Mechanism and engineering of CRISPR-associated endonucleases

by

Samuel Henry Sternberg

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, Chair
Professor Carlos Bustamante
Professor Jamie H. D. Cate
Professor Susan Marqusee

Fall 2014

Abstract

Mechanism and engineering of CRISPR-associated endonucleases

by

Samuel Henry Sternberg

Doctor of Philosophy in Chemistry

University of California, Berkeley

Professor Jennifer A. Doudna, Chair

Bacteria and archaea have evolved multiple defense pathways for protection from invading viruses and plasmids. A recently discovered adaptive immune system relies on specialized genomic loci called CRISPR (clustered regularly interspaced short palindromic repeats), which function together with CRISPR-associated (Cas) proteins to target foreign nucleic acids for degradation. A hallmark feature of CRISPR–Cas immune systems is the use of non-coding RNA transcribed from the CRISPR locus (crRNA) to identify foreign DNA via RNA:DNA base-pairing. Conserved families of Cas enzymes play critical roles both in producing crRNAs and in cleaving DNA sequences targeted with crRNA guides. This work describes the basic functions of two such endonucleases, with a focus on engineering these systems for desired biotechnological applications.

CRISPR loci are initially transcribed as long precursor crRNAs (pre-crRNAs), which must be enzymatically cleaved to generate libraries of mature crRNAs that each target a unique DNA sequence. This processing event typically occurs at the 3' side of a stable RNA stem-loop structure and is catalyzed by Cas6. We show that one Cas6 family member called Csy4 recognizes its RNA substrate with extremely high affinity and exquisite specificity. Binding energy derives exclusively from interactions upstream of the scissile phosphate, allowing Csy4 to retain the cleavage product and sequester the crRNA for subsequent ribonucleoprotein complex formation. Using biochemical assays and three protein–RNA co-crystal structures, we reveal the chemical mechanism of RNA cleavage by Csy4 and identify the catalytic roles of an unusual catalytic dyad comprising histidine and serine residues. Our experiments highlight diverse modes of substrate recognition that enable Csy4 to accurately select CRISPR transcripts for processing while avoiding off-target RNA binding and cleavage.

Following crRNA biogenesis, one or more Cas proteins form large ribonucleoprotein complexes with the crRNA and utilize its sequence content to target complementary nucleic acids. Cas9 is a DNA endonuclease found in some bacteria that uses a dual-guide RNA comprising crRNA and *trans*-activating crRNA (tracrRNA) to identify target DNA sites for cleavage. We unravel the mechanism of DNA interrogation by Cas9:RNA complexes using both single-molecule and bulk biochemical experiments. The target search process is guided by recognition of a short trinucleotide sequence adjacent to potential target sites called the protospacer adjacent motif (PAM), and PAM binding triggers Cas9 catalytic activity. We also present three-dimensional

structures of Cas9 from X-ray crystallography and electron microscopy experiments, which reveal RNA/DNA binding interfaces and the organization of both catalytic domains. Strikingly, RNA binding drives large-scale rearrangements of the Cas9 enzyme to form a central DNA-binding channel. This observation implicates RNA loading as a key step in Cas9 activation.

Cas9:RNA complexes have proven to be extremely effective genome engineering agents in animals and plants. By redesigning the sequence of the crRNA, Cas9 can be programmed to target virtually any desired DNA sequence inside the cell. We reveal that Cas9 can also be programmed to target single-stranded RNA substrates for both high-affinity binding and site-specific cleavage using PAM-presenting oligonucleotides. This approach enables the isolation of specific endogenous mRNA transcripts from cells. We believe that RNA targeting by Cas9 has the potential to transform the study of RNA function, much as site-specific DNA targeting has revolutionized genetic and genomic research.

ACKNOWLEDGEMENTS

I distinctly remember the distress I felt in 2008 when I was forced to choose between graduate programs at different universities that all seemed equally amazing. After picking the University of California, Berkeley, and deferring one year to wrap up my undergraduate research project at Columbia University, I went through a fresh round of agony after learning that Jennifer Doudna was moving her lab to Genentech, where she had accepted a position as Vice President of Discovery Research. (The recently announced 20% campus-wide budget cuts at Berkeley also didn't help in quieting my concerns.) And yet, six years later, here I am writing my thesis after spending four incredible years in Jennifer's lab, and on top of that, having the opportunity to live through the CRISPR-Cas9 genome engineering revolution. Life has a crazy way of working out sometimes.

I couldn't be happier that Jennifer decided to return to Berkeley and accept me into her laboratory as a Ph.D. student. Through her patience and seemingly endless optimism, Jennifer created an environment in which I could truly thrive and blossom as a scientist. Jennifer taught me how to think big-picture about a project and how to organize a set of experiments into a narrative for publication. I improved my presentation skills considerably through her guidance and by observing her speak at conferences, and I gained an appreciation for how to write concisely and to the point. I also learned a lot about how to be a scientist outside of the laboratory, whether at a conference, in a meeting with collaborators, or on the phone with patent lawyers. Perhaps most importantly, Jennifer was incredibly supportive of me throughout my entire Ph.D. I was quite persistent in asking her to attend conferences, and yet I never once received "no" as an answer. When I came to her with the idea to study Cas9 using single-molecule fluorescence in collaboration with a lab at Columbia University, she not only sent me out there for a week to try it out, she ended up paying for me to spend six whole months there. And anytime I had a new idea for an experiment or a collaboration, Jennifer was always on my side and let me pursue my interests with an open mind. My positive experience in the Doudna lab will have a lasting effect on me for the rest of my scientific career.

My successes during my Ph.D. are also intimately linked to the inspiring and stimulating scientists with whom I had the pleasure and honor to work. Acknowledging them satisfactorily here will not be possible, but I'll try nonetheless. Blake Wiedenheft was a true role model in the lab, and he epitomizes the scientific enthusiasm that I also share. Above all, I learned from him how to effectively initiate and foster collaborations, and how to follow my passion in scientific research. Martin Jinek is one of the most brilliant scientists I've had the privilege to work with, and I feel lucky to call him a friend and colleague. Rachel Haurwitz was my first collaborator at Berkeley and helped me hit the ground running. It's been so exciting to see her transition from academia to a very successful career in the private sector. Katie Berry and Dipa Sashital were my closest friends in the lab during my rotation project and nurtured my quantitative, biochemically focused approach to science. Kaihong Zhou has been an incredible manager of the laboratory, and I will never forget her go-getter attitude. She made it easy to perform scientific research to the best of my ability, because I always had everything I needed, and if I didn't, she made sure I got it as soon as possible. If I'm lucky enough to run my own lab someday, finding someone as talented as her to manage it will be my first priority.

I can't stress enough how important my collaborators have been during my Ph.D. First and foremost on the list is Sy Redding, with whom I worked closely for over a year, both in person in New York and remotely. It wasn't until we began working together that I truly appreciated the power of team efforts in science. Back in Berkeley, I worked closely with David Taylor on structures of Cas9 and will never forget the many joyful days we spent together writing, brainstorming, laughing, and even having the occasional cigar. I developed a close friendship with Mitch O'Connell and Ben Oakes through a collaboration that began over a crazy idea, and blossomed into a quick *Nature* paper that is one of my fondest success stories from the lab. And finally, one of the biggest pleasures of my Ph.D. has been mentoring and working together with undergraduate and graduate students. Prashant Bhat is one of the hardest working, most motivated, friendly, and sincere students I think I will ever come across, and I am so happy he approached me during that intro chemistry review session to ask about the Doudna lab. Chantal Guegler and Matias Kaplan worked with me during the summer of 2013, and together we formed a fantastic trio that was even able to experience a repeat run the following summer. There are a number of other students I don't have space to write more about – James Nuñez, Megan Hochstrasser, Brian Castellano, Addison Von Wright, Ben LaFrance, and others – and I have learned immensely from every one of these interactions. And finally I have to thank all the other members of the Doudna lab, and the Stanley Hall 7th floor, that I didn't have the opportunity to specifically mention.

Berkeley is an incredible place to do science, and I've benefitted from my many exchanges with folks from other labs during my time here. Deserving specific mention are Mary Matyskiela, my mentor during my rotation in Andy Martin's lab, and Lacra Bintu and Rodrigo Maillard from Carlos Bustamante's lab. I've also had many productive and rewarding conversations with other students, postdocs, and professors at Berkeley, and all of these have contributed to my development as a scientist. I'd like to especially thank my thesis committee members for their discussions and friendships over the years, and for agreeing to read (or at least, sign-off on) my thesis.

It would be remiss of me if I didn't thank Ruben Gonzalez, my undergraduate advisor at Columbia University. There is nobody that has had, or will have, a more lasting impression and impact on my scientific career. It is because of Ruben that I ever applied to and was accepted into graduate school; he took a chance on a student with no experience, and inspired that student (me) to get where I am today. I learned more from him than I could possibly describe here, but one of the most important things was how to design a properly controlled experiment. Beyond that, Ruben proved to me that you could be a successful scientist and also be a pretty freakin' cool dude. I will never forget my time in his lab and the things I learned there.

My time in Berkeley has predominantly revolved around science, but there are a few things worth mentioning that kept me sane through the years. I developed a passion and obsession for squash that allowed me to expel all the pent up energy from sitting and working in lab all day. I've made a lot of good friends over the years on and off the courts, and look forward to maintaining my ties to the Berkeley squash community in the coming years. Music has played a huge role in my life, but in a very different way in the Bay Area. Although I was trained classically, I became a funk musician through my time with the Funk Revival Orchestra (FRO), a hard hitting, ten-piece funk band. Leaving work every Wednesday to rehearse with the band in Oakland was such a treat, as were the monthly gigs in San Francisco, and I'll never forget the friendships I made in the band. There were all the classic ups and downs that bands experience, and I loved them all. My allegiance to the band was so strong that I even considered thanking F.R.O. for "support" in the acknowledgements section of my 2014 *Nature* article, since my six months spent in New York City to conduct experiments for that paper meant that they had to find a temporary saxophone sub.

Finally, I'd like to thank my friends and family for all their moral support over the years. Many Ph.D. friends shared my high and lows over the years, and I look forward to seeing which path we all follow into the future. I had some fantastic roommates during my time here, and while I wasn't always around that much (too many hours in lab), I always enjoyed the fun times at home when we were all together. I don't have a wife to thank (yet), but a few individuals over the years made my time here that much more special; they know who they are. Berkeley was not my first choice for graduate school for the program alone, but it attracted me for other reasons as well, one of which was the fact that the Bay Area was already home to my one and only brother, Max Sternberg. After spending six years with a lot of distance separating us, I've been lucky enough to call him my roommate for my entire time in Berkeley (and Oakland and El Cerrito), and he's been a great source of support. I'll never forget our epic late nights playing table tennis, competing in backgammon and Wii sports, and frequenting the stoop out in front of our house. Last but not least, I have to thank my parents. I am who I am today because of them, and all my successes are a direct result of their love and dedication in parenting me. In all honesty, I couldn't imagine two better parents than them, and my dream is to one day raise kids as well as they raised me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

RNA-guided genetic silencing systems in bacteria and archaea

† Part of the work presented in this chapter has previously been published in the following review article: Wiedenheft, B., Sternberg, S.H., Doudna, J.A. (2012). RNA-guided genetic silencing systems in bacteria and archaea. Nature *482*, 331–338.

‡ The work presented in this chapter predates many of the discoveries reported in this thesis and serves as background information. For an updated review, including the development of CRISPR-Cas9 applications, please see Chapter 7.

1.1 Abstract

Clustered regularly interspaced short palindromic repeat (CRISPR) are essential components of nucleic-acid-based adaptive immune systems that are widespread in bacteria and archaea. Similar to RNA interference (RNAi) pathways in eukaryotes, CRISPR-mediated immune systems rely on small RNAs for sequence-specific detection and silencing of foreign nucleic acids, including viruses and plasmids. However, the mechanism of RNA-based bacterial immunity is distinct from RNAi. Understanding how small RNAs are used to find and destroy foreign nucleic acids will provide new insights into the diverse mechanisms of RNA-controlled genetic silencing systems.

1.2 Introduction

Bacteria and archaea are the most diverse and abundant organisms on the planet, thriving in habitats that range from hot springs to humans. However, viruses outnumber their microbial hosts in every ecological setting, and the selective pressures imposed by these rapidly evolving parasites has driven the diversification of microbial defense systems (Hoskisson and Smith, 2007; Rodriguez-Valera et al., 2009; Weinbauer, 2004). Historically, our understanding of antiviral immunity in bacteria has focused on restriction-modification systems, abortive-phage phenotypes, toxin–antitoxins and other innate defense systems (Labrie et al., 2010; Stern and Sorek, 2011). More recently, bioinformatic, genetic and biochemical studies have revealed that many prokaryotes use an RNA-based adaptive immune system to target and destroy genetic parasites (reviewed in (Al-Attar et al., 2011; Deveau et al., 2010; Horvath and Barrangou, 2010; Karginov and Hannon, 2010; Makarova et al., 2011b; Marraffini and Sontheimer, 2010a; Sorek et al., 2008)). Such adaptive immunity, previously thought to occur only in eukaryotes, provides an example of RNA-guided destruction of foreign genetic material by a process that is distinct from RNA interference (RNAi) (**Fig. 1.1**).

In response to viral and plasmid challenges, bacteria and archaea integrate short fragments of foreign nucleic acid into the host chromosome at one end of a repetitive element known as CRISPR (clustered regularly interspaced short palindromic repeat) (Andersson and Banfield, 2008; Barrangou et al., 2007; Garneau et al., 2010). These repetitive loci serve as molecular ‘vaccination cards’ by maintaining a genetic record of prior encounters with foreign transgressors. CRISPR loci are transcribed, and the long primary transcript is processed into a library of short CRISPR-derived RNAs (crRNAs) (Carte et al., 2008a; Deltcheva et al., 2011; Gesner et al., 2011; Haurwitz et al., 2010; Sashital et al., 2011; Wang et al., 2011) that each contain a sequence complementary to a previously encountered invading nucleic acid. Each crRNA is packaged into a large surveillance complex that patrols the intracellular environment and mediates the detection and destruction of foreign nucleic acid targets (Brouns et al., 2008; Garneau et al., 2010; Hale et al., 2008; Jore et al., 2011a; Lintner et al., 2011; Wiedenheft et al., 2011a; 2011b).

CRISPRs were originally identified in the *Escherichia coli* genome in 1987, when they were described as an unusual sequence element consisting of a series of 29-nucleotide repeats separated by unique 32-nucleotide ‘spacer’ sequences (Ishino et al., 1987). Repetitive sequences with a similar repeat–spacer–repeat pattern were later identified in phylogenetically diverse bacterial and archaeal genomes, but the function of these repeats remained obscure until many spacer sequences were recognized as being identical to viral and plasmid sequences (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). This observation led to the hypothesis that CRISPRs provide a genetic memory of infection (Bolotin et al., 2005), and the detection of short CRISPR-derived RNA transcripts suggested that there may be functional similarities between CRISPR- based immunity and RNAi (Makarova et al., 2006; Mojica et al., 2005). Here, we review three stages of CRISPR-based adaptive immunity and compare mechanistic aspects of these immune systems to other RNA-guided genetic silencing pathways.

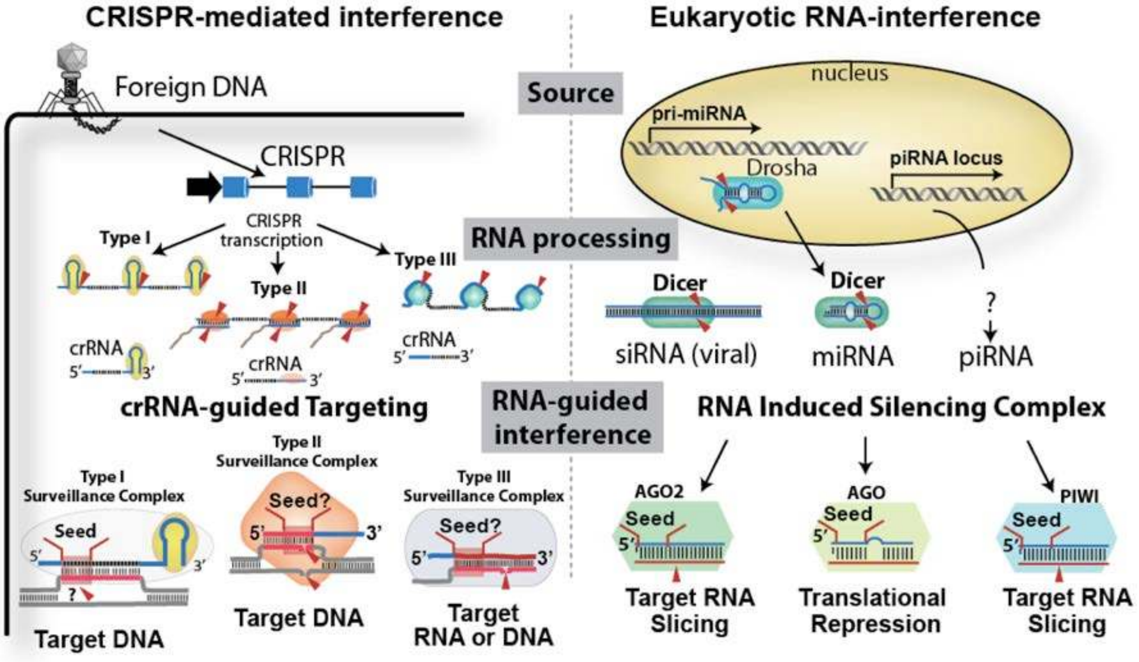


Figure 1.1 | Parallels and distinctions between CRISPR RNA-guided silencing systems and RNAi. CRISPR systems and RNAi recognize long RNA precursors that are processed into small RNAs, which act as sequence-specific guides for targeting complementary nucleic acids. In CRISPR systems, foreign DNA is integrated into the CRISPR locus, and long transcripts from these loci are processed by a CRISPR-associated (Cas) or RNase III family nuclease. The short CRISPR-derived RNAs (crRNAs) assemble with Cas proteins into large surveillance complexes that target destruction of invading genetic material. In some eukaryotes, long double-stranded RNAs are recognized as foreign, and a specialized RNase III family endoribonuclease (Dicer) cleaves these RNAs into short-interfering RNAs (siRNAs) that guide the immune system to invading RNA viruses (Obbard et al., 2009). PIWI-interacting RNAs (piRNAs) are transcribed from repetitive clusters in the genome that often contain many copies of retrotransposons and primarily act by restricting transposon mobility (Aravin et al., 2001; 2007; Obbard et al., 2009). The biogenesis of piRNAs is not yet fully understood. MicroRNAs (miRNAs) are also encoded on the chromosome, and primary miRNA transcripts form stable hairpin structures that are sequentially processed (shown by red triangles) by two RNase III family endoribonucleases (Drosha and Dicer) (Bartel, 2004). miRNAs do not participate in genome defence but are major regulators of endogenous gene expression (Guo et al., 2010). Like crRNAs, eukaryotic piRNAs, siRNAs and miRNAs associate with proteins that facilitate complementary interactions with invading nucleic acid targets. In eukaryotes, the Argonaute proteins pre-order the 5′ region of the guide RNA into a helical configuration, reducing the entropy penalty of interactions with target RNAs (Parker et al., 2009). This high-affinity binding site, called the 'seed' sequence, is essential for target sequence interactions. Recent studies indicate that the CRISPR system may use a similar seed-binding mechanism for enhancing target sequence interactions.

1.3 Architecture and composition of CRISPR loci

The defining feature of CRISPR loci is a series of direct repeats (approximately 20–50 base pairs) separated by unique spacer sequences of a similar length (Grissa et al., 2007a; Marraffini and Sontheimer, 2010a; Rousseau et al., 2009) (**Fig. 1.2**). The repeat sequences within a CRISPR locus are conserved, but repeats in different CRISPR loci can vary in both sequence and length. In addition, the number of repeat–spacer units in a CRISPR locus varies widely within and among organisms (Kunin et al., 2007).

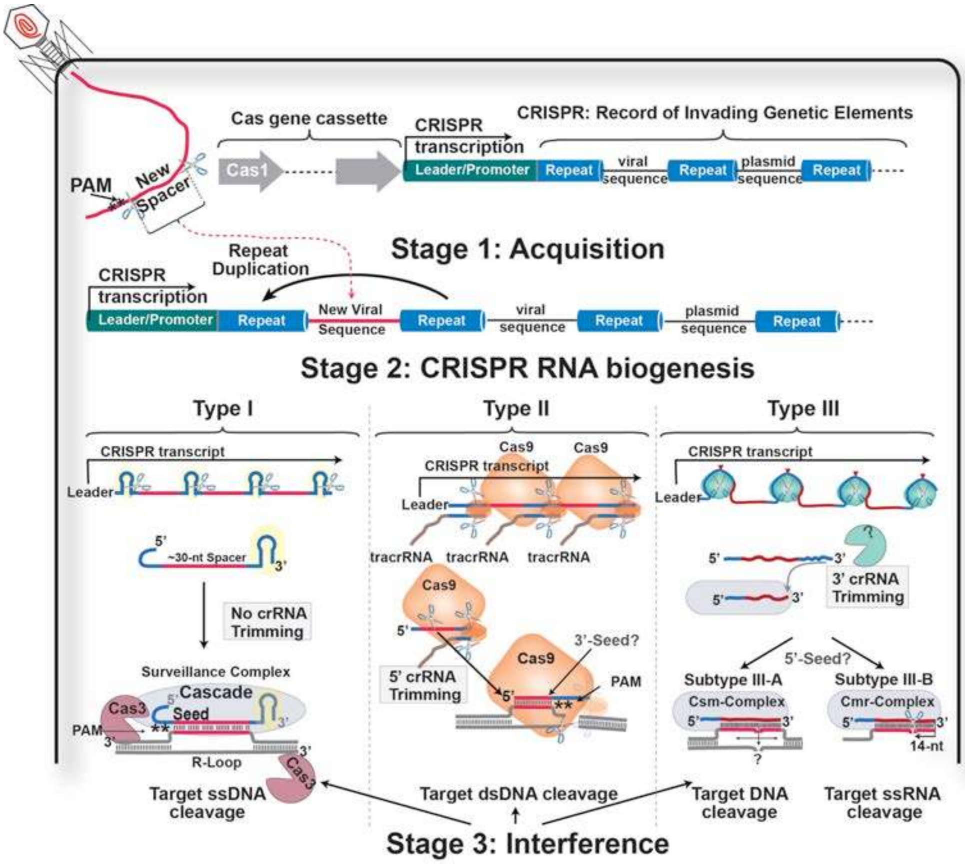


Figure 1.2 | Diversity of CRISPR-mediated adaptive immune systems in bacteria and archaea. A diverse set of CRISPR-associated (*cas*) genes (grey arrows) encode proteins required for new spacer sequence acquisition (Stage 1), CRISPR RNA biogenesis (Stage 2) and target interference (Stage 3). Each CRISPR locus consists of a series of direct repeats separated by unique spacer sequences acquired from invading genetic elements (protospacers). Protospacers are flanked by a short motif called the protospacer adjacent motif (PAM, **) that is located on the 5′ (type I) or 3′ (type II) side in foreign DNA. Long CRISPR transcripts are processed into short crRNAs by distinct mechanisms. In type I and III systems, a CRISPR-specific endoribonuclease (yellow ovals and green circles, respectively) cleaves 8 nucleotides upstream of each spacer sequence. In type III systems, the repeat sequence on the 3′ end of the crRNA is trimmed by an unknown mechanism (green Pacman, right). In type II systems, a *trans*-acting antisense RNA (tracrRNA) with complementarity to the CRISPR RNA repeat sequence forms an RNA duplex that is recognized and cleaved by cellular RNase III (brown ovals). This cleavage intermediate is further processed at the 5′ end resulting in a mature, approximately 40-nucleotide crRNA with an approximately 20-nucleotide 3′-handle. In each system, the mature crRNA associates with one or more Cas proteins to form a surveillance complex (green rectangles). Type I systems encode a Cas3 nuclease (blue Pacman), which may be recruited to the surveillance complex following target binding. A short high-affinity binding site called a seed-sequence has been identified in some type I systems, and genetic experiments suggest that type II systems have a seed sequence located at the 3′ end of the crRNA spacer sequence.

The sequence diversity of these repetitive loci initially limited their detection and obscured their relationship, but computational methods have been developed for detecting repeat patterns rather than related sequences (Bland et al., 2007; Dsouza et al., 1997; Edgar, 2007; Grissa et al., 2007a; Rousseau et al., 2009). One of the first-generation pattern-recognition algorithms identified the repeat–spacer–repeat architecture in phylogenetically diverse bacterial and archaeal genomes, but related structures were not identified in eukaryotic chromosomes (Jansen et al., 2002). Comparative analyses of the sequences adjacent to the CRISPR loci have revealed an (A+T)-rich 'leader' sequence that has been shown to serve as a promoter element for CRISPR transcription (Jansen et al., 2002; Pougach et al., 2010; Pul et al., 2010; Westra et al., 2010). In addition to the leader sequence, Jansen *et al.* (Jansen et al., 2002) identified a set of four CRISPR-associated (*cas)* genes known as *cas*1–4 that are found exclusively in genomes containing CRISPRs. Based on sequence similarity to proteins of known function, Cas3 was predicted to be a helicase and Cas4 a RecB-like exonuclease (Jansen et al., 2002).

Subsequent bioinformatic analyses have shown that CRISPR loci are flanked by a large number of extremely diverse *cas* genes (Haft et al., 2005; Makarova et al., 2006). The *cas1* gene is a common component of all CRISPR systems, and phylogenetic analyses of Cas1 sequences indicate there are several versions of the CRISPR system. Providing additional evidence for the classification of distinct CRISPR types, neighbourhood analysis has identified conserved arrangements of between four and ten *cas* genes that are found in association with CRISPR loci harbouring specific repeat sequences (Kunin et al., 2007).

These distinct immune systems have been divided into three major CRISPR types on the basis of gene conservation and locus organization (Makarova et al., 2011b). More than one CRISPR type is often found in a single organism, indicating that these systems are probably mutually compatible and could share functional components (Makarova et al., 2011b). Despite the variation in number and diversity of *cas* genes, the distinguishing feature of all type I systems is that they encode a *cas3* gene. The Cas3 protein contains an N-terminal HD phosphohydrolase domain and a C-terminal helicase domain (Haft et al., 2005; Jansen et al., 2002; Makarova et al., 2006; Sinkunas et al., 2011). In some type I systems, the Cas3 nuclease and helicase domains are encoded by separate genes (*cas3″* and *cas3′*, respectively), but in each case they are thought to participate in degrading foreign nucleic acids (Brouns et al., 2008; Han and Krauss, 2009; Mulepati and Bailey, 2011; Sinkunas et al., 2011) (**Fig. 1.2**).

Type II CRISPR systems consist of just four *cas* genes, one of which is always *cas9* (formerly referred to as *csn1*). Cas9 is a large protein that includes both a RuvC-like nuclease domain and an HNH nuclease domain. Studies in *Streptococcus pyogenes* and *Streptococcus thermophilus* have indicated that Cas9 may participate in both CRISPR RNA processing and target destruction (Barrangou et al., 2007; Deltcheva et al., 2011; Garneau et al., 2010). Two variations of the type III system have been identified (known as III-A and III-B). This division is supported by the functional differences reported in *Staphylococcus epidermidis* and *Pyrococcus furiosus* (Hale et al., 2009; Marraffini and Sontheimer, 2008). The immune system in *S. epidermidis* (type III-A) targets plasmid DNA *in vivo*, whereas the purified components of the type III-B system in *P. furiosus* have been found to cleave only single-stranded RNA substrates *in vitro*. The functional distinction between these two closely related systems suggests there could be other mechanistic differences between the distinct CRISPR subtypes.

1.4 Integration of new information into CRISPR loci

Acquisition of foreign DNA is the first step of CRISPR-mediated immunity (**Fig. 1.2** & **1.3**). During this stage, a short segment of DNA from an invading virus or plasmid (known as the protospacer) is integrated preferentially at the leader end of the CRISPR locus (Barrangou et al., 2007; Garneau et al., 2010). Although metagenomic studies performed on environmental samples indicate that CRISPRs evolve rapidly in dynamic equilibrium with resident phage populations (Andersson and Banfield, 2008; Snyder et al., 2010; Tyson and Banfield, 2008), the type II system in *S. thermophilus* is currently the only CRISPR system that has been shown to robustly acquire new phage or plasmid sequences in a pure culture. Phage-challenge experiments in *S. thermophilus* have indicated that a small proportion of the cells in a population will typically incorporate a single virus-derived sequence at the leader end of a CRISPR locus (Barrangou et al., 2007; Deveau et al., 2008; Garneau et al., 2010; Horvath et al., 2008). The CRISPR-repeat sequence is duplicated for each new spacer sequence added, thus maintaining the repeat–spacer–repeat architecture. Although the mechanism of spacer integration and replication of the repeat sequence is still unknown, studies in *S. thermophilus* and *E. coli* have indicated that several Cas proteins are involved in the process (Barrangou et al., 2007; Brouns et al., 2008; Garneau et al., 2010; Sapranauskas et al., 2011). Mutational analysis of the *cas* genes in *S. thermophilus* demonstrated that *csn2* (previously known as *cas7*) is required for new spacer sequence acquisition (Barrangou et al., 2007). This gene is not conserved in other CRISPR types, which suggests that either the mechanism of adaptation in *S. thermophilus* is distinct from the other types or that there are functional orthologs of Csn2 in other systems. Furthermore, gene deletion experiments in both *S. thermophilus* and *E. coli* have shown that neither *cas1* nor *cas2* genes are required for CRISPR RNA processing or targeted interference (Babu et al., 2011; Brouns et al., 2008; Sapranauskas et al., 2011). These genetic studies suggest a role for Cas1 and Cas2 in the integration of foreign DNA into the CRISPR.

The role of Cas1 in CRISPR-mediated immunity is still uncertain; however, biochemical and structural data indicate a function for Cas1 in new–spacer–sequence acquisition (Babu et al., 2011; Han et al., 2009; Wiedenheft et al., 2009). Cas1 proteins from *Pseudomonas aeruginosa* (Wiedenheft et al., 2009), *E. coli* (Babu et al., 2011) and *Sulfolobus solfataricus* (Han et al., 2009) have been purified and studied biochemically. The Cas1 protein from *S. solfataricus* has been shown to bind nucleic acids with high affinity (K_d ranging from 20 to 50 nM), but without sequence preference (Han et al., 2009). The Cas1 protein from *E. coli* also binds to DNA with a preference for mismatched or abasic substrates (Chen et al., 2008). This observation is consistent with a recent study showing a physical and genetic interaction between *E. coli* Cas1 and several proteins associated with DNA replication and repair (Babu et al., 2011).

Activity assays with Cas1 from *P. aeruginosa* and *E. coli* indicate that Cas1 is a metal-dependent nuclease. The Cas1 protein from *P. aeruginosa* is a DNA-specific nuclease, whereas the Cas1 protein from *E. coli* had a nuclease activity on a wider range of nucleic acid substrates (Babu et al., 2011; Wiedenheft et al., 2009). These *in vitro* assays suggest that Cas1 proteins interact with nucleic acids in a non-sequence-specific manner.

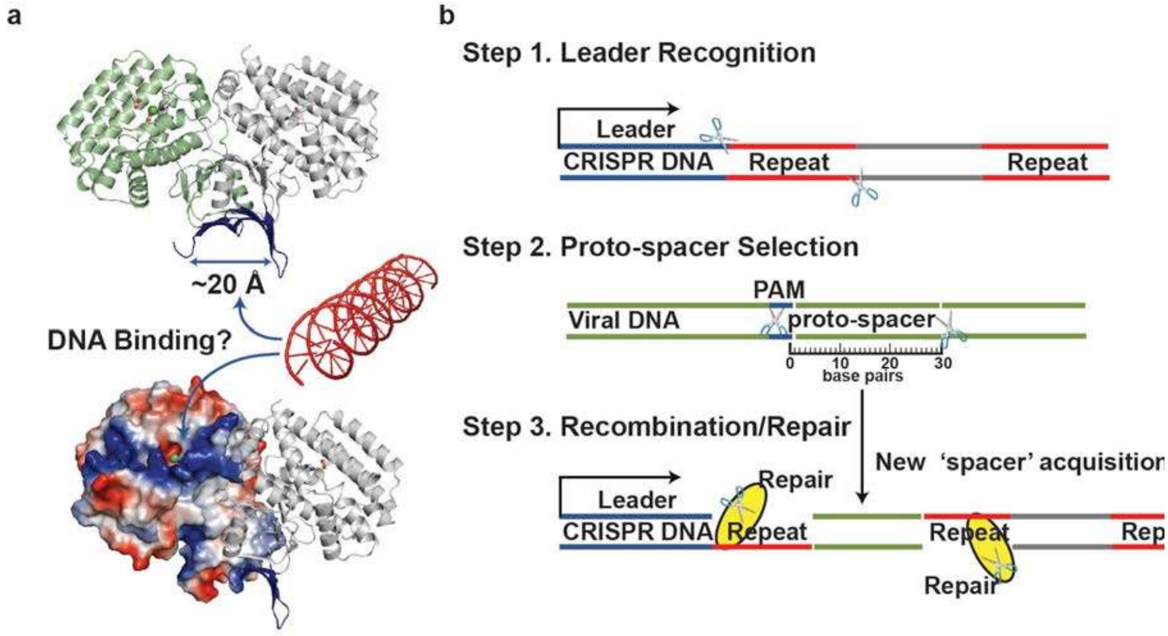


Figure 1.3 | Steps leading to new spacer integration. (a) The Cas1 protein forms a stable homodimer where the two molecules (green and grey) are related by a pseudo-two-fold axis of symmetry (PBD ID: 3GOD). This organization creates a saddle-like structure in the N-terminal domain, in which β-hairpins (blue) from each symmetrically related molecule hang (like stirrups) that are separated by approximately 20 Å, and may interact with the phosphodiester backbone of double-stranded DNA. An electrostatic surface representation (bottom) reveals a cluster of basic residues (blue) that form a positively charged strip across the metal-binding surface of the C-terminal domain. This strip may serve as an electrostatic trap that positions DNA substrates proximally to catalytic metal ions (green sphere). **(b)** CRISPR adaptation occurs by integrating fragments of foreign nucleic acid preferentially at the leader end of the CRISPR, forming new repeat-spacer units in the process. Protospacers are chosen non-randomly and may be selected from regions flanking the protospacer adjacent motif (PAM). Coordinated cleavage of the foreign DNA and integration of the protospacer into the leader-end of the CRISPR occurs through a mechanism that duplicates the repeat sequence and thus preserves the repeat-spacer-repeat architecture of the CRISPR locus. Although the protein components required for this process have not been conclusively identified, Cas1 and other general recombination or repair factors have been implicated (blue ovals).

Crystal structures for five different Cas1 proteins are currently available (Protein Data Bank (PDB) identifiers: 3GOD, 3NKD, 3LFX, 3PV9 and 2YZS) (Babu et al., 2011; Wiedenheft et al., 2009). Although the amino acid sequences for these proteins are extremely diverse (less than 15% sequence identity), their tertiary and quaternary structures are similar. All Cas1 proteins seem to share a two-domain architecture consisting of an N-terminal β-strand domain and a C-terminal α-helical domain (Fig. 3). The C-terminal domain contains a conserved divalent metal-ion binding site, and alanine substitutions of the metal-coordinating residues inhibit Cas1-catalysed DNA degradation (Babu et al., 2011; Wiedenheft et al., 2009). The metal ion is surrounded by a cluster of basic residues that form a strip of positive charge across the surface of the C-terminal domain. This positively charged surface may serve as an electrostatic snare to position nucleic-acid substrates near the catalytic metal ions (Wiedenheft et al., 2009) (**Fig. 1.3**). The Cas1 protein forms a stable homodimer that is formed through interactions between the two β-strand domains, which are related by a pseudo-two-fold axis of symmetry (Babu et al., 2011;

Wiedenheft et al., 2009). This organization creates a saddle-like structure that can be modeled onto double-stranded DNA without steric clashing. β-hairpins, one from each of the two symmetrically related molecules, hang on opposite faces of the double-stranded DNA (like stirrups on a saddle). Although this feature of the Cas1 structure did not initially stand out as a potential DNA-binding site, comparative analysis of the available Cas1 structures reveals a conserved set of positively charged residues along each of the β-hairpins that could contact the phosphate backbone. The two β-hairpins, which are symmetrically related, might participate in sequence-specific interactions with the CRISPR repeat, whereas the large positively charged surface on the C-terminal α-helical domain could account for the high-affinity, non-sequence-specific interactions that have been observed *in vitro*.

