 20 min at 37°C and reactions carried out as described above in the presence of increasing amounts of ssRNA A and ssRNA λ2 (0-1 nM each). In each case, error bars represent the standard deviation from three independent experiments.

Background cleavage in total RNA

LbuC2c2:crRNAλ4 complexes were assembled as previously described for fluorescence RNA detection assay. Complexes were incubated in RNA processing buffer in the presence of 3 ug total RNA with and without 10 nM λ4 ssRNA target. After 2 hr, RNA was isolated by trizol extraction and ethanol precipitation. The RNA fragment size distribution of resuspended samples was resolved using Small RNA Analysis Kit (Agilent) on a Bioanalyzer 2100 (Agilent) using the manufacturer's protocol. Fluorescent intensity curves were normalized in Prism for curve overlay (GraphPad Software).

Table 3.1 Oligonucleotides used in study

Oligo Name	Sequence
Lbu_pre-crRNA_A_SCK314	GGAUUUAGACCACCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lse_pre-crRNA_B_AES484	GGUAAGAGACUACCUCUAUAUGAAAGAGGACUAAAACCAAACAUGAUCUGGGUCAUC
Lsh_pre-crRNA_A_SCK339	GGAUUUAGACCACCCCAAUAUCGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre-crRNA_invert_SCK321	GGAUUUAGACCAGGGGAAGUAAAAACCCCACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre-crRNA_5stem_SCK331	GGAUUUAGACCACCCCCAAAAAUGAAGGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre-crRNA_7bubble_SCK334	GGAUUUAGACCACCCCAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre-crRNA_5bubble_SCK335	GGAUUUAGACCACCCCAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre-crRNA_3stem_SCK342	GGAUUUAGACCACCCCAAAAAUGAAGGGCACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre-cr_5'_mut1_AES497	GGCGUUAGACCACCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre-cr_5'_mut2_AES496	GGAGCUAGACCACCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre-cr_5'_mut3_AES495	GGAUCCAGACCACCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre-cr_5'_mut4_AES477	GGAUUCGGACCACCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre-cr_5'_mut5_AES482	GGAUUUACCCCACCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre-cr_5'_mut6_AES478	GGAUUUAAUCCACCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre-cr_5'_mut7_AES480	GGAUUUAGAAAACCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre_cr_5'_mut8_AES498	GGAUUUAGACCGCCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre_stem_mut1_AES502	GGAUUUAGACCAGCCCAAAAAUGAAGGGCACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre_stem_mut2_AES501	GGAUUUAGACCACCGCAAAAAUGAAGCGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre_stem_mut3_AES500	GGAUUUAGACCACACCAAAAAUGAAGGUGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre_stem_mut4_AES499	GGAUUUAGACCACCACAAAAAUGAAGUGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre_stem_mut5_AES504	GGAUUUAGACCACUCCAAAAAUGAAGGAGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre_cr_3'_mut1_AES505	GGAUUUAGACCACCCCAAAAAUGAAGGGGCAUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre_cr_3'_mut2_AES506	GGAUUUAGACCACCCCAAAAAUGAAGGGGACGCAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre_cr_3'_mut3_AES507	GGAUUUAGACCACCCCAAAAAUGAAGGGGACUAGCACAGGGGCAGAGAUGAUGACCCU
Lbu_pre_cr_3'_mut4_AES508	GGAUUUAGACCACCCCAAAAAUGAAGGGGACUAAAGUAGGGGCAGAGAUGAUGACCCU
crLbu_A_GG_AES432	GGCCACCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
crLbu_B_AES451	GGCCACCCCAAAAAUGAAGGGGACUAAAACACAAACAUGAUCUGGGUCAUC
A.0_target_AES450	GGCACACCCGCAGGGGGGAGCCAAAAGGGUCAUCAUCUCUGCCCCCACAGCAGAAGCCCC
B_target_AES452	GGGAACCCCAAGGCCAACCGCGAGAAGAUGACCCAGAUCAUGUUUGAGACCUUCAACACCCC
crLbu_Lambda2_AES453	GGCCACCCCAAAAAUGAAGGGGACUAAAACAGUGAUAAGUGGAAUGCCAUG
crLbu_Lambda3_MOC410	GGCCACCCCAAAAAUGAAGGGGACUAAAACACUGGUGAACUUCCGAUAGUG
crLbu_Lambda4_MOC411	GGCCACCCCAAAAAUGAAGGGGACUAAAACACAGAUAUAGCCUGGUGGUUC
Lambda2_target_MOC28	GGCUCAAUUUUGACAGCGGUCAUGGCAUUCCACUUAUCACUGGCAUCCUUCCACUC
Lambda3_target_MOC36	GGAAAUCAUUCAACACCCGCACUAUCGGAAGUUCACCAGCCAGCCGCAGCACGUU
Lambda4_target_MOC37	GGCAAUAAAAAUGCGCCGCCUGAACCACCAGGCUAUAUCUGCCACUCAUUGUUGUGA
crLbu_betaActin_1_AES451	GGCCACCCCAAAAAUGAAGGGGACUAAAACACAAACAUGAUCUGGGUCAUC
pre-crLbu_dimer_SCK324	GGAUUUAGACCACCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCUAUUUAGACCACCCCAAAAAUGAAGGGGACUAAAACAGUGAUAAGUGGAAUGCCAUG
crLbu_lambda2_SCK315	GGAUUUAGACCACCCCAAAAAUGAAGGGGACUAAAACAGUGAUAAGUGGAAUGCCAUG
Lbu_pre_cr_5'_4mer1_AES481	GGAUUUAAUAAACCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre_cr_5'_4mer2_AES479	GGAUUCGAUCCACCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre_cr_5'_4mer3_SCK343	GGAUUUAGGAAGCCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
Lbu_pre_cr_5'_4mer4_AES503	GGAUUUAGACCAGGCCAAAAAUGAAGGCCACUAAAACAGGGGCAGAGAUGAUGACCCU
crLbu_GuideWalk1_SCK302	GGCCACCCCAAAAAUGAAGGGGACUAAAACAUUUUGGCUCCCCCCCUGCAAAUGAG
crLbu_GuideWalk2_SCK303	GGCCACCCCAAAAAUGAAGGGGACUAAAACAACCCUUUUGGCUCCCCCCCUGCAAA
crLbu_GuideWalk3_SCK304	GGCCACCCCAAAAAUGAAGGGGACUAAAACAGAUGACCCUUUUGGCUCCCCCCCUG
crLbu_GuideWalk4_SCK305	GGCCACCCCAAAAAUGAAGGGGACUAAAACAAGAUGAUGACCCUUUUGGCUCCCCC
crLbu_GuideWalk5_SCK306	GGCCACCCCAAAAAUGAAGGGGACUAAAACAGCAGAGAUGAUGACCCUUUUGGCU
crLbu_GuideWalk6_SCK307	GGCCACCCCAAAAAUGAAGGGGACUAAAACAGGGGGCAGAGAUGAUGACCCUUUU
crLbu_GuideWalk7_SCK308	GGCCACCCCAAAAAUGAAGGGGACUAAAACACAGAGGGGGCAGAGAUGAUGACCC
crLbu_GuideWalk8_SCK309	GGCCACCCCAAAAAUGAAGGGGACUAAAACAUCAGCAGAGGGGGCAGAGAUGAUG
A.1_target_U_MOC279	GGCUCAUUUGCAGGGGGGAGCCAAAAGGGUCAUCAUCUCUGCCCCCUCUGCUGAUGCCCC
A.2_target_70nt_AES447	GGCCUGACUGCUCUCAUUUGCAGUUGGGAGCCAAAAGGGUCAUCAUCUCUGCCCCCUCUGCUGAUGCCCC
A.3_target_80nt_AES448	GGACCUGUGAAUCCUGACUGCUCUCAUUUGCAGUUGGGAGCCAAAAGGGUCAUCAUCUCUGCCCCCUCUGCUGAUGCCCC
A.4_5'_ts_shift_AES449	GGCACACCCGCAGGGUUUAGCCAAAAGGGUCAUCAUCUCUGCCCCCUCUGCUGAUGCCCC
crLbu_A_16nt_trunc_SCK282	GGCCACCCCAAAAAUGAAGGGGACUAAAACACAGAGAUGAUGACCCU

crLbu_A_24nt_ext_SCK283	GGCCACCCCAAAAAUGAAGGGGACUAAAACAAGAGGGGGCAGAGAUGAUGACCCU
crLbu_A_mature_GA_SCK340	GACCACCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
crLbu_A_mature_GGGA_SCK341	GGGACCACCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
crLbu_A_mature_CCA_AES461	CCACCCCAAAAAUGAAGGGGACUAAAACAGGGGCAGAGAUGAUGACCCU
T7 Forward (DNA)	TAATACGACTCACTATAGG

Appendix I: Supplementary figures for Chapter 2

This work was done collaboratively and was originally published in *Science* Magazine:

Knight, S.C., Xie, L., Deng, W. Guglielmi, B., Witkowsky, L.B., Bosanac, L., Zhang, E.T., El Beheiry, M., Masson, J.-B., Dahan, M., Liu, Z., Doudna, J.A., Tjian, R. Dynamics of CRISPR-Cas9 genome interrogation in living cells. Science *350*, 823-826 (2015).

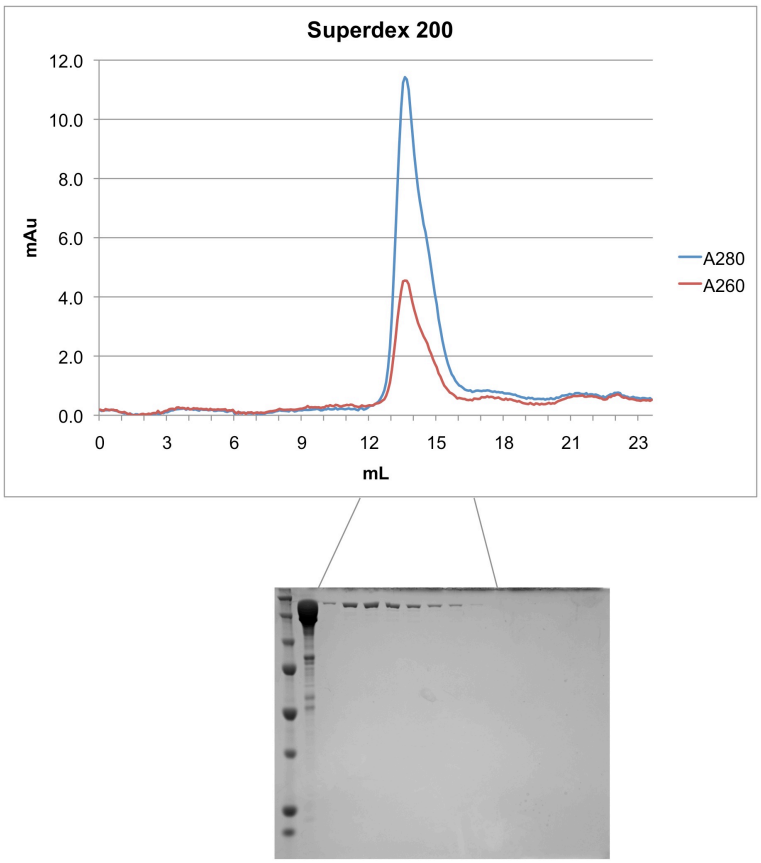


Figure A1.1. Size exclusion purification of His_6-dCas9-HaloTag. Fractions containing dCas9-HaloTag from the ion-exchange column were further purified on a size exclusion Superdex200 column (GE Life Sciences). Pure fractions containing dCas9-HaloTag were pooled and concentrated for subsequent enzymatic assays.

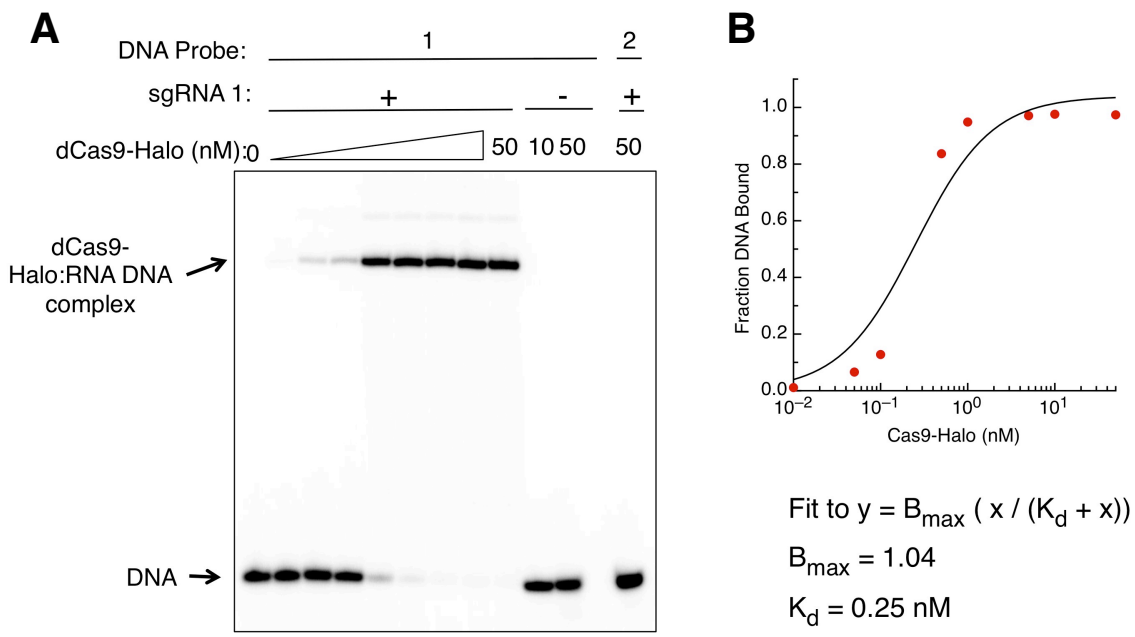


Figure A1.2. DNA binding assays with purified dCas9-HaloTag. (A) Gel-shift assay for determination of the dissociation constant for His_6-dCas9-HaloTag binding to DNA. dCas9-HaloTag was pre-incubated with sgRNA for 10 min prior to performing the EMSA. Assays were also performed in the absence of sgRNA (indicated with a (-) in the sgRNA channel) or with a non-target DNA probe (DNA probe 2 in the DNA channel) as negative controls. Refer to methods for sgRNA and DNA sequences and purification. **(B)** Fitting of EMSA data to extract the dissociation constant (K_d). Bands were quantified in ImageJ (NIH) and fit to the function shown in the figure panel.

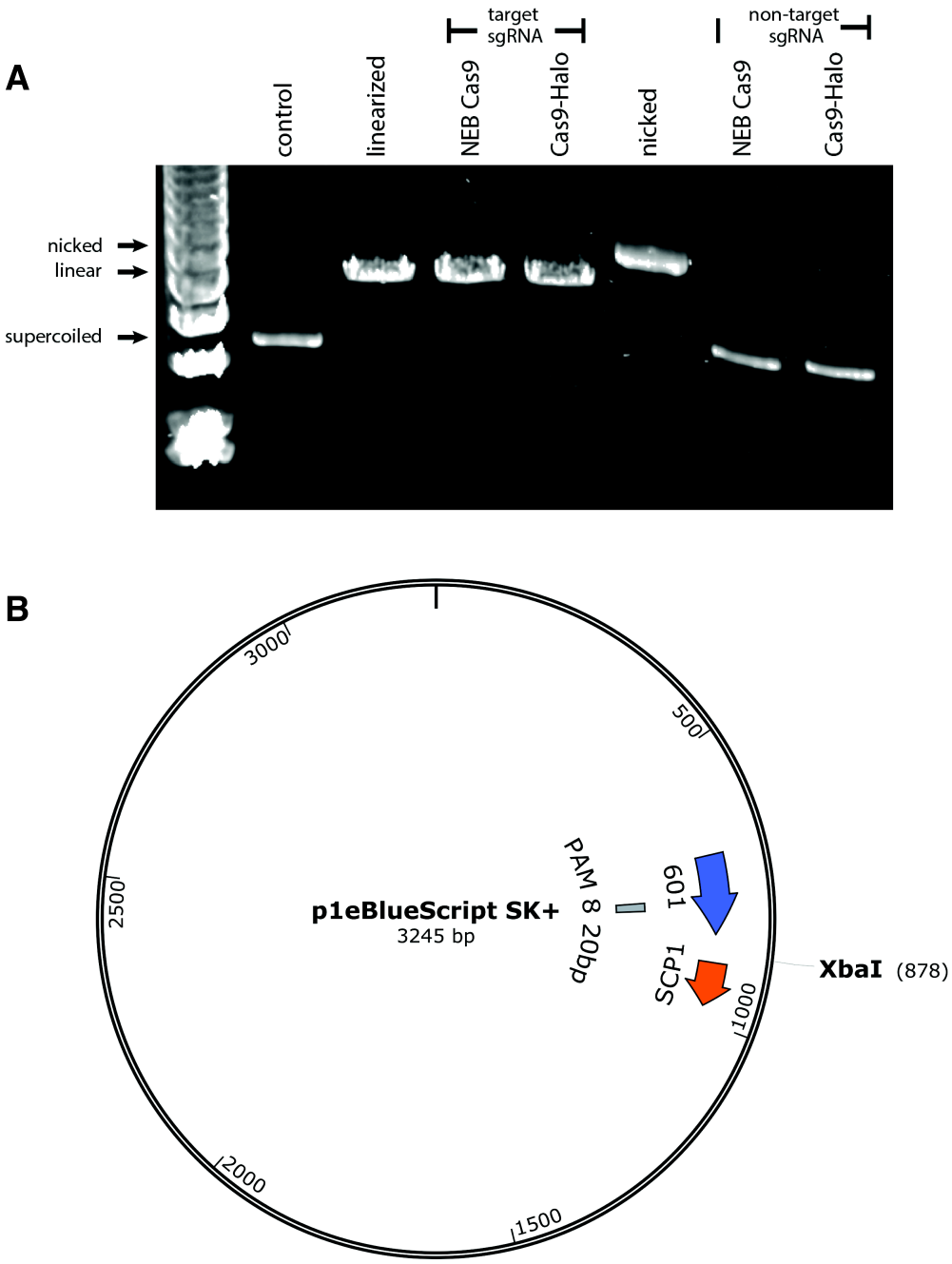


Figure A1.3. Cleavage assays with Cas9-HaloTag. Catalytically active His_6-Cas9-HaloTag protein was pre-incubated with sgRNA for 10 min prior to adding p1eBlueScript SK+ plasmid to a final reaction concentration of 500 nM sgRNA, 100 nM His_6-Cas9-HaloTag, and 20 nM plasmid DNA in reaction buffer. The reaction was incubated at 37 °C for 1 h prior to running on a 0.5% agarose gel (Panel A). Cleavage with XbaI served as a control for the linearized plasmid. No significant cleavage was observed by Cas9-HaloTag loaded with a non-target sgRNA. The plasmid map, including sense sgRNA mapping (designated PAM 8), is shown in Panel B.

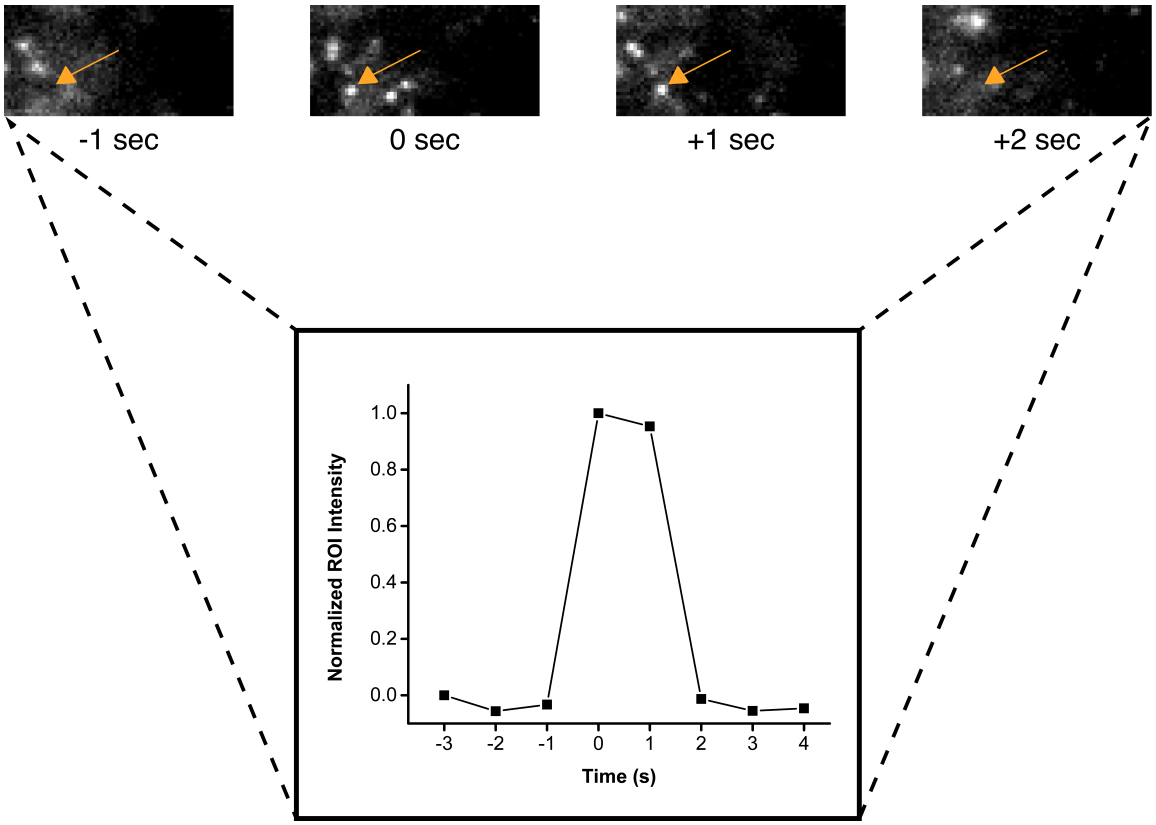


Figure A1.4. The HaloTag domain allows for visualization of single dCas9 molecules in vivo. Apo dCas9-HaloTag molecules were excited using a 561 nm laser and visualized using 2D-epi illumination with a 10 *ms* exposure time and 1 *s* lapse time between frames. The average signal intensity in a region of interest (ROI, indicated with an orange arrow) was quantified before association, during association (times t = 0 and 1 *s*), and after photobleaching/dissociation of a single dCas9-HaloTag molecule. Intensity was internally normalized relative to the signals at times t = –3 *s* (before association) and 0 *s* (first single-molecule association event).