In spite of these structural studies and biochemical results, it is still only possible to speculate on the role of Cas1 in the integration of new spacer sequences, and many steps associated with the integration process still need to be explained. For example, new spacer sequences are inserted preferentially at the leader end of the CRISPR, but the mechanism of leader end recognition is unknown. One simple model suggests that the leader sequence contains a recognition element that recruits the integration machinery. It is equally possible that integration relies on single- stranded regions of the CRISPR DNA that are made available during transcription. Transcription-associated recombination is involved in genome stability (Aguilera, 2002), and a mechanism that couples integration together with transcription would link the process of adaptation to CRISPR RNA expression, ensuring that spacers from the most recent virus or plasmid are transcribed first.

The integration machinery must be able to distinguish foreign DNA from that of the host genome. The molecular cues that are involved in the distinction of 'self ' from 'non-self ' are still unknown, but sequencing of CRISPR loci following phage challenge suggests that spacer sequences are not selected at random (Bolotin et al., 2005; Deveau et al., 2008; Garneau et al., 2010; Horvath et al., 2008; Mojica et al., 2009; Semenova et al., 2011). Mapping spacer sequences onto viral genomes reveals a short sequence motif proximal to the protospacer, which is referred to as the protospacer adjacent motif (PAM). PAM sequences are only a few nucleotides long, and the precise sequence varies depending on the CRISPR system type (Mojica et al., 2009). This variation suggests that one or more of the Cas proteins associated with each immune system is involved in PAM recognition, but the mechanism governing this specificity is unknown.

1.5 CRISPR RNA biogenesis

Spacer acquisition is the first step of immunization, but successful protection from bacteriophage or plasmid challenge requires the CRISPR to be transcribed and processed into short CRISPR-derived RNAs (crRNAs). crRNAs were first detected by small RNA profiling in *Archaeoglobus fulgidus* (Tang et al., 2002) and *S. solfataricus* (Tang et al., 2005). Northern-blot analysis using probes against the repeat sequence of the CRISPR revealed a 'ladder-like' pattern of RNA consistent with a long precursor CRISPR RNA transcript (pre-crRNA) that was processed at approximately 60-nucleotide intervals. In fact, the 3′ ends of cloned crRNAs were mapped to the middle of the CRISPR repeat (Tang et al., 2002), which suggested that the repeat sequence was recognized and cleaved.

The need for crRNAs in CRISPR-mediated defense was demonstrated initially by investigation of a CRISPR-specific endoribonuclease in *E. coli* called Cas6e (formerly known as Cse3 or CasE) (Brouns et al., 2008). Cas6e specifically binds and cleaves within each repeat sequence of the long pre-crRNA, resulting in a library of crRNAs that each contain a unique spacer sequence flanked by fragments of the adjacent repeats. Mutation of a conserved histidine blocks crRNA biogenesis and leaves the cell susceptible to phage infection (Brouns et al., 2008).

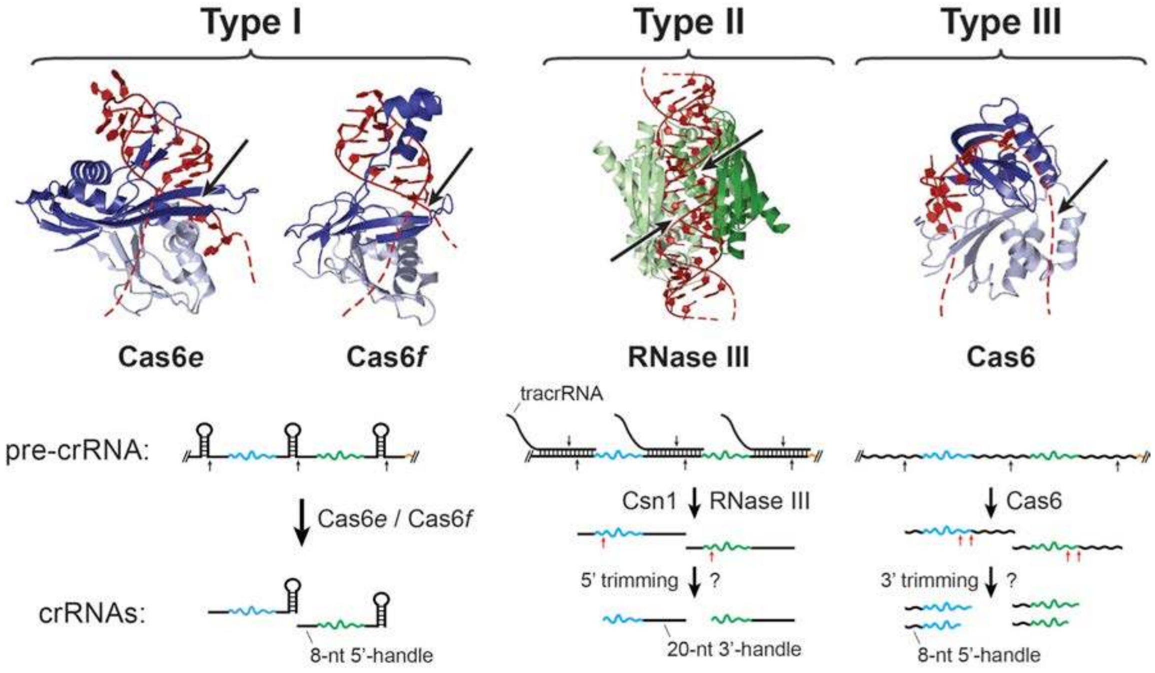


Figure 1.4 | Diverse mechanisms of CRISPR RNA biogenesis. CRISPR RNA repeats are specifically recognized and cleaved by diverse mechanisms. In type I CRISPR systems, Cas6e (PDB ID: 2Y8W) and Cas6f (PDB ID: 2XLK) recognize the major groove of the crRNA stem-loop primarily through electrostatic interactions using a β-hairpin and α-helix, respectively. Cleavage occurs at the double-stranded–single-stranded junction (black arrows), leaving an 8-nt 5′-handle on mature crRNAs. In type II CRISPR systems, tracrRNA hybridizes to the pre-crRNA repeat to form duplex RNAs that are substrates for endonucleolytic cleavage by host RNase III (PDB ID: 2EZ6), an activity that may also require Cas9. Subsequent trimming (red arrows) by an unidentified nuclease removes leftover repeat sequences from the 5′ end. Cas6 (PDB ID: 3PKM) in type III-B CRISPR systems specifically recognizes single-stranded RNA, upstream of the scissile phosphate, on a face of the protein opposite that of the previously identified active site residues. The remainder of the repeat substrate probably wraps around the protein (red dashed line) to allow cleavage 8 nucleotides upstream of the repeat-spacer junction. Subsequent 3′ trimming (red arrows) generates mature crRNAs of two discrete lengths. The N-terminal domain of all Cas 6 family proteins adopts a ferredoxin-like fold (light blue). The C-terminal domain of Cas6 and Cas6e also adopts a ferredoxin-like fold but the C-terminal domain of Cas6f is structurally distinct (dark blue).

The Cas6e protein consists of a double ferredoxin-like fold that selectively associates with specific RNA repeats and does not associate with DNA or CRISPR RNAs containing a non-cognate repeat sequence (Brouns et al., 2008; Ebihara et al., 2006; Gesner et al., 2011; Sashital et al., 2011) (**Fig. 1.4**). Crystal structures of Cas6e bound to a CRISPR RNA repeat reveal a combination of sequence- and structure-specific interactions that explain the molecular mechanism of substrate recognition (Gesner et al., 2011; Sashital et al., 2011). The repeat sequence of the *E. coli* CRISPR is partially palindromic, and the RNA forms a stable (approximately 20-nucleotide) stem loop (Brouns et al., 2008; Kunin et al., 2007). A positively charged β-hairpin in Cas6e interacts with the major groove of the RNA duplex, which positions the 3′ strand of the crRNA stem along a conserved, positively charged cleft on one face of the protein (Gesner et al., 2011; Sashital et al., 2011) (Fig. 4). RNA binding induces a conformational change that disrupts the bottom base pair of the stem and positions the scissile phosphate within the enzyme active site for site-specific cleavage (Sashital et al., 2011). CRISPR RNA cleavage occurs 8 nucleotides upstream of the spacer sequence, which results in 61-nucleotide mature crRNAs consisting of a 32-nucleotide spacer flanked by 8 nucleotides of the repeat sequence on the 5′ end (known as the 5′-handle) and 21 nucleotides of the remaining repeat sequence on the 3′ end (Fig. 4). Cas6e remains tightly bound to the 3′ stem-loop (Sashital et al., 2011) and may serve as a nucleation point for assembly of a large effector complex, Cascade (CRISPR-associated complex for antiviral defense), that is required for phage silencing in the next stage of the immune system (Brouns et al., 2008; Jore et al., 2011a; Wiedenheft et al., 2011a) (discussed later).

Crystal structures of CRISPR-specific endoribonucleases from two other immune systems offer additional insights into the co-evolutionary relationship between these specialized enzymes and their cognate RNAs (Carte et al., 2008a; Haurwitz et al., 2010; Wang et al., 2011) (**Fig. 1.4**). In *P. aeruginosa*, Cas6f (previously known as Csy4) specifically binds and cleaves the CRISPR-RNA-repeat 8 nucleotides upstream of the spacer sequence, which leaves a similar 8-nucleotide 5′-handle on mature crRNAs (Haurwitz et al., 2010). The co-crystal structure of Cas6f bound to its cognate RNA reveals interesting parallels between the method of RNA binding used by Cas6f and Cas6e (Gesner et al., 2011; Haurwitz et al., 2010; Sashital et al., 2011). Like Cas6e, the *P. aeruginosa* Cas6f protein recognizes the sequence and shape of a stable stem-loop in the crRNA repeat sequence by interacting extensively with the major groove of the double-stranded RNA. However, the structural elements responsible for this interaction are distinct between the two proteins (Gesner et al., 2011; Haurwitz et al., 2010; Sashital et al., 2011) (**Fig. 1.4**). The Cas6f protein has a two-domain architecture, which consists of an N-terminal ferredoxin-like fold similar to that in Cas6e, but its C-terminal domain is structurally distinct. An arginine-rich helix in the C-terminal domain of Cas6f inserts into the major groove of the crRNA duplex, and the bottom of the crRNA is positioned for sequence-specific hydrogen-bonding contacts in the RNA major groove. These contacts position the scissile phosphate of the crRNA in the enzyme active site so that cleavage occurs 8 nucleotides upstream of the spacer sequence (Gesner et al., 2011; Haurwitz et al., 2010; Sashital et al., 2011) (**Fig. 1.4**).

Although Cas6f and Cas6e recognize the sequence and shape of the crRNA hairpin in their respective systems, CRISPR RNA repeats in other CRISPR systems are thought to be unstructured (Kunin et al., 2007). For example, the Cas6 protein from *P. furiosus* associates with CRISPR transcripts that are expected to contain unstructured repeats (Carte et al., 2010). The specific recognition of an unstructured RNA repeat requires a distinct mechanistic solution for RNA substrate discrimination. Remarkably, crystallographic studies of the Cas6 protein from *P. furiosus* have revealed the same duplicated ferredoxin-like fold observed in the Cas6e protein, but with a different mode of RNA recognition involving the opposite face of the protein (**Fig. 1.4**). In Cas6, the two ferredoxin-like folds clamp the 5′ end of the single-stranded RNA repeat sequence in place (Wang et al., 2011). Although the RNA in this structure is disordered in the enzyme active site, biochemical studies have shown that cleavage occurs 8 nucleotides upstream of the spacer sequence (Carte et al., 2008a; 2010). While the nucleotide sequences at the cleavage site vary for each of the different Cas6 proteins, all Cas6 family endoribonucleases cleave their cognate RNA 8 nucleotides upstream of the spacer sequence using a metal-ion-independent mechanism.

Despite advances in our understanding of crRNA biogenesis, the diversity of *cas* genes has obscured identification of the protein factors responsible for CRISPR RNA processing in some systems. Type II immune systems consist of four *cas* genes, none of which have a detectable sequence similarity to known CRISPR-specific endoribonucleases. Recently, a different CRISPR RNA processing mechanism has been reported that involves RNase-III-mediated cleavage of double-stranded regions of the CRISPR RNA repeats (Deltcheva et al., 2011). The first indication of this mechanism came from deep sequencing of RNA from *S. pyogenes*. An abundant transcript containing a 25-nucleotide sequence that was complementary to the CRISPR repeat was identified. This RNA, termed tracrRNA (*trans*-activating CRISPR RNA), is coded on the opposite strand and just upstream of the CRISPR locus. Genetic and biochemical experiments demonstrated that tracrRNA and pre-crRNA are co-processed by RNase III, which produces cleavage products with a 2 nucleotide 3′ overhang (Deltcheva et al., 2011). *In vivo* processing of CRISPR RNAs required Cas9 (previously known as Csn1), although a precise role for this enzyme in RNA processing has not yet been defined. The essential role of cellular proteins that are not solely involved in CRISPR-mediated defense, such as RNase III, indicates that different host factors may be involved as ancillary components of these immune systems.

1.6 CRISPR RNA-guided interference

The third stage of CRISPR-mediated immunity is target interference (**Fig. 1.2**). Here crRNAs associate with Cas proteins to form large CRISPR-associated ribonucleoprotein complexes that can recognize invading nucleic acids. Foreign nucleic acids are identified by base-pairing interactions between the crRNA spacer sequence and a complementary sequence from the intruder. Phage- and plasmid-challenge experiments performed in several model systems have demonstrated that crRNAs complementary to either the coding or the non-coding strand of the invading DNA can provide immunity (Barrangou et al., 2007; Brouns et al., 2008; Gudbergsdottir et al., 2011; Manica et al., 2011; Marraffini and Sontheimer, 2008; Semenova et al., 2011). This is indicative of an RNA-guided DNA-targeting system, and indeed a pathway for DNA silencing has recently been demonstrated in *S. thermophilus* (Garneau et al., 2010). DNA sequencing and Southern blots indicated that both strands of the target DNA are cleaved within the region that is complementary to the crRNA spacer sequence (Garneau et al., 2010). This mechanism efficiently eliminates foreign DNA sequences, which have been specified by the spacer region of the crRNA, but avoids targeting the complementary DNA sequences in the CRISPR region of the host chromosome. The mechanism for distinguishing self from non-self is built into the crRNA. The spacer sequence of each crRNA is flanked by a portion of the adjacent CRISPR repeat sequence, and any complementarity beyond the spacer into the adjacent repeat region signals self and prevents the destruction of the host chromosome (Marraffini and Sontheimer, 2010b).

However, not all CRISPR systems target DNA. *In vitro* experiments using enzymes from the type III-B CRISPR system of *P. furiosus* have shown that this system cleaves target RNA rather than DNA (Hale et al., 2009). All DNA targeting systems encode a complementary DNA sequence for each crRNA in the CRISPR locus and therefore require a mechanism for distinguishing self (CRISPR locus) from non-self (invading DNA). In contrast, systems that target RNA may not be required to make this distinction because most CRISPR loci are transcribed only in one direction and thus do not generate complementary RNA targets. CRISPR systems that target RNA may be uniquely capable of defending against viruses that have RNA-based genomes. However, adaptation of the CRISPR in response to a challenge by an RNA-based virus will probably require the invading RNA to be reverse-transcribed into DNA before it can be integrated into the CRISPR locus.

Cas proteins directly participate in target binding. Recent biochemical studies have shown that CRISPR-associated complexes facilitate target recognition by enhancing sequence-specific hybridization between the CRISPR RNA and complementary target sequences (Wiedenheft et al., 2011b). A short high-affinity binding site located at one end of the crRNA spacer sequence governs the efficiency of target binding, and viruses that acquired a single mismatch in this region were able to escape detection by the immune system (Semenova et al., 2011). This high-affinity binding site is functionally analogous to the 'seed' sequence (**Fig. 1.1**) that has been identified in eukaryotic microRNAs (miRNAs) (Bartel, 2004). Structural and biochemical studies have shown that Argonaute proteins facilitate target recognition by pre-ordering the nucleotides at the 5′ end of the miRNA in a helical configuration (Parker et al., 2009). This pre-ordering reduces the entropic penalty that is associated with helix formation and provides a thermodynamic advantage for target binding within this region. A similar mechanism may occur during crRNA target binding, providing an interesting example of convergent evolution between CRISPR-based immunity in prokaryotes and RNAi in eukaryotes (**Fig. 1.1**).

Structural and biochemical studies have been performed on CRISPR-associated complexes isolated from three different type I CRISPR systems (Hale et al., 2009; Jore et al., 2011a; Lintner et al., 2011; Wiedenheft et al., 2011a; 2011b). These complexes seem to share some general morphological features, but the precise special arrangement of the Cas proteins and their interactions with the crRNA have been unclear. Sub-nanometer-resolution structures of the CRISPR-associated complex from *E. coli* (Cascade) have recently been determined using cryo-electron microscopy (Wiedenheft et al., 2011a). This complex is comprised of an unequal stoichiometry of 5 functionally essential Cas proteins and a 61-nucleotide crRNA (Brouns et al., 2008; Jore et al., 2011a; Wiedenheft et al., 2011a). The structure reveals a sea-horse- shaped architecture in which the crRNA is displayed along a helical arrangement of protein subunits that protect the crRNA from degradation (Wiedenheft et al., 2011a). The 5′ and 3′ ends of the crRNA form unique structures that are anchored at opposite ends of the Cascade complex, displaying the 32-nucleotide spacer sequence for base-pairing with complementary targets.

The structure of Cascade bound to a 32-nucleotide target sequence (Wiedenheft et al., 2011a) reveals a concerted conformational change that could be a signal for recruiting Cas3. Cas3 — the *trans*-acting nuclease of type I CRISPR systems — may function as a target 'slicer' in a similar way to Argonaute in RNAi pathways (Beloglazova et al., 2011; Brouns et al., 2008; Mulepati and Bailey, 2011; Sinkunas et al., 2011). Although Cas3 was implicated previously in the process of self versus non-self discrimination, recent studies have demonstrated that Cascade recognizes the PAM directly and that mutations in the PAM decrease Cascade's affinity for the target (Semenova et al., 2011). The importance of the PAM is highlighted by the recovery of phage and plasmid escape mutants, which frequently contain a single mutation in the PAM (Deveau et al., 2008; Garneau et al., 2010; Horvath et al., 2008; Sapranauskas et al., 2011; Semenova et al., 2011). The structure of Cascade indicates that the PAM is positioned near the 'tail' of the sea-horse-shaped complex. High-resolution structures and mutational analysis of the nucleic acid and protein components in this and related systems are needed to determine the mechanisms of target authentication and degradation.

1.7 Applications of CRISPR structure and function

The sequence diversity of CRISPR loci, even within closely related strains, has been used for high-resolution genotyping and forensic medicine. This technique, known as spoligotyping (spacer oligotyping), has been applied successfully to the analysis of human pathogens, including *Mycobacterium tuberculosis* (Groenen et al., 1993), *Corynebacterium diphtheria* (Mokrousov et al., 2007) and *Salmonella enterica* (Liu et al., 2011). Spoligotyping was developed long before the function of CRISPRs was understood, but now that studies have begun to reveal the biological function and mechanism of CRISPR-mediated genetic silencing, new opportunities for creative applications have emerged. Laboratory strains of bacteria are grown in high-density bioreactors for many different applications in the food industry, and they are becoming increasingly important in the production of biofuels. CRISPR systems offer a natural mechanism for adapting economically important bacteria for resistance against multiple phages.

The biochemical activities of various Cas proteins may have useful applications in molecular biology in much the same way that DNA restriction enzymes have revolutionized cloning and DNA manipulation. A wide range of CRISPR-specific endoribonucleases that recognize small RNA motifs with high affinity expand the number of tools available for manipulating nucleic acids. In addition, a crRNA-guided ribonucleoprotein complex in *P. furiosus* was shown to cleave target RNAs (Hale et al., 2009). Site-specific cleavage of target RNA molecules could have a range of uses, from generating homogeneous termini after *in vitro* transcription to targeting a specific intracellular messenger RNA for inactivation in a similar way to RNAi. CRISPRs also provide a new mechanism for limiting the spread of antibiotic resistance or the transfer of virulence factors by blocking horizontal gene transfer (Garneau et al., 2010; Marraffini and Sontheimer, 2008). In addition, CRISPRs participate in a regulatory mechanism that alters biofilm formation in *P. aeruginosa* (Cady and O'Toole, 2011; Zegans et al., 2009). Although the clinical relevance of CRISPRs remains to be demonstrated, the opportunities for creative implementation of this new gene-regulation system are perceivably vast.

1.8 Future directions of CRISPR biology

The discovery of some of the fundamental mechanisms of CRISPR- based adaptive immunity has raised new questions and highlighted the areas with the greatest potential for future research. Although CRISPR RNA processing and targeting steps are now understood in some detail, how and when target sequences are identified during a phage infection or plasmid transformation are still unclear. Furthermore, why DNA or RNA target sequences are chosen, and their fate once they are bound to a crRNA-targeting complex is not well understood. In addition, the mechanisms by which foreign sequences are selected and integrated into CRISPR loci are almost entirely unknown. Some CRISPR loci seem to be considerably more active than others, at least under laboratory conditions, so selection of the model organisms will be important. The diversity and prevalence of CRISPR systems throughout microbial communities ensures that new findings and applications in this field will be forthcoming in the years ahead.

Chapter 2

Mechanism of substrate selection by a highly specific CRISPR endoribonuclease

† Part of the work presented in this chapter has previously been published in the following research article: Sternberg, S.H., Haurwitz, R.E., Doudna, J.A. (2012). Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. RNA *18*, 661–672.

‡ Samuel Sternberg designed experiments and performed all of the Csy4 binding and cleavage assays. Rachel Haurwitz performed the Northern blot analysis and helped design experiments. Jennifer Doudna supervised the project.

2.1 Abstract

Bacteria and archaea possess adaptive immune systems that rely on small RNAs for defense against invasive genetic elements. CRISPR (clustered regularly interspaced short palindromic repeats) genomic loci are transcribed as long precursor RNAs which must be enzymatically cleaved to generate mature CRISPR-derived RNAs (crRNAs) that serve as guides for foreign nucleic acid targeting and degradation. This processing occurs within the repetitive sequence and is catalyzed by a dedicated Cas6 family member in many CRISPR systems. In *Pseudomonas aeruginosa,* crRNA biogenesis requires the endoribonuclease Csy4 (Cas6f), which binds and cleaves at the 3' side of a stable RNA stem-loop structure encoded by the CRISPR repeat. We show here that Csy4 recognizes its RNA substrate with a ~50 pM equilibrium dissociation constant, making it one of the highest-affinity protein:RNA interactions of this size reported to date. Tight binding is mediated exclusively by interactions upstream of the scissile phosphate that allow Csy4 to remain bound to its product and thereby sequester the crRNA for downstream targeting. Substrate specificity is achieved by RNA major groove contacts that are highly sensitive to helical geometry, as well as a strict preference for guanosine adjacent to the scissile phosphate in the active site. Collectively, our data highlight diverse modes of substrate recognition employed by Csy4 to enable accurate selection of CRISPR transcripts while avoiding spurious, off-target RNA binding and cleavage.

2.2 Introduction

Many bacteria and archaea employ small CRISPR-derived RNAs (crRNAs) as molecular sentries that base pair with phage or plasmids and thereby trigger degradation of these foreign nucleic acids by CRISPR-associated (Cas) proteins (Al-Attar et al., 2011; Horvath and Barrangou, 2010; Marraffini and Sontheimer, 2010a). CRISPR-derived precursor transcripts (pre-crRNAs) are processed enzymatically to generate the mature crRNAs that assemble into large ribonucleoprotein effector complexes (Brouns et al., 2008). In type I and type III CRISPR systems, as defined by Makarova and colleagues (Makarova et al., 2011b), a single endoribonuclease from the Cas6 superfamily cleaves pre-crRNAs within each invariant repeat sequence to generate ~60 nucleotide (nt) products in which segments of the repeat sequence flank the target-binding spacer sequence (Brouns et al., 2008; Carte et al., 2008b; Gesner et al., 2011; Haurwitz et al., 2010; Lintner et al., 2011; Sashital et al., 2011). crRNA biogenesis in type II systems requires RNase III, which cleaves double-stranded RNA (dsRNA) substrates formed by base pairing between a small, non-coding RNA (tracrRNA) and the pre-crRNA (Deltcheva et al., 2011). Pre-crRNA processing is a hallmark of the CRISPR-Cas system, and the inactivation of these endoribonucleases results in a complete loss of immune system function (Brouns et al., 2008; Deltcheva et al., 2011; Sapranauskas et al., 2011).

We showed previously that Csy4, recently reclassified as Cas6f (Makarova et al., 2011b), generates crRNAs in type I-F CRISPR systems (formerly the *Yersinia pestis* subtype) by cleaving pre-crRNAs at the bottom of stable stem-loops encoded by the CRISPR repeat (Haurwitz et al., 2010) (**Fig. 2.1a**). The co-crystal structure of Csy4 bound to its pre-crRNA substrate (PDB ID: 2XLK) revealed a diverse set of molecular interactions that mediate RNA recognition (**Fig. 2.1b**). A highly basic α-helix docks into the major groove of the stem-loop and contains multiple arginine residues that form a network of hydrogen bonds with the RNA phosphate backbone along the 5' strand of the stem. In a manner reminiscent of DNA-binding proteins, Csy4 interacts with the bottom two base pairs of the stem-loop through a direct readout mechanism involving formation of base-specific hydrogen bonds between the major groove faces of A19 and G20 and residues Gln104 and Arg102, respectively. The aromatic side chain of Phe155 stacks below the terminal base pair, thereby positioning the scissile phosphate within the active site. Together, these interactions enable Csy4 to recognize and cleave a single repetitive RNA sequence inside the cell, ensuring correct crRNA processing without off-target effects.

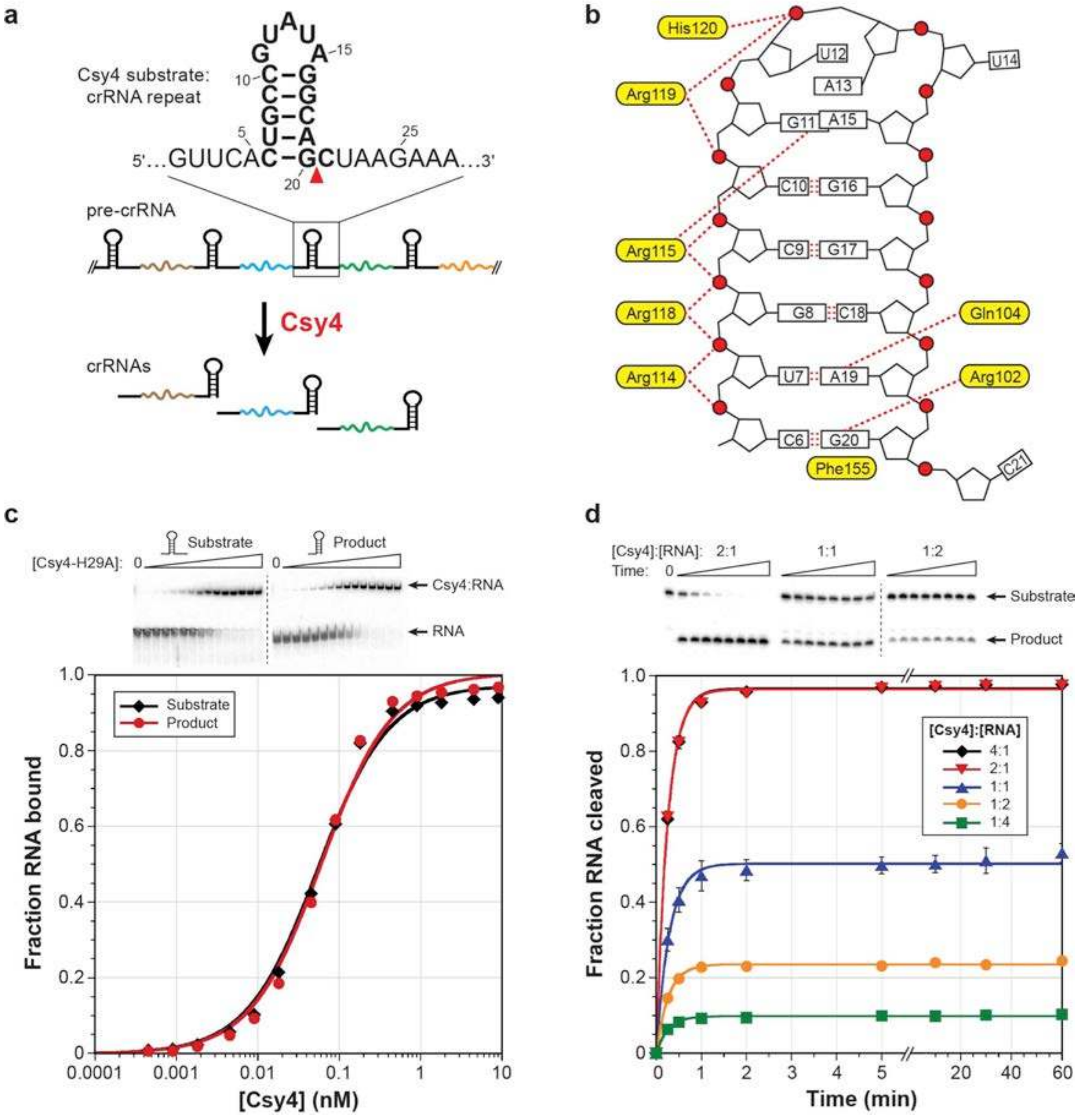


Figure 2.1 | Csy4 binds its substrate and product with high affinity and functions as a single-turnover enzyme. **(a)** Csy4 cleaves within pre-crRNA repeat sequences (black) to generate mature crRNAs that contain a spacer sequence (colored line) flanked by fragments of the repeat. The substrate sequence and cleavage site (red triangle) are indicated above, with the crRNA construct previously used for crystallography shown in bold. **(b)** A schematic depicts protein:RNA contacts revealed by a co-crystal structure of Csy4 bound to a fragment of the crRNA repeat (PDB ID: 2XLK). Important amino acid residues are shown in yellow, and RNA nucleotides are numbered as in (a). Red circles, pentagons, boxes, and red dotted lines denote phosphates, ribose groups, bases, and hydrogen-bonding interactions, respectively. **(c)** EMSAs (top) were performed with Csy4-H29A and the substrate

and product of the cleavage reaction. The resulting data for these and all subsequent binding assays were fit with a standard binding isotherm to yield equilibrium dissociation constants (solid lines; see Materials and Methods), and average K_d and standard error of the mean (SEM) values from at least three independent experiments are reported in **Table 2.1. (d)** RNA cleavage assays were conducted at five different enzyme:substrate molar ratios, and the extent of the reaction at various time points was assessed by denaturing PAGE (top). The resulting data for these and all subsequent cleavage assays were fit with a single exponential to yield first-order rate constants (solid lines; see Materials and Methods), and average k_{obs} and SEM values from three independent experiments are reported in **Table 2.1**. Error bars for each time point represent the standard deviation and are not always visible.

Bioinformatic analyses of Csy4-related Cas proteins together with existing CRISPR databases (Grissa et al., 2007b) have revealed a potentially large number of enzyme variants whose substrate specificities have co-evolved with the RNAs encoded by CRISPR repeats. Gaining a thorough understanding of the selection mechanism by which *Pseudomonas aeruginosa* Csy4 faithfully binds and cleaves its substrate should inform future work aimed at expanding the toolbox of these sequence-specific endoribonucleases. Furthermore, the propensity of many pre-crRNA repeat sequences to form small, stable stem-loops (Kunin et al., 2007) suggests that general principles of substrate recognition employed by Csy4 will be broadly applicable to other Cas6 family members that associate with structured repeats.

To determine the importance of sequence- and shape-specific RNA recognition during pre-crRNA processing, we investigated the relative contributions of substrate base-pair composition and geometry to binding and cleavage by Csy4. Here we show that Csy4 binds its substrate RNA with extremely high affinity ($K_d \approx 50$ pM) and functions as a single-turnover enzyme. Single-stranded RNA (ssRNA) nucleotides that flank the stem-loop contribute negligibly to binding energy, but base-pair changes throughout the double-stranded stem and mutations to the loop sequence result in substantially weaker binding. We find that substrate recognition also involves the precise length of the stem, such that small base-pair insertions cause severe binding and/or cleavage defects due to their effects on helical geometry and substrate positioning. These findings reveal how Csy4 employs a unique set of molecular interactions to achieve highly specific selection of its pre-crRNA substrate while discriminating against similar, non-cognate stem-loop structures.

2.3 Materials and Methods

2.3.1 Protein expression and purification

R102A, Q104A, F155A and H29A Csy4 mutants were purified as described (Haurwitz et al., 2010). R114A/R118A, R118A/R115A and R115A/R119A Csy4 double mutants were generated using site-directed mutagenesis and purified essentially as described previously (Haurwitz et al., 2010), with the following exceptions. Protein genes encoded by the pHGWA vector (Busso et al., 2005) were overexpressed in BL21(DE3) cells. Following the second Ni-NTA affinity purification step, Csy4 mutants were purified by size exclusion chromatography using a single Superdex 75 (16/60) column (GE Healthcare) in 100 mM HEPES (pH_{RT} 7.5), 500 mM KCl, 5% glycerol, 1mM TCEP. Proteins were then concentrated and buffer exchanged into 100 mM HEPES (pH_{RT} 7.5), 150 mM KCl, 5% glycerol, 1mM TCEP, snap frozen in liquid nitrogen and stored at -80 °C.

2.3.2 Northern blot analysis

Total RNA was extracted from cultures of *P. aeruginosa* PAO1, *P. aeruginosa* UCBPP-PA14 and a *csy4* deletion strain of *P. aeruginosa* UCBPP-PA14 (SMC3894) (Zegans et al., 2009) grown to exponential phase using the mirVana kit (Ambion). Duplicate samples of each RNA preparation (6 µg) were separated on adjacent lanes of a 15% denaturing polyacrylamide gel and subsequently transferred to a nylon membrane (Hybond-N+, GE Healthcare) using a semi-dry transfer cell (BioRad). The single membrane was then cut in half to yield two membranes with identical samples. The membranes were pre-treated with ULTRAHyb-Oligo Hybridization Buffer (Ambion) and probed with 5'-[^{32}P]-radiolabeled DNA oligonucleotides corresponding to either the crRNA repeat sequence (5'-GTTCACTGCCGTATAGGCAGCTAAGAAA-3') or the reverse complement of the crRNA repeat (5'-TTTCTTAGCTGCCTATACGGCAGTGAAC-3'). Membranes were washed twice with 2X saline-sodium citrate (SSC) buffer containing 0.5% SDS and visualized by phosphorimaging.

2.3.3 RNA transcription, purification and 5' radiolabeling

The following RNAs were synthesized by Integrated DNA Technologies: the non-cleavable substrate, product RNA (Δ21-28), 5' truncation constructs (Δ1-5, Δ1-4), the 5'-strand (nucleotides 1-12) and 3'-strand (nucleotides 13-28) used to generated the nicked substrate, the G20A mismatched substrate and three substrates containing base-pair substitutions at the bottom of the stem (C6U/G20A, C6G/G20C, U7A/A19U). All other RNAs were transcribed *in vitro* using T7 polymerase and purified using denaturating polyacrylamide gel electrophoresis, according to the following protocol. Synthetic single-stranded DNA templates (Integrated DNA Technologies) containing the reverse complement of the desired crRNA repeat construct were annealed to a 1.5-fold molar excess of an oligonucleotide corresponding to the T7 promoter sequence (5'-TAATACGACTCACTATA-3'). Templates encoded an extra guanosine at the 5' end of all constructs in order to ensure optimal transcription by T7 polymerase. This had no effect on binding affinities but did lead to a slight (~20%) increase in k_{obs} for cleavage of the WT-crRNA repeat substrate. Transcription reactions (100 µl) were incubated at 37 °C for 3-5 hours and contained 1 µM template DNA, 100 µg/mL T7 polymerase, 1 µg/mL pyrophosphatase (Roche), 5 mM NTPs, 30 mM Tris-Cl (pH_{RT} = 8.1), 25 mM $MgCl_2$, 10 mM dithiothreitol (DTT), 2 mM spermidine and 0.01% Triton X-100. Reactions were then treated with 5 units of DNase (Promega) and incubated for an additional 30 minutes at 37 °C before being loaded on a 15% urea-polyacrylamide gel. RNAs were excised from the gel and eluted into DEPC (diethylpyrocarbonate) H_2O overnight at 4 °C. 5' triphosphates were removed by incubating RNAs at 37 °C for 1 hour with 10 units of calf intestinal phosphate (New England Biolabs) in 1X NEBuffer 3, followed by phenol-chloroform extraction and ethanol precipitation. RNAs were resuspended in DEPC H_2O and stored at -20 °C.