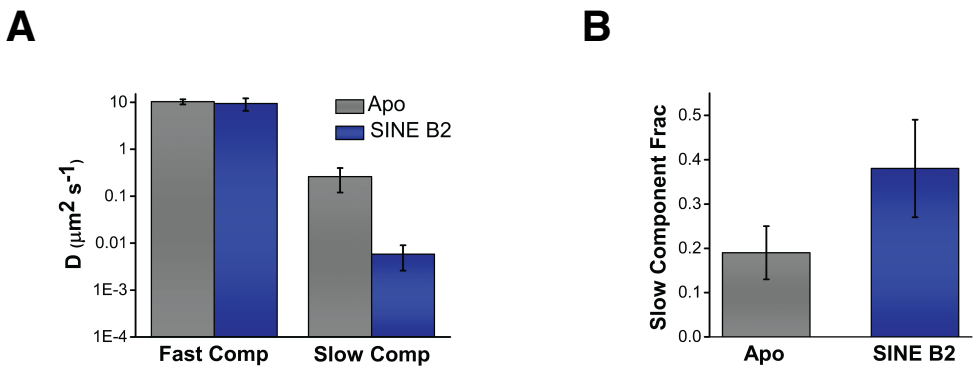


Figure A1.5. Quantification of Cas9 fluorescence correlation data. (A) Bar graphs illustrating the diffusion coefficient magnitudes for the fast and slow components from the FCS curves depicted in Figure 2.1C. **(B)** Fractional contribution of the slow component in the two-exponential fitting of FCS data for apo and Cas9 with B2 sgRNA. FCS measurements were conducted in diffraction-limited spots at random locations within cell nuclei (*N* = 11 cells for each condition).

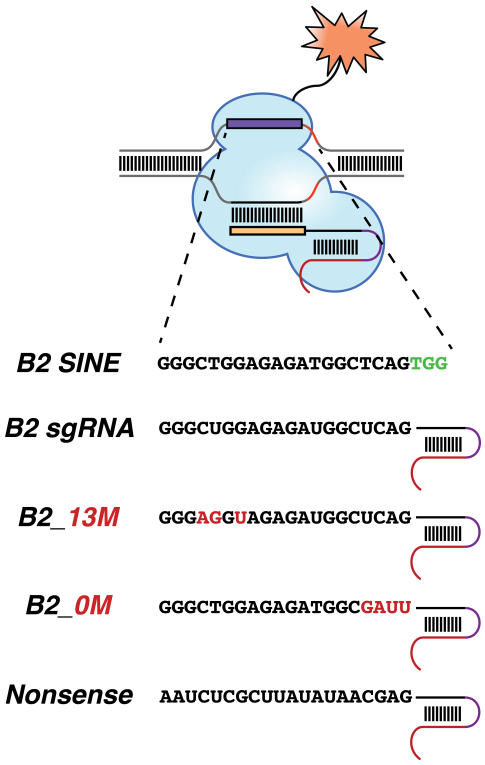
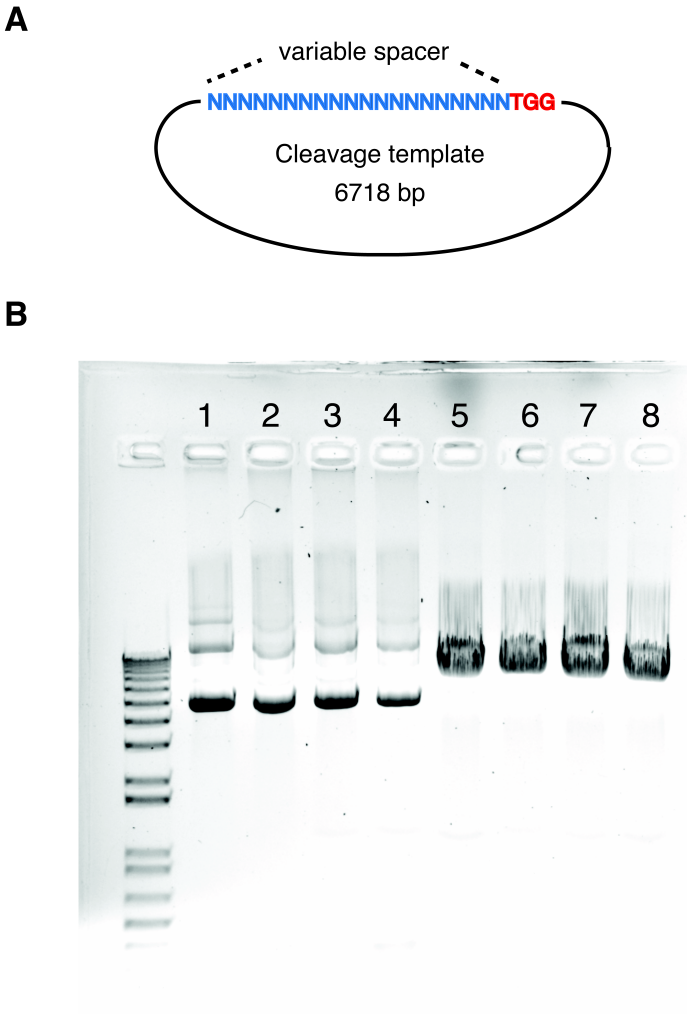


Figure A1.6. Design of the B2- and phage-derived nonsense sgRNAs used in this study. Red denotes positions at which nucleotides have been mutated relative to the homologous SINE B2 sequence. Green denotes the corresponding genomic PAM sequence. Single dCas9-HaloTag molecules were excited using a 561 nm laser and visualized using 2D epi-illumination (10 *ms* exposure time).



Lane Key

1. SINE B2 DNA only
2. SINE B2 DNA + SINE B2 sgRNA
3. SINE B2 DNA + apo Cas9
4. SINE B2 DNA + Nonsense sgRNA + Cas9
5. Nonsense DNA + Nonsense sgRNA + Cas9
6. B2_0M DNA + B2_0M sgRNA + Cas9
7. B2_13M DNA + B2_13M sgRNA + Cas9
8. SINE B2 DNA + SINE B2 sgRNA + Cas9

Figure A1.7. The nonsense and B2-derived sgRNAs are functional for Cas9-HaloTag activity. (A) General design of cleavage templates with variable 20 nt spacers (Nonsense, B2_0M, B2_13M, or SINE B2) and a uniform PAM (TGG) sequence (map available upon request). **(B)** Catalytically active His_6-Cas9-HaloTag protein was pre-incubated with sgRNA for 10 min prior to adding cleavage plasmid to a final reaction concentration of 500 nM sgRNA, 100 nM His_6-Cas9-HaloTag, and 20 nM DNA in reaction buffer (Methods). The reaction was incubated at 37 °C for 30 min prior to running on a 0.7% agarose gel.

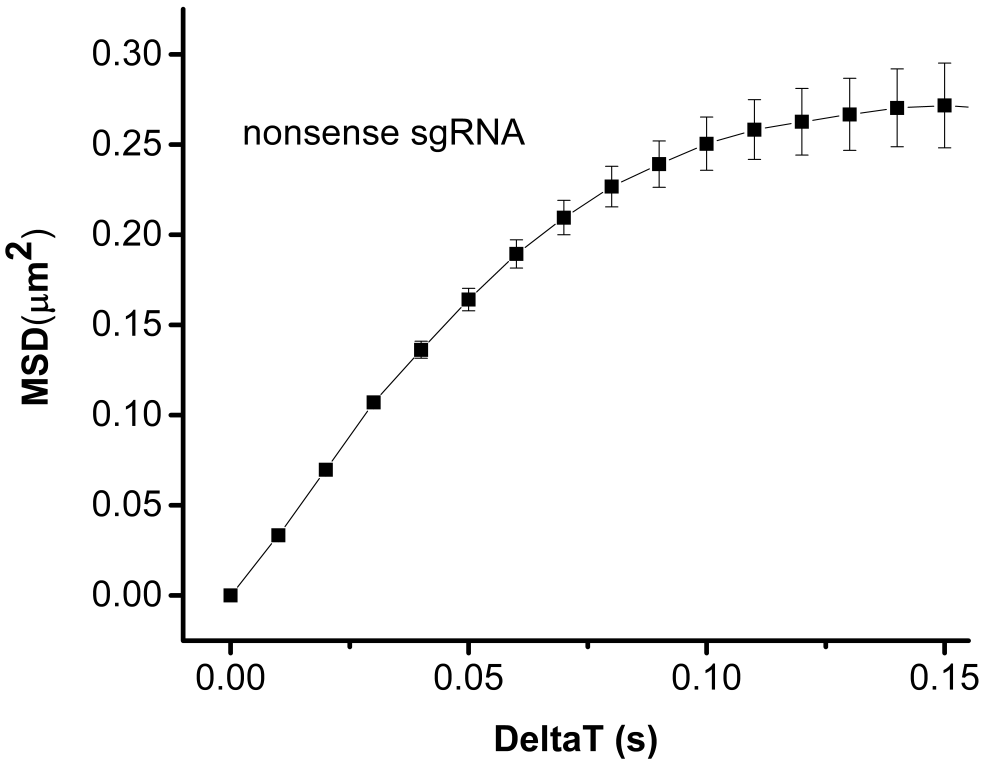


Figure A1.8. Mean square displacement curve for dCas9-HaloTag with nonsense sgRNA. Single dCas9-HaloTag molecules were excited using a 561 nm laser and visualized using 2D epi-illumination at a 10 *ms* exposure time. The mean square displacement for the population of trajectories was calculated for sliding ΔT windows using MSD analyzer.

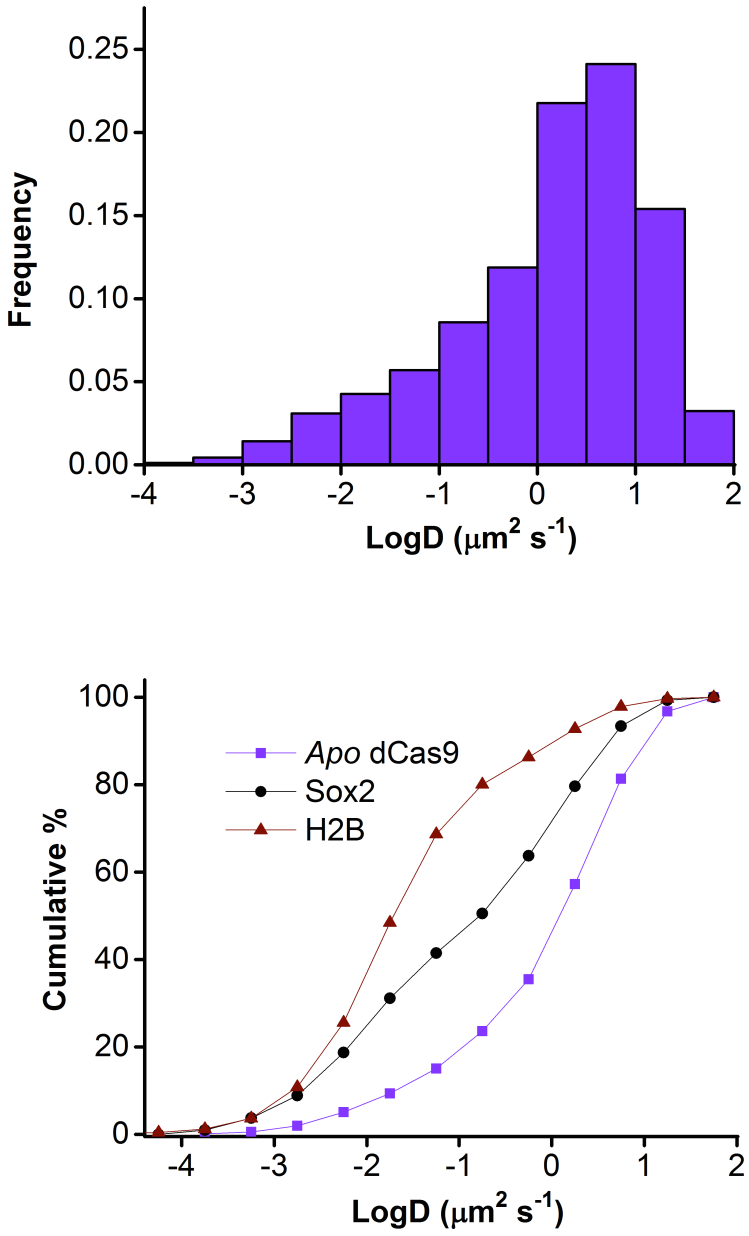


Figure A1.9. 2D diffusion analysis for apo dCas9-HaloTag. Images were collected at 10 *ms* exposure times using 2D epi-illumination. The root mean square displacement of our single-molecule localizations was analyzed to generate the log*D* profile shown in the top panel. The cumulative log*D* distribution is shown versus Sox2 and H2B in the bottom panel. *N* = 6 cells.

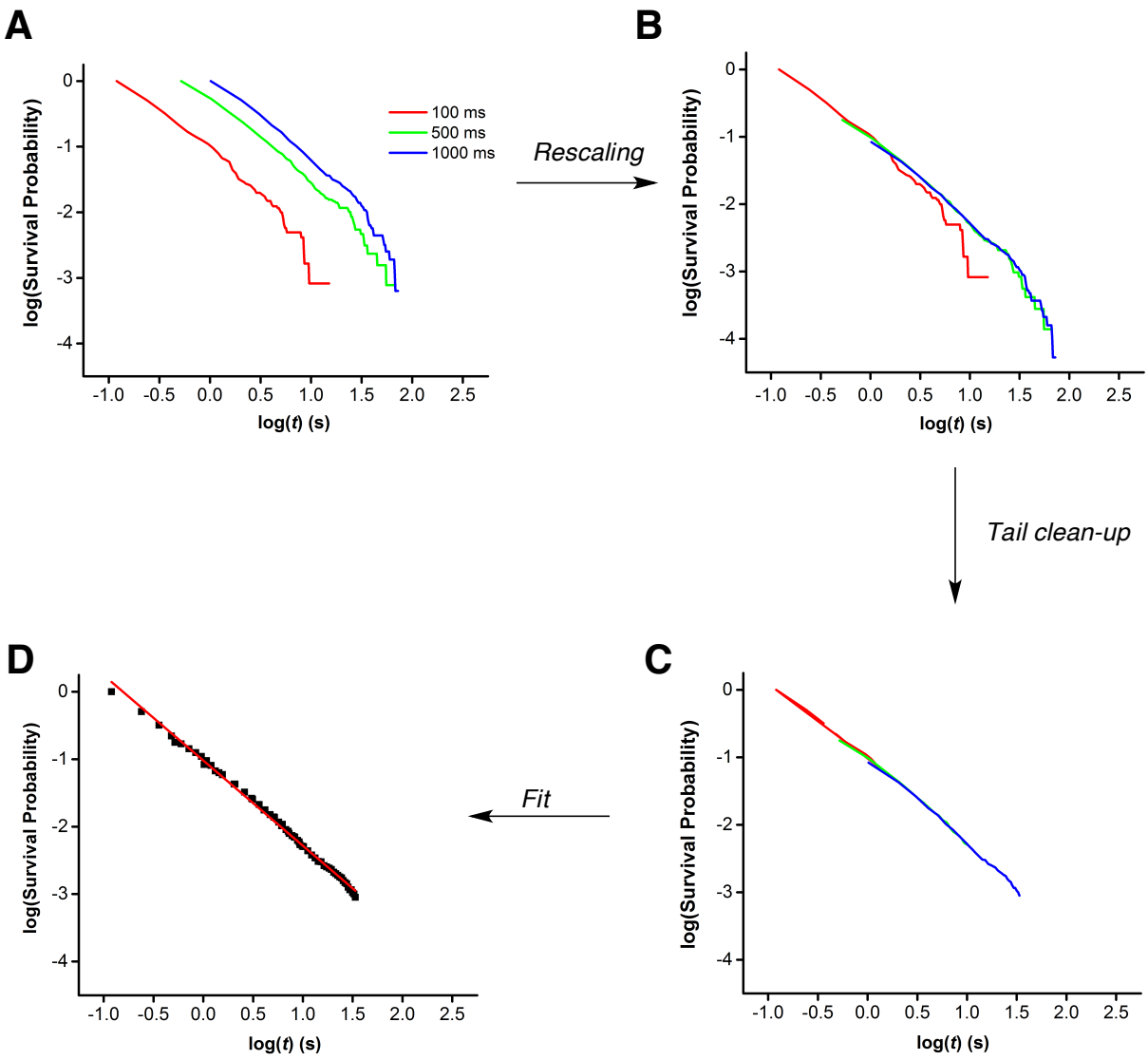


Figure A1.10. Survival probability plot concatenation and extraction of average off-target residence time. Time-lapse residence time measurements were used to construct survival probability plots and extract an averaged k_{off} for the nonsense sgRNA(Normanno et al., 1AD). Briefly, the raw survival probability plots depicted in (**A**) were rescaled and concatenated as in (**B**). The linear region of this concatenated plot (**C**) was fit as in (**D**). The inverse slope of this line corresponds to the average off-target residence time with the nonsense sgRNA (τ_{ns}, 0.75 ± 0.1 s). Importantly, this value is consistent with FRAP measurements of dCas9-eGFP with nonsense sgRNA, which suggest mostly short-lived (milliseconds to seconds) off-target interactions with chromatin.

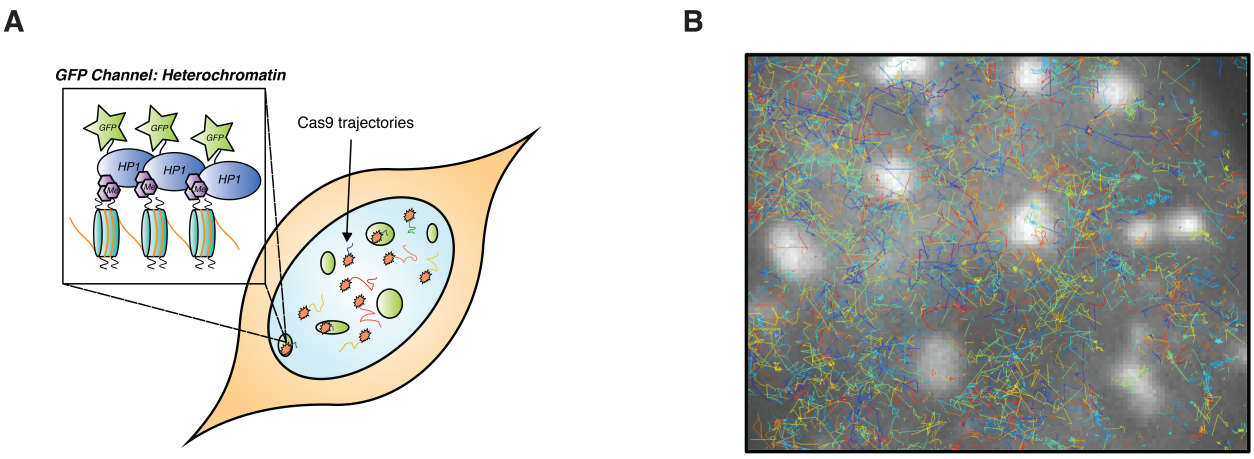


Figure A1.11. Cas9 undersamples heterochromatin while searching for targets within eukaryotic genomes. (A) Schematic illustrating labeling of HP1 with eGFP and tracking of dCas9-HaloTag molecules loaded with nonsense sgRNA within heterochromatic regions (HRs). dCas9-HaloTag molecules were switched to the dark state and stochastically re-excited and tracked using 405/561 nm excitation. **(B)** Sample composite image showing the depletion of dCas9-HaloTag tracks (average 30 ± 9% density reduction) in HRs relative to open chromatin.

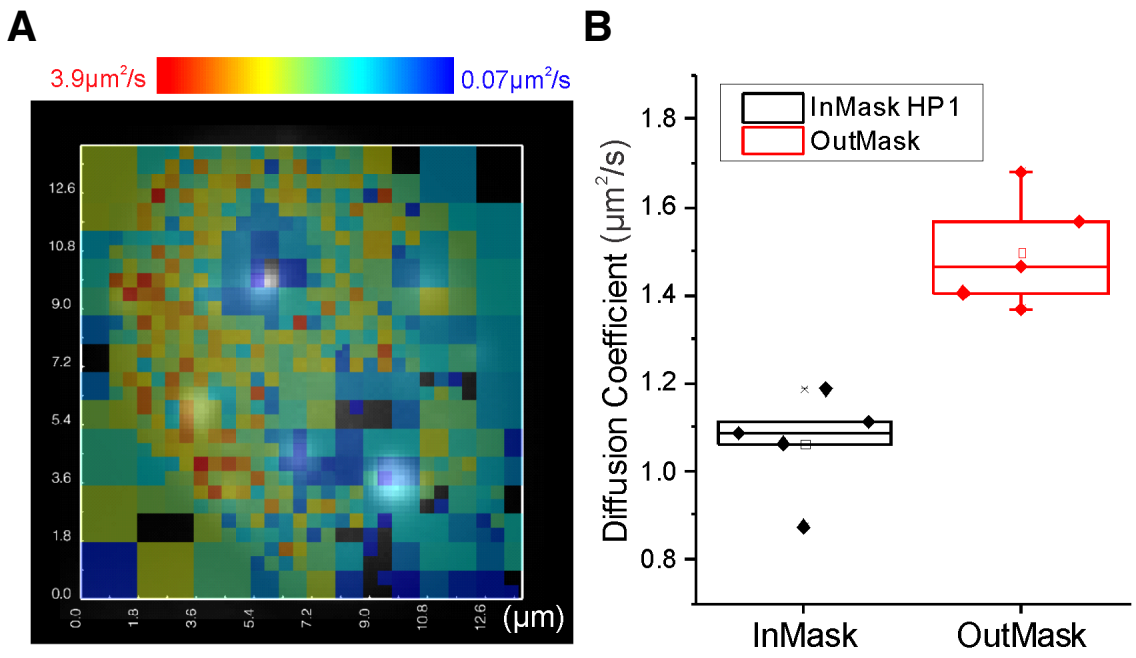


Figure A1.12. Cas9 HR diffusion analysis using Bayesian inference. (A) Diffusion analysis generated by InferenceMAP(Beheiry et al., 2015). The diffusion map is color-coded according to the scale bar on the right, with red corresponding to higher diffusion coefficients. The *x,y* units are in microns. **(B)** Averaged diffusion coefficients in heterochromatic regions (black) versus other regions of the nucleus (red) (*N* = 5 cells). Briefly, the diffusion map from (A) was converted to a grayscale image with pixel intensities corresponding to diffusion coefficient magnitudes. Binary masks were generated based on HP1 intensities in the image, and the pixel-averaged diffusion coefficients were calculated for HP1-enriched (InMask) and HP1-depleted (OutMask) regions. Regions with no calculation by InferenceMAP were excluded from the analysis.

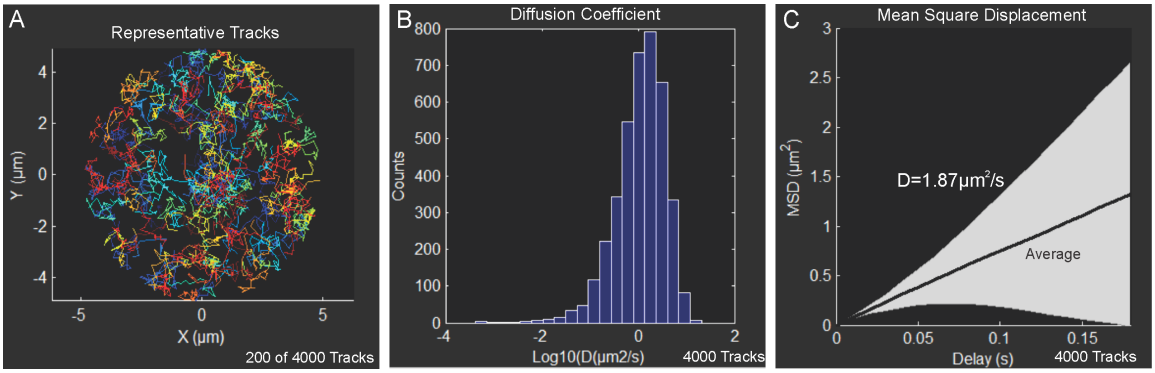


Figure A1.13. Brownian diffusion simulation in the nucleus. (A) Representative tracks of Brownian diffusion simulation in a nucleus (see Methods for details of the track simulation). **(B)** Diffusion coefficient histogram of 4000 simulated tracks. Diffusion coefficients (*x*-axis) are in log scale. **(C)** Mean square displacement (MSD) plot for 4000 simulated tracks. Weighted average values for each delay are presented as the middle line. The grayed area represents the weighted standard deviation over all MSD curves.

Appendix II: Supplementary figures for Chapter 3

This work was done collaboratively and was originally published in *Nature*:

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

*Indicates co-first author

All co-authors have consented to reprinting the original publication for this thesis. Reprinted with permission from Nature Publishing Group. Alexandra East-Seletsky, Spencer C. Knight, and Mitchell O'Connell conceived the study and designed experiments with assistance from Jamie H.D. Cate, Robert Tjian, and Jennifer A. Doudna. David Burstein performed bioinformatic analyses. Alexandra East-Seletsky and Mitchell O'Connell performed primary experiments for the Nature publication with technical assistance from Spencer C. Knight. All authors wrote and discussed the manuscript.

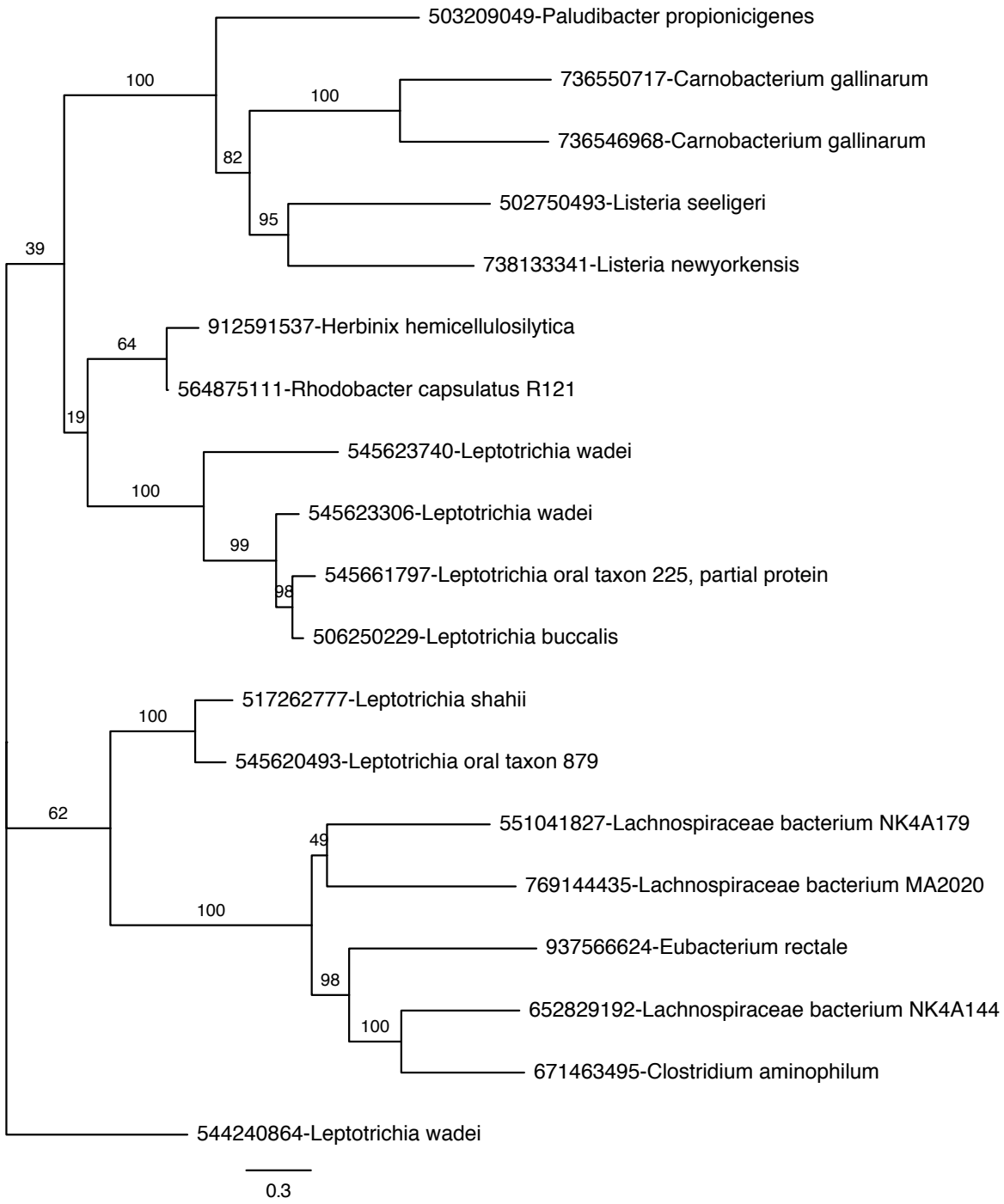


Figure A2.1. Phylogenetic tree of the C2c2 family. C2c2 coding sequences were aligned using a maximum likelihood estimation.

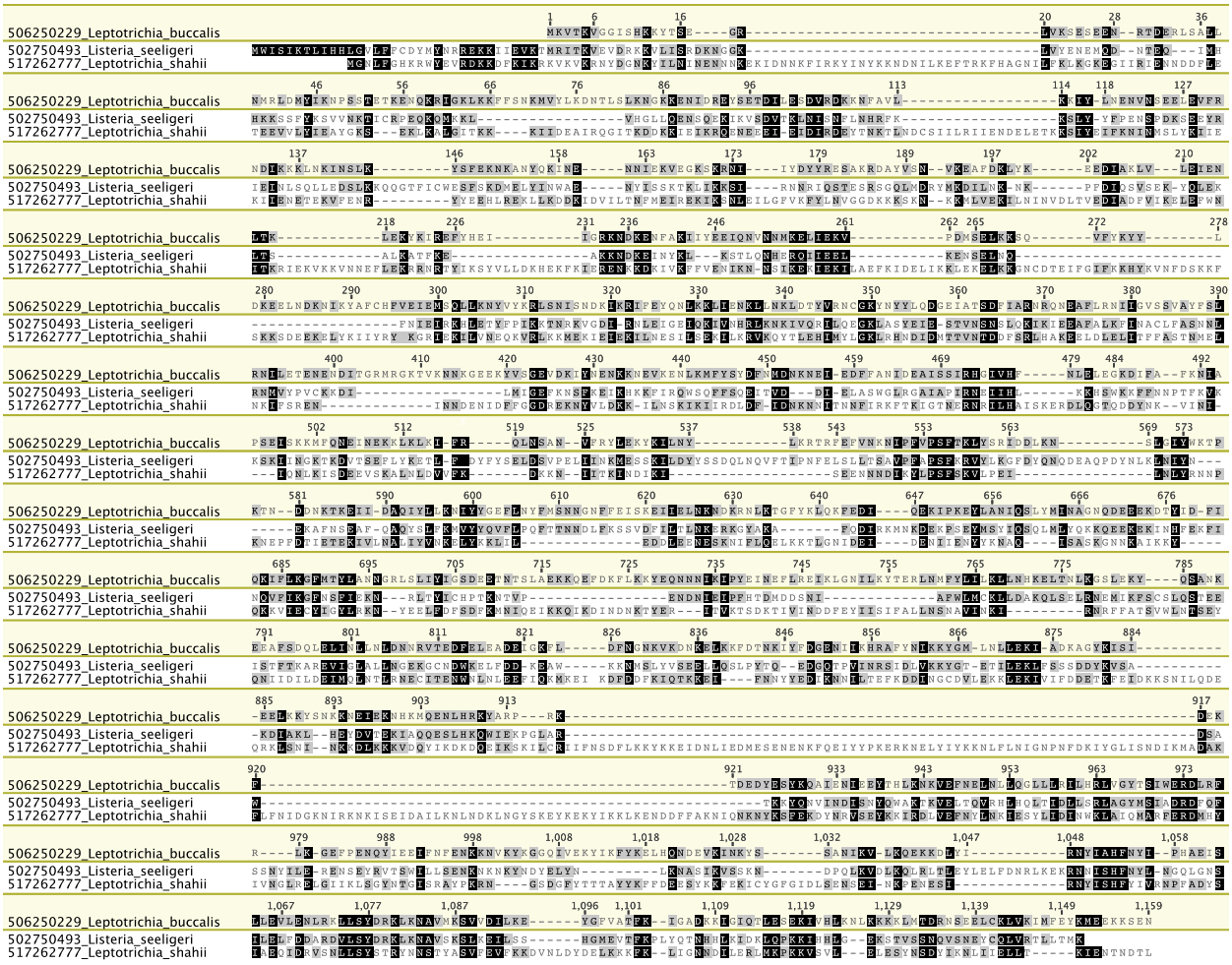


Figure A2.2. Alignment of protein sequences from three C2c2 homologs. Multiple sequence alignment of the three analyzed homologs of C2c2; coordinates are based on LbuC2c2.

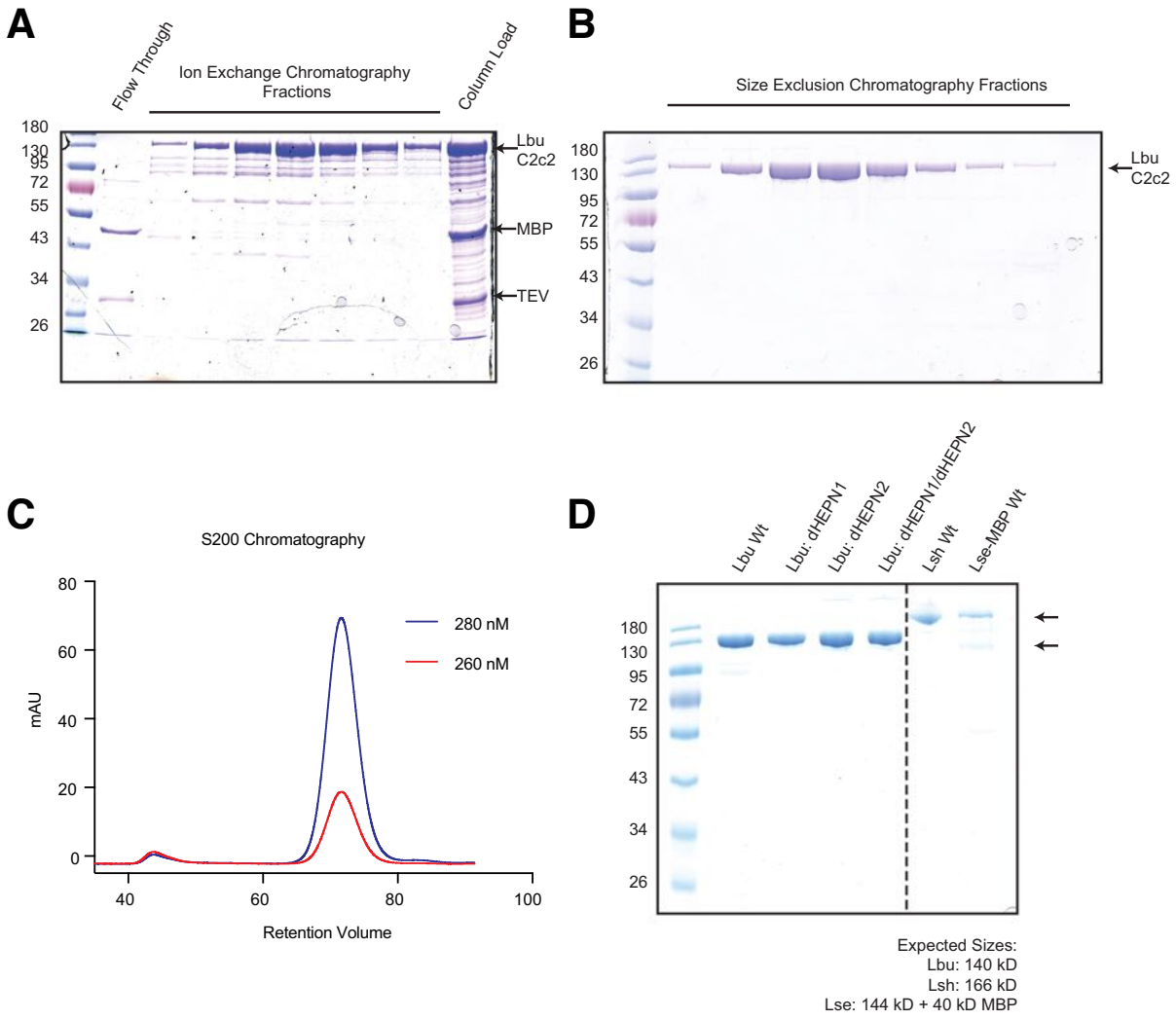


Figure A2.3. Purification and production of C2c2. All C2c2 homologs were expressed in *E. coli* as His-MBP fusions and purified by a combination of affinity, ion exchange and size exclusion chromatography. The Ni^{+} affinity tag was removed by incubation with TEV protease. Representative SDS-PAGE gels of chromatography fractions are shown in **(A, B)**. **(C)** The chromatogram from Superdex 200 (16/60) column demonstrating that C2c2 elutes as a single peak, devoid of nucleic acid. **(D)** SDS PAGE analysis of purified proteins used in this study.

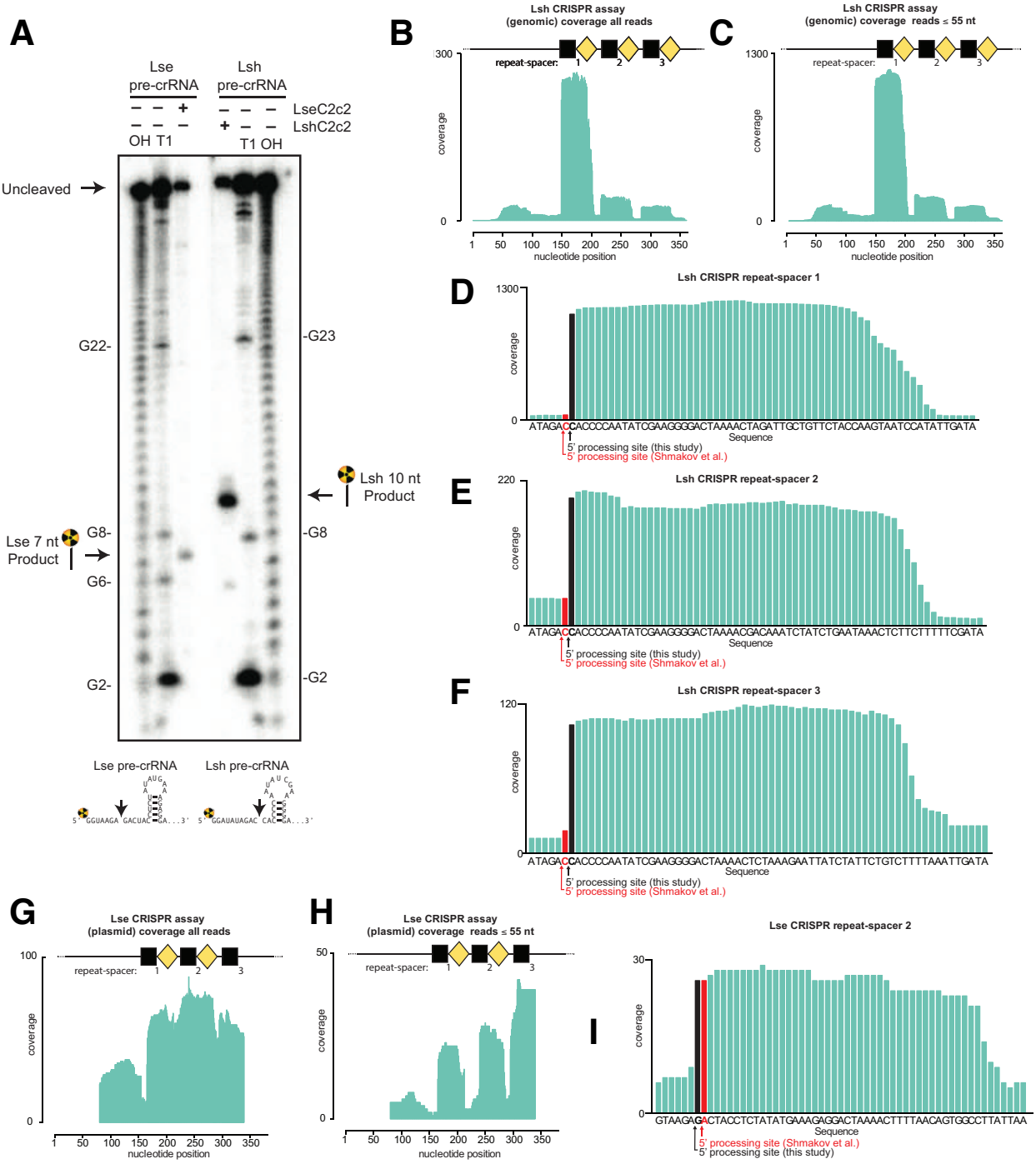

A
Lse pre-crRNA
Lsh pre-crRNA
LseC2c2
LshC2c2
OH T1
T1 OH
Uncleaved
G22-
-G23
Lsh 10 nt Product
Lse 7 nt Product
G8-
-G8
G6-
G2-
-G2
Lse pre-crRNA
Lsh pre-crRNA
B
Lsh CRISPR assay (genomic) coverage all reads
repeat-spacer:
coverage
nucleotide position
C
Lsh CRISPR assay (genomic) coverage reads ≤ 55 nt
D
Lsh CRISPR repeat-spacer 1
ATAGACCACCCCAATATCGAAGGGGACTAAAACTAGATTGCTGTTCTACCAAGTAATCCATATTGATA
Sequence
5' processing site (this study)
5' processing site (Shmakov et al.)
E
Lsh CRISPR repeat-spacer 2
ATAGACCACCCCAATATCGAAGGGGACTAAAACGACAAATCTATCTGAATAAACTCTTCTTTTTCGATA
F
Lsh CRISPR repeat-spacer 3
ATAGACCACCCCAATATCGAAGGGGACTAAAACTCTAAAGAATTATCTATTCTGTCTTTTAAATTGATA
G
Lse CRISPR assay (plasmid) coverage all reads
H
Lse CRISPR assay (plasmid) coverage reads ≤ 55 nt
I
Lse CRISPR repeat-spacer 2
GTAAGAGGACTACCTCTATATGAAAGAGGACTAAAACTTTTAACAGTGGCCTTATTAA

Figure A2.4. Mapping of pre-crRNA processing by C2c2 in vitro and in vivo. (A) Cleavage site mapping of LseC2c2 and LshCc2c2 cleavage of a single cognate pre-crRNA array. OH: alkaline hydrolysis ladder; T1: T1 RNase hydrolysis ladder. Cleavage reactions were performed with 100 nM C2c2 and <1 nM pre-crRNA. b-i, Re-analysis of LshC2c2 (**B-F**) and LseC2c2 (**G-I**) CRISPR array RNA sequencing experiments from Shmakov *et al*. (Fig. S7 and Fig. 5, respectively). All reads (**B,G**) and filtered reads (55 nt or less; as per original Shmakov *et al.* analysis; **C,H**) were stringently aligned to each CRISPR array using Bowtie2 (see Methods). Detailed views of individual CRISPR repeat-spacers are shown for Lsh (**D-F**) and Lse (**I**). Differences in 5' end pre-crRNA processing are indicated by arrows below each sequence. BAM alignment files of our analysis are available in Supplementary Materials. This mapping clearly indicates that the 5' ends of small RNA sequencing reads generated from Lsh pre-crRNAs map to a position 2 nts from the base of the predicted hairpin, in agreement with our in vitro processing data (**A**). This pattern holds for all mature crRNAs detected from both native expression in *L. shahii* and heterologous expression in *E. coli* (data not shown, BAM file available in supplementary methods). Unfortunately, the LseC2c2 crRNA sequencing data (used in **G-I**) is less informative due to low read depth, and each aligned crRNA exhibits a slightly different 5' end with little obvious uniformity. The mapping for one of the processed repeats (repeat-spacer 2; **I**) is in agreement with our data but only with low confidence due to the insufficient read depth.