For biochemical experiments, 10 pmol RNA were 5' radiolabeled by incubating with 5 units T4 polynucleotide kinase (New England Biolabs) and ~3-6 pmol (~0.2-0.4 mCi) [γ-^{32}P]-ATP (Promega) in 1X T4 polynucleotide kinase reaction buffer at 37 °C for 30 minutes, in a 25 µl reaction. After heat inactivation (65 °C for 20 minutes), reactions were spun through an illustra MicroSpin G-25 column (GE Healthcare) to remove ATP. Radiolabeled RNAs were diluted to ~100 nM stock concentrations with DEPC H_2O and stored at -20 °C.

2.3.4 Electrophoretic mobility shift assays

Protein concentrations were determined by taking multiple absorbance spectra using a NanoDrop spectrophotometer (Thermo Scientific), averaging A_{280nm} values and converting to molar concentrations using the calculated Csy4 extinction coefficient (15,470 M^{-1} cm^{-1}). Spectra were also recorded under denaturing conditions (6 M guanidine hydrochloride, 20 mM potassium phosphate buffer, pH 6.5), and absorbance values were within error of those taken under native conditions. Binding experiments were conducted in the following buffer: 20 mM HEPES (pH_{RT} 7.5), 100 mM KCl, 5% glycerol, 0.01% Igepal-630, 1 mM DTT and 0.1 mg/mL yeast tRNA (Sigma-Aldrich) to prevent non-specific binding. After diluting concentrated 5'-[^{32}P]-RNA and Csy4 stock solutions into 1X binding buffer, trace amounts of RNA ($\leq$0.05-0.2 nM, depending on construct and specific activity) were incubated with increasing concentrations of Csy4 in a 15 µl reaction at room temperature (~24 °C) for one hour. 12 µl of each reaction were then loaded on a 10% native polyacrylamide gel containing 0.5X TBE buffer and resolved by running at 12 W for 90-120 minutes at 4 °C in 0.5X TBE running buffer. Phosphor screens were exposed to dried gels and scanned with a Storm imager (GE Healthcare), and the intensities of unbound and Csy4-bound RNA were quantified using ImageQuant (GE Healthcare). The fraction of RNA bound at each Csy4 concentration was plotted as a function of Csy4 concentration, and binding data were fit with a standard binding isotherm using Kaleidagraph (Synergy Software), according to the equation: fraction bound = $A \times [\text{Csy4}] \div (K_d + [\text{Csy4}])$, where A is the amplitude of the binding curve.

Binding experiments with the substrate nicked between U12 and A13 contained ~1 nM radiolabeled 3'-strand (nucleotides 13-28) and a 1,000-fold excess (1 µM) of cold 5'-strand (nucleotides 1-12). For experiments with K_d values in the low pM range, binding data were also fit with the solution of a quadratic equation describing a bimolecular dissociation reaction, as described previously (Maag and Lorsch, 2003), out of concern that [RNA] in these experiments was not sufficiently below the K_d to approximate $[\text{Csy4}]_{\text{total}} = [\text{Csy4}]_{\text{free}}$. This analysis returned values that agreed well with equilibrium dissociation constants determined from the standard binding isotherm equation, so these original values are reported. When fitting binding data with the rc-crRNA repeat, the amplitude was set equal to one because saturation could not be reached. Binding data with the RNA substrate containing a five G–C base-pair insertion showed apparent cooperativity and were fit with a modified binding equation using a variable Hill coefficient ($n \approx 1.5$) and an amplitude fixed at one.

At least one binding experiment for each RNA or Csy4 mutant titrated Csy4 across a concentration range of three orders of magnitude centered around the K_d. Additional replicates typically tested five concentration points centered around the K_d and returned values in excellent agreement with those derived from a more complete titration. K_d values presented in the text and in **Tables 2.1 & 2.2** represent the average and standard error of the mean from at least three independent experiments. The average percent error for all reported K_d values is 10%. ΔΔG values for Csy4 or RNA mutants were calculated according to the equation:

$$\Delta\Delta G = -RT\ln(K_{d,\text{WT}}/K_{d,\text{mutant}})$$

where R is the gas constant, T is temperature (set to 298 K) and $K_{d,\mathrm{WT}}/K_{d,\mathrm{mutant}}$ is the ratio of K_d values for the WT and mutant construct.

2.3.5 RNA cleavage assays

Cleavage assays were conducted at room temperature (~24 °C) in the following buffer: 20 mM HEPES, 100 mM KCl, 1 mM DTT, $\mathrm{pH_{RT}}$ 7.5. Single-turnover cleavage experiments were 55 µl in volume and contained 0.5 nM 5'-[^{32}P]-RNA and a saturating concentration of Csy4 (typically 500 nM). At each desired time point, a 10 µl aliquot was removed and quenched by mixing it with 50 µl phenol:chloroform:isoamyl alcohol 25:24:1, pH 8.0 (Sigma-Aldrich). The aqueous layer was mixed with an equal volume of formamide loading dye, heated to ~80 °C for ~2 minutes and separated on a 15% urea-polyacrylamide gel in 0.5X TBE running buffer. RNA was visualized by phosphorimaging, and the intensities of uncleaved and cleaved RNA were quantified using ImageQuant (GE Healthcare). The fraction of RNA cleaved at each time point was plotted as a function of time, and these data were fit with a single exponential decay curve using Kaleidagraph (Synergy Software), according to the equation: fraction cleaved = $A \times (1 - \exp(-k \times \mathrm{t}))$, where A is the amplitude of the curve, k is the first-order rate constant and t is time. In order to avoid overestimating k in cases where the RNA was not quantitatively cleaved, the amplitude was fixed at one when fitting cleavage data for the substrate containing a G–C substitution at the bottom base pair, G20A and G20C mismatch mutants and for the substrate with two G–C base-pairs inserted below the stem-loop. Cleavage of the WT-crRNA repeat by Csy4-R118A/R115A and Csy4-R115A/R119A revealed biphasic kinetics, and the data were fit to a double exponential decay. The slower kinetic process may reflect a rate-limiting conformational change. Both rate constants are reported in **Table 2.2**.

To ensure that Csy4 concentrations were saturating and that the on-rate for Csy4:RNA binding was not rate-limiting, cleavage experiments were repeated at 5-fold higher enzyme concentrations and analyzed similarly. This analysis frequently returned slightly larger rate constants for RNAs with fast cleavage kinetics, which we attribute to slower quenching rates in the presence of more enzyme. Overall, rate constants for these experiments were generally within ~30% of those measured at the lower enzyme concentration. The precise nature of the rate-limiting step in our single-turnover cleavage assays is not known, and so first-order rate constants are reported as k_{obs}. k_{obs} values presented in the text and in **Tables 2.1** and **2.2** represent the average and standard error of the mean from three independent experiments. The average percent error for all reported k_{obs} values is 4%.

Cleavage experiments with WT-Csy4 and WT-crRNA repeat at variable molar ratios (**Fig. 2.1d**) were conducted at a constant RNA concentration of 10 nM (0.25 nM 5'-radiolabeled RNA, 9.75 nM unlabeled RNA) and varying Csy4 concentrations (40, 20, 10, 5, 2.5 nM) in a final volume of 88 µl. 10 µl aliquots were removed and quenched at 0.25, 0.5, 1, 2, 5, 10, 30 and 60 minutes, and analyzed as described above. In determining the concentration of unlabeled RNA, hypochromicity of the stem-loop was corrected for by first hydrolyzing the RNA to nucleotides by incubating in 3 M NaOH at 50 °C for one hour. Then, absorbance spectra were recorded using a NanoDrop spectrophotometer (Thermo Scientific), and $A_{260\mathrm{nm}}$ values were averaged and converted to molar concentrations using the calculated extinction coefficient (295,900 $\mathrm{M^{-1}\ cm^{-1}}$). The 50% yield observed at an enzyme:substrate molar ratio of 1:1 may reflect Csy4 dimerization (Przybilski et al., 2011) or partial specific activity of purified WT-Csy4.

2.4 Results

2.4.1 Csy4 binds the crRNA repeat stem-loop with high affinity and functions as a single-turnover catalyst

Csy4 is a specialized ribonuclease that selects CRISPR transcripts from the cellular milieu for binding and cleavage. To determine the basis for this selectivity, we first examined the thermodynamic stability of the Csy4:RNA complex and the energetic contributions of protein:RNA interactions observed crystallographically (**Fig. 2.1b**). Using modified RNA substrates and/or Csy4 mutants, equilibrium dissociation constants (K_d) were measured using electrophoretic mobility shift assays (EMSAs). The RNA substrates we tested derive from the invariant 28-nt repeat sequence found within pre-crRNAs generated from *P. aeruginosa* strain UCBPP-PA14 CRISPR locus 2 (Grissa et al., 2007b), herein referred to as the crRNA repeat (**Fig. 2.1a**). We used the catalytically inactive Csy4-H29A mutant (Haurwitz et al., 2010) for experiments focused on analyzing the effects of changes to the RNA substrate, enabling investigation of RNA binding independent of cleavage. Wild-type (WT) Csy4 and Csy4-H29A bind a non-cleavable RNA substrate with affinities that are within 3-fold of each other (**Fig. 2.2**).

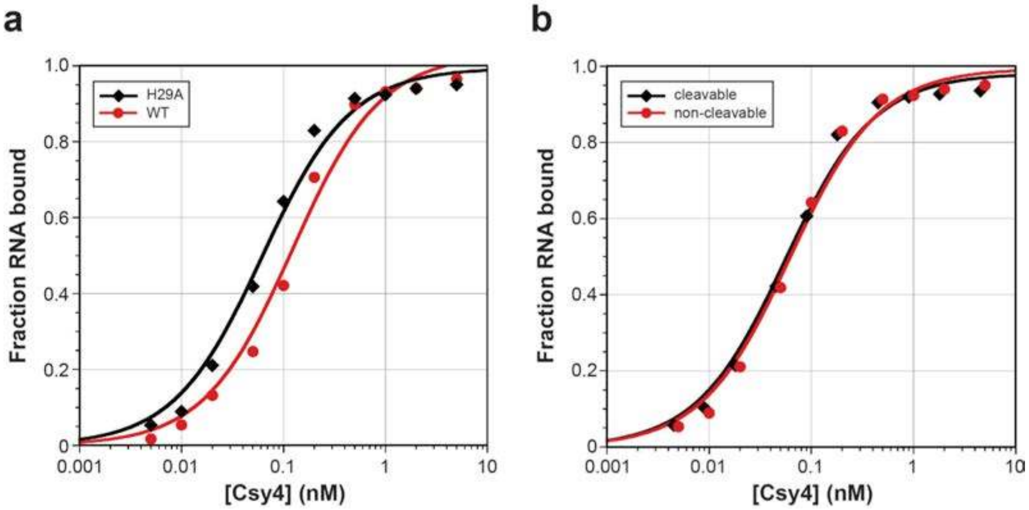


Figure 2.2. | Binding controls with Csy4-H29A and a non-cleavable RNA substrate. (a) Electrophoretic mobility shift assays (EMSAs) were performed with WT-Csy4 or Csy4-H29A and a non-cleavable crRNA repeat substrate containing a deoxyribonucleotide substitution at G20. WT-Csy4 exhibits an apparent binding affinity ~3-fold lower than Csy4-H29A. **(b)** To confirm that the non-cleavable and cleavable crRNA repeat substrates are bound similarly, EMSAs were performed with Csy4-H29A and both RNAs. Binding data with these substrates are indistinguishable. For these and all subsequent binding assays, the data were fit with a standard binding isotherm to yield equilibrium dissociation constants (solid lines; see **Materials and Methods**), and average K_d and standard error of the mean (SEM) values from at least three independent experiments are reported in **Table 2.1**.

Strikingly, Csy4 binds the full-length, WT crRNA repeat substrate with extremely high affinity, characterized by an equilibrium dissociation constant of ~50 pM (**Fig. 2.1c** and **Table 2.1**). Because Csy4 and the mature crRNA form part of the large Csy ribonucleoprotein complex responsible for target recognition (Wiedenheft et al., 2011b), we wondered whether Csy4 also retains high-affinity binding to the cleaved crRNA. Using a synthetic RNA corresponding to the 5' product stem-loop structure, we found that Csy4 binds this RNA indistinguishably from the substrate (**Fig. 2.1c**). Thus, all protein:RNA interactions contributing favorably to binding energy occur upstream of the scissile phosphate. Analysis of substrates truncated in the 5' ssRNA region allowed us to further demonstrate that nucleotides 1-4 of the crRNA repeat are completely dispensable for binding (**Fig. 2.3a**), indicating that the high-affinity interaction we observe requires only the 15-nt stem-loop and one upstream nucleotide. We observed binding defects when A5 was mutated, suggesting that it might be specifically recognized. Indeed, a crystal structure of a Csy4:product RNA complex containing nucleotides 2-20 of the crRNA repeat sequence revealed base-specific hydrogen bonds between the Watson-Crick face of A5 and the peptide backbone of Leu139 (**Fig. 2.3b** and (Haurwitz et al., 2012)).

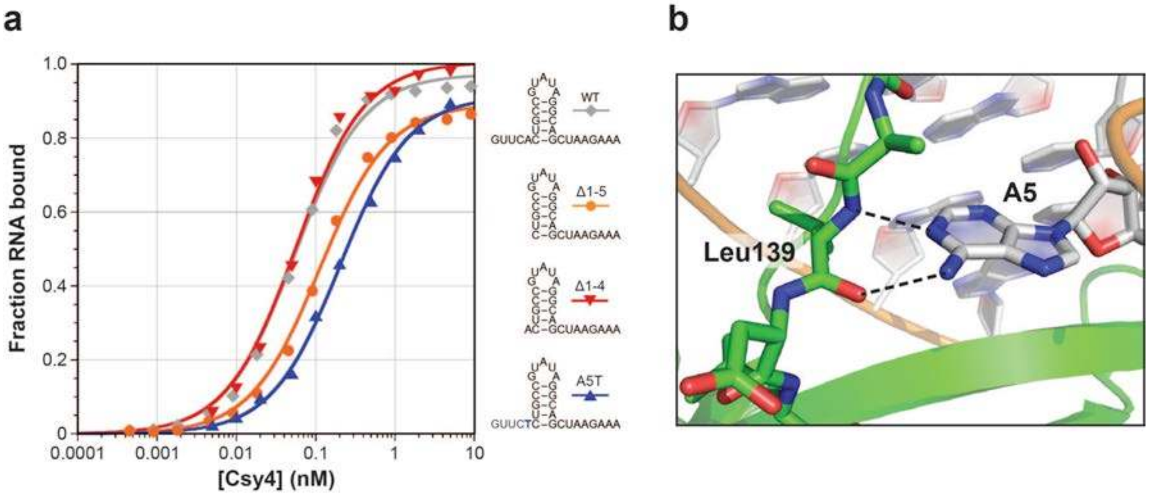


Figure 2.3 | Sequence-specific recognition of A5 by Csy4. (a) To determine the contributions of single-stranded RNA (ssRNA) nucleotides upstream of the stem-loop to overall binding energy, EMSAs were performed with Csy4-H29A and RNA substrates containing either deletions (Δ1-4, Δ1-5) or a mutation (A5T) in the first 5 ssRNA nucleotides. The 2-fold binding defect observed with Δ1-5 RNA relative to the WT-crRNA repeat was abolished when A5 was reintroduced (Δ1-4), indicating that A5 is the only ssRNA nucleotide bound by Csy4. The even larger magnitude binding defect (~5-fold) with a substrate mutated at this position (A5T) suggests that this binding is sequence-specific. **(b)** Csy4-S22C was crystallized bound to a product RNA containing nucleotides 2-20 of the WT-crRNA repeat (Haurwitz et al., 2012). While nucleotides 2-4 are disordered and not visible in the electron density, A5 is seen interacting with the peptide backbone of Leu139 via two base-specific hydrogen bonds (dashed lines) to N1 and N6 at the Watson-Crick edge.

Considering the retention of Csy4 and crRNA in the Csy complex (Wiedenheft et al., 2011b), we speculated that tight association of Csy4 with its product may be an intrinsic mechanistic feature of Csy4 during crRNA biogenesis in type I-F CRISPR systems. To test this hypothesis, we carried out cleavage assays at a range of enzyme:substrate molar ratios and monitored both the rate and yield of product formation. As seen in **Figure 1.1d**, Csy4 completely lacks the ability to engage in multiple-turnover catalysis. The overall yield of the cleavage reaction remained directly proportional to the Csy4 concentration when present in sub-stoichiometric amounts relative to substrate, even with incubation times >200-fold longer than the reaction time constant. All time courses fit well to a single exponential decay and yielded uniform, first-order observed rate constants (k_{obs}; **Table 2.1**), which would only be the case in the absence of multiple turnover behavior under conditions where the on-rate is not rate-limiting. These observations indicate that Csy4 remains product-bound after the reaction and is thereby strongly inhibited from performing additional rounds of RNA cleavage. Interestingly, crRNA repeat cleavage reached only 50% completion at an enzyme:substrate molar ratio of 1:1. A recent study used a two-hybrid system to demonstrate that Csy4 can interact with itself, but this result could not be repeated for all fusion constructs (Przybilski et al., 2011). While we cannot formally exclude the possibility that Csy4 might function as a dimer with one inactive subunit, our gel filtration experiments are consistent with purified Csy4 existing as a monomer (data not shown). Therefore, we speculate that the incomplete cleavage we observe reflects partial specific activity of purified WT-Csy4.

Table 2.1 | Binding and cleavage data for mutant crRNA repeat substrates.

RNA	K_d (nM)[a]	$K_{d,\mathrm{rel}}$[b]	ΔΔG (kcal/mol)[c]	k_{obs} (min^{-1})[d]	$k_{obs,\mathrm{rel}}$[e]
WT (synthetic)	0.050 ± 0.006	1.1	0.1 ± 0.1	3.8 ± 0.1	1.2
WT (transcribed)[†]	0.045 ± 0.009	1.0	NA	4.50 ± 0.06 (500 nM)	1.0
				3.98 ± 0.03 (40 nM)	1.1
				4.0 ± 0.1 (20 nM)	1.1
				3.5 ± 0.2 (10 nM)	1.3
				3.8 ± 0.2 (5 nM)	1.2
				3.90 ± 0.06 (2.5 nM)	1.2
Δ21-28 (product)	0.049 ± 0.006	1.1	0.0 ± 0.1	NA	NA
Δ1-5	0.09 ± 0.01	2.0	0.4 ± 0.1	2.83 ± 0.07	1.6
Δ1-4	0.047 ± 0.005	1.0	0.0 ± 0.1	2.85 ± 0.08	1.6
A5T	0.216 ± 0.009	4.8	0.9 ± 0.1	2.4 ± 0.1	1.9
Reverse complement (rc)	5600 ± 400	120,000	6.9 ± 0.1	0.0057 ± 0.0004	790
GUGUA loop (A13G)	0.05 ± 0.01	1.2	0.1 ± 0.2	4.6 ± 0.2	0.97
UAUAC loop (G11U,U12A,A13U,U14A,A15C)	337 ± 3	7,400	5.3 ± 0.1	2.57 ± 0.08	1.8
UUCG loop (G11U,A13C,U14G,ΔA15)	2000 ± 400	45,000	6.3 ± 0.2	1.90 ± 0.09	2.4

AAAAA loop (G11A,U12A,U14A)	700 ± 100	15,000	5.7 ± 0.2	2.8 ± 0.2	1.6
Nicked (between U12 and A13)	108 ± 4	2,400	4.6 ± 0.1	ND	ND
G–C, 1st base pair (C6G,G20C)	54 ± 3	1,200	4.2 ± 0.1	0.00060 ± 0.00002	7,500
U–A, 1st base pair (C6U,G20A)	0.9 ± 0.1	20	1.8 ± 0.1	0.0272 ± 0.0005	170
A–U, 1st base pair (C6A,G20U)	0.211 ± 0.005	4.7	0.9 ± 0.1	0.037 ± 0.002	120
C–G, 2nd base pair (U7C,A19G)	35 ± 2	760	3.9 ± 0.1	1.64 ± 0.06	2.7
G–C, 2nd base pair (U7G,A19C)	1.5 ± 0.1	33	2.1 ± 0.1	1.17 ± 0.02	3.8
A–U, 2nd base pair (U7A,A19U)	0.60 ± 0.09	13	1.5 ± 0.2	1.20 ± 0.09	3.8
C–G, 3rd base pair (G8C,C18G)	2.9 ± 0.3	64	1.5 ± 0.1	1.50 ± 0.01	3.0
A–U, 3rd base pair (G8A,C18U)	0.12 ± 0.03	2.6	0.6 ± 0.2	4.37 ± 0.07	1.0
U–A, 3rd base pair (G8U,C18A)	0.103 ± 0.002	2.3	0.5 ± 0.1	1.58 ± 0.04	2.8
G–C, 4th base pair (C9G,G17C)	0.63 ± 0.02	14	1.6 ± 0.1	3.87 ± 0.07	1.2
A–U, 4th base pair (C9A,G17U)	0.25 ± 0.01	5.5	1.0 ± 0.1	4.40 ± 0.06	1.0
U–A, 4th base pair (C9U,G17A)	0.15 ± 0.03	3.3	0.7 ± 0.2	4.17 ± 0.09	1.1
G–C, 5th base pair (C10G,G16C)	3.9 ± 0.5	85	2.6 ± 0.1	3.2 ± 0.1	1.4
A–U, 5th base pair (C10A,G16U)	0.40 ± 0.03	8.8	1.3 ± 0.1	2.87 ± 0.09	1.6
U–A, 5th base pair (C10U,G16A)	0.09 ± 0.02	2.0	0.4 ± 0.2	2.05 ± 0.04	2.2
Mutate 3 base pairs, #1 (G8A,C9U,C10U,G16A,G17A,C18U)	0.31 ± 0.02	6.9	1.2 ± 0.1	2.4 ± 0.1	1.9
Mutate 3 base pairs, #2 (G8U,C9A,C10A,G16U,G17U,C18A)	7.0 ± 0.3	160	3.0 ± 0.1	0.54 ± 0.02	8.3
Mutate 3 base pairs, #3 (G8C,C9G,C10G,G16C,G17C,C18G)	217 ± 9	4,800	5.0 ± 0.1	0.312 ± 0.007	14
C6G mismatch	3.8 ± 0.6	85	2.6 ± 0.2	0.128 ± 0.003	35
C6A mismatch	0.057 ± 0.008	1.3	0.1 ± 0.1	0.456 ± 0.002	9.9

G20A mismatch	1.9 ± 0.3	41	2.2 ± 0.2	0.0003 ± 0.0001	14,000
G20C mismatch	3.67 ± 0.09	81	2.6 ± 0.1	0.00020 ± 0.00002	23,000
1 extra G–C, top of stem	70 ± 10	1,600	4.4 ± 0.2	2.74 ± 0.06	1.6
2 extra G–C, top of stem	2200 ± 200	49,000	6.4 ± 0.1	0.28 ± 0.01	16
5 extra G–C, top of stem	4000 ± 1000	91,000	6.8 ± 0.2	0.083 ± 0.001	54
5 extra G–C, top of stem 3' A bulge	253 ± 3	5,600	5.1 ± 0.1	ND	ND
5 extra G–C, top of stem 3' AA bulge	26 ± 2	570	3.8 ± 0.1	ND	ND
5 extra G–C, top of stem 3' AAA bulge	10.5 ± 0.8	230	3.2 ± 0.1	2.78 ± 0.06	1.6
1 extra A–U, bottom of stem	0.061 ± 0.003	1.3	0.2 ± 0.1	2.25 ± 0.02	2.0
2 extra A–U, bottom of stem	0.57 ± 0.02	13	1.5 ± 0.1	2.813 ± 0.009	1.6
1 extra G–C, bottom of stem	9.6 ± 0.8	210	3.2 ± 0.1	0.102 ± 0.002	44
2 extra G–C, bottom of stem	36 ± 1	790	4.0 ± 0.1	0.0028 ± 0.0002	1,600

[a] Reported as the average and standard error of the mean (SEM) from at least three independent experiments. All binding experiments were performed with Csy4-H29A.

[b] Calculated by dividing each K_d value by the K_d for WT-crRNA repeat (0.045 nM).

[c] Reported as the average and SEM, and calculated according to the equation:
$\Delta\Delta G = -RT\ln(K_{d,\mathrm{WT}}/K_{d,\mathrm{mutant}})$

[d] Reported as the average and SEM from three independent experiments. All cleavage experiments were performed with WT-Csy4.

[e] Calculated by dividing k_{obs} for the WT-crRNA repeat (4.50 min^{-1}; at 500 nM Csy4) by the k_{obs} value for each mutant RNA substrate.

[†] k_{obs} values are reported for experiments at all WT-Csy4 concentrations tested.

NA, not applicable. ND, not determined.

2.4.2 Protein determinants of high-affinity crRNA repeat binding and cleavage

The high-affinity interaction between Csy4 and the crRNA repeat substrate is tighter than many protein:RNA complexes studied to date. We were therefore interested in gaining a detailed understanding of the primary sources of binding energy, as informed by interactions identified from our crystal structure. We began by focusing on the bottom of the RNA stem, where the side-chains of Arg102 and Gln104 are each involved in two sequence-specific hydrogen bonds with the major groove faces of G20 and A19, respectively. Using a synthetic, non-cleavable substrate that is bound indistinguishably from the WT-crRNA repeat (**Fig. 2.2b**), EMSAs with Csy4-R102A and Csy4-Q104A mutants revealed that the binding energies contributed by these amino acids are quite distinct. The crRNA repeat binds >2,000-fold more weakly to Csy4-R102A, representing a ΔΔG of 4.6 kcal/mol, whereas RNA binding by Csy4-Q104A is destabilized by only 1.4 kcal/mol relative to WT (**Fig. 2.4a** and **Table 2.2**). This difference may be explained in part by the expected +1 charge on the arginine's guanidinium group at physiological pH. Whereas deletion of an uncharged hydrogen bond typically weakens binding between enzyme and substrate by 0.5-1.8 kcal/mol, charged hydrogen bond generally contribute some 3-6 kcal/mol binding energy (Fersht, 1987), in good agreement with our data.

In addition to its interaction with Arg102, G20 of the crRNA repeat stacks onto the aromatic side-chain of Phe155. Stacking interactions between aromatic amino acids and nucleotides can contribute up to 5.5 kcal/mol of binding energy (Auweter et al., 2006; Nolan et al., 1999), but we were surprised to observe a negligible 1.5-fold binding defect (ΔΔG = 0.2 kcal/mol) with a Csy4-F155A mutant (**Fig. 2.4a**). Given the pre-crRNA processing defects we observed previously with Csy4-F155A (Haurwitz et al., 2010), these data suggest that Phe155 instead plays a role in achieving rapid cleavage kinetics. Indeed, under single-turnover conditions with saturating enzyme concentrations (see **Materials and Methods**), the F155A mutant led to a ~50-fold reduction in the observed cleavage rate constant (**Fig. 2.4b**). Csy4-R102A also exhibited a ~20-fold defect in cleavage kinetics, whereas the rate of cleavage by Csy4-Q104A was within 2.5-fold of WT (**Fig. 2.4b**). Collectively, these data suggest that, independent of their effects on binding energy, Phe155 and Arg102 are important for anchoring the G20 guanine in the active site and may thereby assist in positioning the ribose for subsequent activation of its 2'-OH nucleophile.

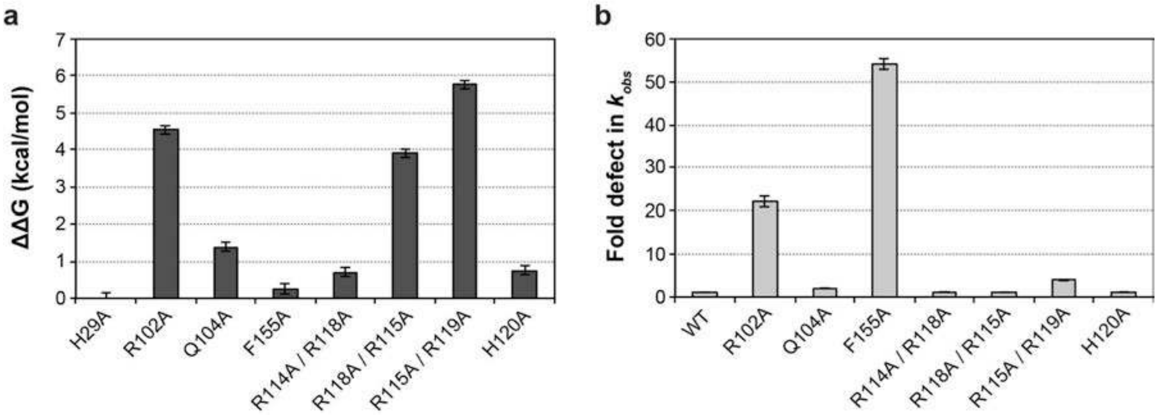


Figure 2.4 | Amino acid contributions to binding energy and cleavage kinetics. (a) Csy4 residues involved in base-pair recognition and phosphate backbone contacts were mutated to alanine in order to assess their energetic contributions to binding. EMSAs were performed with a noncleavable crRNA repeat substrate containing a deoxyribonucleotide substitution at G20, and binding defects relative to Csy4-H29A were determined and converted to ΔΔG values (T = 298 K). Plotted are the average and SEM from at least three independent experiments. **(b)** First-order rate constants (k_{obs}) for WT-crRNA repeat cleavage by each Csy4 mutant were determined. Cleavage data for R118A/R115A and R115A/R119A mutants showed biphasic kinetics and were fit with a double exponential decay to yield two rate constants (**Table 2.2**), the faster of which is shown. Plotted are the average fold defects (relative to WT-Csy4) and SEM from three independent experiments. Average K_d, k_{obs}, and SEM values are reported in **Table 2.2**.

Moving up the crRNA repeat stem, we next focused on interactions observed in the crystal structure between the RNA and residues (**Fig. 2.1b**) found in the α-helix that inserts into the major groove of the double-stranded stem. The guanidinium groups of Arg114, Arg115, Arg118 and Arg119 each present ≥2 hydrogen-bond donors within 3 Å of acceptors in the RNA phosphate backbone, yet their contributions to overall binding energy differ widely, as assessed through double R →A mutations. In particular, Arg114 and Arg118, which contact adjacent phosphates, contribute only 0.7 kcal/mol of binding energy, whereas alanine mutations at Arg115 and Arg119 led to a >15,000-fold binding defect (ΔΔG = 5.8 kcal/mol; **Fig. 2.4a**). While all four residues are positioned to act as arginine forks, in that each side-chain contacts adjacent phosphates (Calnan et al., 1991), only Arg115 and Arg119 may simultaneously utilize all three nitrogen atoms of the guanidinium group as hydrogen bond donors. Arg115 hydrogen bonds to two phosphates in addition to the major groove face of G11, which forms part of the G·A sheared base pair at the bottom of the GUAUA pentaloop, and Arg119 is situated in a unique pocket of the loop where it interacts with phosphates separated by two nucleotides. His120 also interacts with a phosphate at the apex of the loop and contributes 0.8 kcal/mol of binding energy (**Fig. 2.4a**). The specific network of multi-dentate contacts between the arginine-rich helix and the RNA stem-loop suggests that high affinity binding to the crRNA repeat is highly shape-specific, especially with regard to the tertiary structure of the loop. The large magnitude of the binding energy contributed by this protein helix enables Csy4 to maintain a tight grip on the substrate and product, but this interaction is not required for catalytic activity. Cleavage rates for the H120A and R →A double mutants under saturating conditions were within 5-fold of WT-Csy4 (**Fig. 2.4b**).

Table 2.2 | Binding and cleavage data for Csy4 mutants.

Csy4	K_d (nM)[a]	$K_{d,\mathrm{rel}}$[b]	ΔΔG (kcal/mol)[c]	k_{obs} (min^{-1})[d]	$k_{obs,\mathrm{rel}}$[e]
WT	0.132 ± 0.009	2.9	0.6 ± 0.1	4.50 ± 0.06	1.0
H29A	0.045 ± 0.009	1.0	NA	NA	NA
R102A	98 ± 5	2,200	4.6 ± 0.1	0.20 ± 0.01	22
Q104A	0.48 ± 0.02	11	1.4 ± 0.1	2.160 ± 0.006	2.1
F155A	0.068 ± 0.006	1.5	0.2 ± 0.1	0.083 ± 0.002	54
R114A/R118A	0.15 ± 0.01	3.2	0.7 ± 0.1	3.70 ± 0.06	1.2
R118A/R115A†	34 ± 2	740	3.9 ± 0.1	3.9 ± 0.2 (0.31)	1.2
				0.008 ± 0.001 (0.52)	560
R115A/R119A†	780 ± 50	17,000	5.8 ± 0.1	1.14 ± 0.03 (0.78)	3.9
				0.035 ± 0.004 (0.20)	130

H120A	0.16 ± 0.01	3.6	0.8 ± 0.1	3.83 ± 0.03	1.2

[a]Reported as the average and standard error of the mean (SEM) from at least three independent experiments. All binding experiments were performed with the non-cleavable substrate containing a deoxyribonucleotide substitution at G20.

[b]Calculated by dividing each K_d value by the K_d for Csy4-H29A (0.045 nM).

[c]Reported as the average and SEM, and calculated according to the equation:
$\Delta\Delta G = -RT\ln(K_{d,\mathrm{H29A}}/K_{d,\mathrm{mutant}})$

[d]Reported as the average and SEM from three independent experiments. All cleavage experiments were performed with the *in vitro* transcribed WT-crRNA repeat substrate.

[e]Calculated by dividing k_{obs} for WT-Csy4 (4.50 min^{-1}) by the k_{obs} value for each Csy4 mutant.

[†]Cleavage data showed biphasic kinetics and were fit with a double exponential decay. Both rate constants and their respective amplitudes (in parentheses) are reported.

NA, not applicable.

2.4.3 High-affinity crRNA repeat binding is sensitive to the loop structure

The direction of CRISPR loci transcription in *P. aeruginosa* has not been directly analyzed, and a recent report that detected mature crRNAs by Northern blot analysis used dsDNA probes that were not strand-specific (Cady and O'Toole, 2011). Transcription in a direction opposite to that of our own predictions would generate pre-crRNAs containing the reverse complement of the crRNA repeat sequence. To determine whether Csy4 also recognizes and cleaves this potential substrate, we generated the reverse complement crRNA (rc-crRNA) repeat by *in vitro* transcription and tested its affinity for Csy4-H29A. We found that the rc-crRNA repeat binds Csy4 >10^5-fold more weakly than the WT-crRNA repeat (**Fig. 2.5a**) and is cleaved >750-fold more slowly (**Fig. 2.6a**), strongly suggesting that the genuine Csy4 substrate *in vivo* is pre-crRNA transcribed in an orientation consistent with our previous work (Haurwitz et al., 2010). Northern blot analysis using single-stranded probes indeed confirmed the presence of crRNAs in *P. aeruginosa* UCBPP-PA14 with the repeat sequence we define in **Figure 2.1a**, but failed to detect transcripts from the opposite strand (**Fig. 2.7**).

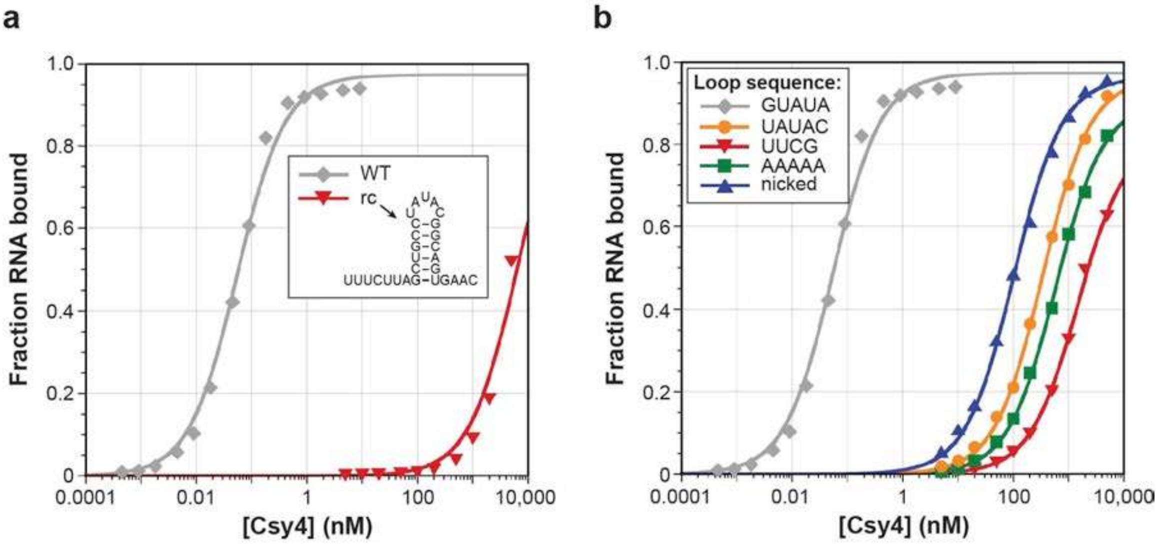


Figure 2.5 | Importance of loop sequence for high-affinity RNA binding. (a) EMSAs demonstrate that Csy4 binds the reverse complement of the crRNA repeat (rc) >10^5-fold more weakly than the WT-crRNA repeat. **(b)** Mutant RNA substrates were generated by changing the WT loop sequence (GUAUA) to a quintuple mutant (UAUAC), the highly stable UUCG tetraloop, or a poly(A) pentaloop, or by removing the loop through use of a substrate nicked between U12 and A13. EMSAs reveal substantial defects associated with binding these mutant RNAs.

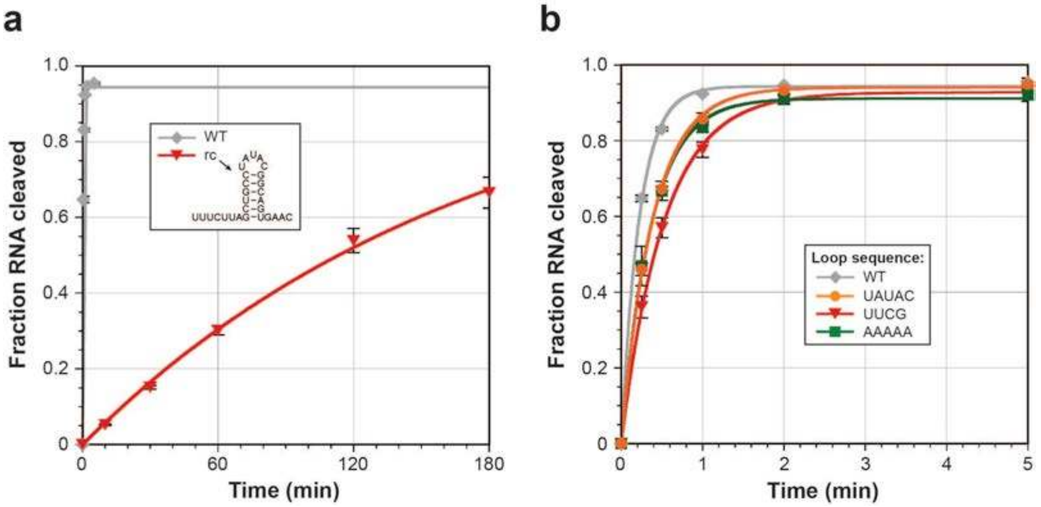


Figure 2.6 | Cleavage of rc-crRNA repeat and loop mutant substrates. Cleavage assays were performed with the reverse complement (rc) crRNA repeat **(a)** and RNA substrates containing mutated loop sequences **(b)**. The rc-crRNA repeat substrate was cleaved >750-fold slower than the WT-crRNA repeat substrate, whereas substrates containing mutations only in the loop sequence were cleaved at rates within 2.5-fold of WT. For these and all subsequent cleavage assays, the data were fit with a single exponential to yield first-order rate constants (solid lines; see **Materials and Methods**), and average k_{obs} and SEM values from three independent experiments are reported in **Table 2.1**.

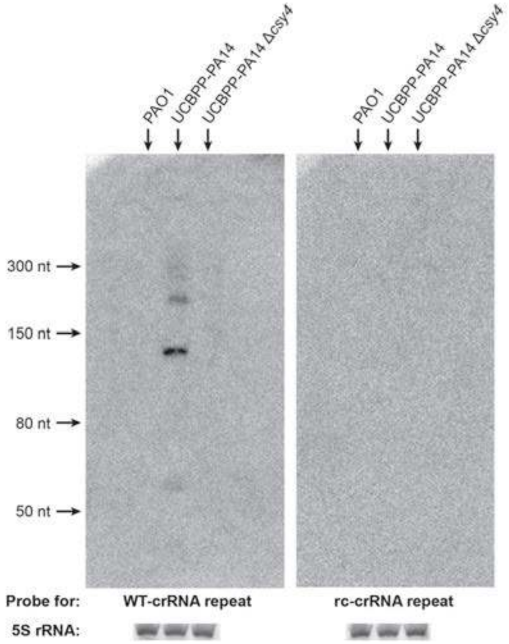


Figure 2.7 | Northern blot analysis of crRNAs in *Psuedomonas aeruginosa.* Total RNA was extracted from strains of *P. aeruginosa* without CRISPRs (PAO1), with a complete CRISPR-Cas locus (UCBPP-PA14) or with a CRISPR-Cas locus harboring a Csy4 gene deletion. Duplicates of each RNA preparation were separated by 15% denaturing PAGE, transferred to nylon membranes and probed with DNA oligonucleotides complementary to either the WT-crRNA repeat (left) or the rc-crRNA repeat (right). The gel was stained with SYBR Gold before transfer, and the 5S RNA band is shown as a loading control. RNAs containing the WT-crRNA repeat but not the reverse complement were detected in *P. aeruginosa* UCBPP-PA14. The laddering pattern is consistent with precursor transcripts that were incompletely processed, thereby yielding multiples of the length of a mature crRNA (60 nucleotides). Hybridization to the mature crRNA is likely to be less efficient than to partially processed species because the probe is complementary to only 20 out of 28 nucleotides. The precursor transcript may be prone to rapid degradation in the absence of Csy4, explaining why the pre-crRNA band in the Δ*csy4* strain is not more prominent.

When comparing the two RNA sequences, rc-crRNA repeat contains the identical five base-pair stem sequence as the WT-crRNA repeat but with an additional predicted G-U wobble base pair below and different loop and flanking ssRNA sequences, indicating that one or more of these regions are specifically recognized by Csy4. Having already demonstrated the negligible binding defects resulting from deletion of flanking ssRNA nucleotides, we suspected that destabilized binding of the rc-crRNA repeat resulted primarily from the inability of Csy4 to interact productively with the UAUAC loop sequence and/or the unique tertiary structure it would impose on the RNA substrate. The GUAUA loop encoded by CRISPR locus 2 in *P. aeruginosa* UCBPP-PA14 forms a GNR(N)A pentaloop structure (Legault et al., 1998), in which U14 flips out of the loop to enable a GNRA tetraloop fold that involves sequential stacking of U12, A13 and A15 on the 3' strand of the stem (Haurwitz et al., 2010). The CRISPR 3 locus encodes a GUGUA loop in the repeat sequence that is predicted to form the same pentaloop structure, and this crRNA structure is bound and cleaved indistinguishably from the substrate with a GUAUA loop (**Fig. 2.8**). We hypothesized that Csy4 specifically recognizes this loop motif, and that other loop sequences unable to conform to a GNRA tetraloop fold bind much more weakly.

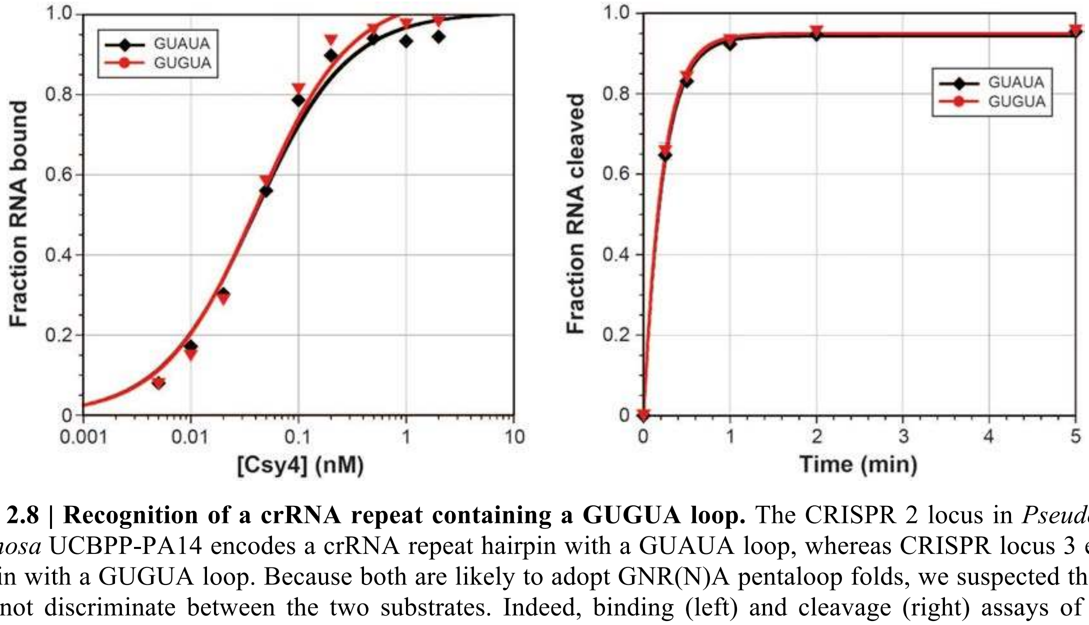


Figure 2.8 | Recognition of a crRNA repeat containing a GUGUA loop. The CRISPR 2 locus in *Pseudomonas aeruginosa* UCBPP-PA14 encodes a crRNA repeat hairpin with a GUAUA loop, whereas CRISPR locus 3 encodes a hairpin with a GUGUA loop. Because both are likely to adopt GNR(N)A pentaloop folds, we suspected that Csy4 would not discriminate between the two substrates. Indeed, binding (left) and cleavage (right) assays of crRNA repeat substrates containing either loop sequence reveal indistinguishable biochemical behaviors.

To test this, we generated a panel of RNA substrates containing mutated loop sequences and tested their affinity for Csy4-H29A. In agreement with our hypothesis, Csy4 bound to each RNA at least 7,000-fold more weakly than WT (**Fig. 2.5b**). Even a nicked RNA substrate formed from two oligonucleotides annealed *in trans* interacted more favorably with Csy4 than those containing a non-GNRA-like loop (**Fig. 2.5b & 2.9**). These experiments confirm that high-affinity Csy4 binding relies in part on a precise substrate tertiary structure in the loop region, independently of base-specific contacts, and that the absence of a loop altogether is less detrimental to binding than the presence of a non-native loop. It is interesting to note that, despite their weakened binding, RNAs with mutated loops were cleaved at rates within 2.5-fold of the WT-crRNA repeat at saturating Csy4 concentrations (**Fig. 2.6b**). This was true even for a substrate containing the same loop (UAUAC) as the rc-crRNA repeat, which had a >750-fold defect in k_{obs}. Since the stacking interaction between the terminal C–G base pair and the aromatic side chain of Phe155 is important for cleavage (**Fig. 2.4b**), we suspected that the additional base pair below the WT stem in the rc-crRNA repeat might impede Csy4 activity (see below).

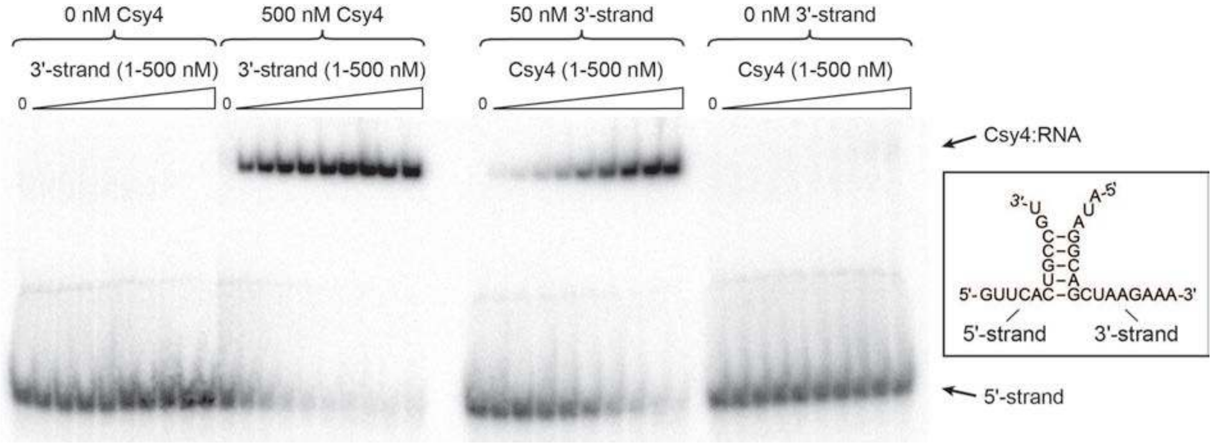


Figure 2.9 | Binding controls with a nicked crRNA repeat substrate. An RNA substrate was generated using two synthetic oligonucleotides constituting nucleotides 1-12 (5'-strand) and 13-28 (3'-strand) of the WT-crRNA repeat substrate (see boxed inset, right), and EMSAs were performed with Csy4-H29A. The 5'-strand was [^{32}P]-radiolabeled for these experiments and present at 0.5 nM in all binding reactions. To confirm that the observed electrophoretic mobility gel shift represented Csy4 bound to the hybridized two-strand duplex, experiments were performed that increased the 5'/3'-strand molar ratio with and without Csy4 present (left), or that increased the Csy4 concentration with and without the 3'-strand present (right). No shift was observed in the absence of Csy4, indicating either that the hybridized substrate does not stably form without Csy4 or that it does not shift relative to the 5'-strand alone. Additionally, Csy4 does not bind the 5'-strand alone at the highest concentrations tested (500 nM). These data indicate that binding requires the double-stranded substrate and that Csy4 may trap a hybridized duplex that is thermodynamically unstable under these experimental conditions.

2.4.4 Specificity within the crRNA repeat stem sequence during binding and cleavage

We were particularly interested in investigating the ability of Csy4 to discriminate between substrates containing the cognate five base pairs in the stem and those with similar but non-cognate sequences. We therefore made all individual Watson-Crick base-pair substitutions at each position in the double-stranded stem and determined the energetic costs associated with binding each mutant RNA substrate relative to the WT-crRNA repeat using EMSAs (**Fig. 2.10a**). The data reveal that base-pair changes throughout the stem result in varying degrees of

Csy4:RNA complex destabilization, ranging from 0.4-4.2 kcal/mol. The largest defects result from G–C and C–G substitutions at the ultimate and penultimate base pairs, respectively, where Arg102 and Gln104 provide a direct readout mechanism of recognition and confer similar degrees of discrimination in spite of their unequal contributions to binding energy. To confirm this, we repeated binding experiments with RNA substrates containing substitutions at the bottom two base pairs using either Csy4-R102A or Csy4-Q104A (**Fig. 2.10a**). As expected, the overall specificity for particular base pairs at either position is lost when the amino acid specificity determinant is absent. The Csy4:RNA co-crystal structure did not reveal sequence-specific contacts with Watson-Crick base pairs in the upper part of the double-stranded stem (Haurwitz et al., 2010), but we observed substantial energetic penalties for binding substrates with base-pair substitutions in this region (**Fig. 2.10a**). Furthermore, the magnitude of these binding defects was highly sequence-dependent; when multiple base-pair substitutions were made in the top three base pairs simultaneously, binding defects ranged from 7- to almost 5,000-fold (**Fig. 2.11**), with the largest destabilization occurring when each C–G pair was mutated to its complement. These results reveal that substrate sequence specificity is mediated by Csy4 via a mechanism that does not rely exclusively on base-specific interactions.

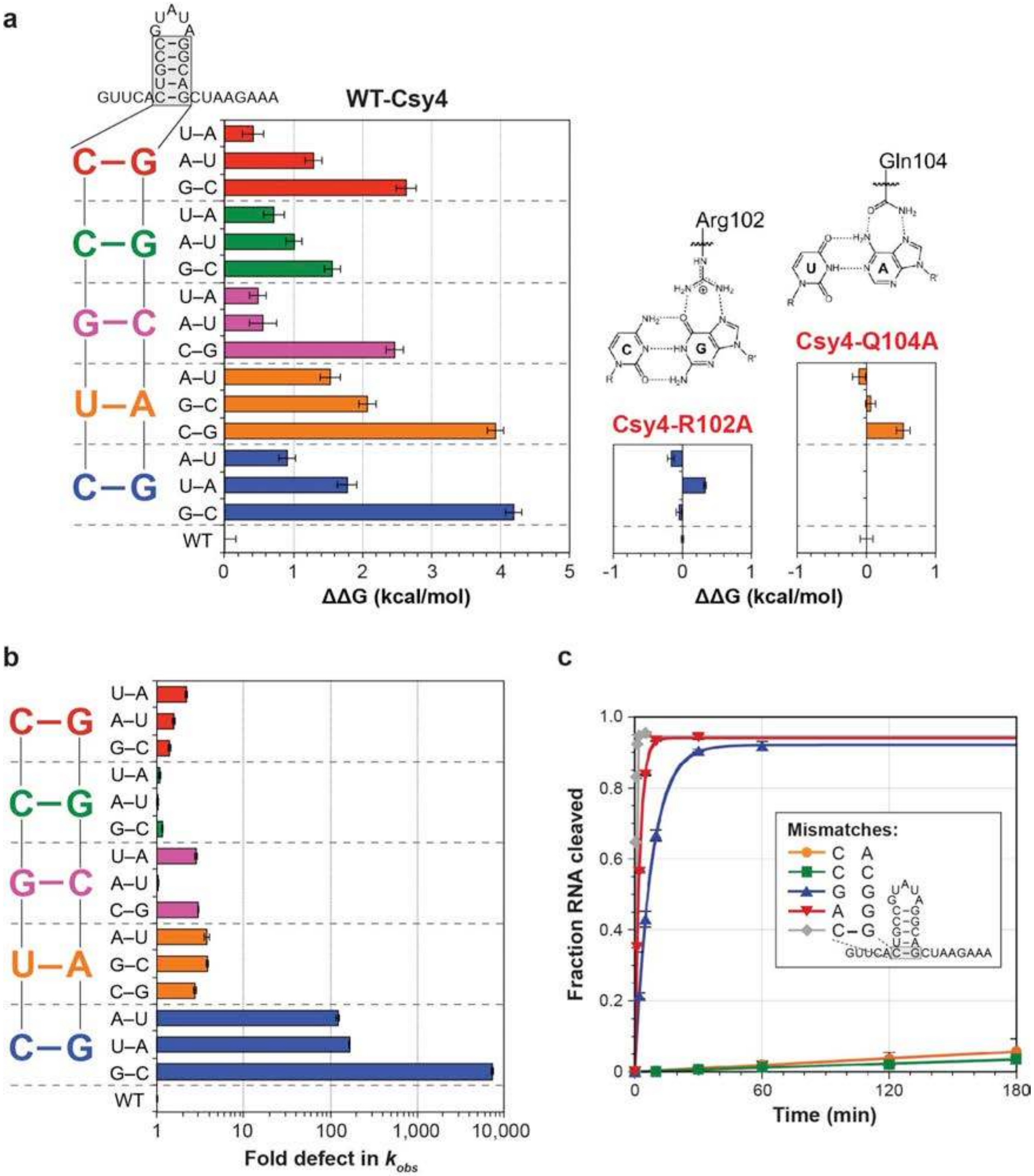


Figure 2.10 | Substrate specificity within the crRNA repeat stem. (a) A library of mutated crRNA repeat substrates was generated containing all possible Watson-Crick base-pair substitutions at each position in the double-stranded stem. EMSAs were performed with Csy4-H29A and these RNA substrates, and the resulting binding defects relative to WT-crRNA repeat were determined and converted to ΔΔG values (T = 298 K). The WT stem sequence is shown at the left, with data for base-pair substitutions at each position color-coded similarly. Binding experiments with RNAs mutated at the bottom C–G or U–A base pair were repeated with Csy4-R102A (middle) or Csy4-Q104A (right), respectively; ΔΔG values were calculated relative to WT-crRNA repeat binding by each Csy4 mutant. Shown above are chemical structures of the interactions made by Arg102 and Gln104 with the WT base pairs. **(b)** Single- turnover cleavage assays were performed with the same library of RNA mutants as in (a), and the resulting defects in k_{obs} relative to WT-crRNA repeat were determined. The data are plotted as in (a). **(c)** To investigate the importance of the terminal C–G base pair during cleavage, mismatched substrates were generated by mutating C6 or G20 individually and single-turnover cleavage assays were performed.

We next investigated whether these specificity determinants also influence the chemical cleavage reaction. To test this, we conducted single-turnover cleavage experiments with WT-Csy4 at saturating concentrations using the same library of RNAs as in **Fig. 2.10a**, and determined the first-order rate constants for RNA cleavage (k_{obs}) relative to WT. In stark contrast to the observed binding specificity, rate constants governing the cleavage of RNA substrates with base-pair substitutions at any position other than the terminal position were within 4-fold of WT (**Fig. 2.10b**). However, any mutation of the terminal C–G base-pair in the stem-loop is detrimental for cleavage of the crRNA repeat, with kinetic defects ranging from ~100- to 7,500-fold. To further dissect the importance of the terminal C–G base pair, we generated a series of RNA substrates containing mismatches at this position by mutating either C6 or G20 independently. Cleavage time courses with these substrates (**Fig. 2.10c**) clearly demonstrate the importance of G20, regardless of whether or not a base pair can form at the terminal position. RNA substrates containing C6A or C6G mutations are cleaved at rates within 40-fold of WT, whereas mutation of G20 to either adenosine or cytosine leads to >10,000-fold defects.

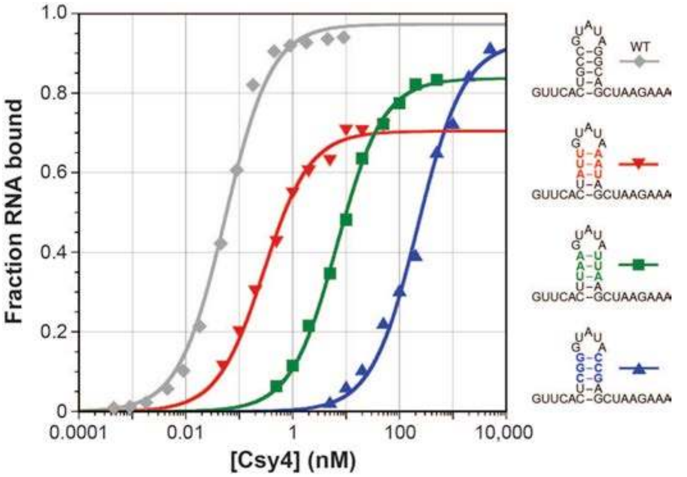


Figure 2.11 | Recognition of base pairs at the top of the stem. In order to investigate the sequence specificity in the upper part of the stem-loop, we generated mutant crRNA repeat substrates (right) that contained consecutive substitutions in the top three base-pairs and performed EMSAs with Csy4-H29A. A construct that maintained the same purine-pyrimidine pattern (red) showed the mildest binding defect, whereas mutating C–G base pairs to their complement resulted in a ~5,000-fold defect.

2.4.5 Csy4 is highly selective for stem-loops of defined length

Having interrogated Csy4 for sequence specificity throughout the crRNA repeat, we also wondered whether Csy4 is sensitive to the length of the crRNA repeat stem. To test this, we inserted one or two base pairs at the top of the duplex region and tested these substrates for binding. Strikingly, just one or two additional G–C base pairs led to 1,600- and 49,000-fold weaker binding affinities, respectively (**Fig. 2.12a**). This was particularly surprising because the crystal structure did not immediately suggest any obvious steric clashes that would result from insertions at the top of the stem. However, given the large energetic contribution of the arginine-rich helix to binding (**Fig. 2.4c**), we suspected that additional base pairs would disrupt protein-loop interactions and prevent stable docking of this helix into the major groove of the double-stranded stem. A-form dsRNA helices have deep and narrow major grooves that are generally inaccessible to proteins (Draper, 1995), but exceptions occur in proximity to helix termini or asymmetric bulges, where the major groove can widen considerably (Weeks and Crothers, 1993). We hypothesized that base-pair insertions cause narrowing of the major groove and thereby disrupt high-affinity interactions between the arginine-rich helix and crRNA repeat.

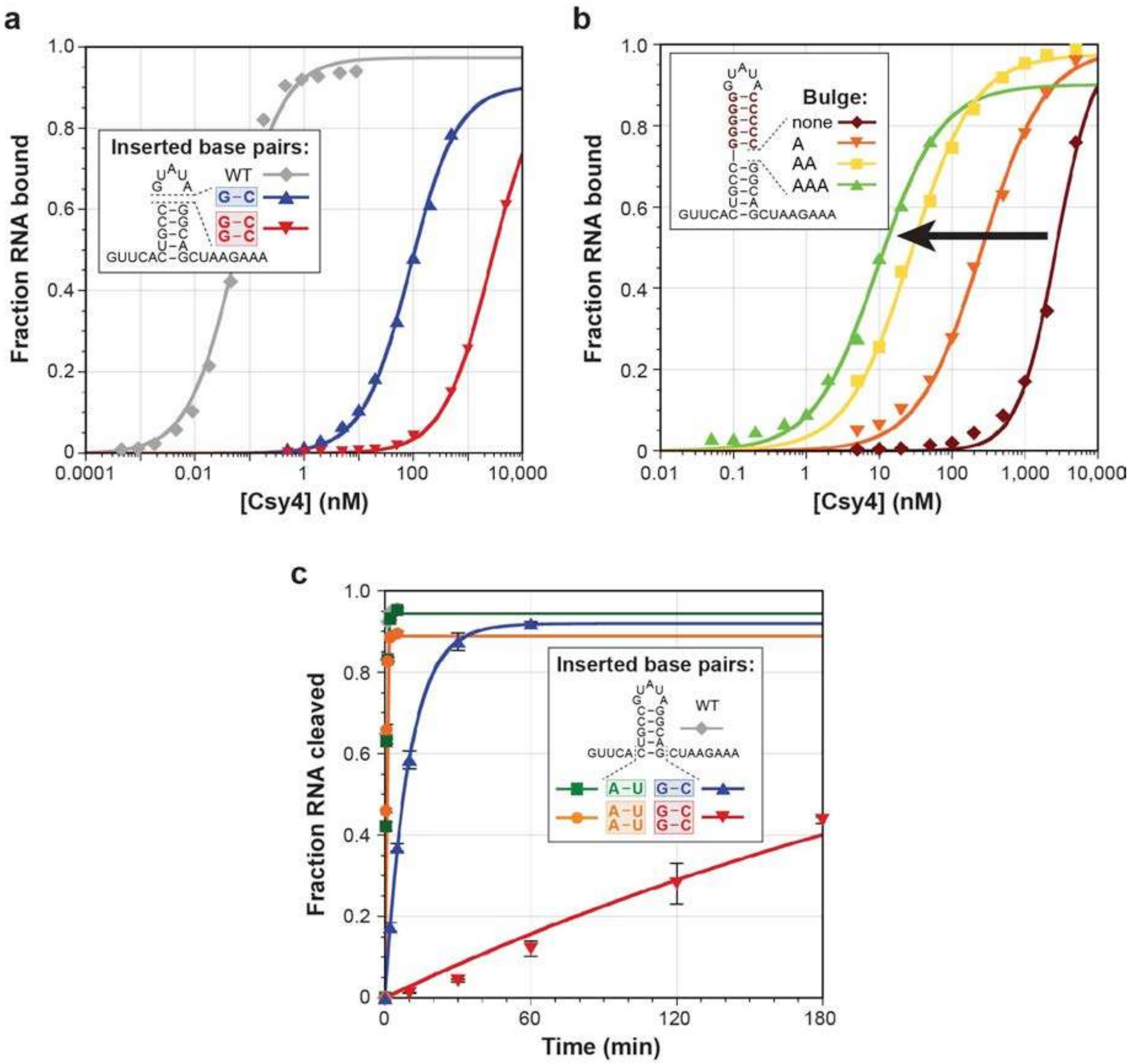


Figure 2.12 | Stem length dependence during substrate binding and cleavage. (a) One or two G–C base pairs were inserted at the top of the stem between the closing C–G base pair and the GUAUA pentaloop, and EMSAs were performed. **(b)** To test the hypothesis that longer stems prevent stable binding of the arginine-rich helix via their effect on major groove accessibility, a substrate was generated that contains five G–C base pairs inserted above the WT stem. Subsequently, asymmetric adenosine bulges were inserted on the 39 side of the duplex between the five-base-pair WT stem and the five-base-pair insertion. EMSAs reveal that binding affinities increase monotonically (black arrow) with bulges of increasing size. **(c)** One or two G–C or A–U base pairs were inserted below the terminal C–G base pair, and cleavage time courses were performed. Additional A–U base pairs have negligible effects on k_{obs}, whereas two additional G–C base pairs result in ≥1500-fold slower kinetics.

To test this idea, we generated an RNA construct that contains five G–C base pairs inserted atop the WT stem sequence while retaining the GUAUA pentaloop. This RNA was bound with an equilibrium dissociation constant of 4 μM (**Fig. 2.12b**), representing nearly a 10^5-fold defect relative to WT. We then introduced adenosine bulges of varying size on the 3' side of the stem, at the junction between the WT five base-pair stem sequence and the five G–C base-pair insertion. These types of asymmetric bulges within perfectly base-paired dsRNA helices have been shown previously to increase major groove accessibility progressively as a function of bulge size, as probed using diethylpyrocarbonate (DEPC) reactivity (Weeks and Crothers, 1993). In excellent agreement with our hypothesis, we found that the binding affinity of Csy4 for these bulged substrates increased in concert with bulges of increasing size (**Fig. 2.12b**), suggesting that major groove widening enables stable docking of the arginine-rich helix. The inability to form favorable protein-loop interactions likely explains why bulged substrates are still bound >200-fold more weakly than the WT-crRNA repeat.

We also investigated the effects of inserting one or two base pairs at the bottom of the stem-loop below the terminal C–G base pair. We observed a range of binding defects, although these were milder than those resulting from insertions at the top of the stem (**Fig. 2.13a**). Cleavage defects at saturating enzyme concentrations were highly dependent on sequence: whereas substrates containing one or two A–U base-pair insertions were cleaved at rates within 2-fold of the WT substrate, one or two G–C base-pair insertions resulted in ~50- and ~1,500-fold lower k_{obs} values, respectively (**Fig. 2.12c**). Partial RNase T1 digestions and RNA hydrolysis ladders revealed that these RNA constructs were cleaved above the inserted base pair(s) and just below the WT C–G base pair (**Fig. 2.13b**). Thus, Csy4-catalyzed cleavage likely requires prior melting of any additional secondary structures below the five base-pair stem, such that the WT stem is correctly positioned in the binding pocket and the guanosine containing the 2'-OH nucleophile can productively interact with Arg102 and Phe155. In support of this interpretation, A–U and U–A base pairs are thermodynamically less stable than G–C and C–G base pairs at the termini of RNA duplexes (Xia et al., 1998) and are likely to be more susceptible to transient fraying (Snoussi and Leroy, 2001), explaining the large magnitude of k_{obs} differences for these distinct insertions.

Collectively, these data indicate that beyond sequence-specific recognition of its crRNA repeat substrate, Csy4 is finely tuned to bind and cleave stem-loop substrates containing just five base pairs within the dsRNA region, through at least two distinct mechanisms. First, binding energy contributed by the arginine-rich helix requires an accessible major groove, which depends on the double-stranded stem being properly spaced between interaction sites at its base (e.g. with Arg102) and the loop sequence. Second, rapid cleavage requires the positioning of a terminal C–G base pair within the active site and prior disruption of any additional secondary structures below.

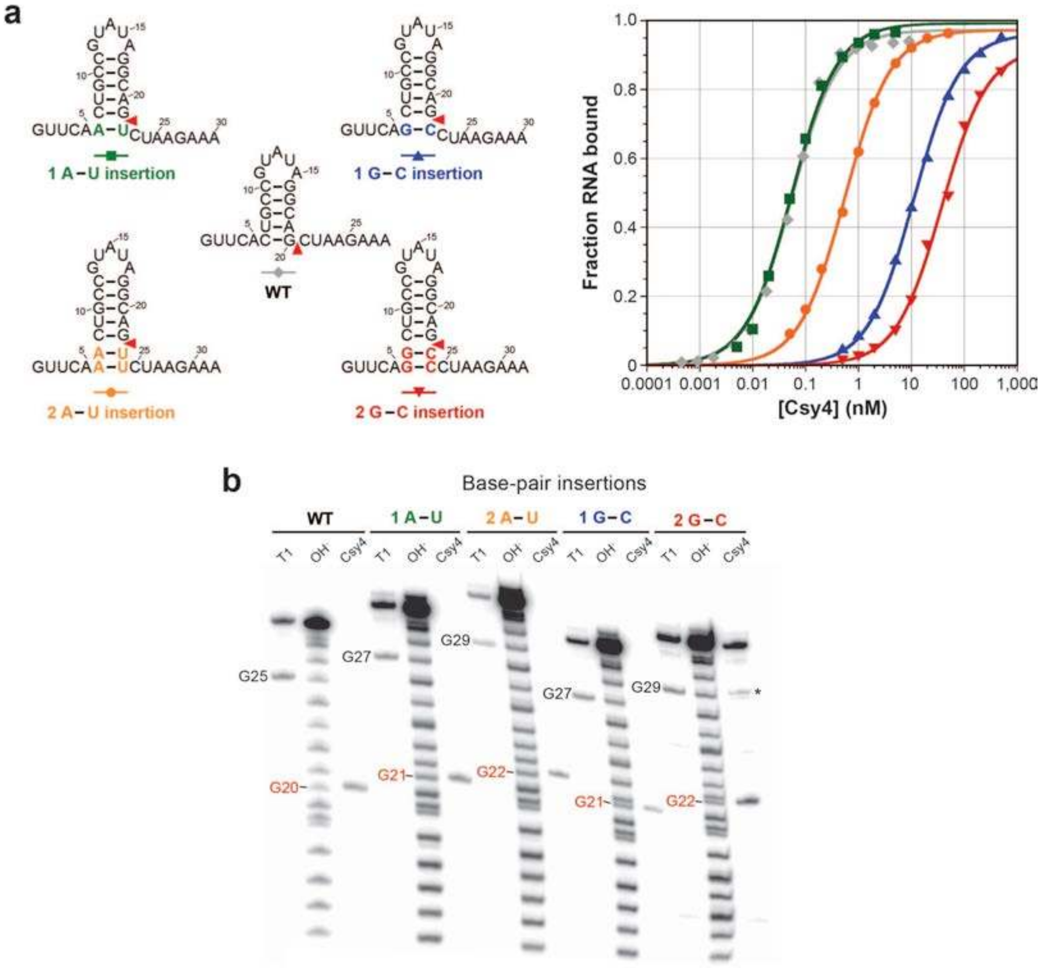


Figure 2.13 | Binding data and cleavage site mapping for base-pair insertion constructs. (a) RNA substrates containing base-pair insertions below the terminal C–G base pair were generated (left), and EMSAs were performed with Csy4-H29A (right). Binding defects were mildest for one or two A–U base-pair insertions (~1- and ~10-fold) and increased to ~200- and ~800-fold for one or two G–C base pairs, respectively. The cleavage site for each RNA substrate is indicated with a red triangle. **(b)** To experimentally determine cleavage sites, partial RNase T1 digestions and hydrolysis ladders were conducted and resolved by denaturing PAGE adjacent to Csy4 cleavage products. Nucleotides are numbered as in (a), and the guanosine residue directly upstream of the scissile phosphate is shown in red. In all cases, cleavage by Csy4 occurs just below the C–G base pair at the bottom of the WT five base-pair stem, above the base-pair insertion(s). * denotes a minor side product.

2.5 Discussion

The CRISPR-Cas adaptive immune system has evolved a sophisticated strategy for generating large libraries of short effector RNAs that target invasive genetic elements for destruction. Rather than requiring each crRNA to be individually transcribed, the repetitive CRISPR architecture allows large precursor transcripts to be successively processed by Cas endoribonucleases (in type I and III CRISPR systems) that are precisely tailored for specific recognition and cleavage of the invariant repeat sequence. Here we have defined the various molecular strategies employed by one such Cas enzyme – Csy4 (Cas6f) from *Pseudomonas aeruginosa* UCBPP-PA14 – to enable an impressive degree of affinity and specificity for its crRNA repeat substrate.

The Csy4:RNA complex is characterized by a ~50 pM equilibrium dissociation constant (K_d) and requires only a 16-nt stem-loop motif for tight binding. For comparison, U1A protein, MS2 coat protein and the N^{λ} protein bind their RNA substrates with K_d values of 50 pM, 2.6 nM and 5 nM, respectively (Cilley and Williamson, 1997; LeCuyer et al., 1995; van Gelder et al., 1993). High-energy interactions are mediated almost exclusively within the major groove of a double-stranded RNA stem-loop, a region of A-form helices that is generally refractory to protein contacts because of its inaccessibility. Prior work used chemical probing to demonstrate that the termini of dsRNA contain uncharacteristically wide major grooves (Weeks and Crothers, 1993), which explains how direct readout of A19 and G20 at their major groove edge is possible. Our data reveal that stable binding of the arginine-rich helix further up the stem is also highly sensitive to major groove accessibility, and that this requirement enables up to ~50,000-fold discrimination against hairpin substrates containing slightly longer stems. Four arginines within this α-helix are precisely positioned to contact multiple phosphates within the RNA backbone and adopt conformations reminiscent of the arginine fork first described for HIV-1 Tat protein by Frankel and co-workers (Calnan et al., 1991). This mode of multi-dentate interaction requires precise interatomic P-P distances, indicating that the network of hydrogen bonds formed by the arginine-rich helix depends on a very specific substrate conformation. Indeed, changes to the loop sequence or to the identity of base pairs in the upper part of the stem result in substantial binding defects, despite the general lack of base-specific contacts in this region. Substrate selection thus proceeds in large part via an indirect readout mechanism, whereby a particular RNA tertiary structure is recognized that is contingent on both primary sequence and the distinct helical geometry it imposes. Similar modes of substrate recognition have been described for a number of dsDNA-binding proteins (Otwinowski et al., 1988; Rohs et al., 2009).