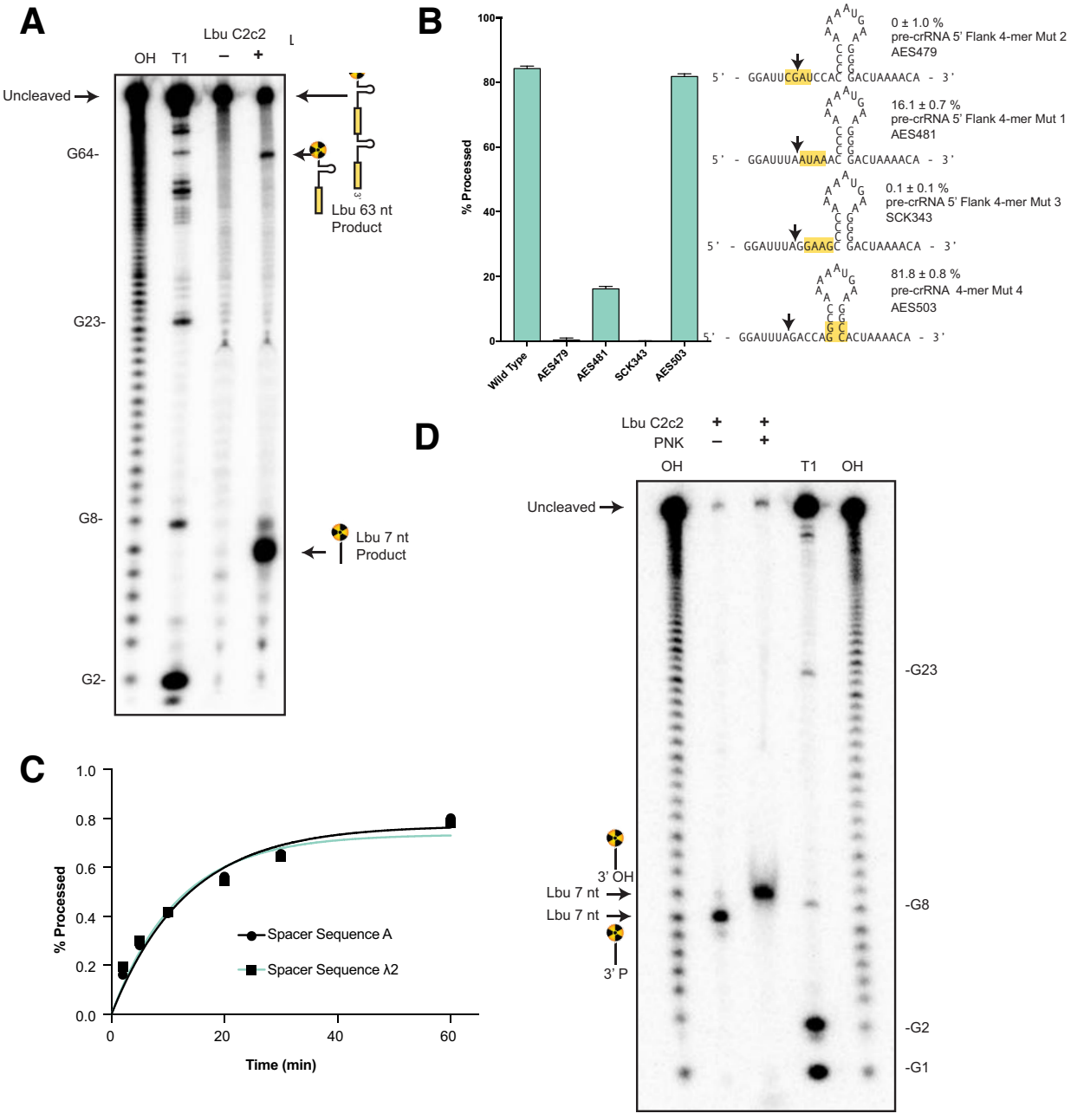

A
Lbu C2c2
OH
T1
–
+
Uncleaved
G64-
G23-
G8-
G2-
Lbu 63 nt Product
Lbu 7 nt Product
B
% Processed
100
80
60
40
20
0
Wild Type
AES479
AES481
SCK343
AES503
0 ± 1.0 %
pre-crRNA 5' Flank 4-mer Mut 2
AES479
5' - GGAUUCGAUCCAC GACUAAAACA - 3'
16.1 ± 0.7 %
pre-crRNA 5' Flank 4-mer Mut 1
AES481
5' - GGAUUUAAUAAAC GACUAAAACA - 3'
0.1 ± 0.1 %
pre-crRNA 5' Flank 4-mer Mut 3
SCK343
5' - GGAUUUAGGAAGC GACUAAAACA - 3'
81.8 ± 0.8 %
pre-crRNA 4-mer Mut 4
AES503
5' - GGAUUUAGACCAG CACUAAAACA - 3'
C
% Processed
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
Time (min)
Spacer Sequence A
Spacer Sequence λ2
D
Lbu C2c2
PNK
OH
T1
OH
Uncleaved
3' OH
Lbu 7 nt
Lbu 7 nt
3' P
-G23
-G8
-G2
-G1

Figure A2.5. Further investigations into the substrate requirements and mechanism of pre-crRNA processing by C2c2. (A) Cleavage site mapping of LbuCc2c2 cleavage of a tandem pre-crRNA array. OH: alkaline hydrolysis ladder; T1: T1 RNase hydrolysis ladder. Cleavage reactions were performed with 100 nM LbuC2c2 and <1 nM pre-crRNA. A schematic of cleavage products is depicted on right, with arrows indicating the mapped C2c2 cleavage products. **(B)** LbuC2c2 4-mer mutant pre-crRNA processing data demonstrating the importance of the 5' single-stranded flanking region for efficient pre-crRNA processing. Percentage of pre-crRNA processing was measured after 60 min (mean ± s.d., n = 3). **(C)** Representative LbuC2c2 pre-crRNA cleavage time-course demonstrating that similar rates of pre-crRNA processing occur independent of crRNA spacer sequence pseudo-first-order rate constants (k_{obs}) (mean ± s.d.) are 0.07 ± 0.04 min^{-1} and 0.08 ± 0.04 min^{-1} for spacer A and spacer λ2, respectively. **(D)** End group analysis of cleaved RNA by T4 polynucleotide kinase (PNK) treatment. Standard processing assay conditions were used to generate cleavage product, which was then incubated with PNK for 1 hr to remove any 2′, 3′-cyclic phosphates/3' monophosphates. Retarded migration of band indicates removal of the charged, monophosphate from the 3' end of radiolabeled 5' product.

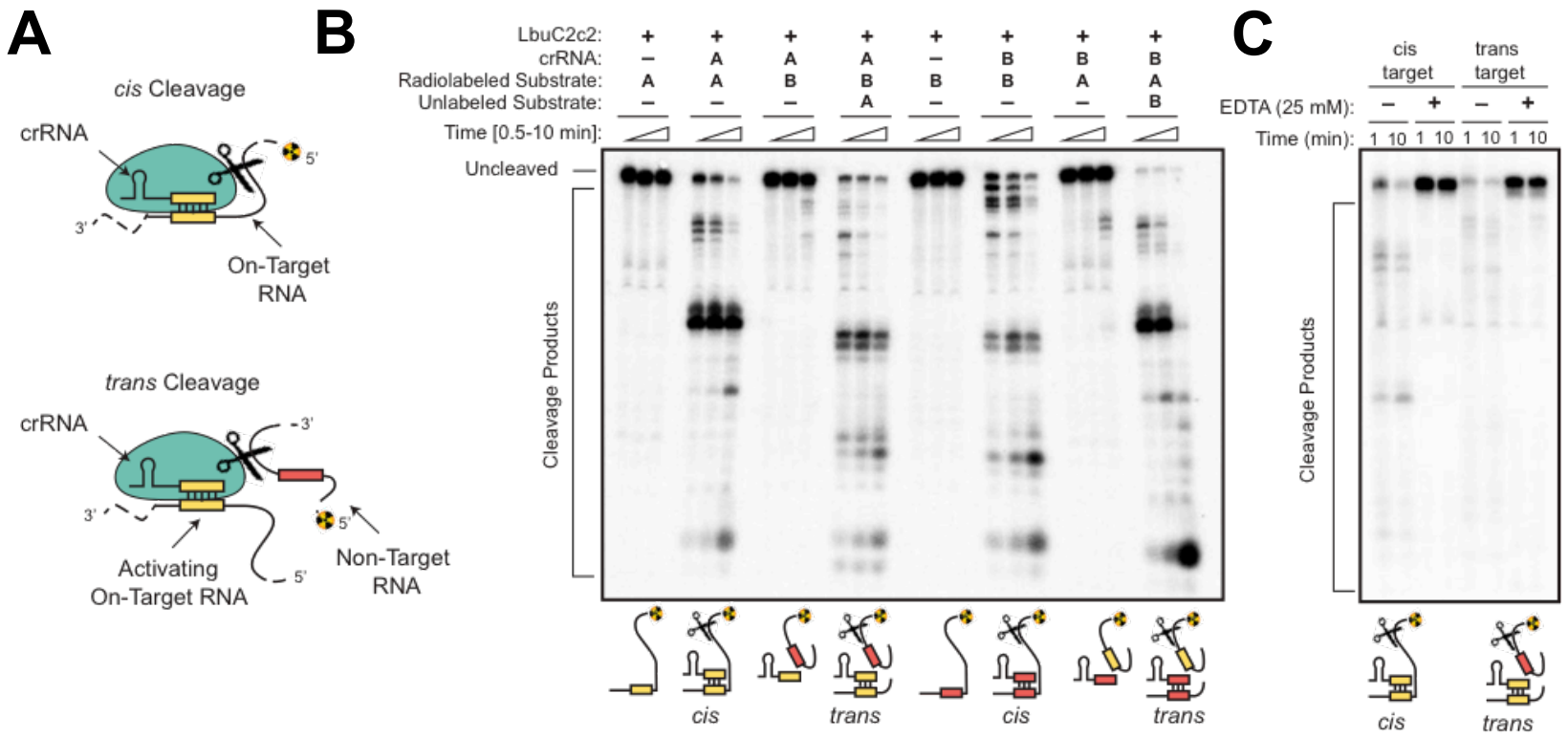
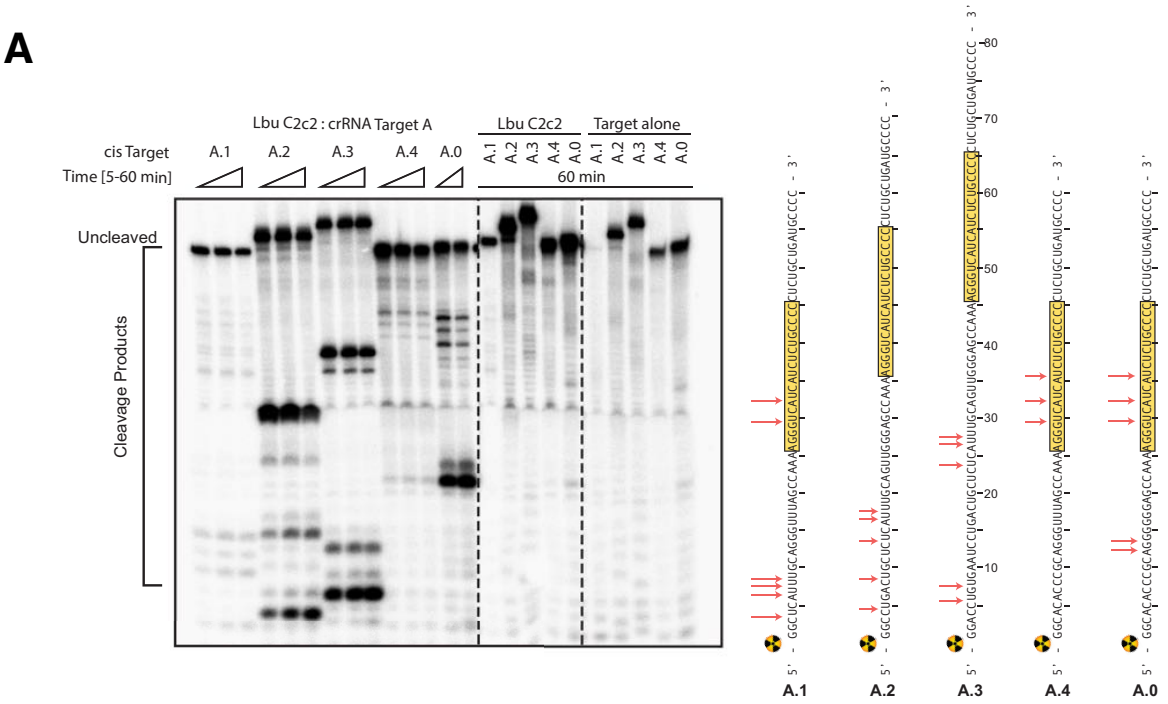
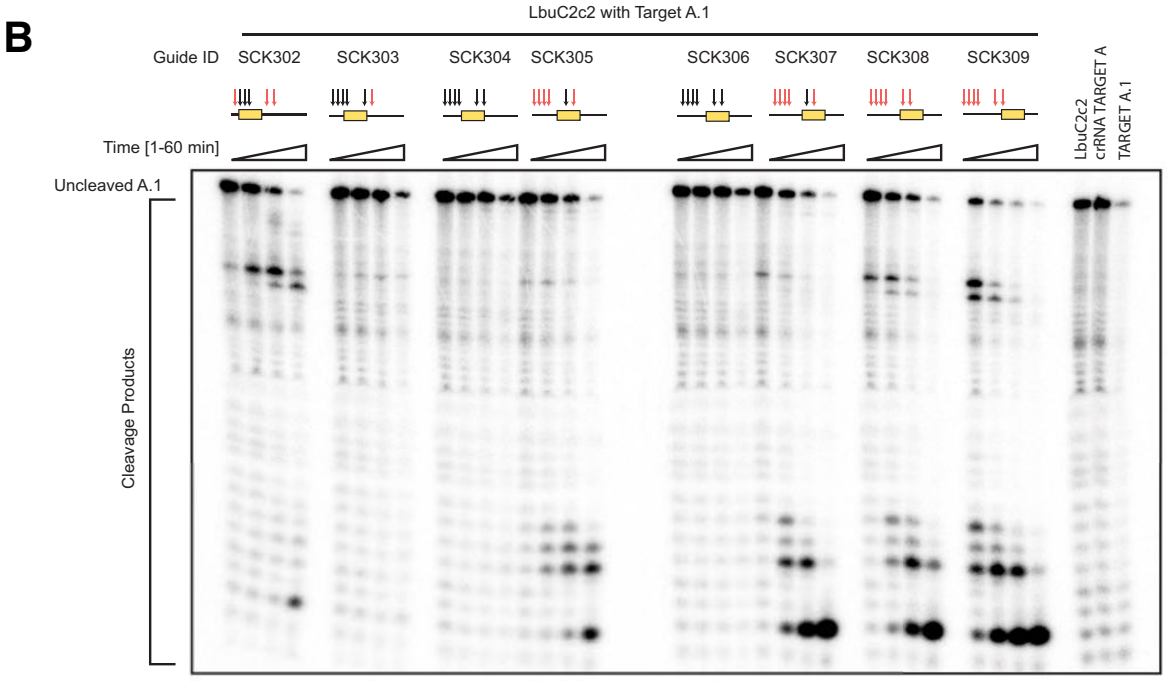

A
cis Cleavage
crRNA
5'
3'
On-Target
RNA
trans Cleavage
crRNA
3'
5'
Activating
On-Target RNA
Non-Target
RNA
B
LbuC2c2:
crRNA:
Radiolabeled Substrate:
Unlabeled Substrate:
Time [0.5-10 min]:
Uncleaved
Cleavage Products
cis
trans
cis
trans
C
cis
target
trans
target
EDTA (25 mM):
Time (min):
1 10 1 10 1 10 1 10
Cleavage Products
cis
trans

Figure A2.6. LbuC2c2 catalyzes guide-dependent ssRNA degradation on *cis* and *trans* targets. **(A)** Schematic of the two modes of C2c2, guide-dependent ssRNA degradation. **(B)** Cleavage of two distinct radiolabeled ssRNA substrates, A and B, by LbuC2c2. Complexes of 100 nM C2c2 and 50 nM crRNA were pre-formed at 37 °C, and reaction was initiated upon addition of <1 nM 5'-labeled target RNA at 25°C. *Trans* cleavage reactions contained equimolar (<1 nM) concentrations of radiolabeled non-guide-complementary substrate, and unlabeled on-target ssRNA. For multiple ssRNA substrates, we observed that LbuC2c2 catalyzed efficient cleavage only when bound to the complementary crRNA, indicating that LbuC2c2:crRNA cleaves ssRNA in an RNA-guided fashion This activity is hereafter referred to as on-target or *cis*-target cleavage. LbuC2c2-mediated *cis* cleavage resulted in a laddering of multiple products, with cleavage preferentially occurring before uracil residues, analogous to LshC2c2**(Abudayyeh et al., 2016)**. We repeated non-target cleavage reactions in the presence of unlabeled, on-target (crRNA-complementary) ssRNA. In contrast to non-target cleavage experiments performed in *cis*, we observed rapid degradation of non-target RNA in *trans.* The similar RNA cleavage rates and near identical cleavage products observed for both *cis* on-target cleavage and *trans* non-target cleavage implicate the same nuclease center in both activities. **(C)** LbuC2c2 loaded with crRNA targeting spacer A was tested for cleavage activity under both *cis* (target A labeled) and *trans* (target B labeled in the presence of unlabeled target A) cleavage conditions in the presence of 25 mM EDTA.





A.1: Spacer A 5' Flank with U's 60 nt target

```
5' - GGCUCAUUUGCAGGGUUUAGCCAAAAGGGUCAUCAUCUCUGCCCCCUCUGCUGAUGCCCC - 3'
SCK302 G
SCK303  C
SCK304   A
SCK305    C
SCK306     C
SCK307      U
SCK308       C
SCK309        U ← PFS
```

Figure A2.7. LbuC2c2 ssRNA target cleavage site mapping. (A) ssRNA target cleavage assay conducted per Methods demonstrating LbuC2c2-mediated 'cis'-cleavage of several radiolabeled ssRNA substrates with identical spacer-complementary sequences but distinct 5' flanking sequences of variable length and nucleotide composition. Sequences of ssRNA substrates are shown to the right with spacer-complementary sequences for crRNA-A highlighted in yellow. Arrows indicate detected cleavage sites. Gel was cropped for clarity. It should be noted that the pattern of cleavage products produced on different substrates (e.g. A.1, A.2, and A.3) indicates that the cleavage site choice is primarily driven by a uracil preference and exhibits an apparent lack of exclusive cleavage mechanism within the crRNA-complementary target sequence, which is in contrast to what is observed for other Class II CRISPR single effector complexes such as Cas9 and Cpf1 **(Fonfara et al., 2016; Jinek et al., 2012)**. Interestingly, the cleavage pattern observed for substrate A.0 hints at a secondary preference for polyG sequences. **(B)** LbuC2c2 ssRNA target cleavage assay as per Methods, using a range of crRNAs that tile the length of the ssRNA target. The sequence of the ssRNA substrates used in this experiment is shown below the gel with spacer-complementary sequences for each crRNA highlighted in yellow. Arrows indicate predicted cleavage sites. Above each set of lanes, a small diagram indicates the location of the spacer sequence along the target (yellow box) and the cleavage products observed (red arrows) or absent (black arrows). Likewise, it should be noted that for every crRNA the cleavage product length distribution is very similar, again indicating an apparent lack of exclusive cleavage within the crRNA-bound sequence. The absence of a several cleavage products in a subset of the reactions might be explained by the presence of bound C2c2:crRNA on the ssRNA target, which could sterically occlude access to uracils by any cis (intramolecular) or trans (intermolecular) LbuC2c2 active sites. While proper analysis for protospacer flanking site (PFS) preference for LbuC2c2 is beyond the scope of this study, minimal impact of the 3' flanking nucleotide was observed. Expected PFS base is noted in diagram next to each guide tested in red.

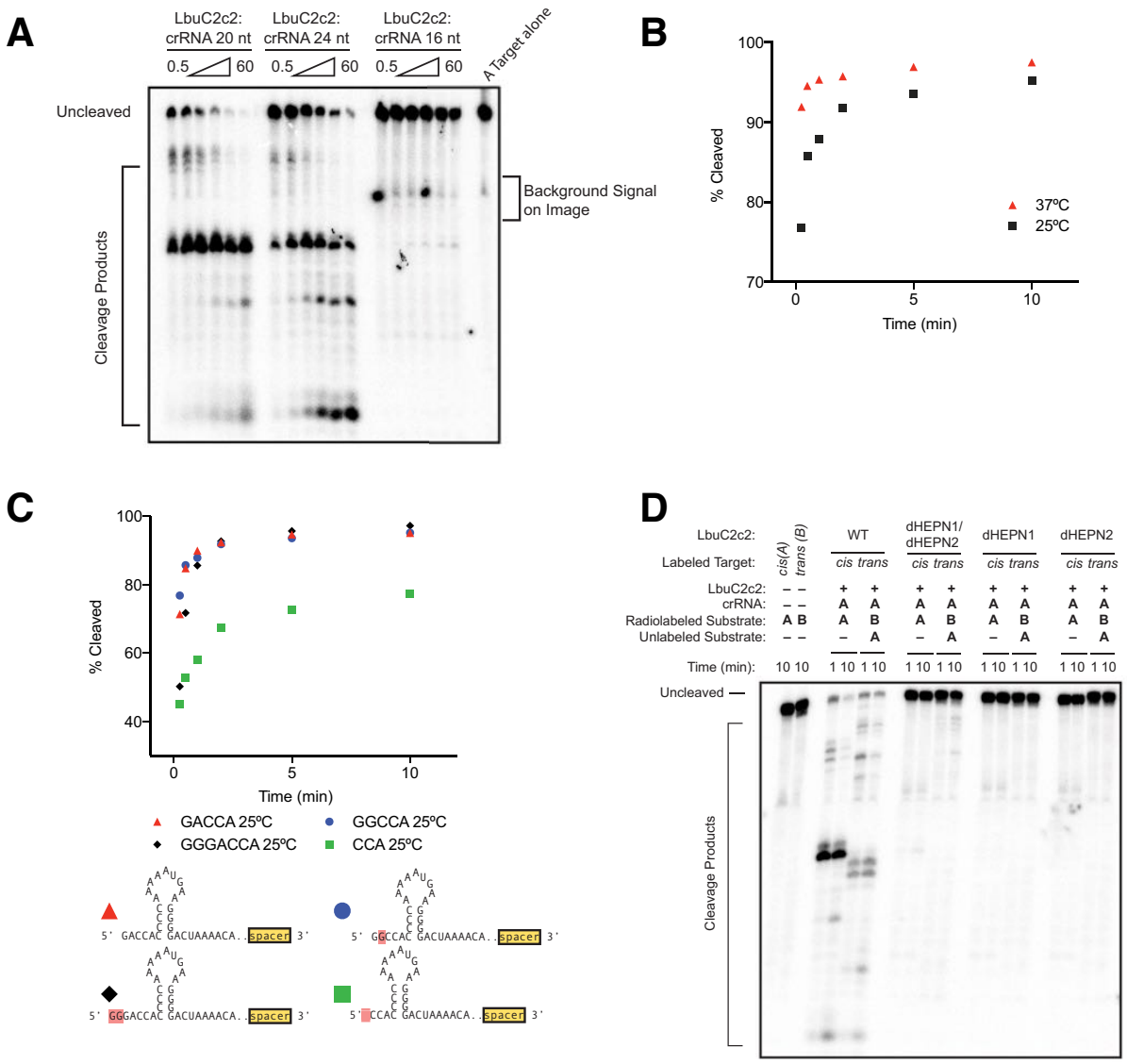


Figure A2.8. Dependence of RNA targeting on crRNA variants, temperature and point mutations. (A) LbuC2c2 ssRNA target cleavage assay carried out, as per Methods with crRNAs possessing 16-nt, 20-nt or 24-nt spacers. **(B)** LbuC2c2 ssRNA target cleavage time-course carried out at either 25°C and 37°C as per methods. **(C)** LbuC2c2 ssRNA target cleavage timecourse carried out as per Methods with crRNAs possessing different 5'-flanking nucleotide mutations. Mutations are highlighted in red. 1-2 nucleotide 5' extensions negligibly impacted cleavage efficiencies. In contrast, shortening the flanking region to 3 nts slowed cleavage rates. **(D)** Impact of point mutations on ribonuclease activity of C2c2 in conserved residue mutants within HEPN motifs for ssRNA targeting.