Csy4 retains the same tight binding for both its substrate and product, and functions as a single-turnover catalyst due to potent product inhibition. These data strongly suggest that crRNA biogenesis in *P. aeruginosa* UCBPP-PA14 requires stoichiometric amounts of the processing endoribonuclease. Cleavage of the crRNA repeat substrate depends critically on the presence of a guanosine upstream of the scissile phosphate, independently of whether or not this nucleotide is base-paired, and is inhibited when additional secondary structure forms below the five base-pair stem. The k_{obs} defects we observed with Csy4-R102A and Csy4-F155A mutants indicate that the G20 base must be tightly locked in place within the enzyme active site in order to rapidly achieve chemical activation of the ribosyl 2'-OH. Other critical active site residues (Tyr176 and Ser148) have also been implicated in properly positioning the G20 ribose in an orientation that is compatible with nucleophilic attack on the downstream phosphodiester bond (Haurwitz et al., 2012).

We recently reported that, together with six copies of Csy3 and single copies of both Csy1 and Csy2, Csy4 and the mature crRNA assemble into a large ribonucleoprotein complex (Csy complex) that is responsible for target recognition during the interference stage of the CRISPR pathway (Wiedenheft et al., 2011b). Our data are consistent with a model where the Csy4-bound crRNA serves as a nucleation point for assembling the remainder of the complex, which does not form independently of RNA (Wiedenheft et al., 2011b). Interestingly, Cse3 (Cas6e), the CRISPR-specific endoribonuclease from type I-E CRISPR systems, also acts as a single-turnover enzyme (Sashital et al., 2011) and forms part of the downstream target recognition effector complex (Cascade) (Brouns et al., 2008; Jore et al., 2011a; Wiedenheft et al., 2011a). It is tempting to speculate that these related enzymes evolved to react stoichiometrically during pre-crRNA cleavage in order to ensure that the mature crRNA is not prematurely released into the cytoplasm but instead remains tightly sequestered by the Cas machinery. While this mechanistic feature may be intrinsic to certain Cas6 family members, it is not generalizable. Cas6 in type III-B CRISPR systems is not a component of the downstream effector complex (Cmr complex) (Hale et al., 2009), and Cas6 from type I-A CRISPR systems remains only loosely associated with the downstream effector complex (archaeal Cascade) (Lintner et al., 2011). Intriguingly, these differences correlate with the thermodynamic stability of hairpin structures encoded by CRISPR repeats typical of each subtype; repeats clustered based on sequence similarity that associate with type I-E and type I-F CRISPR systems encode highly stable RNA secondary structures, whereas those that associate with type I-A and type III-B systems encode RNAs predicted to be unstructured (Kunin et al., 2007).

CRISPR-specific endoribonucleases are unusual in that their biological function involves cleavage of a single, invariant substrate. As such, these enzymes have likely co-evolved with their target crRNA repeats to retain a high degree of substrate specificity, which serves to avoid spurious binding and/or cleavage of non-cognate RNAs inside the cell. The work presented here highlights the diverse molecular strategies exploited by *P. aeruginosa* Csy4 (Cas6f) to generate this selectivity while maintaining an extremely high-affinity interaction with its ligand. The potential benefits of these attributes for molecular biology applications will be exciting to explore further. Finally, future work will be needed to determine whether the underlying principles of RNA stem-loop recognition exhibited by Csy4 are conserved among other Cas6 family members.

Chapter 3

Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA

† Part of the work presented in this chapter has previously been published in the following research article: Haurwitz, R.E., Sternberg, S.H., and Doudna, J.A. (2012). Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA. EMBO J *31*, 2824–2832.

‡ Rachel Haurwitz designed experiments, performed biochemical assays, and solved the crystal structures. Samuel Sternberg performed the pH-rate profile analysis and helped design experiments. Jennifer Doudna supervised the project.

3.1 Abstract

CRISPR-Cas adaptive immune systems protect prokaryotes against foreign genetic elements. crRNAs derived from CRISPR loci base pair with complementary nucleic acids, leading to their destruction. In *Pseudomonas aeruginosa*, crRNA biogenesis requires the endoribonuclease Csy4, which binds and cleaves the repetitive sequence of the CRISPR transcript. Biochemical assays and three co-crystal structures of wild-type and mutant Csy4/RNA complexes reveal a substrate positioning and cleavage mechanism in which a histidine deprotonates the ribosyl 2′-hydroxyl pinned in place by a serine, leading to nucleophilic attack on the scissile phosphate. The active site catalytic dyad lacks a general acid to protonate the leaving group and positively charged residues to stabilize the transition state, explaining why the observed catalytic rate constant is ~10^4-fold slower than that of RNase A. We show that this RNA cleavage step is essential for assembly of the Csy protein-crRNA complex that facilitates target recognition. Considering that Csy4 recognizes a single cellular substrate and sequesters the cleavage product, evolutionary pressure has likely selected for substrate specificity and high-affinity crRNA interactions at the expense of rapid cleavage kinetics.

3.2 Introduction

Many prokaryotes resist viral infection by means of an adaptive immune system that relies on one or more CRISPR (clustered regularly interspaced short palindromic repeats) loci (Al-Attar et al., 2011; Barrangou et al., 2007; Haft et al., 2005; Karginov and Hannon, 2010; Makarova et al., 2006; 2011b; Wiedenheft et al., 2012). CRISPRs contain short virus- or plasmid-derived sequences that are positioned between copies of a repeated sequence (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005; Sorek et al., 2008). Small RNAs generated from the CRISPR locus (crRNAs) assemble with CRISPR-associated (Cas) proteins to form targeting complexes that can base pair with nucleic acids containing complementary sequences, leading to their destruction (Barrangou et al., 2007; Brouns et al., 2008; Garneau et al., 2010; Hale et al., 2009; Marraffini and Sontheimer, 2008).

The production of small RNAs from the CRISPR locus is a hallmark of CRISPR-based immunity (Marraffini and Sontheimer, 2010a; Terns and Terns, 2011). Precursor transcripts encompassing the full-length locus are cleaved within each repeat sequence to generate mature crRNAs that consist of a spacer sequence flanked by portions of the repeat sequence (Marraffini and Sontheimer, 2010a). CRISPR-Cas immune systems fall broadly into three types, in which similar tasks are accomplished using distinct sets of enzymes (Makarova et al., 2011b). In the type II CRISPR system, RNase III cleaves an RNA duplex formed by the CRISPR repeat and a *trans*-activating CRISPR RNA (tracrRNA) (Deltcheva et al., 2011), while in the type I and type III systems, a CRISPR-specific endoribonuclease cleaves the repeat elements in a sequence-specific fashion (Brouns et al., 2008; Carte et al., 2008a; 2010; Gesner et al., 2011; Haurwitz et al., 2010; Lintner et al., 2011; Sashital et al., 2011; Sternberg et al., 2012). We previously demonstrated that Csy4 (also known as Cas6f) is the enzyme responsible for crRNA production in CRISPR subtype I-F (Haurwitz et al., 2010).

Csy4 is a 21.4 kDa protein that recognizes its RNA substrate via sequence- and structure-specific contacts. It cleaves cognate RNAs at the 3′ end of a five-base-pair stem-loop, generating crRNAs comprising a unique spacer sequence flanked by 8 and 20 repeat-derived nucleotides on the 5′ and 3′ ends, respectively. Csy4 has equally tight affinity for both its substrate pre-crRNA and product crRNA, binding both with a 50 pM equilibrium dissociation constant (Sternberg et al., 2012). A single mature crRNA and one copy of Csy4 are components of the large ribonucleoprotein (RNP) Csy targeting complex (Wiedenheft et al., 2011b), but the mechanism of Csy complex assembly is currently unknown.

RNA cleavage by Csy4 is divalent metal ion-independent and requires chemical activation of a ribosyl 2′-hydroxyl for internal nucleophilic attack on the phosphodiester bond (Haurwitz et al., 2010). In the previously reported crystal structures of Csy4 bound to substrate RNA, we used a construct lacking the 2′-hydroxyl nucleophile upstream of the scissile phosphate to abrogate cleavage. The structures revealed three active site-proximal residues: Ser148, His29, and Tyr176. crRNA biogenesis was strongly inhibited by S148C and H29A mutants, while a Y176F mutant exhibited near wild-type activity. This mutational analysis led us to speculate that Ser148 plays a role in activating and/or positioning the 2′-hydroxyl for nucleophilic attack because it is located in close proximity to the 2′ carbon. Based on structural and biochemical evidence, we hypothesized that His29 may act as a proton donor for the 5′-hydroxyl leaving group because mutation of His29 to lysine partially preserved catalytic activity (Haurwitz et al., 2010).

Here we investigated the chemical mechanism of Csy4-catalyzed CRISPR RNA cleavage. Three crystal structures of wild-type and mutant Csy4 bound to product RNAs, coupled with kinetic analyses of mutant Csy4 cleavage rates, suggest a substrate positioning and cleavage mechanism in which Ser148 holds the 2′-hydroxyl nucleophile in place and His29 deprotonates it for attack on the scissile phosphate. The lack of both a general acid and positively charged residues in the active site explains the observed rate constants that are 10^3- to 10^4-fold slower relative to other metal ion-independent ribonucleases. We additionally demonstrate that CRISPR transcript processing by Csy4 is essential for subsequent formation of the Csy complex *in vivo*. Given the essential role Csy4 plays in formation of this targeting complex, slow cleavage rates in conjunction with highly accurate substrate selection likely ensure that cognate pre-crRNA substrates are cleaved with little to no off-target activity on other cellular RNAs.

3.3 Material and Methods

3.3.1 Protein expression and purification

Csy4 and single point mutants were expressed and purified as previously described (Haurwitz et al., 2010) with minor exceptions. Briefly, His_6-MBP-Csy4 or His_6-Csy4 fusion constructs (vectors pHMGWA and pHGWA, respectively (Busso et al., 2005)) were expressed in either *E. coli* BL21(DE3) cells or *E. coli* Rosetta 2(DE3) cells (Novagen). Following batch nickel resin affinity purification, cleavage with TEV protease, and a second nickel resin step, samples were separated on a single Superdex75 (16/60) size exclusion column (GE Healthcare) in 100 mM HEPES pH 7.5, 500 mM potassium chloride, 5% glycerol, and 1 mM TCEP. Proteins were then dialyzed against 100 mM HEPES pH 7.5, 150 mM potassium chloride, 5% glycerol, and 1 mM TCEP; concentrated; and stored at -80 ^{0}C.

3.3.2 RNA cleavage assays

Single-turnover cleavage experiments were performed at 24 ^{0}C in 20 mM HEPES, 100 mM potassium chloride, pH 7.2. Cleavage reactions were carried out in 60 ul volume containing 500 pM [5′-^{32}P]-crRNA repeat (5′-GUUCACUGGCCGUAUAGGCAGCUAAGAAA-3′), 400 nM Csy4, and 72 units RNasin Plus (Promega). At noted time points, 10 ul of the reaction were removed and quenched with 30 ul of acid phenol:chloroform (Ambion). 5 ul of the aqueous layer were mixed with 5 ul of formamide loading buffer and separated on a 15% denaturing polyacrylamide gel in 1X TBE running buffer. Cleaved and uncleaved RNAs were visualized by phosphorimaging and quantified using ImageQuant (GE Healthcare). For each sample, the percentage of RNA cleaved (intensity of cleaved RNA band divided by the sum of the cleaved and uncleaved bands) was plotted as a function of time. Plots were fit to an exponential decay curve using Kaleidagraph (Synergy Software). Rate constants are reported as k_{obs} because the rate-limiting step for cleavage is unknown. All cleavage assays were done in triplicate.

Cleavage reactions for pH-rate profiles were 55 ul in volume, contained 400 nM Csy4 and 500 pM [5′-^{32}P]-crRNA repeat, and were performed in 20 mM buffer, 100 mM potassium chloride, and 1 mM dithiothreitol (DTT). Buffers used were as follows: pH 4.0-6.5 – citric acid; pH 7.0-8.5 – 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES); pH 9.0-9.5 – *N*-cyclohexyl-2-aminoethanesulfonic acid (CHES); and pH 10.0-11.0 – *N*-cyclohexyl-3-aminopropanesulfonic acid (CAPS). Cleavage data were collected and analyzed as described above. pH-rate plots were fit to the following equation using Kaleidagraph (Synergy Software): $k_{obs} = (k_{obs,\text{MAX}} \times K_a) \div (K_a + [\text{H}^+])$, where K_a is an apparent acid dissociation constant and $[\text{H}^+]$ is the proton concentration.

3.3.3 Crystallization

Csy4/RNA complexes were generated and purified as previously described (Haurwitz et al., 2010). All crystals were grown at 18 ^{0}C using the hanging drop vapor diffusion method by mixing equal volumes (1 ul + 1 ul) of protein/RNA sample and reservoir solution. All complexes yielded plate-shaped crystals. Csy4S22C/product complex crystals were grown in 22% PEG4000, 120 mM sodium citrate pH 5.0, and 50 mM magnesium chloride. Csy4S148A/RNA complex crystals were grown in 20% PEG4000, 150 mM sodium citrate pH 5.0, and 100 mM magnesium chloride. Minimal complex crystals were grown in 21% PEG4000, 180 mM sodium citrate pH 5.0, and 100 mM magnesium chloride. Crystals were cryo-protected with reservoir solution containing 25% glycerol and flash frozen in liquid nitrogen. Minimal complex crystals were soaked with mother liquor supplemented with 2 mM ammonium metavanadate for 1.5 hours prior to cryo-protection and flash freezing.

3.3.4 Structure determination

Diffraction data were collected at beam lines 8.2.1 and 8.3.1 of the Advanced Light Source, Lawrence Berkeley National Laboratory. Datasets were processed in XDS (Kabsch, 2010). All three structures were determined using molecular replacement in Phaser (McCoy et al., 2007). Chains A and C (corresponding to protein and RNA, respectively) from the previously solved Csy4/substrate complex (PDB ID 2XLK) were used as search models for the product complex. The Csy4 protein (lacking the arginine-rich helix) and RNA (lacking the A5 nucleotide) models from the product complex were used as search models for the S148A and stem-loop complex structures. The models presented here resulted from iterative rounds of manual rebuilding in COOT (Emsley and Cowtan, 2004) and KiNG (Chen et al., 2009) and refinement in Phenix.refine (Adams et al., 2010). Riding hydrogens were included during refinement. Models were periodically validated using MolProbity (Chen et al., 2010).

All three complexes yielded crystals belonging to the *C*2 monoclinic space group that contained one complex per asymmetric unit. As in one of our previously published substrate structures (PDB ID 2XLI; (Haurwitz et al., 2010)), the RNA stems from neighboring complexes form coaxially stacked helices via an RNA kissing-loop interaction. The RNA helix and the associated arginine-rich alpha helix sit in a large solvent channel and exhibit elevated B factors. In the 2.0 Å resolution product structure, there is clear density for all amino acids in the arginine-rich helix, whereas in the 2.6 Å S148A structure and the 2.3 Å minimal complex structure, there is no density for the arginine-rich helix.

All structure figures were made using PyMol.

3.3.5 Csy complex in vivo reconstitution

The four Csy proteins were co-expressed from a polycistronic expression construct in which Csy3 had a His$_6$ fusion tag along with a synthetic CRISPR locus containing eight repeats and seven identical spacers in *Escherichia coli* BL21(DE3) cells as described previously (Wiedenheft et al., 2011b). Site-directed mutagenesis was used to introduce an alanine substitution at position 29 of the *csy4* gene. Briefly, protein expression was induced with addition of 0.5 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) at an optical cell density at 600 nm of ~0.5, followed by shaking at 18 °C overnight. Samples were lysed and clarified as previously reported (Wiedenheft et al., 2011b). Samples were affinity purified with nickel NTA resin (Qiagen) and incubated overnight with tobacco etch virus (TEV) protease to release the His$_6$ tag. Following a second nickel affinity step, samples were purified on a Superose 6 (10/300) size exclusion column (GE Healthcare) in 20 mM HEPES pH 7.5, 100 mM potassium chloride, and 1 mM TCEP.

3.3.6 Csy complex in vitro reconstitution

Csy3 was recombinantly expressed as a His$_6$-MBP fusion in *E. coli* BL21(DE3) cells. His$_6$-MBP-Csy1 and untagged Csy2 were co-expressed in *E. coli* BL21(DE3) cells. Both protein samples were subject to the same purification steps as described above for Csy4. Mature crRNAs were extracted from in vivo reconstituted Csy complex (see above) by acid phenol:chloroform extraction, chloroform extraction, and ethanol precipitation. Csy1/2, Csy3, Csy4, and crRNA were mixed in 1:6:1:1 molar ratios for a total of 160 ug of sample in 250 ul. Samples were subject to size exclusion chromatography as described in the previous section.

3.4. Results

3.4.1 His29 functions as a general base to activate the 2′-hydroxyl nucleophile

Our previous biochemical analysis of Csy4 implicated a serine residue as the general base or as important for substrate positioning and a histidine residue as a general acid in the transesterification reaction catalyzed by Csy4 (Haurwitz et al., 2010). In our previous experiments, we conducted single time-point (5 minute) reactions. This method may obscure mutants that have severe cleavage defects but nevertheless retain a low level of activity, and so to more accurately investigate the specific involvement of the proposed catalytic dyad and other active site-proximal residues during pre-crRNA cleavage (**Fig. 3.1a**), we performed quantitative single-turnover cleavage assays with various mutants and determined their corresponding first-order rate constants (**Fig. 3.1 & Table 3.1**). Alanine substitution of the active site histidine abolished all activity, indicating that His29 contributes an essential catalytic function. To further investigate its role, we evaluated the pH dependence of Csy4-catalyzed RNA cleavage. The resulting pH-rate profile (**Fig. 3.1d**) exhibits a sigmoidal shape and reveals that cleavage rates increase monotonically with pH. These data are consistent with the catalytic requirement of a single titratable residue having a $pK_a \approx 7.9$ that is active only in its deprotonated state. Consistent with our previous work, a Csy4 mutant with lysine substitution of His29 retains cleavage activity, albeit with ~130-fold slower kinetics than wild-type (**Fig. 3.1c** & **Table 3.1**).

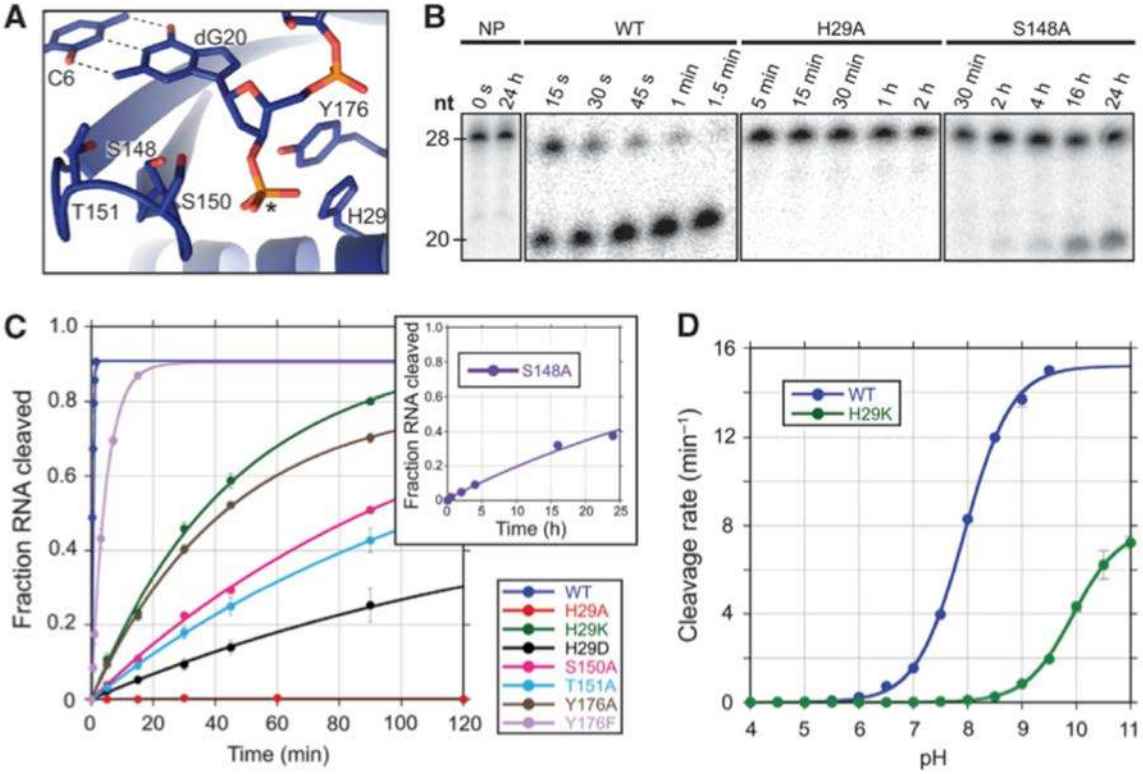
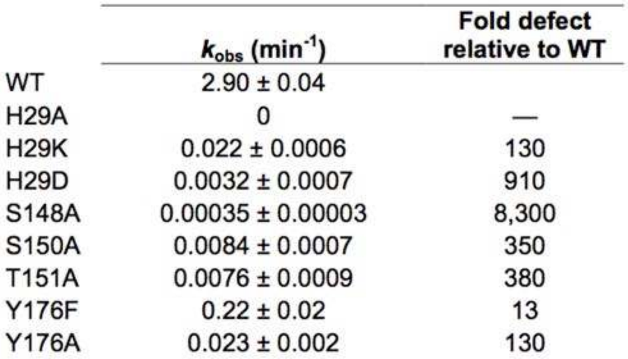


Figure 3.1 | Amino acid contributions to catalysis. (a) Csy4 active site from Csy4/substrate complex (PDB ID 2XLK). Active site residues are shown in stick format and the scissile phosphate is marked with an asterisk. The hydrogen bonds of the base pair between nucleotides C6 and dG20 are shown as dashed lines. **(b)** Representative single-turnover cleavage assays with wild-type and mutant Csy4. No protein (NP) controls shown at left. **(c)** Single-turnover cleavage analysis of wild-type and mutant Csy4. Data plotted are average of triplicate experiments and error bars represent the standard error of the mean (s.e.m.). Solid lines represent fits to an exponential equation. **(d)** pH-rate profile for wild-type and H29K Csy4. Rapid cleavage kinetics above pH 9.5 for wild-type Csy4 prevented accurate determination of the rate. Each data point is an average of three independent experiments and error bars represent the s.e.m. Data were fit according to the equation described in the **Materials and Methods**.

The pH-rate profile for RNA cleavage by the H29K mutant has the same shape as wild-type but is shifted to a higher pH (**Fig. 3.1d**; pK_a ≈ 9.9), in good agreement with the corresponding shift in pK_a of the imidazole and amino side groups of histidine and lysine, respectively. These data strongly suggest that catalytic activity requires His29 to be in its deprotonated form, and that this residue functions as a general base during cleavage by activating the 2′-hydroxyl nucleophile through proton abstraction. Substitution of His29 with aspartate, whose side chain is negatively charged at physiological pH, resulted in a functional enzyme, further supporting the role of His29 as the general base (**Fig. 3.1c**). Direct proton abstraction would require the His29 side chain to be positioned proximally to the G20 2′-hydroxyl, but in the previously published Csy4/substrate structures (Haurwitz et al., 2010), the His29 side chain interacts instead with the scissile phosphate and is not within hydrogen bonding distance of the expected location of the 2′-hydroxyl. Those crystals were grown at acidic pH ranges (~4.6 – 5) where the His29 side chain is likely to be protonated and Csy4 is catalytically defective (**Fig. 3.1d**). Thus, the previously observed interaction between the scissile phosphate and His29 side chain may result artificially from the acidic pH of the crystallization conditions (see below).

Alanine substitution of Ser148 decreased the cleavage rate ~8,000-fold relative to wild-type (**Fig. 3.1** & **Table 3.1**), suggesting that this residue plays a critical role in substrate binding, positioning, or cleavage chemistry (see below). Mutation of Tyr176 to phenylalanine or alanine reduced the cleavage rate only ~13-fold and ~130-fold, respectively (**Fig. 3.1c** & **Table 3.1**). The side chain of Tyr176 points into the active site and stacks on top of the His29 imidazole group; mutation to phenylalanine likely disrupts any role the phenolic hydroxyl plays in substrate binding, whereas mutation to alanine could also disrupt His29 positioning. Alanine substitution of either Ser150 or Thr151, both located in the active site loop, reduced the cleavage rate ~350-fold, suggesting these residues may play a role in either direct binding of the RNA substrate or by forming a network of hydrogen-bonding interactions that orient the side chain of Ser148.

Table 3.1 | Observed cleavage rates for WT and mutant Csy4.



	k_{obs} (min^{-1})	Fold defect relative to WT
WT	2.90 ± 0.04	
H29A	0	—
H29K	0.022 ± 0.0006	130
H29D	0.0032 ± 0.0007	910
S148A	0.00035 ± 0.00003	8,300
S150A	0.0084 ± 0.0007	350
T151A	0.0076 ± 0.0009	380
Y176F	0.22 ± 0.02	13
Y176A	0.023 ± 0.002	130

*Rates are the averages of three independent experiments, and errors represent the standard error of the mean.

3.4.2 The Csy4 active site constrains the G20 ribose in the C2′-endo sugar pucker

To determine how Csy4 interacts with the 2′-hydroxyl nucleophile, we crystallized a Csy4/RNA product complex comprising Csy4S22C and a 19-nucleotide RNA product that was generated by endoribonucleolytic cleavage of a 20-nucleotide substrate RNA (**Fig. 3.2**). Csy4S22C is a mutant of Csy4 that retains wild-type activity and yields better diffracting crystals (Haurwitz et al., 2010). Crystals of this complex diffracted x-rays to 2.0 Å resolution, and the structure was solved by molecular replacement using the previous substrate complex structure (PDB ID 2XLK) as a search model (**Table 3.2**). The structure of Csy4 in this product complex is similar to that observed in the previously published substrate complex (PDB ID 2XLK; RMSD = 0.580 Å for 158 residues) (**Fig. 3.2 & 3.3**). Additionally, the crRNA hairpins of the product and substrate RNAs are bound to Csy4 in the same location and align with an RMSD of 0.495 Å. We observed clear density for a 3′-phosphate (**Fig. 3.4**), consistent with previous mass spectrometry results that identified the termini of Csy4 cleavage products as a 5′-hydroxyl and 3′-phosphate (Wiedenheft et al., 2011b). Additionally, we observe that nucleotide A5, a single-stranded nucleotide immediately upstream of the stem-loop, makes two hydrogen-bonding contacts in a base-specific fashion with the peptide backbone of Leu139 (Sternberg et al., 2012).

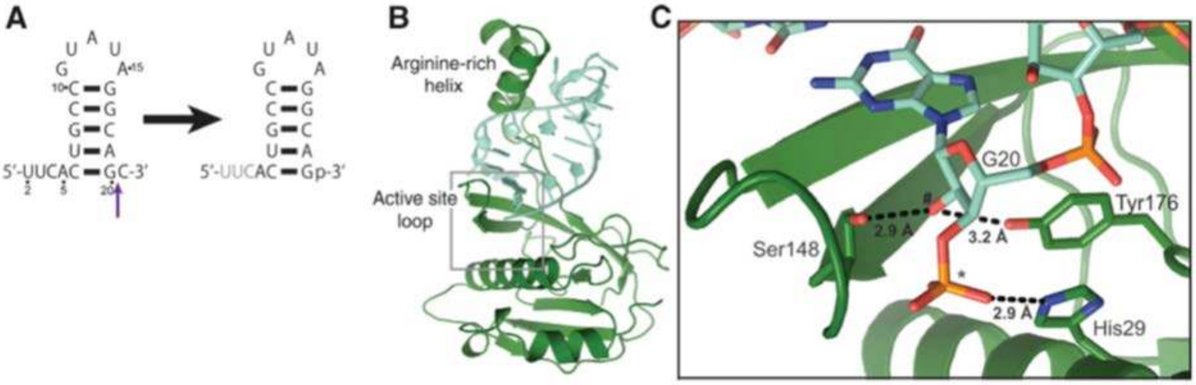


Figure 3.2 | Crystal structure of Csy4/product RNA complex at 2.0 Å resolution. (a) Shown at left is the substrate RNA used to generate the protein/RNA complex. Cleavage by Csy4 (purple arrow) produces the product RNA (right) present in the crystal structure. Gray lettering denotes nucleotides for which there was no corresponding electron density and therefore could not be modeled. **(b)** Overall structure of Csy4S22C (dark green) bound to product RNA (light green). Electron density was well-defined for all 187 amino acids of Csy4 and 16 of the 19 nucleotides in the product RNA. (C) Detailed view of the Csy4 active site (gray box, in B). The 2'-hydroxyl nucleophile is marked with a pound sign and the scissile phosphate is marked with an asterisk. RNA/protein hydrogen-bonding interactions are marked with dashes.

Unique to the product complex structure is the presence of the 2′-hydroxyl nucleophile in the active site (**Fig. 3.2c**), which was readily apparent in the molecular replacement solution (**Fig. 3.4a**). Upon modeling a ribonucleotide into the active site, we observed that the electron density was inconsistent with a ribose in the C3′-endo conformation but was fit well with a ribose in the C2′-endo form (**Fig. 3.4**). The 2′-hydroxyl nucleophile is positioned between the side chains of Ser148 and Tyr176, both of which are within hydrogen-bonding distance (2.8 Å and 3.2 Å) (**Fig. 3.2c**), suggesting that these interactions may force the G20 ribose to adopt the C2′-endo sugar pucker observed in the crystal structure. In-line attack of a 2′-hydroxyl nucleophile on the adjacent scissile phosphate requires a locally extended RNA backbone (Yang,

2011) and does not proceed when the sugar pucker is C3′-endo. The observation of a C2′-endo sugar pucker in the Csy4 active site is therefore representative of the extended conformation that would be required for cleavage to proceed.

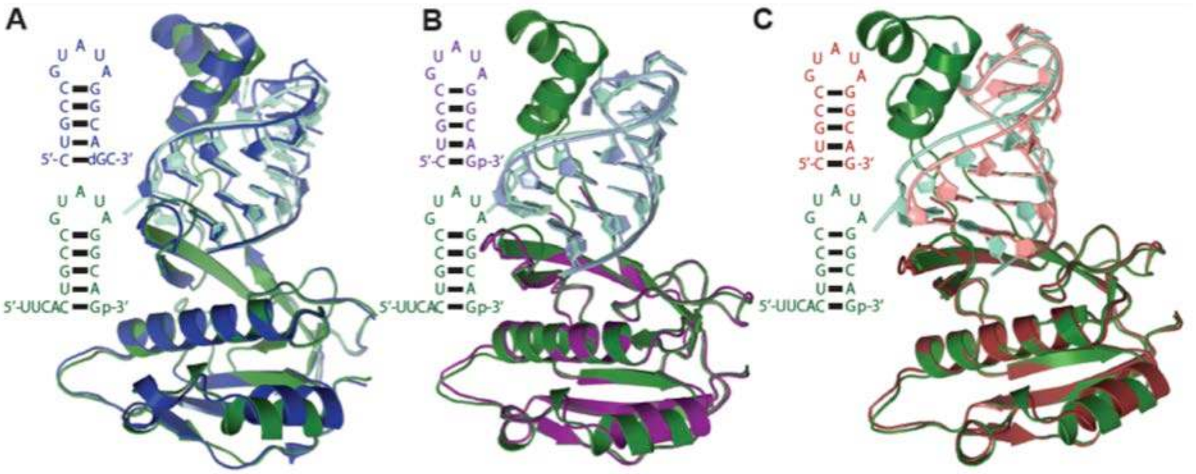


Figure 3.3 | The overall folds of the Csy4/product complexes are highly similar to each other and the previously published Csy4/substrate complex. (a) Chains A and C from the substrate complex (dark blue, PDB ID 2XLK) align with the protein and RNA molecules from the product complex (dark and light green) with an RMSD of 0.431 Å and 0.519 Å over 811 and 214 atoms, respectively. Depicted at left is the RNA content of the crystal structures. **(b)** The protein and RNA molecules from the product and S148A mutant structures (dark and light purple) align with an RMSD of 0.309 Å and 0.526 Å over 815 and 270 atoms, respectively. Depicted at left is the RNA content of the crystal structures. **(c)** The protein and RNA molecules from the product and minimal complex structures (dark red and pink) align with an RMSD of 0.346 Å and 0.499 Å over 843 and 263 atoms, respectively. The RNA stem from the minimal structure exhibits a rigid body rotation with respect to the Csy4 molecule as compared to the other structures shown. Depicted at left is the RNA content of the crystal structures.

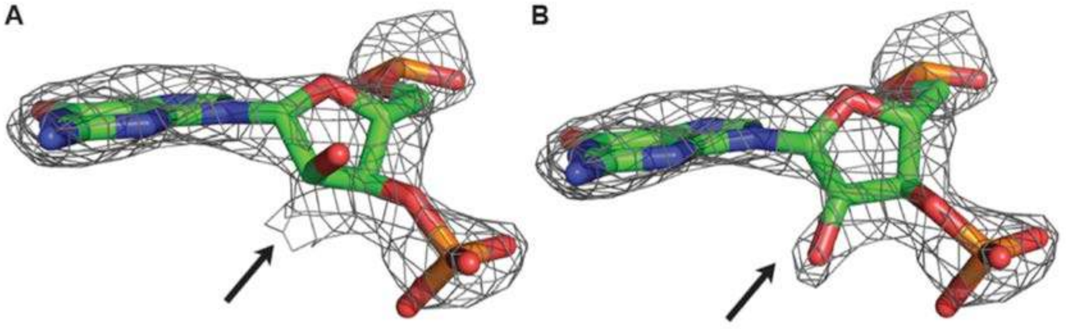


Figure 3.4 | The G20 ribose adopts the C2′-endo conformation in the active site of the product complex. We performed a molecular replacement experiment on the product complex native dataset using the A and C chains (protein and RNA, respectively) from the previously determined Csy4-substrate complex (PDB ID 2XLK) as search models. We removed the G20 nucleotide and C21 phosphate from the RNA search model in order to minimize model bias. A $2F_O$-F_C electron density map calculated from the molecular replacement solution phases contoured at 1 σ is displayed in gray mesh. Density for the G20 nucleotide and C21 phosphate is readily apparent. We manually built the G20 nucleotide into the electron density using a model with either a C3′-endo **(a)** or C2′-endo sugar pucker **(b)**. The nucleotide modeled with a C3′-endo sugar pucker does not accurately account for all of the observed density (black arrow), whereas the nucleotide modeled with a C2′-endo sugar pucker agrees well with the observed electron density.

3.4.3 Ser148 positions the RNA for cleavage

Our cleavage assays demonstrated that the S148A mutation is far more deleterious to catalysis than the Y176A mutation, suggesting that Ser148 is the primary residue responsible for positioning the 2′-hydroxyl and maintaining the requisite extended phosphate backbone conformation. The Tyr176 side chain likely plays a redundant role in stabilization of the C2′-endo conformation and may be more important for positioning His29. To test this hypothesis, we crystallized a complex of Csy4S148A and a 16-nucleotide substrate RNA (**Fig. 3.5**). The resulting 2.6 Å structure (**Fig. 3.5b** & **Table 3.2**), solved by molecular replacement, likely contained a mixture of substrate and product RNAs (16- and 15-nucleotides in length, respectively) due to the slow rate of Csy4S148A-catalyzed cleavage. The C21 nucleotide, immediately downstream of the scissile phosphate, is disordered when present and electron density for this nucleotide is therefore not observed (Haurwitz et al., 2010). The Csy4S148A protein structure is similar to that of wild-type Csy4 (RMSD = 0.403 Å over 134 residues), and the RNA hairpin is bound to the S148A mutant in the same location as observed in the product structure (RMSD = 0.425 Å; **Fig. 3.3b**). However, the active site ribose adopts a C3′-endo sugar pucker in this case, thereby repositioning the 2′-hydroxyl nucleophile 5.5 Å away from the Tyr176 side chain (**Fig. 3.5c**). We conclude that the Tyr176 side chain is insufficient to maintain the C2′-endo sugar pucker in the absence of Ser148, suggesting that the large catalytic defect for the S148A mutant may result from the Csy4 active site relying on the inherent sugar pucker interconversion rate in order for the substrate phosphate backbone to be properly extended for cleavage.