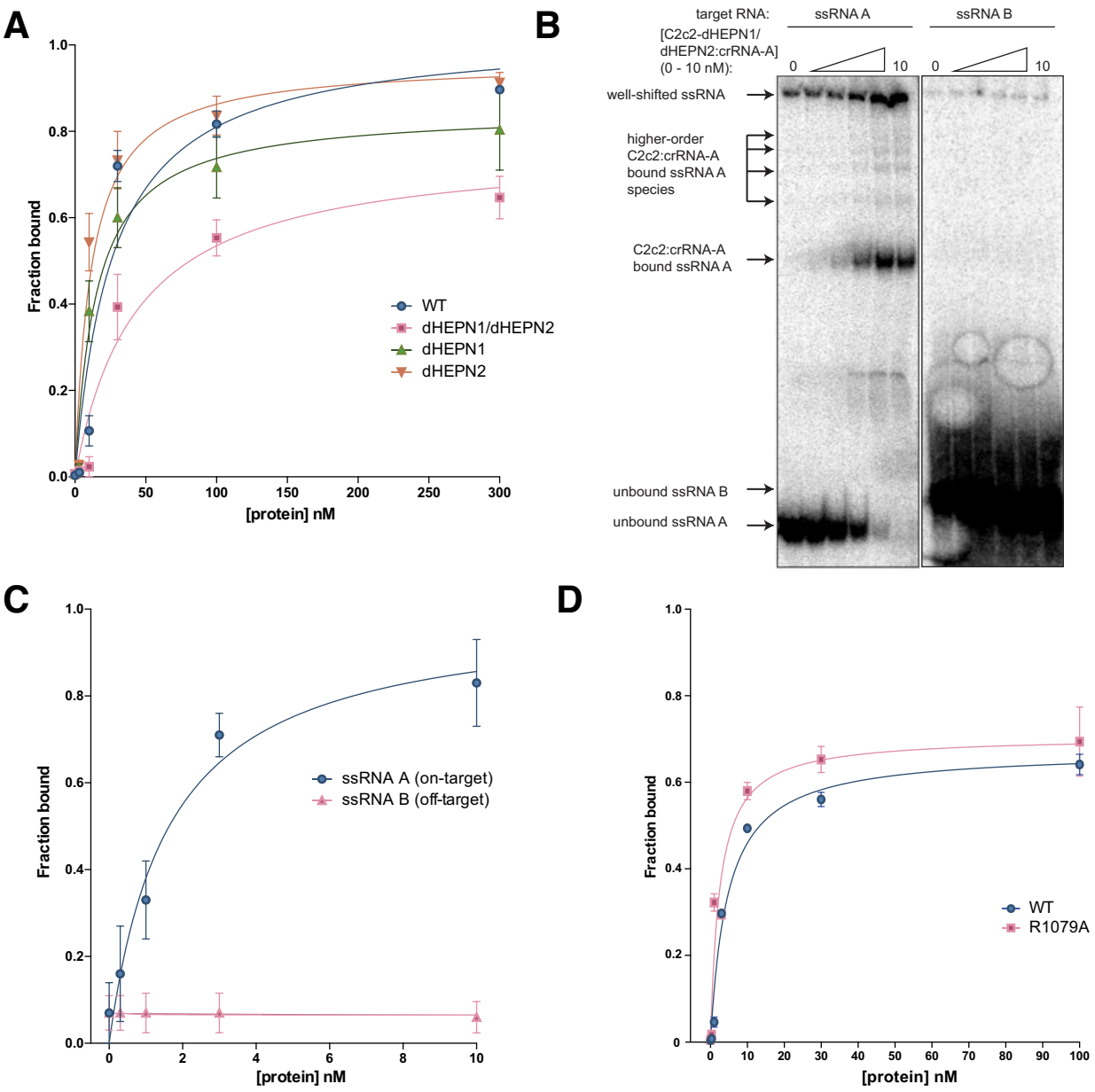
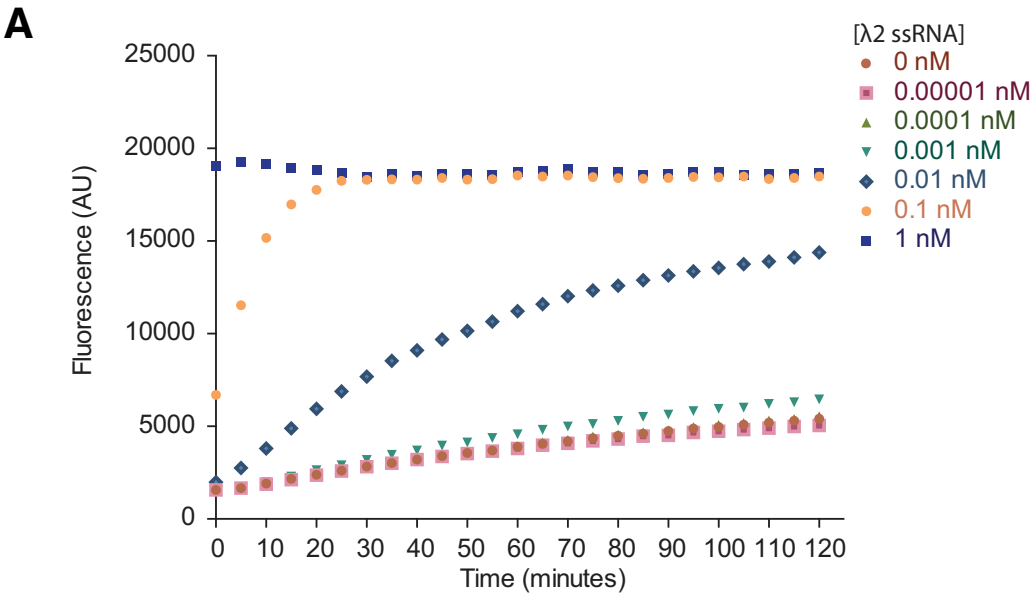

A
Fraction bound
[protein] nM
WT
dHEPN1/dHEPN2
dHEPN1
dHEPN2
B
target RNA:
ssRNA A
ssRNA B
[C2c2-dHEPN1/
dHEPN2:crRNA-A]
(0 - 10 nM):
0
10
0
10
well-shifted ssRNA
higher-order
C2c2:crRNA-A
bound ssRNA A
species
C2c2:crRNA-A
bound ssRNA A
unbound ssRNA B
unbound ssRNA A
C
Fraction bound
[protein] nM
ssRNA A (on-target)
ssRNA B (off-target)
D
Fraction bound
[protein] nM
WT
R1079A

Figure A2.9. Binding data for LbuC2c2 to mature crRNA and target ssRNA. (A) Filter binding assays were conducted as described in the Methods to determine the binding affinity of mature crRNA-A_GG to LbuC2c2-WT, LbuC2c2-dHEPN1, LbuC2c2-dHEPN2, or LbuC2c2-dHEPN1/dHEPN2. The quantified data were fit to standard binding isotherms. Error bars represent the standard deviation from three independent experiments. Measured dissociation constants from three independent experiments (mean ± sd) were 27.1 ± 7.5 nM (LbuC2c2-WT), 15.2 ± 3.2 nM (LbuC2c2-dHEPN1), 11.5 ± 2.5 nM (LbuC2c2-dHEPN2), and 43.3 ± 11.5 nM (LbuC2c2- dHEPN1/dHEPN2). **(B)** Representative electrophoretic mobility shift assay for binding reactions between LbuC2c2-dHEPN1/dHEPN2: crRNA-A_GG and either 'on-target' A ssRNA or 'off-target' B ssRNA, as indicated. Three independent experiments were conducted as described in the Methods. The gel was cropped for clarity. **(C)** Quantified binding data from (**B**) were fitted to standard binding isoforms. Error bars represent the standard deviation from three independent experiments. Measured dissociation constants from three independent experiments (mean ± sd) were 1.62 ± 0.43 nM for ssRNA A and N.D (>>10 nM) for ssRNA B. **(D)** Filter binding assays were conducted as described in the Methods to determine the binding affinity of mature crRNA-A_GA to LbuC2c2-WT and LbuC2c2-R1079A. The quantified data were fit to standard binding isotherms. Error bars represent the standard deviation from three independent experiments. Measured dissociation constants from three independent experiments (mean ± sd) were 4.65 ± 0.6 nM (LbuC2c2-WT) and 2.52 ± 0.5 nM (LbuC2c2-R1079A). It is of note that these binding affinities differ from panel a. This difference is accounted for in a slight difference in the 5` sequence of the guide with panel a guides beginning with a 5`-GGCCA… and panel d 5`-GACCA. While the native sequence guide (5`-GACCA) binds tighter to LbuC2c2, no difference is seen in the RNA targeting efficiencies of these guide variants.


A
25000
20000
15000
10000
5000
0
Fluorescence (AU)
0 10 20 30 40 50 60 70 80 90 100 110 120
Time (minutes)
[λ2 ssRNA]
0 nM
0.00001 nM
0.0001 nM
0.001 nM
0.01 nM
0.1 nM
1 nM

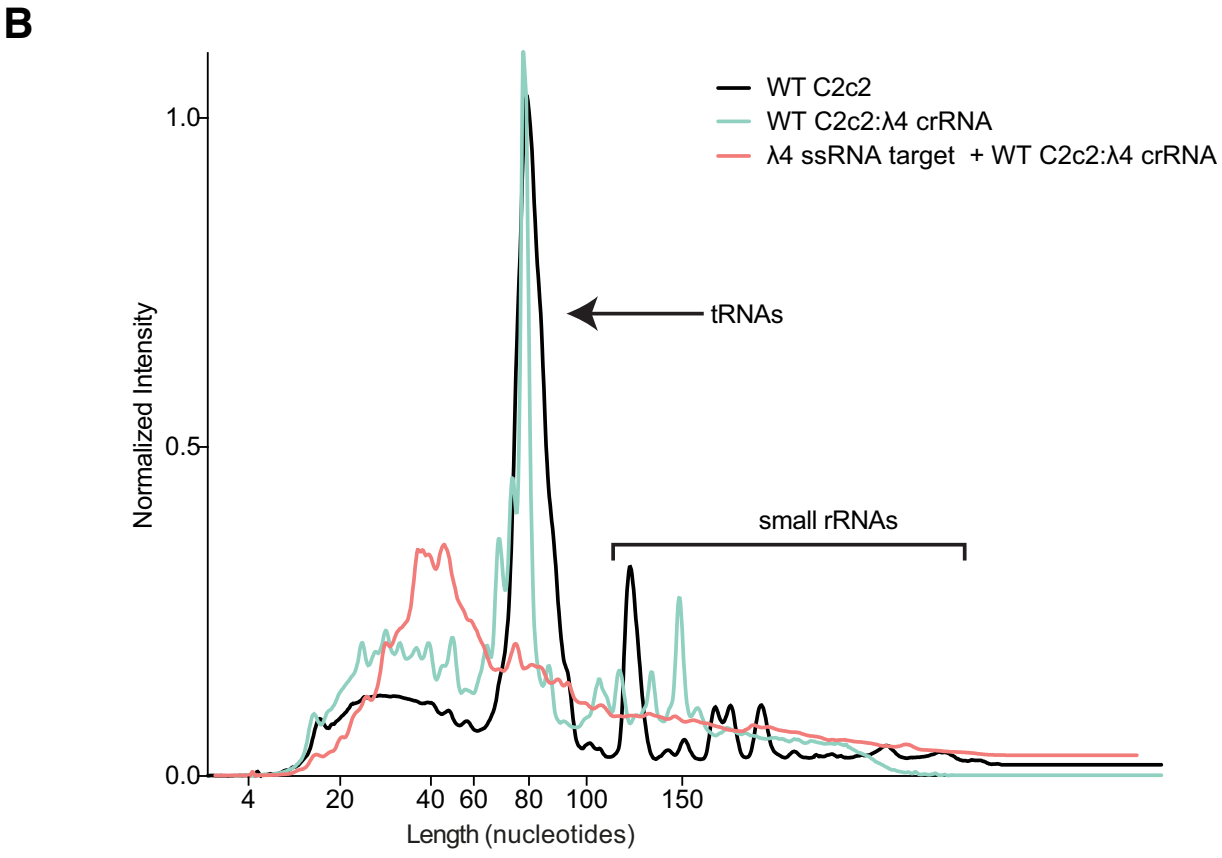

B
1.0
0.5
0.0
Normalized Intensity
4 20 40 60 80 100 150
Length (nucleotides)
WT C2c2
WT C2c2:λ4 crRNA
λ4 ssRNA target + WT C2c2:λ4 crRNA
tRNAs
small rRNAs

Figure A2.10. RNase detection assay λ2-ssRNA time-course and background RNA cleavage. (A) LbuC2c2:crRNA-λ2 was incubated with RNAase-Alert substrate (Thermo-Fisher)) and 100 ng HeLa total RNA in the presence of increasing amounts of λ2 ssRNA (0-1 nM) for 120 min at 37°C. Fluorescence measurements were taken every 5 min. The 1 nM λ2 ssRNA reaction reached saturation before the first time point could be measured. Error bars represent the standard deviation from three independent experiments. **(B)** LbuC2c2:crRNA-λ4 or apo LbuC2c2 was incubated in HeLa total RNA for 2 hours in the presence or absence of on-target activating λ4 ssRNA. Degradation of background small RNA was resolved on a small RNA chip in a Bioanalyzer 2100 as per Methods. Small differences are seen in the fragment profile of between apo LbuC2c2 and LbuC2c2:crRNA-λ4. In contrast, upon addition of the on-target ssRNA to the reaction, a drastic broadening and shifting of the tRNA peak reveals extensive degradation of other structured and nonstructured RNA's present in the reaction upon activation of LbuC2c2 *trans* activity.

Appendix III: A deep learning framework for genomic sequence classification

The recent development of natural language processing (NLP) has greatly enhanced our ability to classify and extract information from text in an automated fashion. This Appendix explores the application of deep learning and NLP concepts to the classification of genomic sequences. We focus specifically on the problem of pervasive translational readthrough in *Drosophila melanogangster*, and present a convolutional neural network (CNN) model that accurately distinguishes readthrough versus non-readthrough genes based solely on genomic sequence. Importantly, our model conservatively predicts dozens of new readthrough candidates across the entire *D. melanogangster* genome. We envision that the CNN pipeline presented herein could be more generally applied to a range of applications requiring binary classification of DNA sequences.

This work was done collaboratively with Hervé Marie-Nelly and Johannes Freitag as part of a larger effort to characterize the molecular mechanisms underpinning stop codon translational readthrough in eukaryotic cells.

Introduction

Translation– the production of proteins from mRNA– is tightly regulated by initiation factors, tRNA levels, chemical modifications, and regulatory sequence motifs (Jackson et al., 2010; Sonenberg and Dever, 2003; Sonenberg and Hinnebusch, 2009). Despite these regulatory elements, a number of modes of unconventional translation have been discovered in recent years. Internal ribosomal entry sites (IRES) allow for translation initiation in the absence of a Kozak sequence (Filbin and Kieft, 2009); deliberate frameshifting allows for multiple polypeptide products to be synthesized from a single mRNA (Caliskan et al., 2015); upstream open reading frames (uORFs) facilitate protein synthesis from the 5'-untranslated region (5'-UTR) of the mRNA; and stop codons (UAA, UAG, or UGA) can in some cases be read through by the ribosome to generate extended polypeptide products (Firth and Brierley, 2012).

Originally discovered in the context of viral gene expression, translational readthrough has been observed across a plethora of higher order eukaryotes (Dunn et al., 2013; Firth and Brierley, 2012; Freitag et al., 2013; Jungreis et al., 2011; Loughran et al., 2014). Hundreds of genes in *Drosophila melanogangster* engage in significant levels of stop codon readthrough, and in fungi readthrough has been implicated in cryptic peroxisomal targeting of functional dehydrogenases. In the context of *Homo sapiens*, a few dozen readthrough candidates have been identified, although there is little overlap across different studies and experimental methods. To date, the biological significance and pervasiveness of stop codon readthrough in humans remains largely unknown.

Recently, natural language processing (NLP) and convolutional neural networks (CNNs) have emerged as powerful tools for extracting quantitative information from text data (Kim, 2014; Mikolov et al., 2013). In contrast to other machine learning algorithms, CNNs allow for solutions to supervised classification problems in the absence of prior knowledge of feature importance. These networks have been successfully applied to sentiment analysis of free text, sentence completion, and grammar checking applications. In the context of translational readthrough, we hypothesized that a CNN could be used to identify sequence motifs underpinning stop codon readthrough in eukaryotes. Such a tool could be applied to predict novel readthrough candidates across a range of species based on genomic sequence, thereby expanding our mechanistic understanding of translational biology.

Results and Discussion

We developed a convolutional neural network (CNN) pipeline to differentiate between readthrough versus non-readthrough DNA sequences from the *D. melanogangster* genome (Fig. A3.1). To train the network, we used the PhyloCSF dataset with a 70-30 train-test split. This dataset consists of 256 candidate readthrough genes, several of which have been experimentally verified via GFP reporter assays.

As a first pass, the network was trained on the [-20,+40] sequence region of the genes relative to the position of the canonical stop codon. The 60-nucleotide sequences were divided into N-grams of variable length (1-6), generating sentences of length 10-60 "words" for each sequence. The vector embeddings of these sentences formed the raw matrix input for the network. Network performance showed a strong dependence on N-

gram size, with optimal performance observed for 3-nucleotide words (89% accuracy overall, Fig. A3.2).

To test whether our model was driven by the 5'-flanking coding sequence or the 3'-UTR sequence, we systematically varied the amount of sequence information incorporated in both the coding and UTR regions (Figs. A3.3-A3.4). We observed that model performance correlated positively with the amount of 3'-UTR information included in the input data, while 5'-coding information had little effect on overall accuracy. Taken together, these data suggest that translational readthrough in *D. melanogangster* is driven more by signatures within the 3'-UTR than by the terminal coding region of the canonical gene product.

Given the driving nature of the 3'-UTR in translational readthrough, we next asked whether the 3'-UTRs of readthrough genes might closely resemble conventional coding sequences. To test this, we trained a neural network to discriminate between coding regions and random genomic regions within the *D. melanogangster* genome and applied that network to the classification of the [-20,+80] regions specified above. While our network was able to accurately discriminate between coding versus random genomic sequences with >98% accuracy, it performed close to random in classifying readthrough genes as coding versus random (Fig. A3.5). This suggests that the nature of readthrough sequences in *D. melanogangster* is distinct from both coding and random genomic sequences.

We next sought to determine if classification could be made more accurate by filtering based on softmax regression score. Because class assignment requires discretization of a continuous-valued distribution function, entities closer to the decision boundary tend to be misclassified more frequently. In agreement with this, we found that readthrough misclassifications were disproportionately enriched near the decision boundary (0.0) of the softmax distribution (Fig. A3.6). Furthermore, CNN models performed better with increased cutoff stringency, achieving 100% accuracy at the 25^{th} (75% exclusion) percentile (Fig A3.7). Even after eliminating 75% of the data, our model predicts 63 new pervasive readthrough candidates in *D. melanogangster*.

Conclusion

The model presented here provides a general framework for classifying and identifying new translational readthrough candidates in *D. melanogangster*. We identified the 3'-UTR as a driving factor in predicting translational readthrough, while the 5'-coding region was less informative. Importantly, the CNN pipeline presented here can be broadly applied to a plethora of applications requiring discrete classification based on DNA sequence.

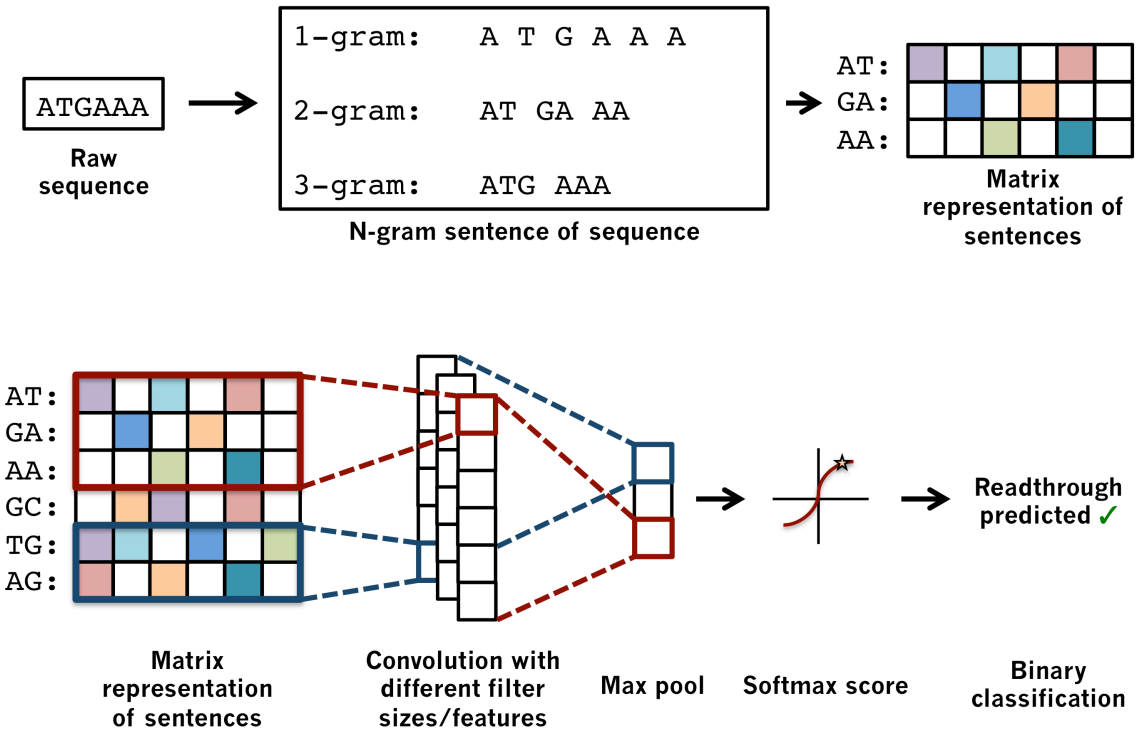


Figure A3.1. Architecture of the convolutional neural network used for genomic classification. (*Top*) Raw sequences were first converted to sentences of variable N-gram length, which were used to generate matrix representations of genomic sequences using Word2Vec N-gram embeddings. (*Bottom*) Convolutional neural networks were trained from a 70-30 split of the genomic data. Once trained, networks could classify genomic sequences *de novo* based on matrix convolution, max pooling, softmax scoring, and binary thresholding.

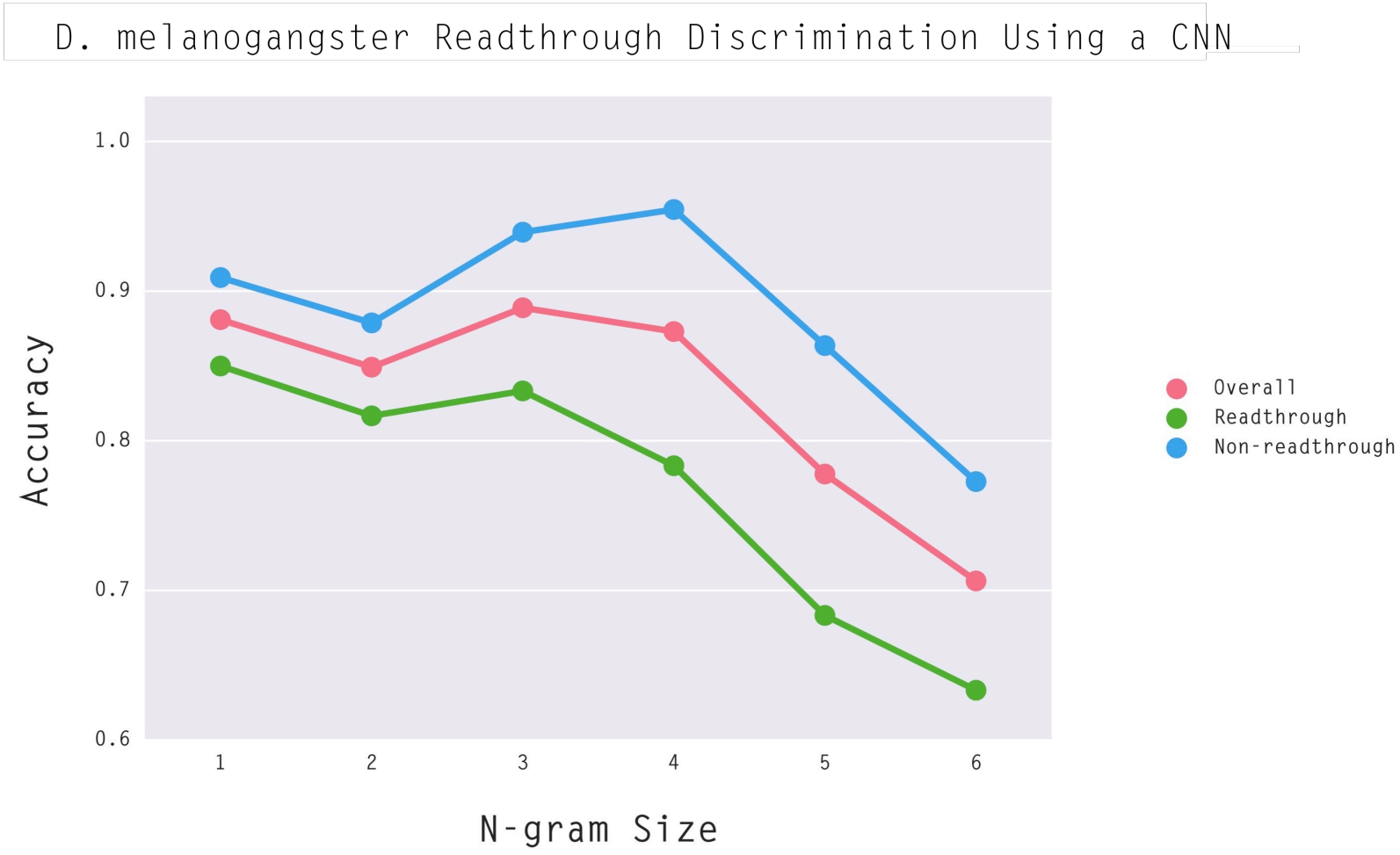

D. melanogangster Readthrough Discrimination Using a CNN
Accuracy
1.0
0.9
0.8
0.7
0.6
1
2
3
4
5
6
N-gram Size
Overall
Readthrough
Non-readthrough

Figure A3.2. Network performance dependence on N-gram size. Neural networks were trained on a 70-30 split of the readthrough/non-readthrough dataset, with equal representation of each class. The model performed with 89% overall accuracy on the test data.

D. melanogangster Readthrough Discrimination v. UTR Length

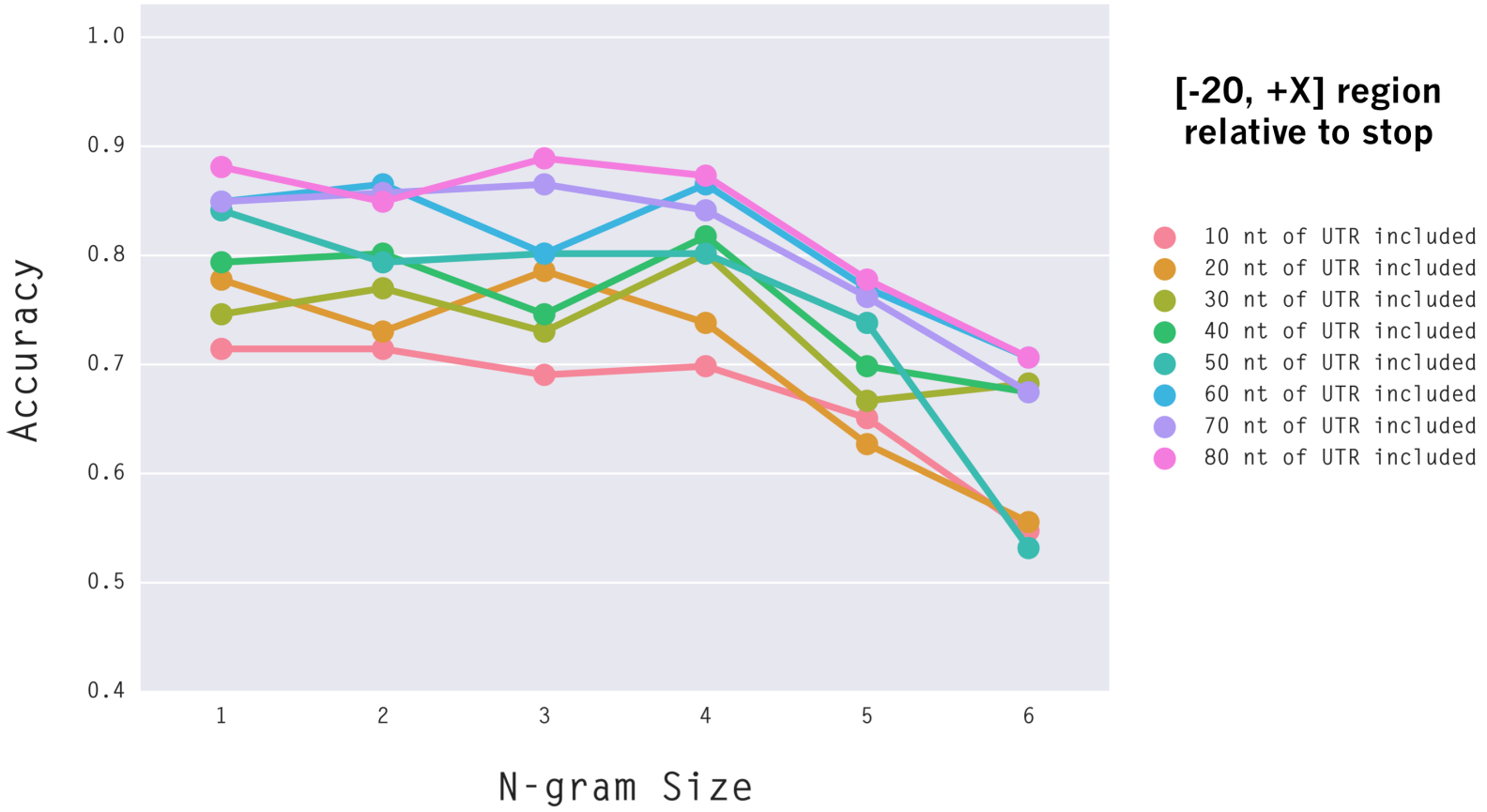


Figure A3.3. Network performance dependence on 3'-UTR. Neural networks were trained on genomic sequences with variable incorporation of 3'-UTR information while keeping the 5'-flanking coding region constant. Overall accuracy across both classes is plotted as a function of N-gram size.

D. melanogangster Readthrough Discrimination Using a CNN

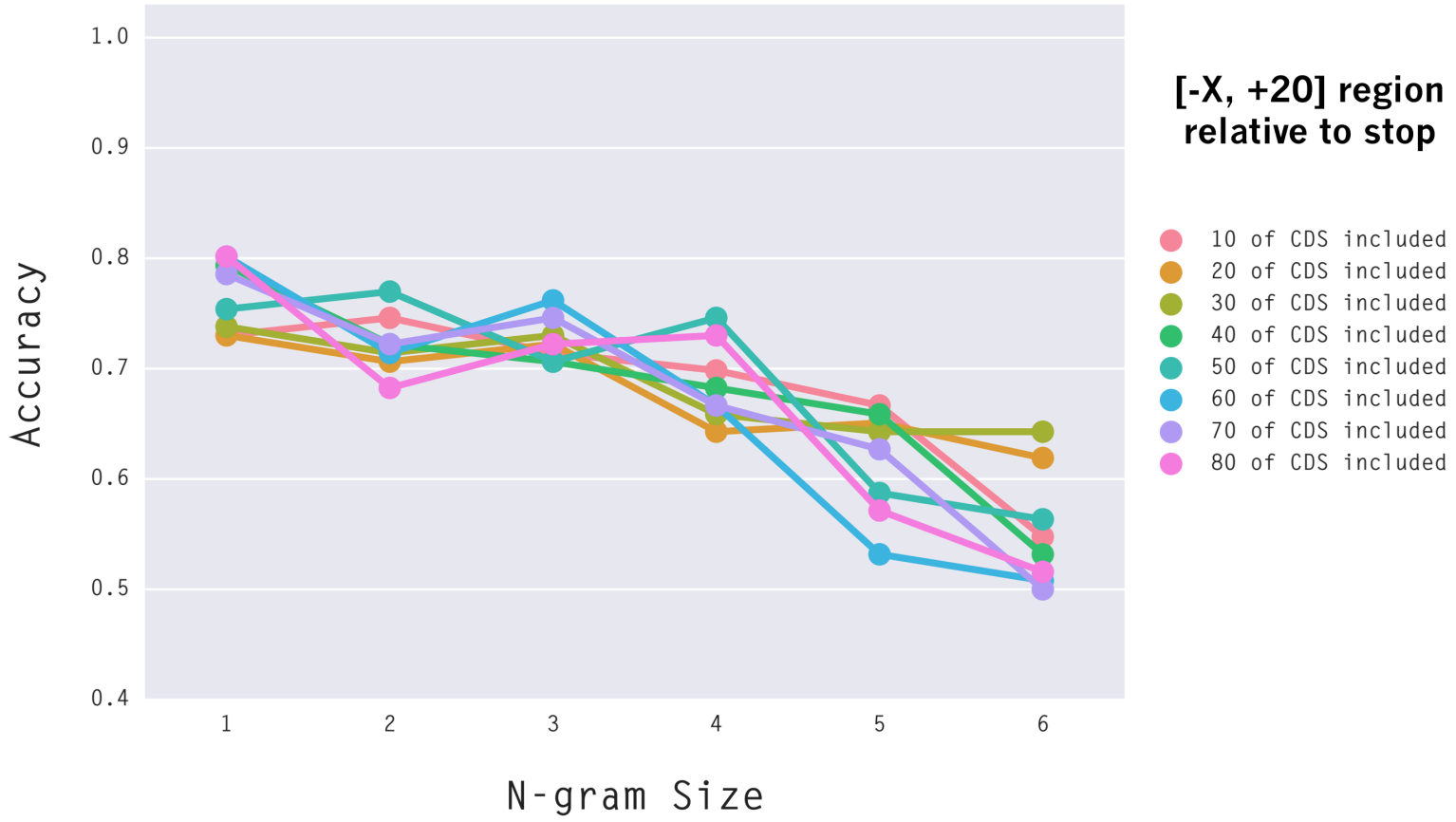


Figure A3.4. Network performance dependence on 5'-flanking coding region. Neural networks were trained on genomic sequences with variable incorporation of 5'-flanking coding information while keeping the 3'-UTR region constant. Overall accuracy across both classes is plotted as a function of N-gram size.

D. melanogangster Readthrough Discrimination Using a CNN

(+) examples:
Coding sequences

(–) examples:
Random genomic sequences

Figure A3.5. Coding signature of readthrough candidates. Two-class CNNs were trained on the *D. melanogangster* genome, with coding and random genomic regions constituting the positive and negative class examples, respectively. The network was then applied to the classification of [-20,+40] sequence region of known readthrough genes. For reference, a purely random classifier is expected to perform with ~50% accuracy.

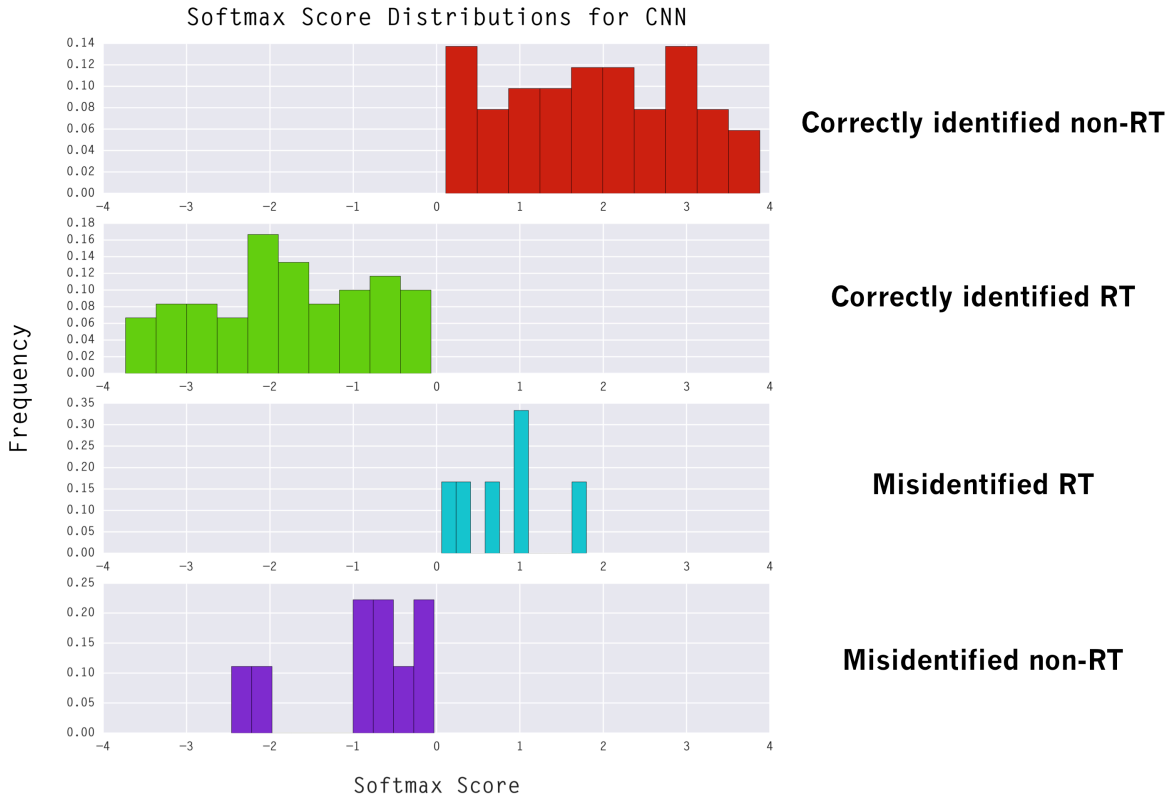


Figure A3.6. Softmax distribution of misclassified sequences. Softmax score distributions are plotted for four categories: correctly identified non-readthrough genes (red), correctly identified readthrough genes (green), readthrough genes misclassified as non-readthrough (teal), and non-readthrough genes misclassified as readthrough (purple). Data are normalized to a total probability density of 1.

D. melanogangster Readthrough Discrimination v. Percentile Cutoff

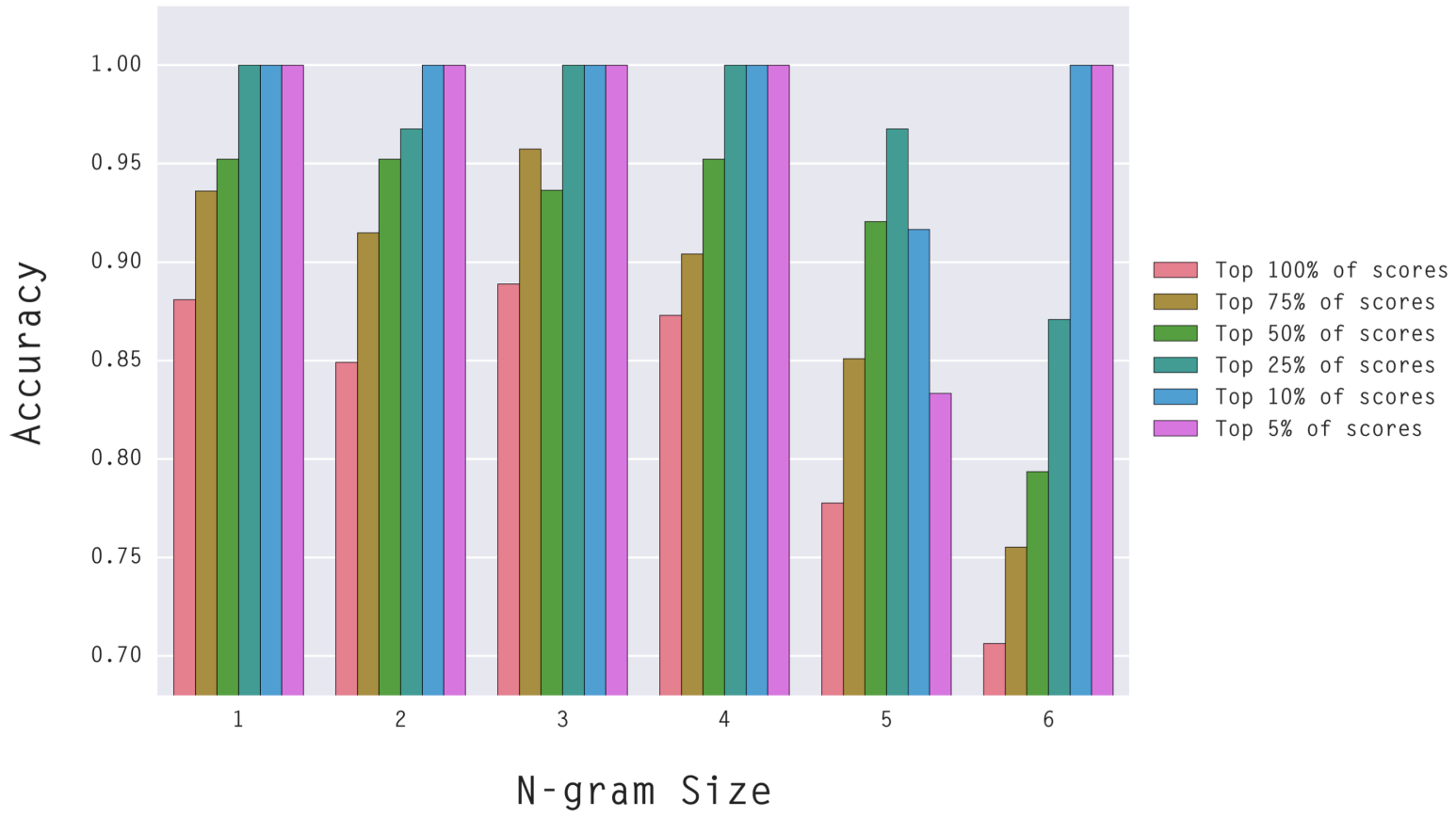


Figure A3.7. Model accuracy as a function of softmax filtering. Networks were trained on a 70-30 split of the data with variable N-gram sizes. Classification accuracy is plotted as a function of softmax tail percentile cutoff.

Materials and Methods

Genomic Sequence Extraction and Embedding

Drosophila reference transcriptomes were downloaded from FlyBase (flybase.org), and the sequence region flanking the canonical stop codon was isolated for both readthrough and non-readthrough genes based on previously published coordinate data (Jungreis et al., 2011). Sequences were tokenized into variable length N-grams using a homebuilt Python script and then embedded into matrix representations using the Word2Vec function native to TensorFlow(Mikolov et al., 2013).

Network architecture and training

A convolutional network was constructed as previously described with the following modifications: a single static channel was used for the input layer, and a dropout layer was incorporated to reduce overfitting (Kim, 2014; Mikolov et al., 2013). The network was generated as a Python class object using TensorFlow. Sample code is shown below:

```
import pandas as pd
import tensorflow as tf
import numpy as np

class TextCNN(object):
    """
    A CNN for text classification.
    Uses an embedding layer, followed by a convolutional, max-pooling and softmax
layer.
    """
    def __init__(
      self, sequence_length, num_classes, vocab_size,
      embedding_size, filter_sizes, num_filters, l2_reg_lambda=0.0):

        # Placeholders for input, output and dropout
        self.input_x    =    tf.placeholder(tf.int32,    [None,    sequence_length],
name="input_x")
        self.input_y = tf.placeholder(tf.float32, [None, num_classes], name="input_y")
        self.dropout_keep_prob = tf.placeholder(tf.float32, name="dropout_keep_prob")

        # Keeping track of l2 regularization loss (optional)
        l2_loss = tf.constant(0.0)

        # Embedding layer
        with tf.device('/cpu:0'), tf.name_scope("embedding"):
            W = tf.Variable(
                tf.random_uniform([vocab_size, embedding_size], -1.0, 1.0),
                name="W")
            self.embedded_chars = tf.nn.embedding_lookup(W, self.input_x)
            self.embedded_chars_expanded = tf.expand_dims(self.embedded_chars, -1)

        # Create a convolution + maxpool layer for each filter size
        pooled_outputs = []
        for i, filter_size in enumerate(filter_sizes):
            with tf.name_scope("conv-maxpool-%s" % filter_size):
                # Convolution Layer
                filter_shape = [filter_size, embedding_size, 1, num_filters]
                W    =    tf.Variable(tf.truncated_normal(filter_shape,    stddev=0.1),
name="W")
                b = tf.Variable(tf.constant(0.1, shape=[num_filters]), name="b")
```

```
                conv = tf.nn.conv2d(
                    self.embedded_chars_expanded,
                    W,
                    strides=[1, 1, 1, 1],
                    padding="VALID",
                    name="conv")
                # Apply nonlinearity
                h = tf.nn.relu(tf.nn.bias_add(conv, b), name="relu")
                # Maxpooling over the outputs
                pooled = tf.nn.max_pool(
                    h,
                    ksize=[1, sequence_length - filter_size + 1, 1, 1],
                    strides=[1, 1, 1, 1],
                    padding='VALID',
                    name="pool")
                pooled_outputs.append(pooled)

        # Combine all the pooled features
        num_filters_total = num_filters * len(filter_sizes)
        self.h_pool = tf.concat(3, pooled_outputs)
        self.h_pool_flat = tf.reshape(self.h_pool, [-1, num_filters_total])

        # Add dropout
        with tf.name_scope("dropout"):
            self.h_drop = tf.nn.dropout(self.h_pool_flat, self.dropout_keep_prob)

        # Final (unnormalized) scores and predictions
        with tf.name_scope("output"):
            W = tf.get_variable(
                "W",
                shape=[num_filters_total, num_classes],
                initializer=tf.contrib.layers.xavier_initializer())
            b = tf.Variable(tf.constant(0.1, shape=[num_classes]), name="b")
            l2_loss += tf.nn.l2_loss(W)
            l2_loss += tf.nn.l2_loss(b)
            self.scores = tf.nn.xw_plus_b(self.h_drop, W, b, name="scores")
            self.predictions = tf.argmax(self.scores, 1, name="predictions")

        # CalculateMean cross-entropy loss
        with tf.name_scope("loss"):
            losses        =        tf.nn.softmax_cross_entropy_with_logits(self.scores,
self.input_y)
            self.loss = tf.reduce_mean(losses) + l2_reg_lambda * l2_loss

        # Accuracy
        with tf.name_scope("accuracy"):
            correct_predictions  =  tf.equal(self.predictions,  tf.argmax(self.input_y,
1))
            self.accuracy   =   tf.reduce_mean(tf.cast(correct_predictions,   "float"),
name="accuracy")
```

References

Abrahamsson, S., Chen, J., Hajj, B., Stallinga, S., Katsov, A.Y., Wisniewski, J., Mizuguchi, G., Soule, P., Mueller, F., Darzacq, C.D., et al. (2012). Fast multicolor 3D imaging using aberration-corrected multifocus microscopy. Nat Meth *10*, 60–63.