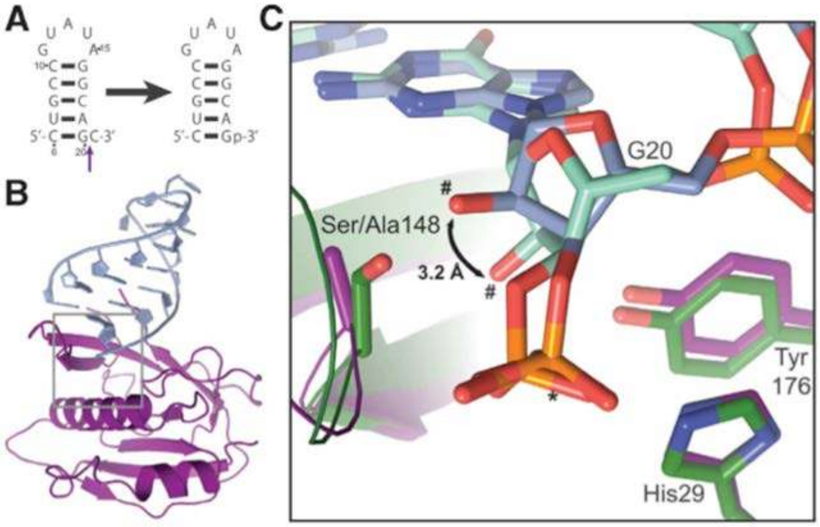
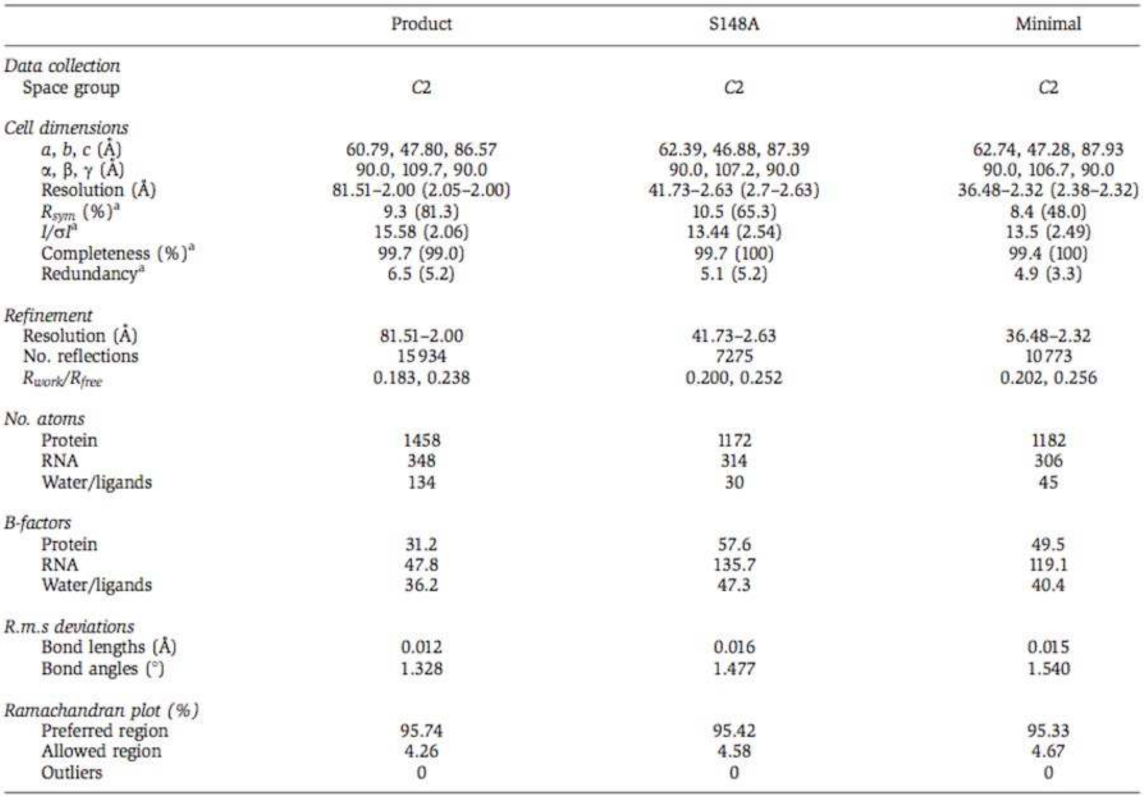


Figure 3.5 | Crystal structure of Csy4S148A/RNA complex at 2.6 Å resolution. (a) Shown at left is the substrate RNA used to generate the protein/RNA complex. Cleavage by Csy4 (purple arrow) produces product RNA (right). Because of the slow cleavage rate of the S148A mutant, crystals likely contained a mixed population of substrate and product RNAs. **(b)** Overall structure of Csy4S148A (dark purple) and RNA (light purple). 153/187 amino acids and 14/15 nucleotides could be modeled into the electron density. The amino acids composing the arginine-rich helix are among those for which there is little to no electron density. **(c)** Superposition and close-up of product complex (green) and S148A complex (purple) active sites (gray box, in (b)). The double-headed black arrow highlights the 3.2 Å change in 2'-hydroxyl location between the two structures. The two 2'-hydroxyl nucleophiles are labeled with pound signs and the scissile phosphates are indicated with an asterisk.

Table 3.2 | Data collection and refinement statistics.



	Product	S148A	Minimal
Data collection			
Space group	C2	C2	C2
Cell dimensions			
a, *b*, *c* (Å)	60.79, 47.80, 86.57	62.39, 46.88, 87.39	62.74, 47.28, 87.93
α, β, γ (Å)	90.0, 109.7, 90.0	90.0, 107.2, 90.0	90.0, 106.7, 90.0
Resolution (Å)	81.51–2.00 (2.05–2.00)	41.73–2.63 (2.7–2.63)	36.48–2.32 (2.38–2.32)
R_{sym} (%)[a]	9.3 (81.3)	10.5 (65.3)	8.4 (48.0)
$I/\sigma I$[a]	15.58 (2.06)	13.44 (2.54)	13.5 (2.49)
Completeness (%)[a]	99.7 (99.0)	99.7 (100)	99.4 (100)
Redundancy[a]	6.5 (5.2)	5.1 (5.2)	4.9 (3.3)
Refinement			
Resolution (Å)	81.51–2.00	41.73–2.63	36.48–2.32
No. reflections	15934	7275	10773
R_{work}/R_{free}	0.183, 0.238	0.200, 0.252	0.202, 0.256
No. atoms			
Protein	1458	1172	1182
RNA	348	314	306
Water/ligands	134	30	45
B-factors			
Protein	31.2	57.6	49.5
RNA	47.8	135.7	119.1
Water/ligands	36.2	47.3	40.4
R.m.s deviations			
Bond lengths (Å)	0.012	0.016	0.015
Bond angles (°)	1.328	1.477	1.540
Ramachandran plot (%)			
Preferred region	95.74	95.42	95.33
Allowed region	4.26	4.58	4.67
Outliers	0	0	0

[a]Values in parentheses denote highest resolution shell.

3.4.4 His29 may interact directly with the 2′-hydroxyl nucleophile

As described above, all of the Csy4/RNA crystal structures result from crystals grown at pH 4.6 – 5. To determine what interactions His29 may make in the absence of the potentially pH-induced interaction with the scissile phosphate, we crystallized a complex of Csy4 and a 15-nucleotide RNA composed of only the crRNA hairpin with a 3′-hydroxyl terminus (**Fig. 3.6**). The 2.3 Å resolution structure of this complex (hereafter called the minimal structure) once again revealed a Csy4 conformation similar to that observed previously (RMSD = 0.466 Å over 140 residues; RNA superposition RMSD = 0.483 Å) (**Fig. 3.6b, Fig. 3.3c** & **Table 3.2**). While the locations of the Tyr176 and His29 side chains are nearly identical between the product and minimal structures, the G20 nucleotide and the active site loop that contains Ser148 shift 3.4 Å and 2.5 Å between the two structures, respectively (**Fig. 3.6c**). The G20 ribose is in the C2′-endo conformation, and the 2′-hydroxyl nucleophile is 3.7 Å away from both the His29 and Tyr176 side chains (**Fig. 3.6d**). The lack of a 3′-phosphate results in significant disorder in the active site loop as is evidenced by a lack of density for residue 149 and for the side chains of nearly all of the active site loop residues (**Fig. 3.6d**). This structure provides evidence that there is flexibility in the location of RNA within the Csy4 active site because in previous structures, the

His29 sidechain is greater than 5 Å from the G20 2′-hydroxyl. This flexibility likely facilitates His29 activating the 2′-hydroxyl nucleophile via proton abstraction.

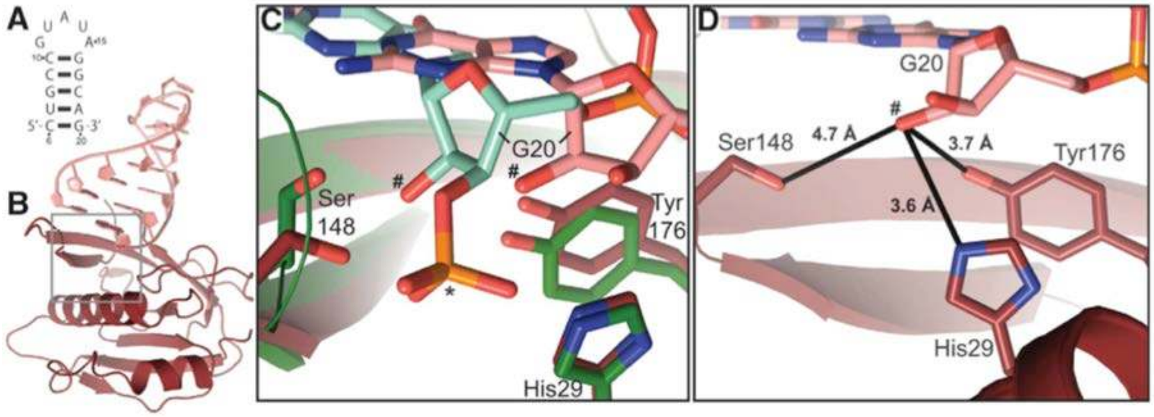


Figure 3.6 | Crystal structure of Csy4/RNA minimal complex at 2.3 Å resolution. (a) The stem-loop RNA used for co-crystallography lacks a 3'-phosphate. **(b)** Overall structure of Csy4 (dark red) and stem-loop RNA (pink). 151/187 amino acids and all 15 RNA nucleotides could be modeled into the electron density. Electron density for the active site loop is severely broken, and a dashed line indicates its approximate location. There is no electron density for the arginine-rich helix. **(c)** Superposition and detailed view of product complex (green) and minimal complex (red) active sites (gray box, in (b)). The scissile phosphate belonging to the product complex is marked with an asterisk and the two 2'-hydroxyl nucleophiles are marked with pound signs. (D) Magnified view of the minimal complex active site. Black lines indicate the distances between active site residues and the 2'-hydroxyl nucleophile.

3.4.5 Csy complex formation requires Csy4-catalyzed cleavage of CRISPR transcripts

Recent work has demonstrated that Csy4 associates with three other Cas proteins (Csy1-3) and a single copy of crRNA to form the Csy complex, which targets complementary nucleic acids (Wiedenheft et al., 2011b). To determine whether pre-crRNA cleavage by Csy4 is necessary for complex formation, we co-expressed Csy1-3 and a pre-crRNA with either wild-type Csy4 or the catalytically inactive mutant, Csy4H29A (Haurwitz et al., 2010), in *Escherichia coli* BL21(DE3) cells. The Csy complex was affinity purified via a 6X-histidine tag appended to the N-terminus of Csy3, followed by size exclusion chromatography. Co-expression of the wild-type proteins and pre-crRNA yielded an RNP with an estimated molecular mass of ~350 kilodaltons (**Fig. 3.7**), in agreement with our previous work (Wiedenheft et al., 2011b). Substitution of catalytically inactive Csy4 in the co-expression experiment resulted in the purification of only Csy3, which was not associated with a crRNA (**Fig. 3.7**). Csy3 over-expressed on its own in *E. coli* BL21(DE3) cells purifies as both a large oligomeric complex containing non-specific RNA and as a nucleic acid-free monomer (unpublished observations), similar to the two peaks observed for Csy3 co-expressed with mutant Csy4. To ensure that Csy4H29A is defective only in catalysis and not in its ability to interact with other Csy complex components, we mixed together Csy complex components that were individually recombinantly purified and evaluated the mixtures by size exclusion chromatography. Adding either wild-type or H29A Csy4 to Csy1-3 and a mature crRNA resulted in Csy complex formation (**Fig. 3.8**), suggesting that the Csy4H29A mutant is defective only for catalysis and not for interaction with other Csy complex components, and that catalysis is a necessary precursor to complex formation.

Taken together with previous work demonstrating that Csy complex assembly does not proceed in the absence of RNA (Wiedenheft et al., 2011b), we conclude that Csy4-catalyzed biogenesis of mature crRNAs with fully processed termini is necessary for stable Csy complex formation.

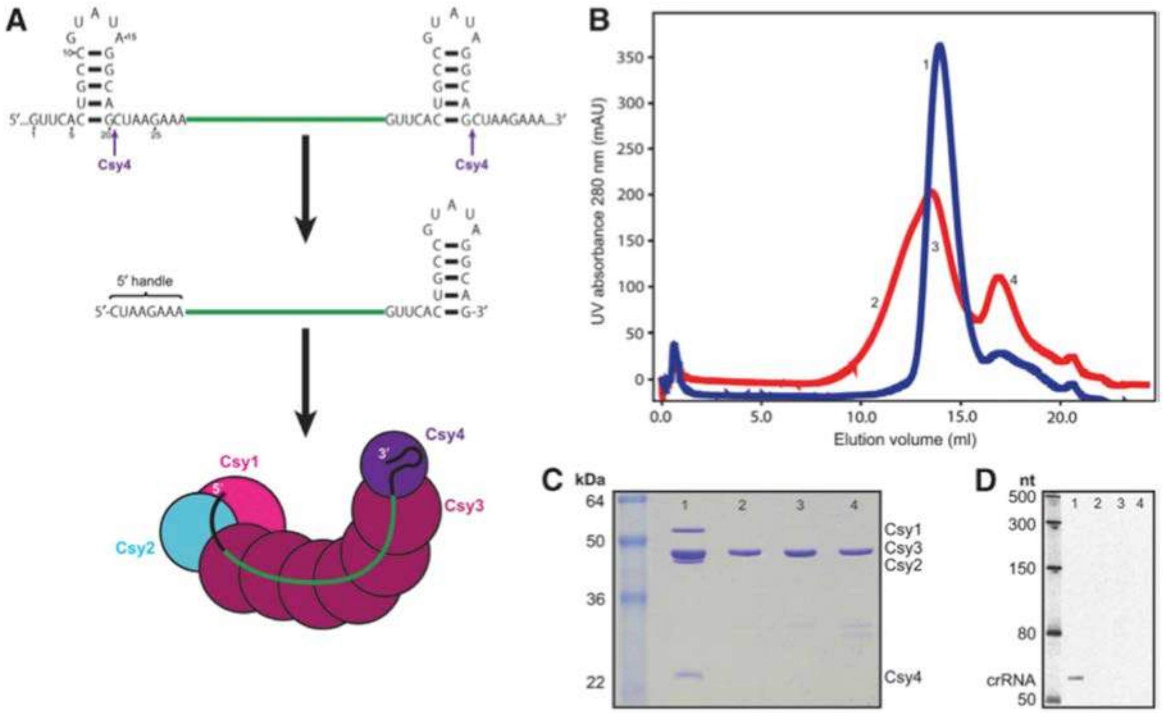


Figure 3.7 | Csy4 cleavage of pre-crRNA is required for Csy complex formation. (a) Schematic depicting pre-crRNA cleavage by Csy4 and formation of the Csy CRISPR ribonucleoprotein (crRNP) complex. The CRISPR repeat and spacer sequence are in black and green, respectively. Cleavage sites are denoted with purple arrows. **(b)** Superose 6 gel filtration column elution profiles of affinity-purified Csy1, Csy2, His$_6$-Csy3, and pre-crRNA co-expressed with wild-type (blue) or H29A (red) Csy4. **(c)** Coomassie blue-stained 12% SDS–PAGE showing protein components of the superose 6 fractions for wild-type (lane 1) and H29A (lanes 2–4, as noted in (b)) Csy4 co-expression assays. **(d)** SYBR Gold-stained 15% denaturing PAGE showing phenol:chloroform extracted nucleic acids from superose 6 fractions (from (b)).

3.5 Discussion

The production of crRNAs is central to CRISPR-mediated adaptive immunity in prokaryotes. The three crystal structures of Csy4/RNA complexes and quantitative cleavage assays presented here reveal an unexpected endoribonuclease active site in which a serine residue constrains the nucleophile-containing ribose in the C2′-endo sugar pucker and a histidine residue serves as the general base to activate the 2′-hydroxyl nucleophile. Unlike RNase A and other well-studied metal ion-independent nucleases, the Csy4 active site lacks a general acid and positively charged residues near the active site that would lower the energetic barrier to the transition state, resulting in correspondingly slow cleavage rates. We propose that upon binding a pre-crRNA substrate, the Ser148 residue rearranges the G20 ribose into the C2′-endo conformation, providing the correct geometry for His29 to abstract a proton from the 2′-hydroxyl nucleophile and enable nucleophilic attack of the scissile phosphate. The resulting 2′,3′-cyclic phosphate terminus is likely opened to a 3′-phosphate via hydrolysis by a water. Csy4 then retains its crRNA product (Sternberg et al., 2012) and serves as the nucleation point for Csy complex formation.

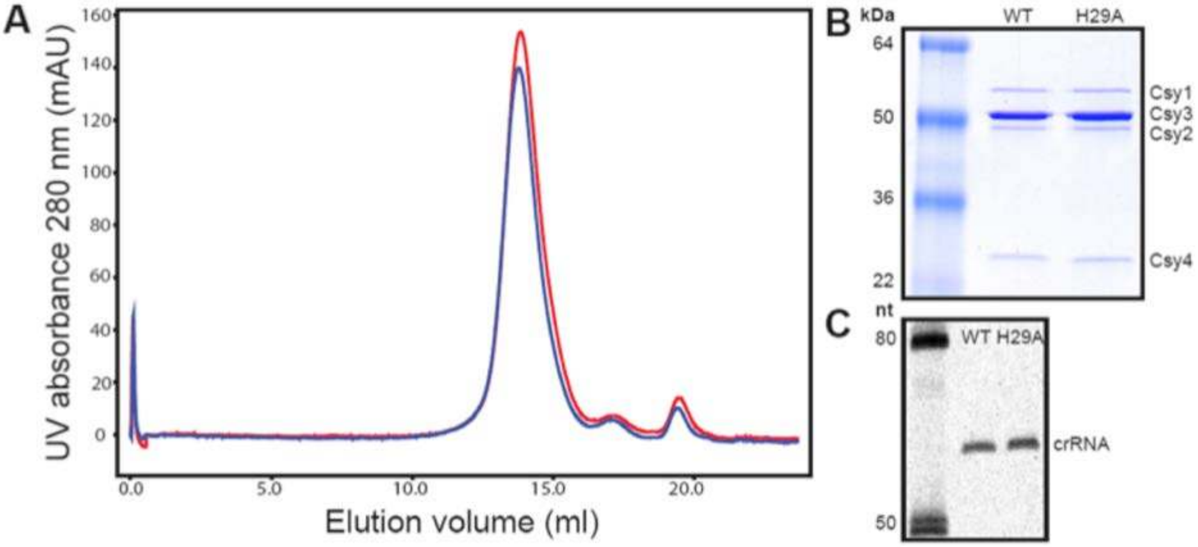


Figure 3.8 | Csy4H29A is competent for assembly into the Csy complex. (a) Superose 6 gel filtration column elution profiles of recombinantly assembled Csy complex containing individually purified Csy1/Csy2 heterodimer, Csy3 monomer, mature crRNA, and wild-type (blue) or H29A (red) Csy4. **(b)** Coomassie blue-stained 12% SDS-PAGE showing protein components of the superose 6 fractions for wild-type (WT) and H29A mutant (H29A) Csy4 *in vitro* assembly assays (as noted in (a)). **(c)** SYBR Gold-stained 15% denaturing PAGE showing phenol:chloroform extracted nucleic acids from superose 6 fractions (from (a)).

We observe that the G20 ribose in the wild-type Csy4 active site adopts the C2′-endo sugar pucker. The C2′-endo conformation is generally rare in double-stranded RNA but is overrepresented in catalytic active sites and RNA tertiary interactions (Mortimer and Weeks, 2009). In the Csy4 active site, Ser148 and Tyr176 likely interact directly with the 2′-hydroxyl nucleophile via hydrogen bonding, restraining the ribose ring in the C2′-endo conformation. Mutation of Ser148 to alanine slows cleavage nearly 8,000-fold and allows the G20 ribose to retain the C3′-endo conformation. We propose that this significant cleavage rate defect may arise from a particularly slow rate of C2′-endo/C3′-endo interconversion at the G20 ribose in the absence of the Ser148 side chain. While most RNA sugars interconvert between the C2′- and C3′-endo conformations on a microsecond to millisecond time scale (Johnson and Hoogstraten, 2008), a discrete set of C2′-endo nucleotides has been observed to experience local dynamics with half-lives on the order of 10-100 seconds, significantly slower than other local RNA conformational changes (Gherghe et al., 2008; Mortimer and Weeks, 2009). For example, the folding rate of bacterial RNase P RNA is limited by the sugar pucker interconversion of a single RNA nucleotide from C3′-endo to C2′-endo, which occurs at a rate of ~0.24 min^{-1} (Mortimer and Weeks, 2009). Consistent with the observation that members of this class of slow interconverting C2′ endo-containing ribonucleotides are partially constrained by hydrogen-bonding or base-stacking interactions (Gherghe et al., 2008), the G20 nucleotide base pairs with C6, hydrogen-bonds with Arg102 on the major groove face, and stacks below A19 and above Phe155. We hypothesize that G20 belongs to this unusual class of C2′-endo containing nucleotides and propose that the ~8,000-fold defect in observed cleavage rate of the S148A mutant is due in large part to the extremely slow sugar pucker interconversion dynamics of the G20 nucleotide. However, we cannot rule out that the hydrogen bonding interaction between Ser148 and the 2′-hydroxyl also contributes to nucleophile activation.

The observed rate of cleavage for wild-type Csy4 (~3 min^{-1}) is orders of magnitude slower than that of other well-characterized RNases. For example, RNase A enzymes from a variety of organisms cleave RNA substrates with apparent single-turnover rate constants of 910 to 40,500 min^{-1} (Katoh et al., 1986), and the colicin E5 ribonuclease from *E. coli* cleaves minimal substrates with a k_{cat} of ~5,000 min^{-1} (Ogawa et al., 2006). In fact, Csy4 has an observed cleavage rate similar to ribozyme-catalyzed RNA cleavage rate constants, which are typically <2 min^{-1} (Zamel et al., 2004). Ribozymes perform the same transesterification reaction as protein RNases (Cochrane and Strobel, 2008), but are thought to be significantly slower because they typically lack general acids and bases with pK_a values close to neutral pH (Yang, 2011). The well characterized metal-independent RNase families of RNase A, RNase T1, and RNase T2 contain catalytic cores composed of a histidine pair; a glutamate and histidine; and a glutamate, lysine, and three histidines, respectively (Yang, 2011). Like many of these protein RNases, the Csy4 active site contains a histidine general base, but it appears to lack a general acid as there is no chemically appropriate residue positioned proximal to the 5′-hydroxyl leaving group. Consistent with this observation is the sigmoidal shape of the Csy4 pH-rate profile (**Fig. 3.1d**). Whereas RNase A exhibits a bell-shaped pH-rate profile indicative of a cleavage mechanism that relies on two titratable residues (Raines, 1998), the Csy4 pH-rate profile is consistent with only a single titratable residue that is likely to be His29.

An additional hallmark of metal ion-independent RNases is stabilization of the pentacovalent transition state by one or more positively charged residues (Cochrane and Strobel, 2008). Like ribozymes, which lack functional groups that are positively charged at a neutral pH, Csy4 does not have any positively charged residues in or surrounding the active site. We hypothesize that Csy4 compensates for a lack of stabilizing positive charges by making additional hydrogen bonds to the transition state, analogous to the hairpin ribozyme, which makes 2-3 more contacts to the transition state than precursor or product RNAs (Cochrane and Strobel, 2008; Rupert and Ferré-D'Amaré, 2001; Rupert et al., 2002). This is consistent with the ~350-fold effect on cleavage observed for alanine substitution of Ser150 or Thr151, which lie in the active site loop and participate in a hydrogen bonding network that can include Ser148 and the scissile phosphate (**Fig. 3.9**). Through this network, Ser150 and Thr151 may aid in the stabilization of the pentacovalent transition state.

Using an *in vivo* assembly assay, we found that crRNA processing by the endoribonuclease Csy4 is essential to the stable formation of crRNA-containing targeting complexes that bind to complementary nucleic acids and trigger their degradation. Because Csy complexes do not stably form on unprocessed pre-crRNA, we hypothesize that the formation of the mature Csy crRNP requires a free 5′ terminus generated by Csy4-catalyzed cleavage. Mature crRNAs across multiple CRISPR types contain 8-nts of repeat-derived sequence at the 5’ end (Brouns et al., 2008; Carte et al., 2008a; Hale et al., 2009; Marraffini and Sontheimer, 2008), and it has been proposed that these sequences, termed the 5’ handle, may serve as Cas protein binding sites (Terns and Terns, 2011; Wiedenheft et al., 2012). For example, the 5' handle forms a hook-like structure in the crRNP from *E. coli* K12 (Cascade) that correlates with termination of the ribonucleoprotein filament (Wiedenheft et al., 2011a). We speculate that the 5′ handle of the mature crRNA in *P. aeruginosa* recruits one or more Csy proteins to the nascent RNP. The requirement for a free crRNA 5′ terminus during complex formation would therefore point to specific recognition of the 5′ handle in the assembly of Cas protein complexes.

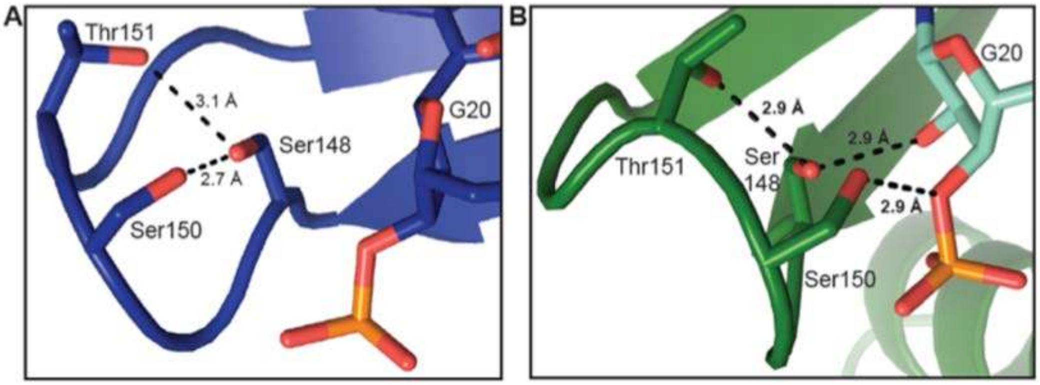


Figure 3.9 | Active site loop residues have the potential to form a hydrogen bonding network with one another and the bound RNA. Detailed view of the active site loops from the **(a)** substrate (PDB ID 2XLK) and **(b)** product complexes. Dashed lines indicate hydrogen-bonding interactions.

These observations, along with recent work demonstrating a very tight crRNA binding affinity by Csy4 (50 pM) (Sternberg et al., 2012), have led us to conclude that Csy4 evolved as a finely tuned RNA binding protein while retaining only modest cleavage kinetics. Similarly, the CRISPR type I-E endoribonuclease (referred to as Cas6e, Cse3, or CasE) exhibits relatively slow cleavage kinetics (~5 min^{-1}) and tight substrate and product binding ($K_\mathrm{d} \approx$ 3nM) (Sashital et al., 2011). Both Csy4 and Cse3 retain their crRNA products and are members of the crRNPs that target invading nucleic acids. Evolution of these two CRISPR systems has likely selected for CRISPR endoribonucleases whose highly accurate substrate selection ensures incorporation of the appropriate RNA into the targeting complex, while the lack of substrate turnover has contributed no selective pressure for rapid cleavage kinetics.

Chapter 4

DNA interrogation by the CRISPR RNA-guided endonuclease Cas9

† Part of the work presented in this chapter has previously been published in the following research article: 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. (*These authors contributed equally.)

‡ Samuel Sternberg designed experiments, performed biochemical assays and single-molecule experiments, and assisted in single-molecule data analysis. Sy Redding conducted single-molecule experiments and data analysis, and assisted in the design and interpretation of biochemical assays. Martin Jinek assisted with the design and interpretation of initial single-molecule experiments. Eric Greene and Jennifer Doudna supervised the project.

4.1 Abstract

The CRISPR-associated enzyme Cas9 is an RNA-guided endonuclease that uses RNA:DNA base-pairing to target foreign DNA in bacteria. Cas9:guide RNA complexes are also effective genome engineering agents in animals and plants. Here we use single-molecule and bulk biochemical experiments to determine how Cas9:RNA interrogates DNA to find specific cleavage sites. We show that both binding and cleavage of DNA by Cas9:RNA require recognition of a short trinucleotide protospacer adjacent motif (PAM). Non-target DNA binding affinity scales with PAM density, and sequences fully complementary to the guide RNA but lacking a nearby PAM are ignored by Cas9:RNA. Competition assays provide evidence that DNA strand separation and RNA:DNA heteroduplex formation initiate at the PAM and proceed directionally towards the distal end of the target sequence. Furthermore, PAM interactions trigger Cas9 catalytic activity. These results reveal how Cas9 employs PAM recognition to quickly identify potential target sites while scanning large DNA molecules, and to regulate double-stranded DNA scission.

4.2 Introduction

RNA-mediated adaptive immune systems in bacteria and archaea rely on Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) and CRISPR-associated (Cas) proteins that function together to provide protection from invading viruses and plasmids(Wiedenheft et al., 2012). Bacteria harboring CRISPR-Cas loci respond to viral and plasmid challenge by integrating short fragments of the foreign nucleic acid (protospacers) into the host chromosome at one end of the CRISPR array (Barrangou et al., 2007). Transcription of the CRISPR array followed by enzymatic processing yields short CRISPR RNAs (crRNAs) that direct Cas protein-mediated cleavage of complementary target sequences within invading viral or plasmid DNA (Brouns et al., 2008; Deltcheva et al., 2011; Garneau et al., 2010). In Type II CRISPR-Cas systems, Cas9 functions as an RNA-guided endonuclease that uses a dual-guide RNA consisting of crRNA and *trans*-activating crRNA (tracrRNA) for target recognition and cleavage by a mechanism involving two nuclease active sites that together generate double-stranded DNA breaks (DSBs) (Gasiunas et al., 2012; Jinek et al., 2012).

RNA-programmed Cas9 has proven to be a versatile tool for genome engineering in multiple cell types and organisms (Bassett et al., 2013; Cong et al., 2013; Friedland et al., 2013; Gratz et al., 2013; Hwang et al., 2013; Jinek et al., 2013; Li et al., 2013; Mali et al., 2013c; Nekrasov et al., 2013; Shan et al., 2013; Wang et al., 2013b; Xie and Yang, 2013). Guided by either a natural dual-RNA complex or a chimeric single-guide RNA (Jinek et al., 2012), Cas9 generates site-specific DSBs that are repaired either by non-homologous end joining (NHEJ) or homologous recombination (HR), providing a facile means of modifying genomic information. In addition, catalytically inactive Cas9 alone or fused to transcriptional activator or repressor domains can be used to alter transcription levels at sites targeted by guide RNAs (Bikard et al., 2013; Gilbert et al., 2013; Maeder et al., 2013; Mali et al., 2013a; Perez-Pinera et al., 2013; Qi et al., 2013). Despite the remarkable ease in applying this technology, the fundamental mechanism that enables Cas9:RNA to locate specific 20 base-pair (bp) DNA targets within the vast sequence space of bacterial and eukaryotic genomes remains unknown.

4.3. Materials and Methods

4.3.1 Cas9 and RNA preparation

Wild-type Cas9 and D10A/H840A dCas9 from *S. pyogenes* were purified as described (Jinek et al., 2012), and a 3x-FLAG tag was cloned onto the C-terminus of Cas9 for single-molecule experiments. crRNAs (42 nucleotides in length) were either ordered synthetically (Integrated DNA Technologies) or transcribed in vitro with T7 polymerase using single-stranded DNA templates (**Extended Data Table 1**), as described (Sternberg et al., 2012). tracrRNA was also transcribed in vitro and contained nucleotides 15–87 following the numbering scheme used previously (Jinek et al., 2012). crRNA:tracrRNA duplexes were prepared by mixing equimolar concentrations of each RNA in Hybridization Buffer (20 mM Tris-HCl pH 7.5, 100 mM KCl, 5 mM $MgCl_2$), heating to 95 °C for 30 seconds, and slow-cooling.

4.3.2 DNA curtains post-steady state binding measurements

Post steady-state binding assays were performed with single-tethered DNA curtains (Fazio et al., 2008; Visnapuu and Greene, 2009). First, 100 nM 3x-FLAG-tagged dCas9 was reconstituted with 1 μM crRNA:tracrRNA targeting the desired region of λ-DNA by incubating for ~10 min at 37 °C in Reaction Buffer (20 mM Tris-HCl pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 5% glycerol, 1 mM DTT). 10 nM dCas9:RNA was then incubated with λ-DNA (100 pM) for ~15 min at 37 ˚C in 40 mM Tris-HCl pH 7.5, 25 mM KCl, 1 mg mL^{-1} BSA, 1 mM $MgCl_2$, and 1 mM DTT, before being diluted to 1 nM and injected into the flow cell. The flow cell was then washed with 3–5 mL of Imaging Buffer containing 40 mM Tris-HCl, 25 mM KCl, 1mg mL^{-1} BSA, 1 mM $MgCl_2$, 1 mM DTT, 0.75 nM YOYO1 (Life Technologies), 0.8% glucose, and 0.2X glucose oxidase/catalase. Finally, 0.5 nM anti-FLAG antibody-coated QDs were incubated in the flow cell for 5 min, followed by a wash of 1–2 mL of Imaging Buffer. Curtains were imaged and dCas9:RNA positions determined by fitting a 2D Gaussian to individual molecules (Gorman et al., 2012; Wang et al., 2013a). The data from all six dCas9:RNA complexes (λ1–λ6) were combined, and error bars for the combined set were generated by bootstrap methods (Gorman et al., 2012; Wang et al., 2013a).

4.3.3 DNA curtains equilibrium binding measurements

Binding position and lifetime measurements were performed using the λ2 crRNA:tracrRNA and double-tethered DNA curtains (Gorman et al., 2012; Wang et al., 2013a). Cas9 was reconstituted with a 10X excess of crRNA:tracrRNA and incubated with anti-FLAG antibody-coated QDs for ~10 min. Cas9:RNA was then diluted to 2 nM in Imaging Buffer containing 0–100 mM KCl as indicated, and injected into the flow cell. The approximate ionic strength for Imaging Buffer containing 0, 25, and 100 mM KCl is 32, 57, and 132 mM, respectively, given the expected ionization of Tris-HCl at pH 7.5 and presence of 1 mM $MgCl_2$. Videos were recorded at 50, 25, or 10 Hz, and the position of each binding event was determined from the y-coordinate within kymographs generated for each DNA molecule. The lifetime of each binding event was defined as the difference between the first frame and last frame in which the QD-tagged Cas9:RNA was observed. To analyze lifetimes, all binding events were synchronized, and the probability that a binding event survived up to a particular time was determined as the number of Cas9:RNA complexes bound at time *t* divided by the number initially bound. Position data were pooled to generate the binding distribution histogram, which was binned at 1,078 bp per bin (Visnapuu and Greene, 2009). Error bars for the binding distributions and survival probabilities were determined by bootstrap methods (Gorman et al., 2012; Wang et al., 2013a).

To test for cleavage in our single molecule assays, Cas9:RNA bound to double-tethered DNA was exposed to a 7 M urea wash in the presence of flow. Similar experiments were conducted on single-tethered curtains in the absence of YOYO1, and 42% (N = 150) of target-bound molecules remained bound to the upstream DNA product following the urea wash. The remaining 58% (N = 206) either remained bound to the downstream product containing the PAM, dissociated from the DNA altogether, or remained bound to the upstream fragment but lost their florescent tag.