Abudayyeh, O.O., Gootenberg, J.S., Konermann, S., Joung, J., Slaymaker, I.M., Cox, D.B.T., Shmakov, S., Makarova, K.S., Semenova, E., Minakhin, L., et al. (2016). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science *353*, aaf5573–11.

Anantharaman, V., Makarova, K.S., Burroughs, A.M., Koonin, E.V., and Aravind, L. (2013). Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. Biology Direct *8*, 1–1.

Anders, C., Niewoehner, O., Duerst, A., and Jinek, M. (2015). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature *513*, 569–573.

Arap, W., Kolonin, M., Trepel, M., Lahdenranta, J., Cardo-Vila, M., and Pasqualini, R. (2002). Steps toward mapping the human vasculature by phage display. Nat Med *8*, 121–128.

Axelrod, D. (2001). Total Internal Reflection Fluorescence Microscopy in Cell Biology. Meth. Enzymol. *361*, 1–33.

Baltimore, D. et al. (2015). A prudent path forward for genomic engineering and germline gene modification. Science *348*, 36–38.

Bartholomew, B. (2014). Regulating the Chromatin Landscape: Structural and Mechanistic Perspectives. Annu. Rev. Biochem. *83*, 671–696.

Beheiry, El, M., Dahan, M., and Masson, J.-B. (2015). InferenceMAP: mapping of single-molecule dynamics with Bayesian inference. Nat Rev Microbiol *12*, 594–595.

Betzig, E. (1996). Proposed method for molecular optical imaging. Optics Letters *20*, 237–239.

Betzig, E., Patterson, G.H., Sougrat, R., Lindwasser, O.W., Olenych, S., Bonifacino, J.S., Davidson, M.W., Lippincott-Schwartz, J., and Hess, H.F. (2006). Imaging Intracellular Fluorescent Proteins at Nanometer Resolution. Science *313*, 1642–1645.

Bibikova, M., Carroll, D., Segal, D.J., Trautman, J.K., Smith, J., Kim, Y.G., and Chandrasegaran, S. (2001). Stimulation of Homologous Recombination through Targeted Cleavage by Chimeric Nucleases. Molecular and Cellular Biology *21*, 289–297.

Bibikova, M., Golic, M., Golic, K.G., and Carroll, D. (2002). Targeted Chromosomal

Cleavage and Mutagenesis in Drosophila Using Zinc-Finger Nucleases. Genetics *161*, 1169–1175.

Bishop, A.C., Buzko, O., and Shokat, K.M. (2001). Magic bullets for protein kinases. Trends in Cell Biology *11*, 167–173.

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., and Bonas, U. (2009). Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. Science *326*, 1509–1512.

Bosley, K.S., Botchan, M., Bredenoord, A.L., Carroll, D., Charo, R.A., Charpentier, E., Cohen, R., Corn, J., Doudna, J., Feng, G., et al. (2015). CRISPR germline engineering—the community speaks. Nat Rev Microbiol *33*, 478–486.

Braun, A.P., and Schulman, H. (1995). The Multifunctional Calcium/Calmodulin-Dependent Protein Kinase: From Form to Function. Annu. Rev. Physiol. *57*, 417–445.

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E.V., and van der Oost, J. (2008). Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. Science *321*, 960–964.

Browning, D.F., and Busby, S.J.W. (2004). The regulation of bacterial transcription initiation. Nat Rev Microbiol *2*, 57–65.

Burma, S., Chen, B.P.C., and Chen, D.J. (2006). Role of non-homologous end joining (NHEJ) in maintaining genomic integrity. DNA Repair *5*, 1042–1048.

Buxbaum, A.R., Haimovich, G., and Singer, R.H. (2014). In the right place at the right time: visualizing and understanding mRNA localization. Nat Rev Microbiol 1–15.

Caliskan, N., Peske, F., and Rodnina, M.V. (2015). Changed in translation: mRNA recoding by −1 programmed ribosomal frameshifting. Trends in Biochemical Sciences *40*, 265–274.

Campbell, R.E., Tour, O., Palmer, A.E., Steinbach, P.A., Baird, G.S., Zacharias, D.A., and Tsien, R.Y. (2002). A monomeric red fluorescent protein. Proc Natl Acad Sci USA *99*, 7877–7882.

Carroll, D. (2016). A Perspective on the State of Genome Editing. Molecular Therapy *24*, 412–413.

Carte, J., Wang, R., Li, H., Terns, R.M., and Terns, M.P. (2008a). Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes Dev. *22*, 3489–3496.

Carte, J., Wang, R., Li, H., Terns, R.M., and Terns, M.P. (2008b). Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes Dev. *22*, 3489–3496.

Cech, T.R., and Steitz, J.A. (2014). The Noncoding RNA Revolution— Trashing Old Rules to Forge New Ones. Cell *157*, 77–94.

Celera Genomics (2001). The Sequence of the Human Genome. Science *291*, 1304–1354.

Charpentier, E., Richter, H., van der Oost, J., and White, M.F. (2015). Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. FEMS Microbiology Reviews *39*, 428–441.

Chen, B., Gilbert, L.A., Cimini, B.A., Schnitzbauer, J., Zhang, W., Li, G.-W., Park, J., Blackburn, E.H., Weissman, J.S., Qi, L.S., et al. (2013). Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. Cell *155*, 1479–1491.

Chen, B., Guan, J., and Huang, B. (2016a). Imaging Specific Genomic DNA in Living Cells. Annu. Rev. Biophys. *45*, 1–23.

Chen, B., Hu, J., Almeida, R., Liu, H., Balakrishnan, S., Covill-Cooke, C., Lim, W.A., and Huang, B. (2016b). Expanding the CRISPR imaging toolset with Staphylococcus aureusCas9 for simultaneous imaging of multiple genomic loci. Nucleic Acids Research *44*, e75–e75.

Chen, B., Hu, J., Almeida, R., Liu, H., Balakrishnan, S., Covill-Cooke, C., Lim, W.A., and Huang, B. (2016c). Expanding the CRISPR imaging toolset with Staphylococcus aureusCas9 for simultaneous imaging of multiple genomic loci. Nucleic Acids Research *44*, e75–e75.

Chen, J., Zhang, Z., Li, L., Chen, B.-C., Revyakin, A., Hajj, B., Legant, W., Dahan, M., Lionnet, T., Betzig, E., et al. (2014). Single-Molecule Dynamics of Enhanceosome Assembly in Embryonic Stem Cells. Cell *156*, 1274–1285.

Chi, P., Allis, C.D., and Wang, G.G. (2010). Covalent histone modifications — miswritten, misinterpreted and mis-erased in human cancers. 1–13.

Chockalingam, K., Chen, Z., Katzenellenbogen, J.A., and Zhao, H. (2005). Directed evolution of specific receptor–ligand pairs for use in the creation of gene switches. Proc Natl Acad Sci USA *102*, 5691–5696.

Choi, U.Y., Kang, J.-S., Hwang, Y.S., and Kim, Y.-J. (2015). Oligoadenylate synthase-like (OASL) proteins: dual functions and associations with diseases. *47*, e144–e146.

Christian, M., Cermak, T., Doyle, E.L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove, A.J., and Voytas, D.F. (2010). Targeting DNA Double-Strand Breaks with TAL Effector Nucleases. Genetics *186*, 757–761.

Cohen, B., Colas, P., and Brent, R. (1998). An artificial cell-cycle inhibitor isolated from a combinatorial library. Proc Natl Acad Sci USA *95*, 14272–14277.

Conaway, R.C., and Conaway, J.W. (1993). General Initiation Factors for RNA Polymerase II. Annu. Rev. Biochem. 161–190.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. Science *339*, 819–823.

Cordray, M.S., and Richards-Kortum, R.R. (2012). Emerging Nucleic Acid-Based Tests for Point-of-Care Detection of Malaria. American Journal of Tropical Medicine and Hygiene *87*, 223–230.

Cravatt, B.F., Wright, A.T., and Kozarich, J.W. (2008). Activity-Based Protein Profiling: From Enzyme Chemistry to Proteomic Chemistry. Annu. Rev. Biochem. *77*, 383–414.

Crick, F. (1970). Central Dogma. Nature *227*, 561–563.

Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J., and Charpentier, E. (2012). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature *471*, 602–607.

Deng, W., Shi, X., Tjian, R., Lionnet, T., and Singer, R.H. (2015). CASFISH: CRISPR/Cas9-mediated in situ labeling of genomic loci in fixed cells. Proc Natl Acad Sci USA *112*, 11870–11875.

DiCarlo, J.E., Norville, J.E., Mali, P., Rios, X., Aach, J., and Church, G.M. (2013). Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Research *41*, 4336–4343.

Dietzl, G., Chen, D., Schnorrer, F., Su, K.-C., Barinova, Y., Fellner, M., Gasser, B., Kinsey, K., Oppel, S., Scheiblauer, S., et al. (2007). A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. Nature *448*, 151–156.

Diezmann, von, A., Shechtman, Y., and Moerner, W.E. (2017). Three-Dimensional Localization of Single Molecules for Super-Resolution Imaging and Single-Particle Tracking. Chem. Rev. acs.chemrev.6b00629–acs.chemrev.6b00632.

Doudna, J.A., and Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. Science *346*, 1258096–1258096.

Dunn, J.G., Foo, C.K., Belletier, N.G., Gavis, E.R., and Weissman, J.S. (2013). Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogangster. eLife *2*, e01179–32.

Dyba, M., Jakobs, S., and Hell, S.W. (2003). Immunofluorescence stimulated emission depletion microscopy. Nature Biotechnology *21*, 1303–1304.

Dynan, W.S., and Tjian, R. (1983). The Promoter-Specific Transcription Factor Sp 1 Binds to Upstream Sequencesin the SV40 Early Promoter. Cell *35*, 79–87.

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

Eissenberg, J.C., and Elgin, S.C.R. (2015). The HP1 protein family: getting a grip on chromatin. Current Opinion in Genetics Development *10*, 204–210.

Elf, J., Li, G.-W., and Xie, S. (2007). Probing Transcription Factor Dynamics at the Single-Molecule Level in a Living Cell. Science *316*, 1191–1194.

Espinoza, C.A., Goodrich, J.A., and Kugel, J.F. (2007). Characterization of the structure, function, and mechanism of B2 RNA, an ncRNA repressor of RNA polymerase II transcription. Rna *13*, 583–596.

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

Evitt, N.H., Mascharak, S., and Altman, R.B. (2015). Human Germline CRISPR-Cas Modification: Toward a Regulatory Framework. The American Journal of Bioethics *15*, 25–29.

Falkenberg, K.J., and Johnstone, R.W. (2014). Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders. Nat Rev Microbiol *13*, 673–691.

Filbin, M.E., and Kieft, J.S. (2009). Toward a structural understanding of IRES RNA function. Current Opinion in Structural Biology *19*, 267–276.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. (1998). Potent and specificgenetic interference bydouble-stranded RNA inCaenorhabditis elegans. Nature *391*, 806–811.

Firth, A.E., and Brierley, I. (2012). Non-canonical translation in RNA viruses. Journal of General Virology *93*, 1385–1409.

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

Freitag, J., Ast, J., and Bölker, M. (2013). Cryptic peroxisomal targeting via alternative splicing and stop codon read-through in fungi. Nature *485*, 522–525.

Frock, R.L., Hu, J., Meyers, R.M., Ho, Y.-J., Kii, E., and Alt, F.W. (2014). Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. Nature Biotechnology *33*, 179–186.

Frommer, W.B., Davidson, M.W., and Campbell, R.E. (2009). Genetically encoded biosensors based on engineered fluorescent proteins. Chem. Soc. Rev. *38*, 2833–10.

Gabriel, R., Lombardo, A., Arens, A., Miller, J.C., Genovese, P., Kaeppel, C., Nowrouzi, A., Bartholomae, C.C., Wang, J., Friedman, G., et al. (2011). An unbiased genome-wide analysis of zinc-finger nuclease specificity. Nature Biotechnology *29*, 816–823.

Gantz, V.M., Jasinskiene, N., Tatarenkova, O., Fazekas, A., Macias, V.M., Bier, E., and James, A.A. (2015). Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito Anopheles stephensi. Proc Natl Acad Sci USA *112*, E6736–E6743.

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

Garside, E.L., Schellenberg, M.J., Gesner, E.M., Bonanno, J.B., Sauder, J.M., Burley, S.K., Almo, S.C., Mehta, G., and MacMillan, A.M. (2012). Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. Rna *18*, 2020–2028.

Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunityin bacteria. Proc Natl Acad Sci USA *109*, E2579–E2586.

Gebhardt, J.C.M., Suter, D.M., Roy, R., Zhao, Z.W., Chapman, A.R., Basu, S., Maniatis, T., and Xie, X.S. (2013). Single-molecule imaging of transcription factor binding to DNA in live mammalian cells. Nat Meth *10*, 421–426.

Gilbert, L.A., Horlbeck, M.A., Adamson, B., Villalta, J.E., Chen, Y., Whitehead, E.H., Guimaraes, C., Panning, B., Ploegh, H.L., Bassik, M.C., et al. (2014). Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. Cell *159*, 647–661.

Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., et al. (2013). CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. Cell *154*, 442–451.

Gratz, S.J., Cummings, A.M., Nguyen, J.N., Hamm, D.C., Donohue, L.K., Harrison, M.M., Wildonger, J., and OConnor-Giles, K.M. (2013). Genome Engineering of Drosophila with the CRISPR RNA-Guided Cas9 Nuclease. Genetics *194*, 1029–1035.

Grawunder, U., Zimmer, D., Fugmann, S., Schwarz, K., and Lieber, M.R. (1998). DNA Ligase IV Is Essentialfor V(D)J Recombination and DNA Double-Strand Break Repair in Human Precursor Lymphocytes. Molecular Cell *2*, 477–484.

Grimm, J.B., English, B.P., Chen, J., Slaughter, J.P., Zhang, Z., Revyakin, A., Patel, R., Macklin, J.J., Normanno, D., Singer, R.H., et al. (2015). A general method to improve fluorophores for live-cell and single-molecule microscopy. Nat Meth *12*, 244–250.

Grimm, J.B., English, B.P., Choi, H., Muthusamy, A.K., Mehl, B.P., Dong, P., Brown, T.A., Lippincott-Schwartz, J., Liu, Z., Lionnet, T., et al. (2016). Bright photoactivatable

fluorophores for single-molecule imaging. Nat Meth *13*, 985–988.

Guan, J., Liu, H., Shi, X., Feng, S., and Huang, B. (2017). Tracking Multiple Genomic Elements Using Correlative CRISPR Imaging and Sequential DNA FISH. Biophys. J. *112*, 1077–1084.

Gurskaya, N.G., Verkhusha, V.V., Shcheglov, A.S., Staroverov, D.B., Chepurnykh, T.V., Fradkov, A.F., Lukyanov, S., and Lukyanov, K.A. (2006). Engineering of a monomeric green-to-red photoactivatable fluorescent protein induced by blue light. Nature Biotechnology *24*, 461–465.

Gutschner, T., Haemmerle, M., Genovese, G., Draetta, G.F., and Chin, L. (2016). Post-translational Regulation of Cas9 during G1 Enhances Homology-Directed Repair. CellReports *14*, 1555–1566.

Haber, J.E. (2000). Partners and pathways: repairing a double-strand break. Trends in Genetics *16*, 259–264.

Hai, T., Teng, F., Guo, R., Li, W., and Zhou, Q. (2014). One-step generation of knockout pigs by zygote injection of CRISPR/Cas system. Cell Res *24*, 372–375.

Hale, C.R., Zhao, P., Olson, S., Duff, M.O., Graveley, B.R., Wells, L., Terns, R.M., and Terns, M.P. (2009). RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. Cell *139*, 945–956.

Hall, M.N., Gabay, J., Debarbouille, M., and Schwartz, M. (1982). A role for mRNA secondary structure in the control of translation initiation. Nature *295*, 616–618.

Hammond, A., Galizi, R., Kyrou, K., Simoni, A., Siniscalchi, C., Katsanos, D., Gribble, M., Baker, D., Marois, E., Russell, S., et al. (2015). A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector Anopheles gambiae. Nature Biotechnology *34*, 78–83.

Haurwitz, R.E., Jinek, M., Wiedenheft, B., Zhou, K., and Doudna, J.A. (2010). Sequence- and Structure-Specific RNA Processing by a CRISPR Endonuclease. Science *329*, 1355–1358.

Haurwitz, R.E., Sternberg, S.H., and Doudna, J.A. (2012). Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA. The EMBO Journal *31*, 2824–2832.

Hell, S.W., and Wichmann, J. (1994). Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. Optics Letters *19*, 780–782.

Hilton, I.B., D'Ippolito, A.M., Vockley, C.M., Thakore, P.I., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nature Biotechnology *33*, 510–517.

Hinnebusch, A.G. (2014). The Scanning Mechanism of Eukaryotic Translation Initiation. Annu. Rev. Biochem. *83*, 779–812.

Hochstrasser, M.L., Taylor, D.W., Bhat, P., Guegler, C.K., Sternberg, S.H., Nogales, E., and Doudna, J.A. (2014). CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. Proc Natl Acad Sci USA *111*, 6618–6623.

Hochstrasser, M.L., and Doudna, J.A. (2015). Cutting it close: CRISPR-associated endoribonuclease structure and function. Trends in Biochemical Sciences *40*, 58–66.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014a). Development and Applications of CRISPR-Cas9 for Genome Engineering. Cell *157*, 1262–1278.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014b). Development and Applications of CRISPR-Cas9 for Genome Engineering. Cell *157*, 1262–1278.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nature Biotechnology *31*, 827–832.

Huo, Y., Nam, K.H., Ding, F., Lee, H., Wu, L., Xiao, Y., Farchione, M.D., Zhou, S., Rajashankar, K., Kurinov, I., et al. (2014). Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. Nat Struct Mol Biol *21*, 771–777.

Hwang, W.Y., Fu, Y., Reyon, D., Maeder, M.L., Tsai, S.Q., Sander, J.D., Peterson, R.T., Yeh, J.-R.J., and Joung, J.K. (2013). Efficient genome editing in zebrafish using a CRISPR-Cas system. Nature Biotechnology *31*, 227–229.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Iyama, T., and Wilson, D.M., III (2013). DNA repair mechanisms in dividing and non-dividing cells. DNA Repair *12*, 620–636.

Izeddin, I., Récamier, V., Bosanac, L., Cissé, I.I., Boudarene, L., Dugast-Darzacq, C., Proux, F., Bénichou, O., Voituriez, R., Bensaude, O., et al. (2014). Single-molecule tracking in live cells reveals distinct target-search strategies of transcription factors in the nucleus. eLife *3*, 1–27.

Jackson, A.L., Bartz, S.R., Schelter, J., Kobayashi, S.V., Burchard, J., Mao, M., Li, B., Cavet, G., and Linsley, P.S. (2003). Expression profiling reveals off-target gene regulation by RNAi. Nature Biotechnology *21*, 635–638.

Jackson, R.N., Golden, S.M., van Erp, P.B.G., Carter, J., Westra, E.R., Brouns, S.J.J., van der Oost, J., Terwilliger, T.C., Read, R.J., and Wiedenheft, B. (2014). Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. Science *345*, 1473–1479.

Jackson, R.J., Hellen, C.U.T., and Pestova, T.V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. 1–15.

Jaqaman, K., Loerke, D., Mettlen, M., Kuwata, H., Grinstein, S., Schmid, S.L., and Danuser, G. (2008). Robust single-particle tracking in live-cell time-lapse sequences. Nat Meth *5*, 695–702.

Jiang, F., Taylor, D.W., Chen, J.S., Kornfield, J.E., Zhou, K., Thompson, A.J., Nogales, E., and Doudna, J.A. (2016). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. Science *351*, 867–870.

Jiang, F., Zhou, K., Ma, L., Gressel, S., and Doudna, J.A. (2015). A Cas9-guide RNA complex preorganized for target DNA recognition. Science *348*, 1477–1481.

Jinek, M., Jiang, F., Taylor, D.W., Sternberg, S.H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., et al. (2014). Structures of Cas9 Endonucleases Reveal RNA-Mediated Conformational Activation. Science *343*, 1247997–1247997.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science *337*, 816–821.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013). RNA-programmed genome editing in human cells. eLife *2*, e00471–e00479.

Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, Ü., Wurm, R., Wagner, R., et al. (2011). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nature Structural &Amp; Molecular Biology *18*, 529–536.

Jungreis, I., Lin, M.F., Spokony, R., Chan, C.S., Negre, N., Victorsen, A., White, K.P., and Kellis, M. (2011). Evidence of abundant stop codon readthrough in Drosophila and other metazoa. Genome Research *21*, 2096–2113.

Jurka, J., Kohany, O., Pavlicek, A., Kapitonov, V.V., and Jurka, M.V. (2005). Clustering, duplication and chromosomal distribution of mouse SINE retrotransposons. Cytogenet Genome Res *110*, 117–123.

Kadonaga, J.T. (1998). Eukaryotic Transcription: An Interlaced Review Network of Transcription Factorsand Chromatin-Modifying Machines. Cell *92*, 307–313.

Kamath, R.S., Fraser, A.G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., Kanapin, A., Bot, N.L., Moreno, S., Sohrmann, M., et al. (2003). Systematic functional analysis of the Caenorhabditis elegansgenome using RNAi. *421*, 231–237.

Kapp, L.D., and Lorsch, J.R. (2004). The Molecular Mechanics of Eukaryotic Translation. Annu. Rev. Biochem. *73*, 657–704.

Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution *30*, 772–780.

Keppler, A., Gendreizig, S., Gronemeyer, T., Pick, H., Vogel, H., and Johnsson, K. (2002). A general method for the covalent labeling of fusion proteins with small molecules in vivo. Nature Biotechnology *21*, 86–89.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Emnlp 1–6.