4.3.4 Bulk binding and cleavage experiments

The plasmid DNA substrate contained a λ2 target sequence cloned into the EcoRI and BamHI sites on pUC19. Oligoduplex DNA substrates were 55-bp in length and were prepared by mixing together complementary synthetic oligonucleotides (Integrated DNA Technologies) in Hybridization Buffer, heating to 95 °C for 1–2 min followed by slow-cooling, and purifying on a 5% native polyacrylamide gel (0.5X TBE buffer with 5 mM $MgCl_2$) run at 4 °C. When assayed directly, DNA substrates were 5'-radiolabeled using [γ-^{32}P]-ATP (Promega) and T4 polynucleotide kinase (New England Biolabs) or 3'-radiolabeled using [α-^{32}P]-dATP (Promega) and terminal transferase (New England Biolabs). In some cases, substrates were prepared by 5'-radiolabeling only the target strand, hybridizing it to a 10X excess of the indicated unlabeled complementary strand, and gel purifying the partial/full duplex by 10% native gel electrophoresis.

Cas9:RNA complexes were reconstituted prior to cleavage and binding experiments by incubating Cas9 and the crRNA:tracrRNA duplex for 10 min at 37 °C in Reaction Buffer. Binding experiments used dCas9 (except as indicated) and either equimolar crRNA:tracrRNA or a 10X molar excess of crRNA:tracrRNA over dCas9. Binding reactions contained 0.1–1 nM DNA and increasing apo-dCas9 or dCas9:RNA concentrations, and were incubated at 37 °C for one hour before being resolved by 5% native polyacrylamide gel electrophoresis (0.5X TBE buffer with 5 mM $MgCl_2$) run at 4 °C. DNA was visualized by phosphorimaging, quantified with ImageQuant (GE Healthcare), and analyzed with Kaleidagraph (Synergy Software).

Cleavage assays were conducted in Reaction Buffer at room temperature and analyzed by 1% agarose gel electrophoresis and ethidium bromide staining or 10% denaturing polyacrylamide gel electrophoresis and phosphorimaging. Aliquots were removed at each time point and quenched by the addition of gel loading buffer supplemented with 25 mM EDTA (at 1X). Reactions contained ~1 nM radiolabeled DNA substrate and 10 nM Cas9:RNA (competition experiments) or 100 nM Cas9:RNA. Competition experiments used λ1 target DNA and were supplemented with 500 nM unlabeled competitor DNAs or an extended concentration range of competitor DNAs. All oligoduplex DNA cleavage experiments were visualized by phosphorimaging and quantified with ImageQuant (GE Healthcare)

4.3.5 Analysis of cleavage competition assays

The competition experiments were analyzed to determine the survival probability of the radiolabeled target DNA, S(t). In principle, the survival probability should begin at 1 and go to 0, but in practice the reaction rarely proceeds to completion. Therefore, we conditioned the survival against the probability of a particular amount of target DNA being cleaved. This conditional survival probability, S*(t), relates to the survival probability as follows:

$$S^*(t) = \frac{S(t) - S(\infty)}{1 - S(\infty)}$$

All reactions in the presence of competitor DNA that reached ~90% completion were conditioned against their final values, whereas reactions uncompleted after 2 hrs were conditioned to the reaction in the absence of competitor DNA. For each reaction, we then obtained the change in the survival probability of the target DNA, ΔP_s(t), in the presence of competitor DNA. Finally, ΔP_s(t) was integrated over the 2 hr reaction time. For reactions that reached completion in the presence of competitor within 2 hrs, this analysis yields the change in the mean relaxation time of the reaction (or the inverse of the average reaction rate). In cases where the reaction did not reach completion by the 2 hr time point, this analysis instead yields a mean time spent on competitor DNA during the 2 hr reaction. Notably, this analysis makes no assumptions about the nature of the reaction or the dynamic changes in the reactive species.

The reduction in cleavage rate in the presence of competitor DNA is directly proportional to the time that Cas9:RNA spends bound to each competitor. In each reaction, Cas9:RNA encounters competitor DNA on average more frequently than the target DNA, and the time Cas9:RNA spends interrogating a competitor has the cumulative effect of slowing the overall reaction. The presented models merely state that the amount of time spent on competitor DNA will be proportional to the "observed" complementarity between crRNA and bound DNA, i.e. the number of canonical Watson-Crick base-pairs that can be formed. It then directly follows that in the case where the R-loop is randomly nucleated (regardless of nucleation size), the time bound to competitor DNA will simply scale with the total amount of complementarity between competitor DNA and crRNA. However, in the case where the R-loop is nucleated from a particular site, i.e. the 3' end of the target sequence directly adjacent to the PAM, the time bound to competitor DNA will scale proportionally to the length of contiguous complementarity between the crRNA and DNA beginning from the nucleation site.

4.4. Results

4.4.1 Single-molecule visualization of Cas9

To determine how Cas9:RNA complexes locate targets, we used a single-tethered DNA curtain assay and total internal reflection fluorescence microscopy (TIRFM) to visualize the binding site distribution of single Cas9:RNA molecules on λ-DNA substrates (48,502 bp) (**Fig. 4.1a**) (Fazio et al., 2008; Visnapuu and Greene, 2009). We purified *S. pyogenes* Cas9 containing a C-terminal 3x-FLAG tag that enabled fluorescent labeling using anti-FLAG antibody-coated quantum dots (QDs) (Ilya J Finkelstein, 2010; Visnapuu and Greene, 2009), and generated guide RNAs (dual crRNA:tracrRNA) bearing complementarity to six different 20-bp sites within the λ-

DNA (**Fig. 4.1b** & **Table 4.1**). Control experiments confirmed that neither the 3x-FLAG tag nor QD inhibited DNA cleavage by Cas9:RNA, and that all guide RNAs were functional (**Fig. 4.2**). Initial experiments were conducted with a nuclease-inactive version of Cas9 containing D10A/H840A mutations (dCas9) that binds but does not cleave DNA (Jinek et al., 2012). QD-tagged dCas9:RNA localized almost exclusively to the expected target site in the DNA curtain assay (**Fig. 4.1c**). Furthermore, Cas9 could be directed to any desired region of the phage DNA by redesigning the RNA guide sequence (**Fig. 4.1d** & **4.3**), as anticipated (Cong et al., 2013; Jinek et al., 2012; Mali et al., 2013c). These results demonstrate that DNA targeting by Cas9:RNA is faithfully recapitulated in the DNA curtain assays.

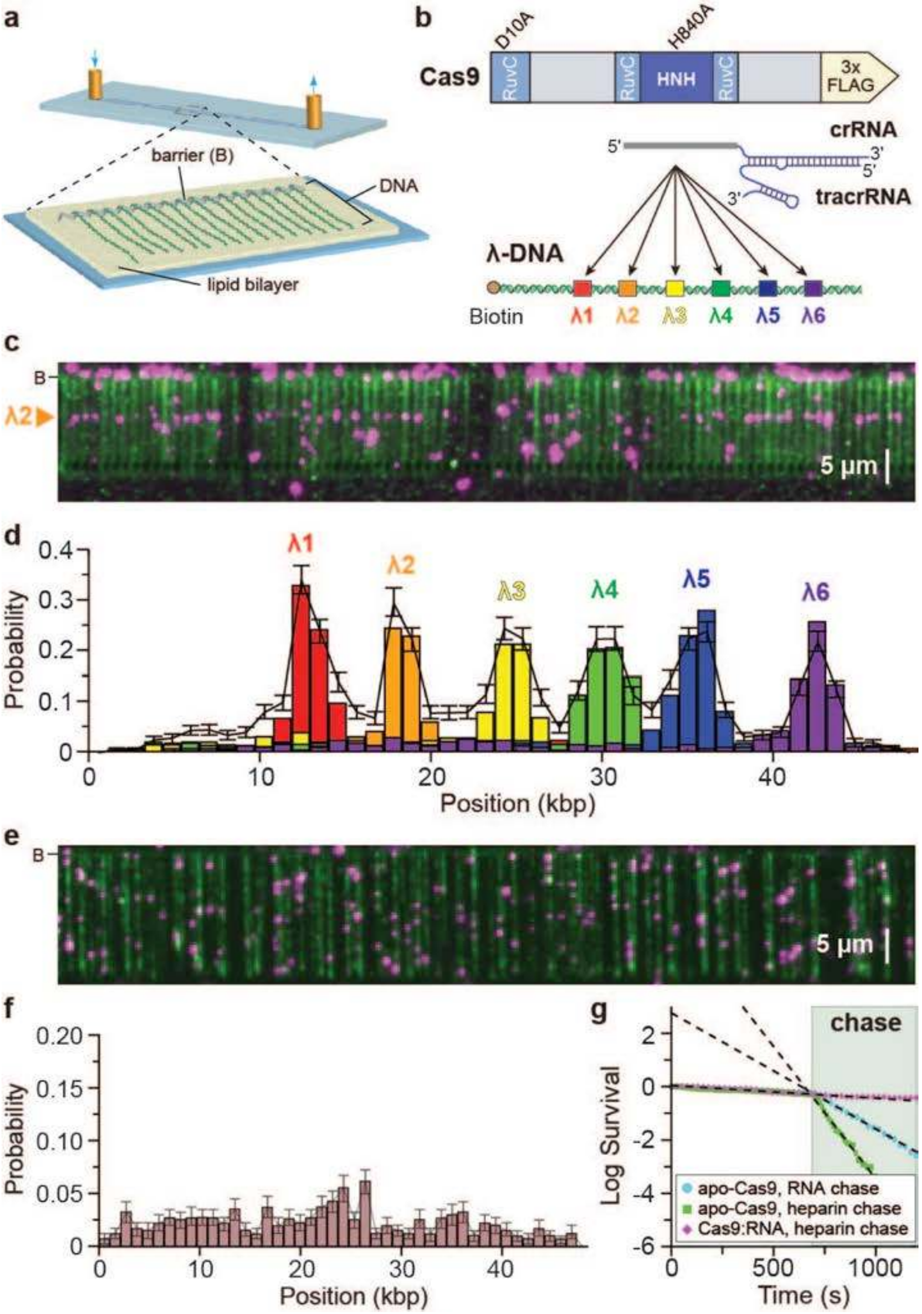


Figure 4.1 | DNA curtains assay for target binding by Cas9:RNA. (a) Schematic of a single-tethered DNA curtain. **(b)** Wild-type Cas9 or dCas9 was programmed with crRNA–tracrRNA targeting one of six sites. **(c)** YOYO1-stained DNA (green) bound by QD-tagged dCas9 (magenta) programmed with λ2 guide RNA. **(d)** dCas9:RNA binding distributions; error bars represent 95% confidence intervals obtained through bootstrap analysis. **(e)** Image of apo-Cas9 bound to DNA curtains. **(f)** Binding distribution of apo-Cas9 (n = 467); error bars represent 95% confidence intervals. g, Lifetimes of DNA-bound apo-Cas9 (n = 205) and Cas9:RNA (n = 104) after injection of 10 mg ml^{-1} heparin, and of apo-Cas9 (n = 233) after injection of 100 nM λ2 crRNA–tracrRNA.

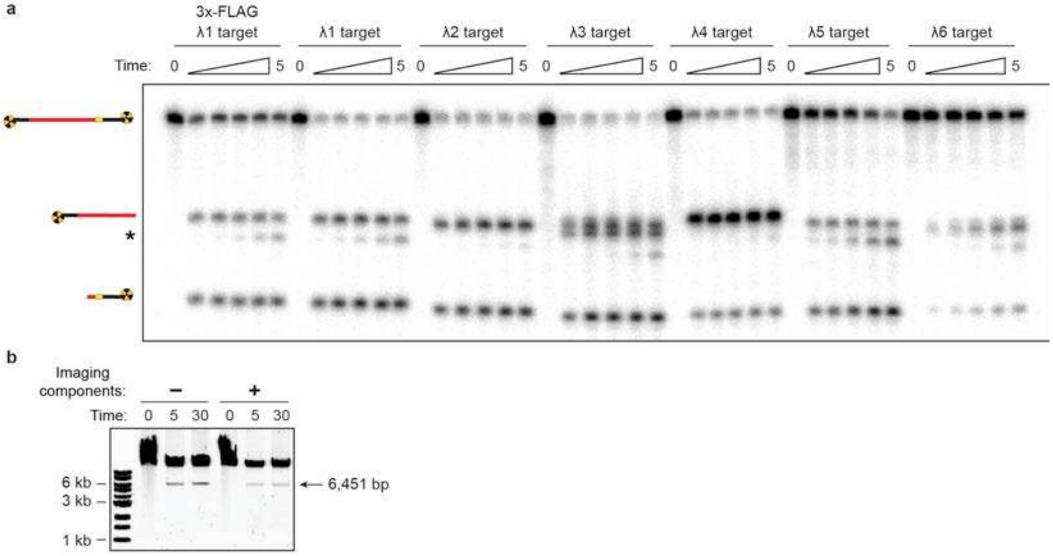


Figure 4.2 | Activity assays of reagents used in single-molecule experiments. (a) Cleavage assays were conducted using radiolabeled 55-base-pair (bp) DNA substrates that contained the six λ-DNA sequences targeted in **Fig. 4.1d**. Each DNA substrate (~1 nM) was incubated with 100 nM Cas9:RNA complex reconstituted using the corresponding guide RNA, and reaction products were resolved by 10% denaturing polyacrylamide gel electrophoresis (PAGE). Reactions contained 3x-FLAG-tagged Cas9 (where indicated) or untagged, wild-type Cas9. * denotes further trimming of the non-target strand. **(b)** Cleavage assay of λ-DNA under conditions identical to those used in single-molecule experiments. Full-length λ-DNA (25 ng µl^{-1}) was incubated with 10 nM Cas9:RNA reconstituted using the λ6 guide RNA, and reaction products were resolved by agarose gel electrophoresis. Successful cleavage is expected to generate DNA products that are 42,051 and 6,451 bp in length. When present, imaging components included anti-FLAG antibody-coated quantum dots, YOYO1, BSA, glucose, and glucose oxidase/catalase.

We next conducted controls using apo-Cas9 protein to verify that the binding observed in DNA curtain assays was due to Cas9:RNA and not apo-Cas9 lacking guide RNA. Interestingly, apo-Cas9 also bound DNA but exhibited no apparent sequence specificity (**Fig. 4.1e,f**). Attempts to measure the dissociation rate of DNA-bound apo-Cas9 were hampered by their exceedingly long lifetimes, placing a lower limit of at least 45 min on the actual lifetime (**Fig. 4.1g**). Biochemical experiments revealed an upper limit of ~25 nM for the equilibrium dissociation constant (K_d) of this apo-Cas9:DNA complex, compared to ~0.5 nM for the Cas9:RNA complex bound to a *bona fide* target site (**Fig. 4.4**).

We next asked whether DNA-bound apo-Cas9 could be distinguished from Cas9:RNA based on a differential response to chases with free guide RNAs. To test this, we measured the lifetime of apo-Cas9 on DNA curtains before and after injection of crRNA:tracrRNA or heparin.

Apo-Cas9 rapidly dissociated from non-specific sites in the presence of either competitor, and this result was verified with bulk biochemical assays **(Fig. 4.1g** & **4.4**). In contrast, target-bound Cas9:RNA was unaffected by heparin or excess crRNA:tracrRNA (**Fig. 4.1g** and **Fig. 4.4**). These findings show that non-specifically bound apo-Cas9 has properties distinct from those of Cas9:RNA complexes bound to their cognate targets.

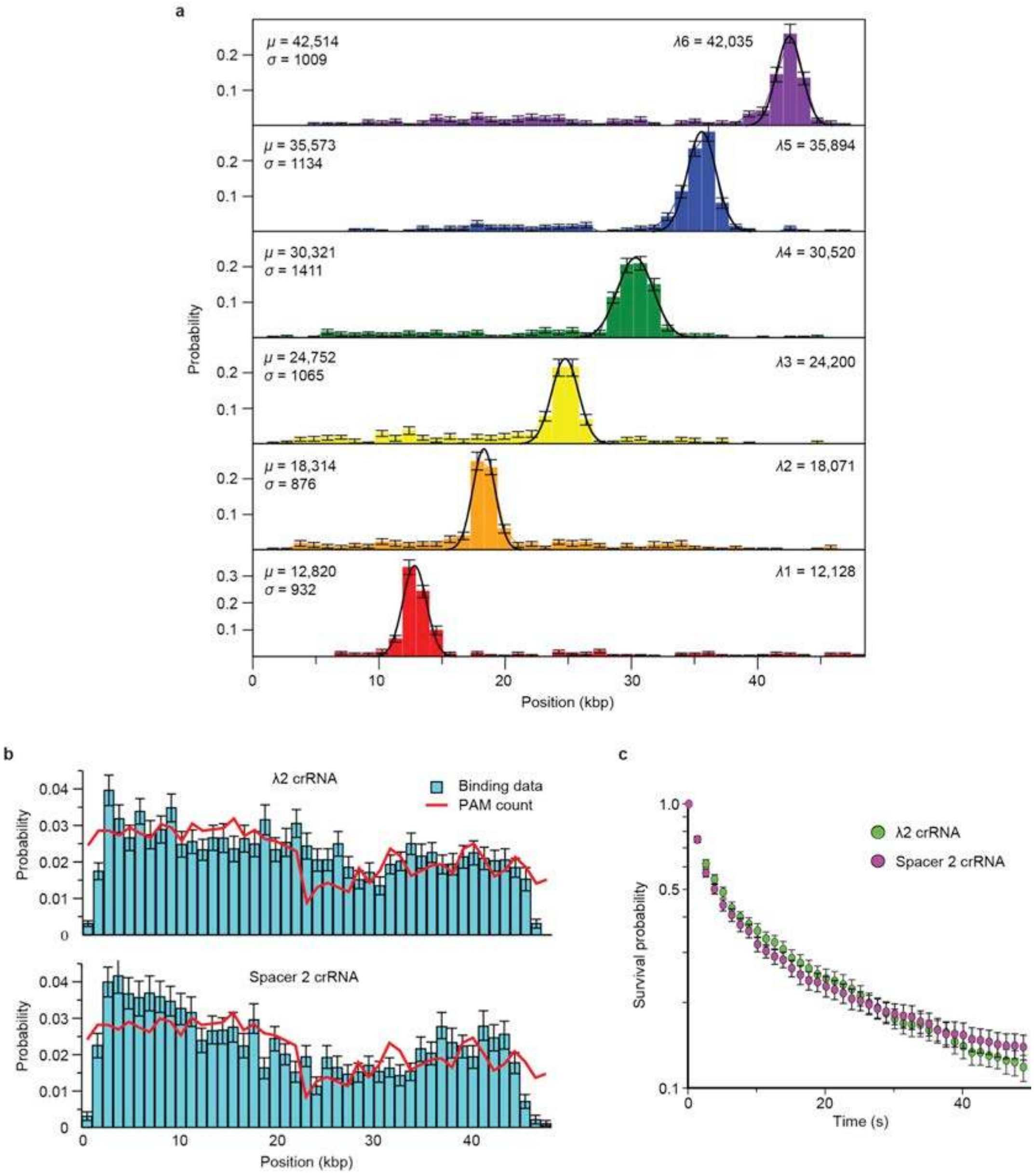


Figure 4.3 | Binding histograms and Gaussian fits for λ-DNA target binding, and analysis of off-target binding. (a) Binding distributions for dCas9 programmed with λ1-λ6 guide RNAs were measured as described in the **Materials and Methods**, and the data for each individual experiment was then bootstrapped and fit with a Gaussian curve. Shown in number of base pairs is the mean, μ, and standard deviation, σ, obtained from each fit, as well as the expected location of each target site in λ-DNA. **(b)** Distribution of Cas9:RNA binding events for λ2 crRNA (n = 2,330, top) and spacer 2 crRNA (n = 2,190, bottom). The density of PAM sites throughout the λ-DNA substrate is shown in red. **(c)** Survival probabilities for non-target binding events with λ2 and spacer 2 crRNA. Data were collected at 25 mM KCl.

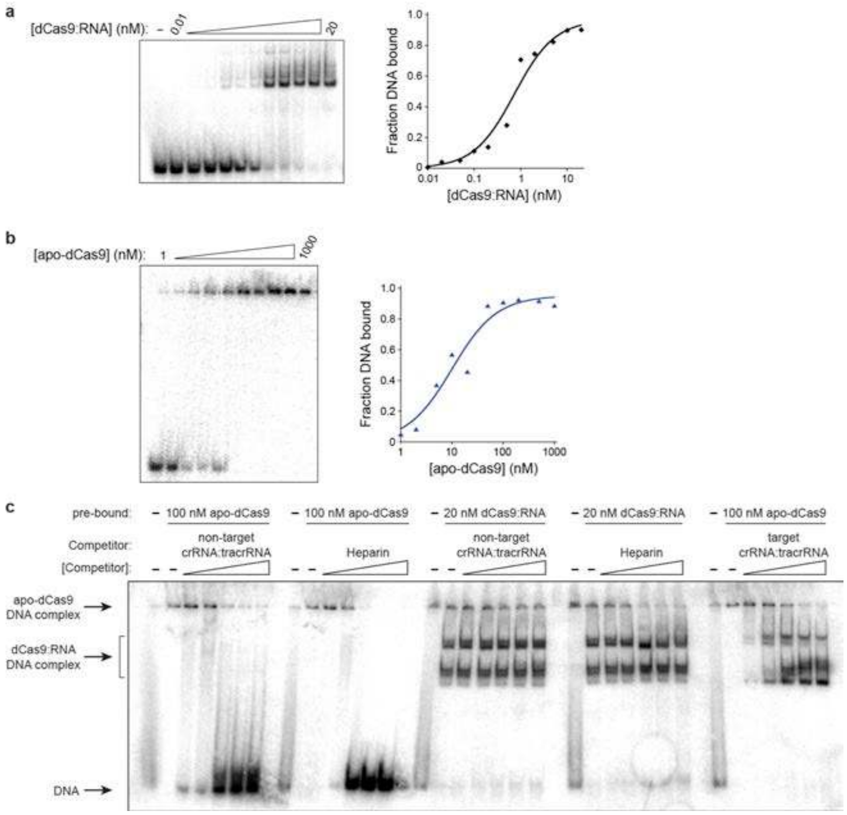


Figure 4.4 | DNA binding by apo-dCas9 and dCas9:RNA. (a) Electrophoretic mobility gel shift assay (left) with radiolabeled 55-bp target DNA and increasing concentrations of dCas9:RNA, using a 10X excess of crRNA:tracrRNA over dCas9. The quantified data (right) were fit with a standard binding isotherm (solid line), and data from three such experiments yielded a equilibrium dissociation constant (K_d) of 0.49 ± 0.21 nM. **(b)** Results for apo-dCas9 shown as in (a). Data from three independent experiments yielded a K_d of 26 ± 15 nM. **(c)** crRNA:tracrRNA duplex and heparin dissociate apo-dCas9 bound to non-specific DNA, but not dCas9:RNA complexes bound to target DNA. 55-bp DNA substrates were pre-incubated with the indicated reagent for 15 minutes at 37 °C, at which point non-targeting crRNA:tracrRNA duplex (10–1000 nM) or heparin (0.01–100 µg mL^{-1}) was added. Reactions were incubated an additional 15 minutes at 37 °C and then resolved by 5% native PAGE. Reactions at the far right show that apo-dCas9 pre-bound to target DNA can be dissociated by complementary crRNA:tracrRNA and re-bind the same DNA in complex with RNA. Note the distinct mobilities of DNA in complex with apo-dCas9 versus DNA in complex with dCas9:RNA.

Our initial experiments used catalytically inactive dCas9 to avoid DNA cleavage. Surprisingly, experiments performed with wild-type Cas9 also failed to reveal DNA cleavage. Rather, Cas9:RNA molecules remained bound to their target sites, yielding identical results to those obtained using dCas9:RNA (**Fig. 4.5a**). We confirmed that the imaging conditions did not inhibit Cas9:RNA cleavage activity (**Fig. 4.2**). These results suggested that Cas9:RNA might cleave DNA but remain tightly bound to both cleavage products, a hypothesis that was confirmed with biochemical gel shift assays using 72-bp duplex DNA substrates (**Fig. 4.6**). To determine whether stable product binding would prevent Cas9:RNA from performing multiple turnover cleavage, we conducted plasmid DNA cleavage assays at varying molar ratios of Cas9:RNA and target DNA and measured the rate and yield of product formation. Surprisingly, the amount of product rapidly plateaued at a level proportional to the molar ratio of Cas9:RNA to DNA, indicating that Cas9:RNA does not follow Michaelis-Menten kinetics (**Fig. 4.5b**). Control experiments indicated that turnover also does not occur with short duplex DNA substrates and is not stimulated by either elevated temperature or an excess of free crRNA:tracrRNA (**Fig. 4.7**).

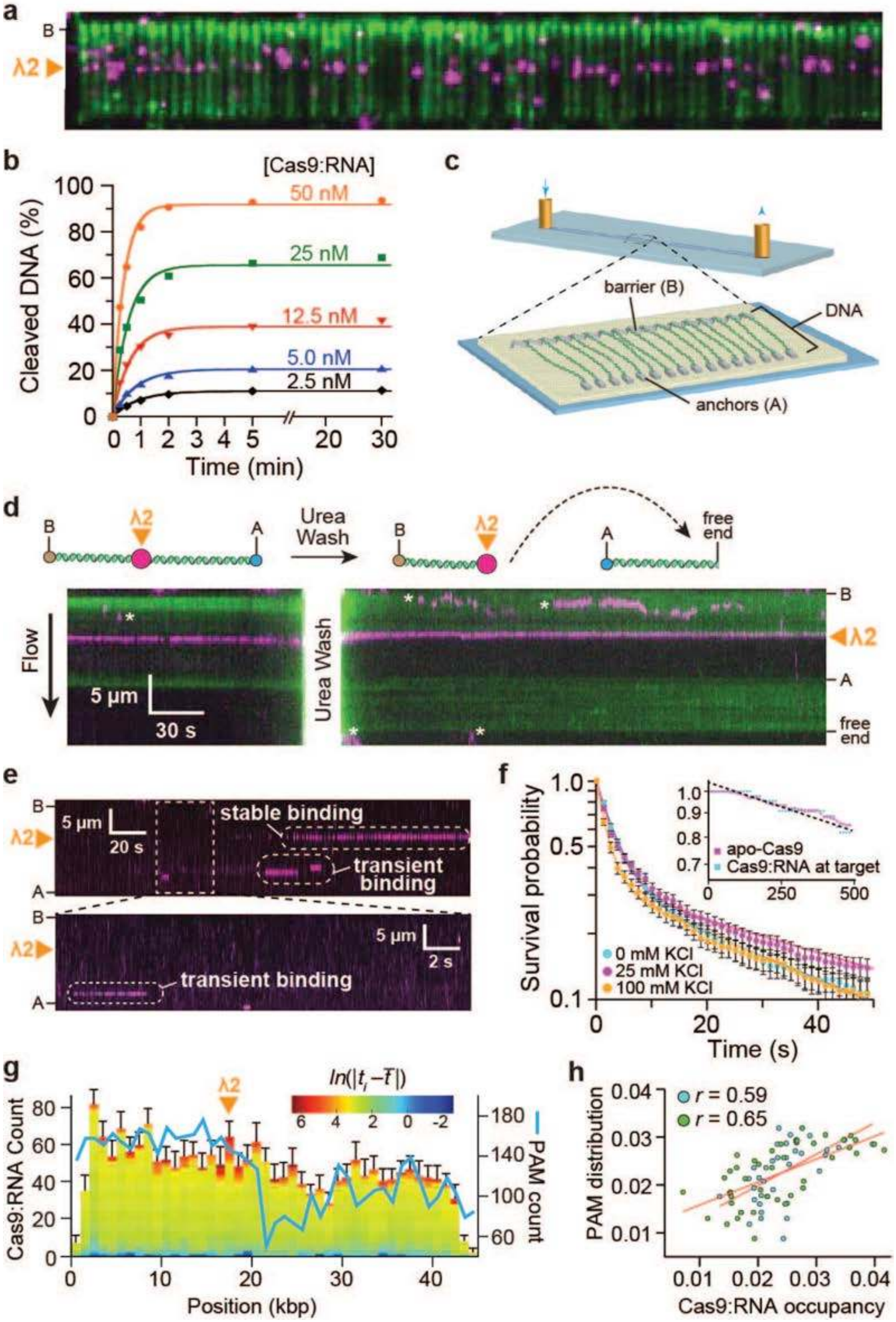


Figure 4.5 | Cas9:RNA remains bound to cleaved products and localizes to PAM-rich regions during the target search. (a) Wild-type Cas9:RNA bound to DNA curtains. **(b)** Cleavage yield of 25 nM plasmid DNA is proportional to [Cas9:RNA]. **(c)** Schematic of a double-tethered DNA curtain. **(d)** Liberation of the cleaved DNA with 7 M urea; asterisks denote QDs that are attached to the lipid bilayer but not bound to the DNA. **(e)** Kymographs illustrating distinct binding events. **(f)** Survival probabilities for non-target binding events; solid lines represent double-exponential fits. Inset: survival probabilities of DNA-bound apo-Cas9 and target DNA-bound Cas9:RNA. **(g)** Distribution of Cas9:RNA binding events (n = 2,330) and PAM density. Color-coding reflects the binding dwell time (t_i) relative to the mean dwell time ($\bar{t}$). **(h)** Correlation of PAM distribution and non-target Cas9:RNA binding for λ2 (blue) and spacer 2 (green) guide RNAs.

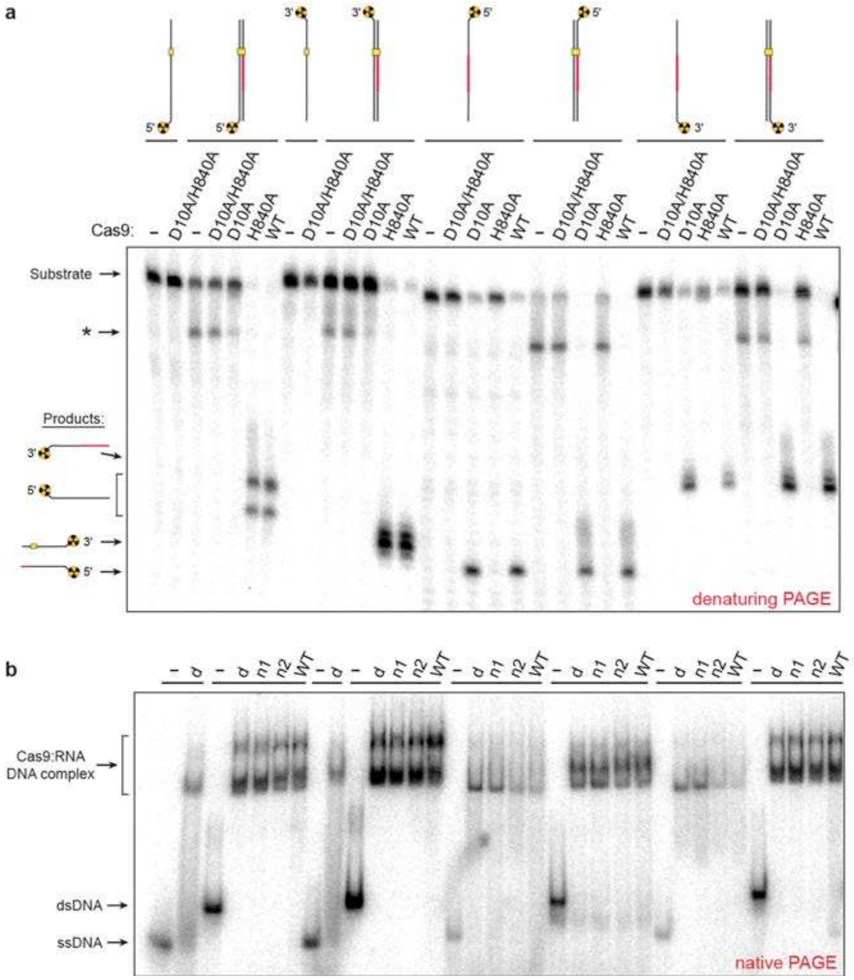


Figure 4.6 | Target DNA cleavage products remain bound to Cas9:RNA. (a,b) DNA substrates 72 nucleotides (nt) in length were radiolabeled at either their 5' or 3' ends and annealed to an unlabeled complementary strand, where indicated (top). The non-target strand contains the PAM (yellow box), whereas the target strand contains the sequence complementary to crRNA (red). Each DNA substrate (~1 nM) was incubated with 100 nM Cas9:RNA complex for 30 minutes at room temperature, using nuclease-inactive D10A/H840A Cas9 (d), both nickase mutants (D10A, n1; H840A, n2), and wild-type (WT). Half the reaction volume was quenched in formamide gel-loading buffer containing 25 mM EDTA and analyzed by 10% denaturing PAGE to verify the expected cleavage pattern of each sample (a). The other half of each reaction was analyzed by 5% native PAGE to determine whether the radiolabeled DNA fragment remained bound to Cas9:RNA (b). Aside from an apparent reduced affinity for the single-stranded target strand after cleavage, WT Cas9:RNA shows an affinity for all four possible radiolabeled DNA products that is indistinguishable from the affinity of dCas9:RNA for uncleaved DNA substrates. Note that the order of samples in (a) and (b) is identical. The additional band present for double-stranded DNA substrates in (a) results from incomplete denaturation and partial migration of intact duplex into the gel (*).

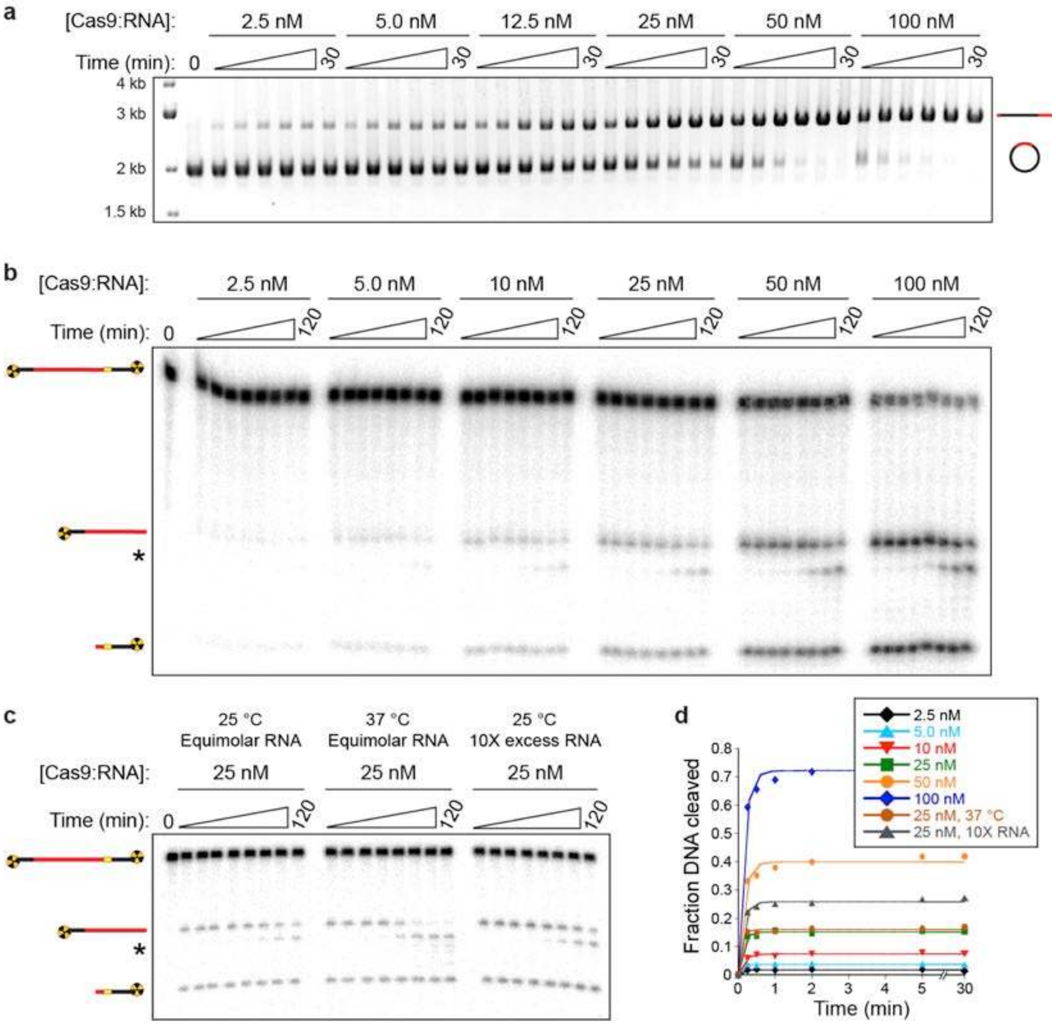


Figure 4.7 | Cas9:RNA acts as a single-turnover enzyme. (a) Agarose gel electrophoresis (1%, TBE buffer) was used to assess cleavage of plasmid DNA containing a λ2 target sequence as a function of Cas9:RNA concentration. DNA (25 nM) was incubated with the indicated concentration of Cas9:RNA, and aliquots were removed at each time point and quenched with gel loading buffer containing 25 mM EDTA. The gel was stained with ethidium bromide, and the quantified data is presented in **Fig. 4.5b. (b)** Similar turnover experiments were conducted with 25 nM radiolabeled λ2 oligoduplex substrates and increasing concentrations of Cas9:RNA. Cleavage data were visualized by phosphorimaging, and * denotes further trimming of the non-target strand. **(c)** Turnover experiments with 25 nM Cas9:RNA were repeated at 37 °C and with a 10X excess of crRNA:tracrRNA over Cas9. Neither condition significantly stimulates turnover. **(d)** Quantified data from experiments in (b) and (c) show that each reaction reaches its maximum yield after ~1 minute and does not increase with further incubation time, demonstrating that Cas9:RNA exhibits single-turnover activity. Note that the observed requirement for a slight stoichiometric excess of Cas9:RNA over DNA to reach reaction completion is likely a result of our enzyme preparations not being 100% active. While modest turnover (2.5-fold) was observed at a single enzyme:substrate stoichiometry in Jinek *et al.* (Jinek et al., 2012), our results clearly demonstrate that the reaction yield remains proportional to the molar ratio between Cas9:RNA and DNA across a range of concentrations.