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

Knight, S.C., Xie, L., Deng, W., Guglielmi, B., Witkowsky, L.B., Bosanac, L., Zhang, E.T., Beheiry, El, M., Masson, J.-B., Dahan, M., et al. (2015). Dynamics of CRISPR-Cas9 genome interrogation in living cells. Science *350*, 823–827.

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

Koo, T., Lee, J., and Kim, J.-S. (2015). Measuring and Reducing Off-Target Activities of Programmable Nucleases Including CRISPR-Cas9. Molecules and Cells *38*, 475–481.

Kotterman, M.A., and Schaffer, D.V. (2014). Engineering adeno-associated viruses for clinical gene therapy. Nat Rev Microbiol *15*, 445–451.

Kozak, M. (2005). Regulation of translation via mRNA structure in prokaryotes and eukaryotes. Gene *361*, 13–37.

Kuscu, C., Arslan, S., Singh, R., Thorpe, J., and Adli, M. (2014). Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. Nature Biotechnology *32*, 677–682.

Lanphier, E., Urnov, F., Ehlen, S., Werner, M., and Smolenski, J. (2015). Don’t edit the human germ line. Nature *519*, 410–411.

Larson, M.H., Gilbert, L.A., Wang, X., Lim, W.A., Weissman, J.S., and Qi, L.S. (2013). CRISPR interference (CRISPRi) for sequence-specific control of gene expression. Nat Protoc *8*, 2180–2196.

Lavis, L.D., and Raines, R.T. (2008). Bright Ideas for Chemical Biology. ACS Chem. Biol. *3*, 142–155.

Lee, D.J., Minchin, S.D., and Busby, S.J.W. (2012). Activating Transcription in Bacteria. Annu. Rev. Microbiol. *66*, 125–152.

Levine, M., and Tjian, R. (2003). Transcription regulation and animal diversity. Nature *424*, 147–151.

Levine, M., Cattoglio, C., and Tjian, R. (2014). Looping Back to Leap Forward: Transcription Enters a New Era. Cell *157*, 13–25.

Li, H. (2015). Structural Principles of CRISPR RNA Processing. Structure/Folding and Design *23*, 13–20.

Li, L., Liu, H., Dong, P., Li, D., Legant, W.R., Grimm, J.B., Lavis, L.D., Betzig, E., Tjian, R., and Liu, Z. (2016). Real-time imaging of Huntingtinaggregates diverting target search andgene transcription. eLife *5*, e17056.

Li, T., Liu, B., Spalding, M.H., Weeks, D.P., and Yang, B. (2012). High-efficiency TALEN-based gene editing produces disease-resistant rice. Nature Biotechnology *30*, 390–392.

Liang, F., Han, M., Romanienko, P.J., and Jasin, M. (1998). Homology-directed repair is a major double-strand break repair pathway in mammalian cells. Proc Natl Acad Sci USA *95*, 5172–5177.

Lichtman, J.W., and Conchello, J.-A. (2005). Fluorescence microscopy. Nat Meth *2*, 910–919.

Lin, S., Staahl, B.T., Alla, R.K., and Doudna, J.A. (2014). Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. eLife *3*, 1–13.

Lippincott-Schwartz, J., Snapp, E., and Kenworthy, A. (2001). Studying Protein Dynamics in Living Cells. Nat Rev Mol Cell Biol *2*, 444–456.

Liu, S.J., Horlbeck, M.A., Cho, S.W., Birk, H.S., Malatesta, M., He, D., Attenello, F.J., Villalta, J.E., Cho, M.Y., Chen, Y., et al. (2017a). CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Science *355*, eaah7111–eaah7116.

Liu, T.Y., Iavarone, A.T., and Doudna, J.A. (2017b). RNA and DNA Targeting by a Reconstituted Thermus thermophilus Type III-A CRISPR-Cas System. PLoS ONE *12*, e0170552–20.

Liu, X.S., Wu, H., Ji, X., Stelzer, Y., Wu, X., Czauderna, S., Shu, J., Dadon, D., Young, R.A., and Jaenisch, R. (2016). Editing DNA Methylation in the Mammalian Genome. Cell *167*, 233–235.e17.

Liu, Z., Lavis, L.D., and Betzig, E. (2015). Imaging Live-Cell Dynamics and Structure at the Single-Molecule Level. Molecular Cell *58*, 644–659.

Liu, Z., Legant, W.R., Chen, B.-C., Li, L., Grimm, J.B., Lavis, L.D., Betzig, E., and Tjian,

R. (2014). 3D imaging of Sox2 enhancer clusters in embryonic stem cells. eLife *3*, 1–29.

Los, G.V., Encell, L.P., McDougall, M.G., Hartzell, D.D., Karassina, N., Zimprich, C., Wood, M.G., Learish, R., Ohana, R.F., Urh, M., et al. (2008). HaloTag: A Novel Protein Labeling Technology for Cell Imaging and Protein Analysis. ACS Chem. Biol. *3*, 373–382.

Loughran, G., Chou, M.-Y., Ivanov, I.P., Jungreis, I., Kellis, M., Kiran, A.M., Baranov, P.V., and Atkins, J.F. (2014). Evidence of efficient stop codon readthrough in four mammalian genes. Nucleic Acids Research *42*, 8928–8938.

Ma, H., Naseri, A., Reyes-Gutierrez, P., Wolfe, S.A., Zhang, S., and Pederson, T. (2015). Multicolor CRISPR labeling of chromosomal loci in human cells. Proc Natl Acad Sci USA *112*, 3002–3007.

Ma, H., Tu, L.-C., Naseri, A., Huisman, M., Zhang, S., Grunwald, D., and Pederson, T. (2016). Multiplexed labeling of genomic loci with dCas9 and engineered sgRNAs using CRISPRainbow. Nature Biotechnology *34*, 528–530.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J.J., Charpentier, E., Haft, D.H., et al. (2015). An updated evolutionary classification of CRISPR–Cas systems. Nat Rev Microbiol *13*, 722–736.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J., Norville, J.E., and Church, G.M. (2013). RNA-Guided Human Genome Engineering via Cas9. Science *339*, 823–826.

Manley, S., Gillette, J.M., Patterson, G.H., Shroff, H., Hess, H.F., Betzig, E., and Lippincott-Schwartz, J. (2008). High-density mapping of single-molecule trajectories with photoactivated localization microscopy. Nat Meth *5*, 155–157.

Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA. Science *322*, 1843–1845.

McIlwain, D.R., Berger, T., and Mak, T.W. (2013). Caspase Functions in Cell Death and Disease. Cold Spring Harbor Perspectives in Biology *5*, a008656–a008656.

Mehta, A., and Haber, J.E. (2014). Sources of DNA Double-Strand Breaks and Models of Recombinational DNA Repair. Cold Spring Harbor Perspectives in Biology *6*, a016428–a016428.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. Nips 1–9.

Miller, J.C., Tan, S., Qiao, G., Barlow, K.A., Wang, J., Xia, D.F., Meng, X., Paschon, D.E., Leung, E., Hinkley, S.J., et al. (2010). A TALE nuclease architecture for efficient genome editing. Nature Biotechnology *29*, 143–148.

Moffat, J., Grueneberg, D.A., Yang, X., Kim, S.Y., Kloepfer, A.M., Hinkle, G., Piqani, B., Eisenhaure, T.M., Luo, B., Grenier, J.K., et al. (2006). A Lentiviral RNAi Library for Human and Mouse Genes Applied to an Arrayed Viral High-Content Screen. Cell *124*, 1283–1298.

Moscou, M.J., and Bogdanove, A.J. (2009). A Simple Cipher Governs DNA Recognition by TAL Effectors. Science *326*, 1501–1501.

Mueller, F., Mazza, D., Stasevich, T.J., and McNally, J.G. (2010). FRAP and kinetic modeling in the analysis of nuclear protein dynamics: what do we really know? Current Opinion in Cell Biology *22*, 403–411.

Mueller, F., Senecal, A., Tantale, K., Marie-Nelly, H., Ly, N., Collin, O., Basyuk, E., Bertrand, E., Darzacq, X., and Zimmer, C. (2013). FISH-quant: automatic counting of transcripts in 3D FISH images. Nat Rev Microbiol *10*, 277–278.

Mulepati, S., Heroux, A., and Bailey, S. (2014). Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. Science *345*, 1479–1484.

Myasnikov, A.G., Simonetti, A., Marzi, S., and Klaholz, B.P. (2009). Structure–function insights into prokaryotic and eukaryotic translation initiation. Current Opinion in Structural Biology *19*, 300–309.

Nagai, T., Ibata, K., Park, E.S., Kubota, M., Mikoshiba, K., and Miyawaki, A. (2001). A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. Nature Biotechnology *20*, 87–90.

Naito, Y., Hino, K., Bono, H., and Ui-Tei, K. (2015). CRISPRdirect: software for designing CRISPR/ Cas guide RNA with reduced off-target sites. Bioinformatics *31*, 1120–1123.

Nam, K.H., Haitjema, C., Liu, X., Ding, F., Wang, H., DeLisa, M.P., and Ke, A. (2012). Cas5d Protein Processes Pre-crRNA and Assembles into a Cascade-like Interference Complex in Subtype I-C/Dvulg CRISPR-Cas System. Structure *20*, 1574–1584.

Nelles, D.A., Fang, M.Y., O'Connell, M.R., Xu, J.L., Markmiller, S.J., Doudna, J.A., and Yeo, G.W. (2016). Programmable RNA Tracking in Live Cells with CRISPR/Cas9. Cell *165*, 488–496.

Newton, A.C. (1995). Protein Kinase C: Structure, Function, and Regulation. J. Biol. Chem. *270*, 28495–28498.

Niewoehner, O., and Jinek, M. (2016). Structural basis for the endoribonuclease activity of the type III-A CRISPR-associated protein Csm6. Rna *22*, 318–329.

Nilsen, T.W. (2013). Gel Purification of RNA. Cold Spring Harbor Protocols *2013*, pdb.prot072942–pdb.prot072942.

Nishimasu, H., and Nureki, O. (2017). Structures and mechanisms of CRISPR RNA-guided effector nucleases. Current Opinion in Structural Biology *43*, 68–78.

Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014). Crystal Structure of Cas9 in Complex with Guide RNA and Target DNA. Cell *156*, 935–949.

Niu, Y., Bin Shen, Cui, Y., Chen, Y., Wang, J., Wang, L., Kang, Y., Zhao, X., Si, W., Li, W., et al. (2014). Generation of Gene-Modified Cynomolgus Monkey via Cas9/RNA-Mediated Gene Targeting in One-Cell Embryos. Cell *156*, 836–843.

Normanno, D., ne, L.B.E., Dugast-Darzacq, C., Chen, J., Richter, C., Proux, F., nichou, O.B.E., Voituriez, R.E.L., Darzacq, X., and Dahan, M. (1AD). Probing the target search of DNA-binding proteins in mammalian cells using TetR as model searcher. Nature Communications *6*, 1–10.

O'Geen, H., Henry, I.M., Bhakta, M.S., Meckler, J.F., and Segal, D.J. (2015). A genome-wide analysis of Cas9 binding specificity using ChIP-seq and targeted sequence capture. Nucleic Acids Research *43*, 3389–3404.

Ormo, M., Cubitt, A.B., Kallio, K., Gross, L.A., Tsien, R.Y., and Remington, S.J. (1996). Crystal Structure of the *Aequorea victoria* Green Fluorescent Protein. Science *273*, 1392–1395.

O'Connell, M.R., Oakes, B.L., Sternberg, S.H., East-Seletsky, A., Kaplan, M., and Doudna, J.A. (2015). Programmable RNA recognition and cleavage by CRISPR/Cas9. Nature *516*, 263–266.

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

Pattanayak, V., Ramirez, C.L., Joung, J.K., and Liu, D.R. (2011). Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection. Nat Meth *8*, 765–770.

Patterson, G.H., and Lippincott-Schwartz, J. (2002). A Photoactivatable GFP for Selective Photolabeling of Proteins and Cells. Science *297*, 1873–1877.

Pavletich, N.P., and Pabo, C.O. (1991). Zinc Finger-DNA Recognition: Crystal Structure of a Zif268-DNA Complex at 2.1 Å. Science *252*, 809–817.

Perrimon, N., Ni, J.Q., and Perkins, L. (2010). In vivo RNAi: Today and Tomorrow. Cold Spring Harbor Perspectives in Biology *2*, a003640–a003640.

Prelich, G. (2012). Gene overexpression: uses, mechanisms, and interpretation. Genetics *190*, 841–854.

Qin, P., Parlak, M., Kuscu, C., Bandaria, J., Mir, M., Szlachta, K., Singh, R., Darzacq, X., Yildiz, A., and Adli, M. (2017). Live cell imaging of low- and non-repetitive chromosome loci using CRISPR-Cas9. Nature Communications *8*, 1–10.

Richardson, C.D., Ray, G.J., DeWitt, M.A., Curie, G.L., and Corn, J.E. (2016). Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. Nature Biotechnology *34*, 339–344.

Rockman, M.V., and Kruglyak, L. (2006). Genetics of global gene expression. Nature Reviews Genetics *7*, 862–872.

Rohrman, B.A., Leautaud, V., Molyneux, E., and Richards-Kortum, R.R. (2012). A Lateral Flow Assay for Quantitative Detection of Amplified HIV-1 RNA. PLoS ONE *7*, e45611–e45618.

Rouillon, C., Zhou, M., Zhang, J., Politis, A., Beilsten-Edmands, V., Cannone, G., Graham, S., Robinson, C.V., Spagnolo, L., and White, M.F. (2013). Structure of the CRISPR Interference Complex CSM Reveals Key Similarities with Cascade. Molecular Cell *52*, 124–134.

Rust, M.J., Bates, M., and Zhuang, X. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). Nat Meth *3*, 793–796.

Sachs, A.B., Sarnow, P., and Hentze, M.W. (1997). Starting at the Beginning, Middle, and End: Translation Initiation in Eukaryotes. Cell *89*, 831–839.

Samai, P., Pyenson, N., Jiang, W., Goldberg, G.W., Hatoum-Aslan, A., and Marraffini, L.A. (2015). Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. Cell *161*, 1164–1174.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. Nature *473*, 337–342.

Schwille, P., Meyer-Almes, F.J., and Rigler, R. (1997). Dual-color fluorescence cross-correlation spectroscopy for multicomponent diffusional analysis in solution. Biophys. J. *72*, 1878–1886.

Sergé, A., Bertaux, N., Rigneault, H., and Marguet, D. (2008). Dynamic multiple-target tracing to probe spatiotemporal cartography of cell membranes. Nat Meth *5*, 687–694.

Shalem, O., Sanjana, N.E., and Zhang, F. (2015). High-throughput functional genomics using CRISPR–Cas9. Nat Rev Microbiol *16*, 299–311.

Shan, Q., Wang, Y., Li, J., Zhang, Y., Chen, K., Liang, Z., Zhang, K., Liu, J., Xi, J.J., Qiu, J.-L., et al. (2013). Targeted genome modification of crop plants using a CRISPR-Cas system. Nature Biotechnology *31*, 686–688.

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

Sheppard, N.F., Glover, C.V.C., III, Terns, R.M., and Terns, M.P. (2016). The CRISPR-associated Csx1 protein of Pyrococcus furiosusis an adenosine-specific endoribonuclease. Rna *22*, 216–224.

Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., Severinov, K., et al. (2015). Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. Molecular Cell *60*, 385–397.

Singh, R., Kuscu, C., Quinlan, A., Qi, Y., and Adli, M. (2015). Cas9-chromatin binding information enables more accurate CRISPR off-target prediction. Nucleic Acids Research *43*, e118–e118.

Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. The EMBO Journal *30*, 1335–1342.

Slaymaker, I.M., Gao, L., Zetsche, B., Scott, D.A., Yan, W.X., and Zhang, F. (2015). Rationally engineered Cas9 nucleases with improved specificity. Science *351*, 84–88.

Smith, J., Bibikova, M., Whitby, F.G., Reddy, A.R., Chandrasegaran, S., and Carroll, D. (2000). Requirements for double-strand cleavage by chimeric restriction enzymes with zinc finger DNA-recognition domains. Nucleic Acids Research *28*, 3361–3369.

Sonenberg, N., and Dever, T.E. (2003). Eukaryotic translation initiation factors and regulators. Current Opinion in Structural Biology *13*, 56–63.

Sonenberg, N., and Hinnebusch, A.G. (2009). Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. Cell *136*, 731–745.

Sprague, B.L., Pego, R.L., Stavreva, D.A., and McNally, J.G. (2004). Analysis of Binding Reactions by Fluorescence Recovery after Photobleaching. Biophys. J. *86*, 3473–3495.

Staals, R.H.J., Agari, Y., Maki-Yonekura, S., Zhu, Y., Taylor, D.W., van Duijn, E., Barendregt, A., Vlot, M., Koehorst, J.J., Sakamoto, K., et al. (2013). Structure and Activity of the RNA-Targeting Type III-B CRISPR-Cas Complex of Thermus thermophilus. Molecular Cell *52*, 135–145.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Sternberg, S.H., Haurwitz, R.E., and Doudna, J.A. (2012). Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. Rna *18*, 661–672.

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

Szczelkun, M.D., Tikhomirova, M.S., Sinkunas, T., Gasiunas, G., Karvelis, T., Pschera, P., Siksnys, V., and Seidel, R. (2014). Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. Proc Natl Acad Sci USA *111*, 9798–9803.

Takahashi, K., and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell *126*, 663–676.

Tarantino, N., Tinevez, J.Y., Crowell, E.F., Boisson, B., Henriques, R., Mhlanga, M., Agou, F., Israel, A., and Laplantine, E. (2014). TNF and IL-1 exhibit distinct ubiquitin requirements for inducing NEMO-IKK supramolecular structures. The Journal of Cell Biology *204*, 231–245.

Taylor, D.W., Zhu, Y., Staals, R.H.J., Kornfeld, J.E., Shinkai, A., van der Oost, J., Nogales, E., and Doudna, J.A. (2015). Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning. Science *348*, 581–585.

Terns, R.M., and Terns, M.P. (2014). CRISPR-based technologies: prokaryotic defense weapons repurposed. Trends in Genetics *30*, 111–118.

Traxler, E.A., Yao, Y., Wang, Y.-D., Woodard, K.J., Kurita, R., Nakamura, Y., Hughes, J.R., Hardison, R.C., Blobel, G.A., Li, C., et al. (2016). A genome-editing strategy to treat β-hemoglobinopathies that recapitulates a mutation associated with a benign genetic condition. Nat Med *22*, 987–990.

Tsai, S.Q., Zheng, Z., Nguyen, N.T., Liebers, M., Topkar, V.V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A.J., Le, L.P., et al. (2014). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nature Biotechnology *33*, 187–197.

Tsien, R.Y. (1998). The Green Fluorescent Protein. Annu. Rev. Biochem. *67*, 509–544.

Urnov, F.D., Rebar, E.J., Holmes, M.C., Zhang, H.S., and Gregory, P.D. (2010). Genome editing with engineered zinc finger nucleases. Nat Rev Microbiol *11*, 636–646.

van den Bosch, M., Bree, R.T., and Lowndes, N.F. (2003). The MRN complex: coordinating and mediating the response to broken chromosomes. EMBO Rep *4*, 844–849.

van der Oost, J., Westra, E.R., Jackson, R.N., and Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of CRISPR–Cas systems. Nat Rev Microbiol *12*, 479–492.

Voss, T.C., and Hager, G.L. (2013). Dynamic regulation of transcriptional states by chromatin and transcription factors. Nat Rev Microbiol *15*, 69–81.

Wang, H., Yang, H., Shivalila, C.S., Dawlaty, M.M., Cheng, A.W., Zhang, F., and Jaenisch, R. (2013). One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering. 1–9.

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

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic Screens in Human Cells Using the CRISPR-Cas9 System. Science *343*, 80–84.

Webb, R.H. (1996). Confocal optical microscopy. Rep. Prog. Phys. *59*, 427–471.

Wilson, R.C., and Doudna, J.A. (2013). Molecular Mechanisms of RNA Interference. Annu. Rev. Biophys. *42*, 217–239.

Wood, A.J., Lo, T.-W., Zeitler, B., Pickle, C.S., Ralston, E.J., Lee, A.H., Amora, R., Miller, J.C., Leung, E., Meng, X., et al. (2011). Targeted Genome Editing Across Species Using ZFNs and TALENs. Science *333*, 307–307.

Wright, A.V., Nuñez, J.K., and Doudna, J.A. (2016). Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. Cell *164*, 29–44.

Wu, X., Scott, D.A., Kriz, A.J., Chiu, A.C., Hsu, P.D., Dadon, D.B., Cheng, A.W., Trevino, A.E., Konermann, S., Chen, S., et al. (2014). Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. Nature Biotechnology *32*, 670–676.

Xiao-Jie, L., Hui-Ying, X., Zun-Ping, K., Jin-Lian, C., and Li-Juan, J. (2015). CRISPR-Cas9: a new and promising player in gene therapy. J. Med. Genet. *52*, 289–296.

Yamano, T., Nishimasu, H., Zetsche, B., Hirano, H., Slaymaker, I.M., Li, Y., Fedorova, I., Nakane, T., Makarova, K.S., Koonin, E.V., et al. (2016). Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. Cell *165*, 949–962.

Yang, L., Guell, M., Niu, D., George, H., Lesha, E., Grishin, D., Aach, J., Scrhock, E., Xu, W., Poci, J., et al. (2015). Genome-wide inactivation of porcine endogenous retroviruses (PERVs). Molecular Systems Biology *350*, 1101–1104.

Ye, L., Wang, J., Tan, Y., Beyer, A.I., Xie, F., Muench, M.O., and Kan, Y.W. (2016). Genome editing using CRISPR-Cas9 to create the HPFH genotype in HSPCs: An approach for treating sickle cell disease and β-thalassemia. Proc Natl Acad Sci USA *113*, 10661–10665.

Yeh, J.E., Toniolo, P.A., and Frank, D.A. (2013). Targeting transcription factors. Current Opinion in Oncology *25*, 652–658.

Yin, H., Kanasty, R.L., Eltoukhy, A.A., Vegas, A.J., Dorkin, J.R., and Anderson, D.G. (2014). Non-viral vectors for gene-based therapy. Nat Rev Microbiol *15*, 541–555.

Zalatan, J.G., Lee, M.E., Almeida, R., Gilbert, L.A., Whitehead, E.H., La Russa, M., Tsai, J.C., Weissman, J.S., Dueber, J.E., Qi, L.S., et al. (2015). Engineering Complex Synthetic Transcriptional Programs with CRISPR RNA Scaffolds. Cell *160*, 339–350.

Zamore, P.D., Tuschl, T., Sharp, P.A., and Bartel, D.P. (2000). RNAi: Double-Stranded RNA Directs the ATP-Dependent Cleavage of mRNA at 21 to 23 Nucleotide Intervals. Cell *101*, 25–33.

Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., et al. (2015). Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell *163*, 759–771.