We next used a double-tethered DNA curtain (**Fig. 4.5c**) (Gorman et al., 2010a; 2010b; Wang et al., 2013a) to confirm that Cas9:RNA catalyzed DNA cleavage in the single-molecule assays. Remarkably, when bound to target sites on λ-DNA, Cas9:RNA failed to dissociate from the DNA even in the presence of heparin (10 μg ml^{-1}) (**Fig. 4.1g**) or up to 0.5 M NaCl (not shown). However, injection of 7 M urea caused Cas9:RNA to release the downstream end of the cleaved DNA containing the PAM, confirming that the DNA was cleaved at the expected target site (**Fig. 4.5d**). These findings show that Cas9:RNA remains tightly bound to both ends of the cleaved DNA, thus acting as a single-turnover enzyme.

4.4.2 Cas9:RNA locates targets by 3D diffusion

To determine how Cas9:RNA locates DNA targets, we visualized the target search process using double-tethered DNA curtains. Site-specific DNA-binding proteins can locate target sites by three-dimensional (3D) collisions or through facilitated diffusion processes including one-dimensional (1D) sliding, hopping, and/or intersegmental transfer (Hippel and Berg, 1989), and these mechanisms can be distinguished by single-molecule imaging (Gorman et al., 2010b; 2012; Wang et al., 2013a). For these assays, Cas9 programmed with λ2 guide RNA was injected into the sample chamber, buffer flow was terminated, and reactions were visualized in real-time. These experiments revealed long-lived binding events at the target site and transient binding events at other sites on the DNA (**Fig. 4.5e,f**). We saw no evidence of Cas9:RNA associating with target sites through mechanisms involving facilitated diffusion (either 1D sliding and/or hopping); instead, all target association appeared to occur directly from solution through 3D collisions (**Fig. 4.5e**).

The shorter-lived, non-specific binding events exhibited complex dissociation kinetics, and the simplest model that describes the data is double-exponential decay with lifetimes of ~3.3 and ~58 seconds (at 25 mM KCl) (**Fig. 4.5f**). These lifetimes were readily distinguished from the long lifetimes of apo-Cas9 (**Fig. 4.5f, inset**). Furthermore, the experiments were conducted in the presence of a saturating (10-fold) molar excess of crRNA:tracrRNA to exclude contamination from apo-Cas9. This result indicates that at least two and possibly more binding intermediates exist on the pathway towards cognate target recognition. Non-specific DNA binding typically involves electrostatic interactions with the phosphate backbone, and therefore non-specific lifetimes tend to decrease rapidly with increasing ionic strength (Hippel and Berg, 1986). Interestingly, the lifetimes of Cas9:RNA bound at non-specific DNA sites were not appreciably affected by salt concentration (**Fig. 4.5f**). One remarkable implication of this finding is that the Cas9:RNA non-target binding events have characteristics more commonly attributed to site-specific association (Hippel and Berg, 1986; Rohs et al., 2010).

To gain further insight into the nature of the Cas9:RNA target search mechanism, we measured the locations and corresponding lifetimes of all binding events (**Fig. 4.5g**). The off-target binding lifetime distributions did not vary substantially at different regions of the DNA. However, the number of observed binding events was not uniformly distributed along the substrate, suggesting that some underlying feature of the λ-DNA might be influencing the target search. The λ phage genome contains a total of 5,677 PAM sites (~1 PAM per 8.5 bp), but it also has an unusual polar distribution of A/T- and G/C-rich sequences (Visnapuu and Greene, 2009), which leads to an asymmetric distribution of PAMs (5'-NGG-3' for *S. pyogenes* Cas9) (**Fig. 4.5g**). Pearson correlation analysis revealed that the Cas9:RNA binding site distribution was positively correlated with the PAM distribution (r = 0.59, P < 0.05) (**Fig. 4.5h**). When we repeated this experiment using a guide RNA having no complementary target sites within λ-DNA (spacer 2 crRNA, as described previously (Jinek et al., 2012)), we found no change in the observed binding lifetimes and an even stronger correlation with the PAM distribution (**Fig. 4.5h** & **4.3b,c**). These results, together with the insensitivity of short-lived binding events to ionic strength, suggested that Cas9:RNA might bind specifically to PAMs and minimize interactions with non-PAM DNA while searching for potential targets.

4.4.3 A PAM is required for DNA interrogation

To test the hypothesis that Cas9:RNA uses PAM recognition as an obligate precursor to interrogation of flanking DNA for potential guide-RNA complementarity, we used competition assays to monitor the rate of Cas9:RNA-mediated DNA cleavage (**Fig. 4.8a,b**). From these data we could extract the average amount of time that Cas9:RNA spends sampling each competitor DNA prior to locating and cleaving a radiolabeled substrate (**Fig. 4.9**). In control experiments, reaction kinetics were not perturbed by the presence of an unlabeled competitor DNA lacking PAMs and bearing no sequence relationship to the crRNA, whereas a competitor containing a PAM and fully complementary target sequence substantially reduced the rate at which the radiolabeled substrate was cleaved (**Fig. 4.8b**).

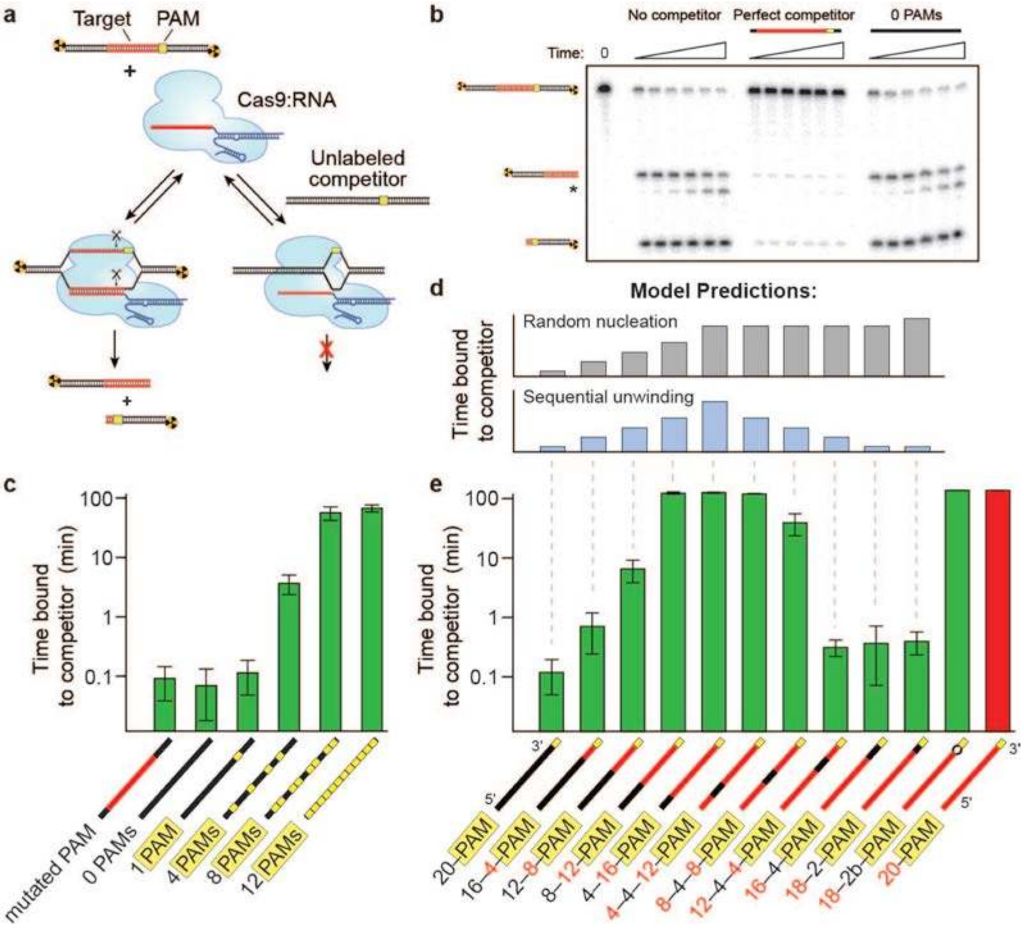


Figure 4.8 | Cas9:RNA searches for PAMs and unwinds double-stranded DNA in a directional manner. (a) Schematic of the competition cleavage assay. **(b)** Cleavage assay with and without competitor DNAs. **(c)** Quantitation of competition data (mean ± s.d.). Competitor cartoon representations show PAMs (yellow) and regions complementary to the crRNA (red). **(d)** Predicted data trends for the random nucleation or sequential

unwinding models aligned with the corresponding data in (e). **(e)** Competition assays using substrates with variable degrees of crRNA complementarity, shown as in (c). Numeric descriptions of the competitor DNAs indicate the regions of complementarity (red) or mismatches (black) to the crRNA sequence.

Next, a series of competitors were tested that bore no complementarity to the crRNA guide sequence (**Table 4.1**) but contained increasing numbers of PAMs (**Fig. 4.8c**). There was a direct correspondence between the number of PAMs and the ability of a DNA competitor to interfere with target cleavage, indicating that the lifetime of Cas9:RNA on competitor DNA increased with PAM density (**Fig. 4.8c**). This result held true over a range of competitor DNA concentrations (**Fig. 4.9**), and the same pattern of competition was observed for DNA binding by dCas9:RNA (**Fig. 4.10**). These results demonstrate that the residence time of Cas9:RNA on non-target DNA lacking PAMs is negligible, and support the hypothesis that the transient, non-target DNA binding events observed on the DNA curtains likely occurred at PAM sequences. While Cas9:RNA complexes undoubtedly sample DNA lacking PAMs, these rapid binding events are neither detectable in single-molecule assays and bulk binding experiments (**Fig. 4.10**), nor do they appreciably influence overall reaction kinetics in bulk biochemical assays.

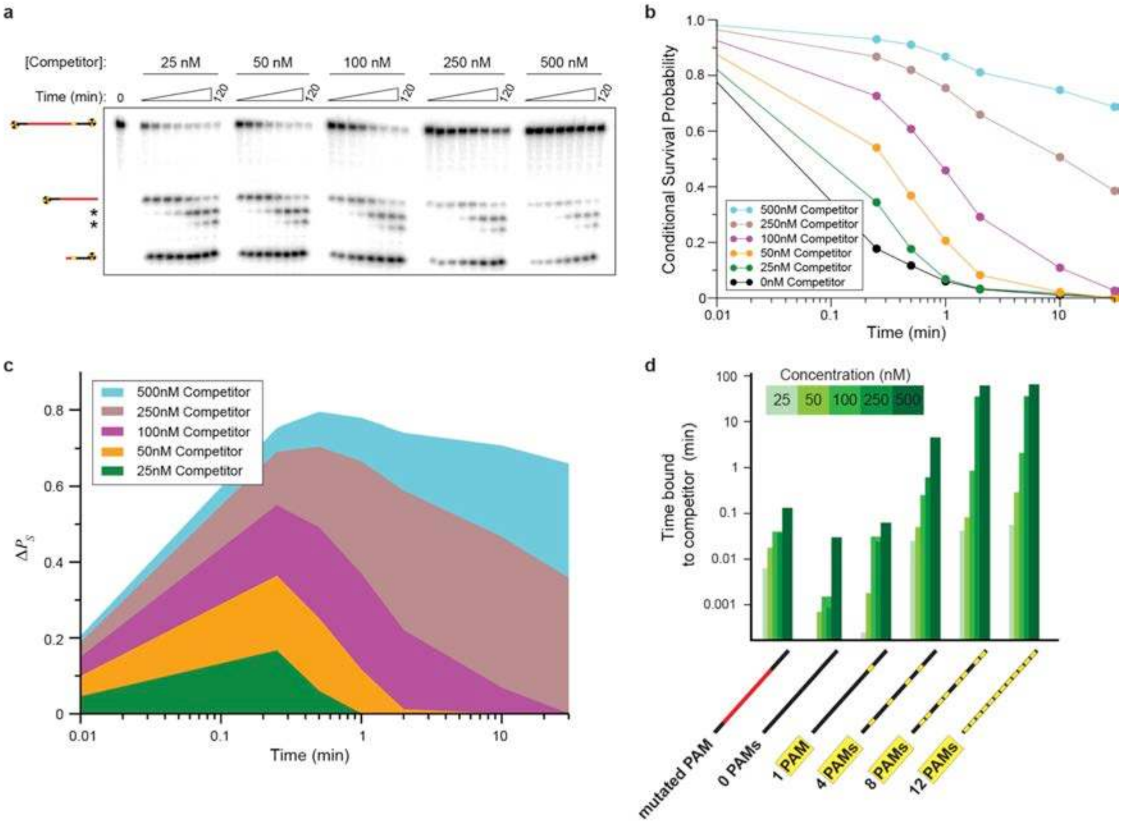


Figure 4.9 | Analysis of competition cleavage assays. (a) Representative cleavage assays as a function of competitor DNA concentration, using a competitor containing 12 PAM sites. Radiolabeled λ1 target DNA (1 nM) was incubated with 10 nM Cas9:RNA and increasing concentrations of the competitor, and reaction products at each time point were resolved by 10% denaturing PAGE. Cleavage data were visualized by phosphorimaging, and * denotes further trimming of the non-target strand. **(b)** Shown are the conditional survival probabilities for the radiolabeled target DNA at each concentration of 12-PAM competitor. **(c)** Shown is the change in survival probability of the target DNA, $\Delta P_s(t)$, as a function of 12-PAM competitor concentration. The area under each curve represents the amount of time that Cas9:RNA spent on the competitor DNA during the reaction. (**d**) Competition

data with a panel of substrates that have no complementarity to the guide RNA and variable numbers of PAMs, and a perfect target sequence with single base-pair mutation in the PAM. The data are presented similarly to **Fig. 4.8c**, but the time bound to competitor is shown for all five concentrations of competitor tested.

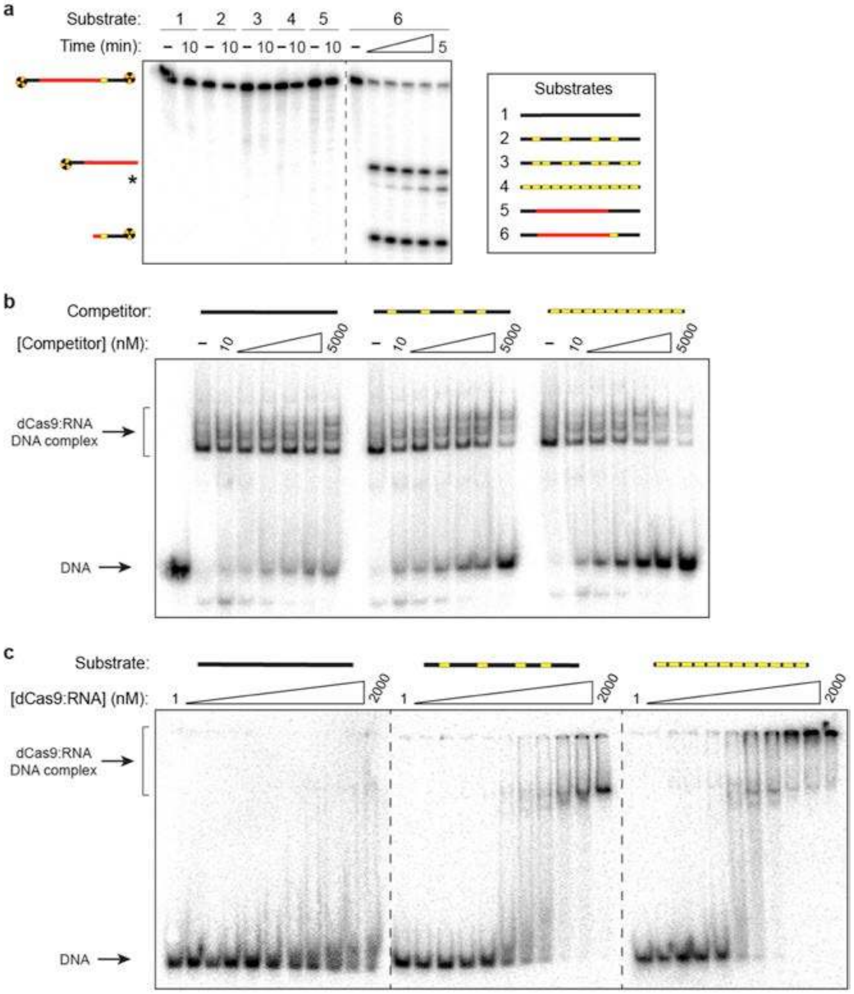


Figure 4.10 | PAM sites in non-target DNA are bound specifically by dCas9:RNA. (a) None of the competitors from **Fig. 4.8c** can be cleaved, including one that bears full complementarity to the crRNA but contains a single base-pair mutation in the PAM. Radiolabeled competitor DNAs and target DNA (1 nM) were incubated with 100 nM WT Cas9:RNA for the indicated time, and reaction products were assessed by 10% denaturing PAGE. * denotes further trimming of the non-target strand. **(b)** PAM-rich competitor DNAs interfere with target DNA binding by dCas9:RNA. The same radiolabeled 55-bp target DNA from **Fig. 4.8b,c** was pre-mixed with increasing concentrations of the indicated competitor DNA and then incubated with 10 nM dCas9:RNA for 60 minutes at 37 °C. Binding reactions were resolved by 5% native PAGE. **(c)** dCas9:RNA has increased affinity for non-target DNA containing multiple PAM sequences. The indicated radiolabeled DNA substrates (~0.02 nM) were incubated with increasing concentrations of dCas9:RNA for 60 minutes at 37 °C, and reactions were resolved by 5% native PAGE. The observed well-shifting at high concentrations may result from multiple dCas9:RNA molecules binding the same DNA substrate.

We next repeated the competition assay with a competitor bearing perfect complementarity to the crRNA, but with a single point mutation in the adjacent PAM (5'-T<u>C</u>G-3'). Like similarly mutated substrates (Jinek et al., 2012), this competitor cannot be cleaved by

Cas9:RNA (**Fig. 4.10**). This competitor failed to inhibit DNA cleavage by Cas9:RNA and behaved comparably to the non-target competitor DNA lacking PAMs, despite the fact that it contained perfect complementarity to the crRNA (**Fig. 4.8c**). Together, these results demonstrate that PAM recognition is an obligate first step during target recognition by Cas9:RNA, as was previously hypothesized (Jinek et al., 2012).

Table 4.1 | RNA and DNA substrates used in this study.

Description	Sequence[a]
λ1 target sequence: 12,128 bp in λ-DNA	5'-GGCGCATAAAGATGAGACGC**TGG**-3' 3'-**CCGCGTATTTCTACTCTGCG**ACC-5'
λ2 target sequence: 18,071 bp in λ-DNA	5'-GTGATAAGTGGAATGCCATG**TGG**-3' 3'-**CACTATTCACCTTACGGTAC**ACC-5'
λ3 target sequence: 24,200 bp in λ-DNA	5'-CTGGTGAACTTCCGATAGTG**CGG**-3' 3'-**GACCACTTGAAGGCTATCAC**GCC-5'
λ4 target sequence: 30,520 bp in λ-DNA	5'-CAGATATAGCCTGGTGGTTC**AGG**-3' 3'-**GTCTATATCGGACCACCAAG**TCC-5'
λ5 target sequence: 35,894 bp in λ-DNA	5'-GGCAATGCCGATGGCGATAG**TGG**-3' 3'-**CCGTTACGGCTACCGCTATC**ACC-5'
λ6 target sequence: 42,035 bp in λ-DNA	5'-GGTGTGAAAGAACACCAACA**GGG**-3' 3'-**CCACACTTTCTTGTGGTTGT**CCC-5'
Oligo for preparing dsDNA T7 promoter, in vitro transcription	5'-**TAATACGACTCACTATA**-3'
ssDNA T7 template[b]: λ1-targeting crRNA	5'-CAAAACAGCATAGCTCTAAAACGCGTCTCATCTTTATGCGTC**TATAGTGAGTCGTATTA**-3'
λ1-targeting crRNA[c]	5'-**GACGCAUAAAGAUGAGACGC**GUUUUAGAGCUAUGCUGUUUUG-3'
λ2-targeting crRNA	5'-**GUGAUAAGUGGAAUGCCAUG**GUUUUAGAGCUAUGCUGUUUUG-3'
λ3-targeting crRNA	5'-**CUGGUGAACUUCCGAUAGUG**GUUUUAGAGCUAUGCUGUUUUG-3'
λ4-targeting crRNA	5'-**CAGATATAGCCTGGTGGTTC**GUUUUAGAGCUAUGCUGUUUUG-3'
ssDNA T7 template[b]: λ5-targeting crRNA	5'-CAAAACAGCATAGCTCTAAAACCTATCGCCATCGGCATTGCC**TATAGTGAGTCGTATTA**-3'
λ5-targeting crRNA	5'-**GGCAAUGCCGAUGGCGAUAG**GUUUUAGAGCUAUGCUGUUUUG-3'
ssDNA T7 template[b]: λ6-targeting crRNA	5'-CAAAACAGCATAGCTCTAAAACTGTTGGTGTTCTTTCACACC**TATAGTGAGTCGTATTA**-3'
λ6-targeting crRNA	5'-**GGUGUGAAAGAACACCAACA**GUUUUAGAGCUAUGCUGUUUUG-3'
ssDNA T7 template[b]: tracrRNA	5'-AAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATGCTGT**CCTATAGTGAGTCGTATTA**-3'
tracrRNA (nt 15-87)	5'-GGACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUU-3'

λ1 target duplex[d]	5'-AGCAGAAATCTCTGCTGACGCATAAAGATGAGACGC**TGG**AGTACAAACGTCAGCT-3' 3'-TCGTCTTTAGAGACGA**CTGCGTATTTCTACTCTGCG**ACCTCATGTTTGCAGTCGA-5'
λ2 target duplex	5'-GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATG**TGG**GCTGTCAAAATTGAGC-3' 3'-CTCACCTTCCTACGGT**CACTATTCACCTTACGGTAC**ACCCGACAGTTTTAACTCG-5'
λ3 target duplex	5'-AACGTGCTGCGGCTGGCTGGTGAACTTCCGATAGTG**CGG**GTGTTGAATGATTTCC-3' 3'-TTGCACGACGCCGACC**GACCACTTGAAGGCTATCAC**GCCCACAACTTACTAAAGG-5'
λ4 target duplex	5'-TCACAACAATGAGTGGCAGATATAGCCTGGTGGTTC**AGG**CGGCGCATTTTTATTG-3' 3'-AGTGTTGTTACTCACC**GTCTATATCGGACCACCAAG**TCCGCCGCGTAAAAATAAC-5'
λ5 target duplex	5'-GAATGAACGATGCAGAGGCAATGCCGATGGCGATAG**TGG**GTATCATGTAGCCGCT-3' 3'-CTTACTTGCTACGTCT**CCGTTACGGCTACCGCTATC**ACCCATAGTACATCGGCGA-5'
λ6 target duplex	5'-AATCGATGGTGTCTCCGGTGTGAAAGAACACCAACA**GGG**GTGTTACCACTACCGC-3' 3'-TTAGCTACCACAGAGG**CCACACTTTCTTGTGGTTGT**CCCCACAATGGTGATGGCG-5'
λ2 target duplex: cloned into pUC19[e]	5'-AATTGAAAGTGATAAGTGGAATGCCATG**TGG**AAAC -3' 3'- CTTT**CACTATTCACCTTACGGTAC**ACCTTTGCTAG-5'
λ1 Competitor: Mutated PAM	5'-AGCAGAAATCTCTGCTGACGCATAAAGATGAGACGC**TCG**AGTACAAACGTCAGCT-3' 3'-TCGTCTTTAGAGACGA**CTGCGTATTTCTACTCTGCG**AGCTCATGTTTGCAGTCGA-5'
Competitor: 0 PAMs	5'-AGCTGCATAACGCGAAAAAATATATTTATCTGCTTGATCTTCAAATGTTGTATTG-3' 3'-TCGACGTATTGCGCTTTTTTATATAAATAGACGAACTAGAAGTTTACAACATAAC-5'
Competitor: 1 PAM	5'-AGCAGAAATCTCTGCTCTGCGTATTTCTACTCTGCG**TGG**AGTACAAACGTCAGCT-3' 3'-TCGTCTTTAGAGACGAGACGCATAAAGATGAGACGCACCTCATGTTTGCAGTCGA-5'
Competitor: 4 PAMs	5'-AGCTGCATAACG**CGG**GAAAATCCATTTATCTGCTTGATCTT**CGG**ATGTTCCATTG-3' 3'-TCGACGTATTGCGCCCTTTTA**GGT**AAATAGACGAACTAGAAGCCTACAA**GGT**AAC-5'
Competitor: 8 PAMs	5'-**GG**CTGCACCACG**CGG**GAAAATCCATTT**AGG**TGCTTCCTCTT**CGG**ATGTTCCATTG-3' 3'-CCGACGT**GGT**GCGCCCTTTTA**GGT**AAATCCACGAA**GGA**GAAGCCTACAA**GGT**AAC-5'
Competitor: 12 PAMs	5'-**GG**C**TGG**ACCACG**CGG**GAAAATCCACCT**AGG**TGGTTCCTCTT**CGG**ATGTTCCATCC-3' 3'-CCGACCT**GGT**GCGCCCTTTTA**GGTGGA**TCCACCAA**GGA**GAAGCCTACAA**GGT**A**GG**-5'
λ1 Competitor: **16**–**4**–PAM	5'-AGCAGAAATCTCTGCTCTGCGTATTTCTACTCACGC**TGG**AGTACAAACGTCAGCT-3' 3'-TCGTCTTTAGAGACGA**GACGCATAAAGATGAGTGCG**ACCTCATGTTTGCAGTCGA-5'
λ1 Competitor: **12**–**8**–PAM	5'-AGCAGAAATCTCTGCTCTGCGTATTTCTTGAGACGC**TGG**AGTACAAACGTCAGCT-3' 3'-TCGTCTTTAGAGACGA**GACGCATAAAGAACTCTGCG**ACCTCATGTTTGCAGTCGA-5'
λ1 Competitor: **8**–**12**–PAM	5'-AGCAGAAATCTCTGCTCTGCGTATAAGATGAGACGC**TGG**AGTACAAACGTCAGCT-3' 3'-TCGTCTTTAGAGACGA**GACGCATATTCTACTCTGCG**ACCTCATGTTTGCAGTCGA-5'
λ1 Competitor: **4**–**16**–PAM	5'-AGCAGAAATCTCTGCTCTGCCATAAAGATGAGACGC**TGG**AGTACAAACGTCAGCT-3' 3'-TCGTCTTTAGAGACGA**GACGGTATTTCTACTCTGCG**ACCTCATGTTTGCAGTCGA-5'
λ1 Competitor: **4**–**4**–**12**–PAM	5'-AGCAGAAATCTCTGCTGACGGTATAAGATGAGACGC**TGG**AGTACAAACGTCAGCT-3' 3'-TCGTCTTTAGAGACGA**CTGCCATATTCTACTCTGCG**ACCTCATGTTTGCAGTCGA-5'
λ1 Competitor: **8**–**4**–**8**–PAM	5'-AGCAGAAATCTCTGCTGACGCATATTCTTGAGACGC**TGG**AGTACAAACGTCAGCT-3' 3'-TCGTCTTTAGAGACGA**CTGCGTATAAGAACTCTGCG**ACCTCATGTTTGCAGTCGA-5'
λ1 Competitor: **12**–**4**–**4**–PAM	5'-AGCAGAAATCTCTGCTGACGCATAAAGAACTCACGC**TGG**AGTACAAACGTCAGCT-3' 3'-TCGTCTTTAGAGACGA**CTGCGTATTTCTTGAGTGCG**ACCTCATGTTTGCAGTCGA-5'

λ1 Competitor: 16–**4**–PAM	5'-AGCAGAAATCTCTGCTGACGCATAAAGATGAGTGCG**TGG**AGTACAAACGTCAGCT-3' 3'-TCGTCTTTAGAGACGA**CTGCGTATTTCTACTCACGC**ACCTCATGTTTGCAGTCGA-5'
λ1 Competitor: 18–**2**–PAM	5'-AGCAGAAATCTCTGCTGACGCATAAAGATGAGACCG**TGG**AGTACAAACGTCAGCT-3' 3'-TCGTCTTTAGAGACGA**CTGCGTATTTCTACTCTGGC**ACCTCATGTTTGCAGTCGA-5'
λ1 Competitor: 18–**2b**–PAM	5'-AGCAGAAATCTCTGCTGACGCATAAAGATGAGAC^{GC}**TGG**AGTACAAACGTCAGCT-3' 3'-TCGTCTTTAGAGACGA**CTGCGTATTTCTACTCTGGC**ACCTCATGTTTGCAGTCGA-5'
λ2 target (55-nt): dsDNA	5'-GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATG**TGG**GCTGTCAAAATTGAGC-3' 3'-CTCACCTTCCTACGGT**CACTATTCACCTTACGGTAC**ACCCGACAGTTTTAACTCG-5'
λ2 target (55-nt): ssDNA target strand	3'-CTCACCTTCCTACGGT**CACTATTCACCTTACGGTAC**ACCCGACAGTTTTAACTCG-5'
λ2 target (55-nt): partial dsDNA without PAM	5'- GCTGTCAAAATTGAGC-3' 3'-CTCACCTTCCTACGGT**CACTATTCACCTTACGGTAC**ACCCGACAGTTTTAACTCG-5'
λ2 target (55-nt): partial dsDNA with PAM	5'- **TGG**GCTGTCAAAATTGAGC-3' 3'-CTCACCTTCCTACGGT**CACTATTCACCTTACGGTAC**ACCCGACAGTTTTAACTCG-5'
λ2 target (72-nt): dsDNA	5'-TTCGAAAGAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATG**TGG**GCTGTCAAAATTGAGCAGACCAAAGA-3' 3'-AAGCTTTCTCACCTTCCTACGGT**CACTATTCACCTTACGGTAC**ACCCGACAGTTTTAACTCGTCTGGTTTCT-5'
λ2 target (72-nt): ssDNA	3'-AAGCTTTCTCACCTTCCTACGGT**CACTATTCACCTTACGGTAC**ACCCGACAGTTTTAACTCGTCTGGTTTCT-5'
λ2 target (72-nt): partial dsDNA without PAM	5'-GCTGTCAAAATTGAGCAGACCAAAGA-3' 3'-AAGCTTTCTCACCTTCCTACGGT**CACTATTCACCTTACGGTAC**ACCCGACAGTTTTAACTCGTCTGGTTTCT-5'
λ2 target (72-nt): partial dsDNA with PAM	5'-**TGG**GCTGTCAAAATTGAGCAGACCAAAGA-3' 3'-AAGCTTTCTCACCTTCCTACGGT**CACTATTCACCTTACGGTAC**ACCCGACAGTTTTAACTCGTCTGGTTTCT-5'

[a] Guide crRNA sequences and complementary DNA target strand sequences are shown in red. PAM sites (5'-NGG-3') are highlighted in yellow on the non-target strand when adjacent to the target sequence, except for PAM competitors in which case all PAMs are highlighted.

[b] The reverse complement of the T7 promoter is indicated in **bold**.

[c] The second nucleotide of the λ1-targeting crRNA was mutated from G to A to match the λ1 target duplex sequence.

[d] The underlined base-pairs were mutated relative to the wild-type λ-DNA sequence in order to remove all PAM sites from the substrate other than the PAM immediately adjacent to the target sequence. The crRNA was mutated accordingly, as were the λ1 competitor DNAs.

[e] The duplex was cloned into EcoRI and BamHI sites on pUC19.

NA, not applicable.

4.4.4 Mechanism of RNA:DNA heteroduplex formation

After PAM recognition, Cas9:RNA must destabilize the adjacent duplex and initiate strand separation to enable base-pairing between the target DNA strand and the crRNA guide sequence. Because Cas9 has no energy-dependent helicase activity, the mechanism of local DNA unwinding has been enigmatic, but must rely upon thermally available energy. One possibility is that PAM binding could induce a general destabilization of the duplex along the length of the entire target sequence, leading to random nucleation of the RNA:DNA heteroduplex (**Fig. 4.8d,**

top). Alternatively, PAM binding may cause only local melting of the duplex, with the RNA:DNA heteroduplex nucleating at the 3' end of the target sequence next to the PAM and proceeding sequentially towards the distal 5' end of the target sequence (**Fig. 4.8d, bottom**).

To distinguish between these two models, we conducted cleavage assays with a panel of DNA competitors in which the length and position of complementarity to the guide RNA was systematically varied (**Table 4.1**). These competitors were designed to distinguish between the random nucleation and sequential unwinding models for heteroduplex formation based upon the predicted patterns of cleavage inhibition for each model (**Fig. 4.8d**). The ability of a competitor DNA to inhibit substrate cleavage by Cas9:RNA increased as the extent of complementarity originating at the 3' end of the target sequence adjacent to the PAM increased (**Fig. 4.8e**). Inhibition increased dramatically when 12 or more base-pairs were complementary to the crRNA guide sequence, which agrees with the requirement for an 8-12 nucleotide seed sequence for the Cas9:RNA DNA cleavage reaction (Jiang et al., 2013; Jinek et al., 2012). Strikingly, although competitors containing mismatches to the crRNA at the 5' end of the target sequence competed effectively for Cas9:RNA binding, competitors containing mismatches to the crRNA at the extreme 3' end immediately adjacent to the PAM were completely inert to binding (**Fig. 4.8e**). This was true even with a 2-bp mismatch followed by 18 bp of contiguous sequence complementarity to the crRNA. Therefore, when mismatches to the crRNA are encountered within the first two nucleotides of the target sequence, Cas9:RNA loses the ability to interrogate and recognize the remainder of the DNA. The pattern of inhibition observed with the different competitor DNAs indicates that sequence homology adjacent to the PAM is necessary to initiate target duplex unwinding until the reaction has proceeded sufficiently far (~12 bp, approximately one turn of an A-form RNA:DNA helix), such that the energy necessary for further propagation of the RNA:DNA heteroduplex falls below the energy needed for the reverse reaction. These findings suggest that formation of the RNA:DNA heteroduplex initiates at the PAM and proceeds through the target sequence by a sequential, step-wise unwinding mechanism consistent with a Brownian ratchet (Abbondanzieri et al., 2005).

As a further test of this model, we used a DNA competitor that contained mismatches to the crRNA at positions 1-2 but was itself mismatched at the same two positions, forming a small bubble in the duplex. Despite the absence of sequence complementarity to the crRNA within the DNA bubble, this substrate was a robust competitor and bound Cas9:RNA with an affinity nearly indistinguishable from that of an ideal substrate (**Fig. 4.8e** & **4.11**). Remarkably, this DNA could also be cleaved with near wild-type rates (**Fig. 4.11**). We speculate that the presence of the DNA bubble allowed Cas9:RNA to bypass the mismatches and reinitiate nucleation of the RNA:DNA heteroduplex downstream of the bubble, thereby propagating strand separation through the remainder of the target.

4.4.5 The PAM triggers Cas9 nuclease activity

The results presented above indicate that PAM recognition plays a central role in target recognition, and that introduction of a small bubble in the DNA target eliminates the need for RNA:DNA heteroduplex formation immediately adjacent to the PAM. One might expect PAM recognition to be dispensable for Cas9:RNA-mediated recognition and cleavage of a single-stranded DNA (ssDNA) target. Surprisingly, however, a ssDNA substrate was cleaved more than two orders of magnitude slower than a double-stranded DNA (dsDNA) substrate (**Fig. 4.12a,b**), despite the fact that dCas9:RNA bound both the dsDNA and ssDNA substrates with similar affinities (**Fig. 4.12b** & **4.13**).

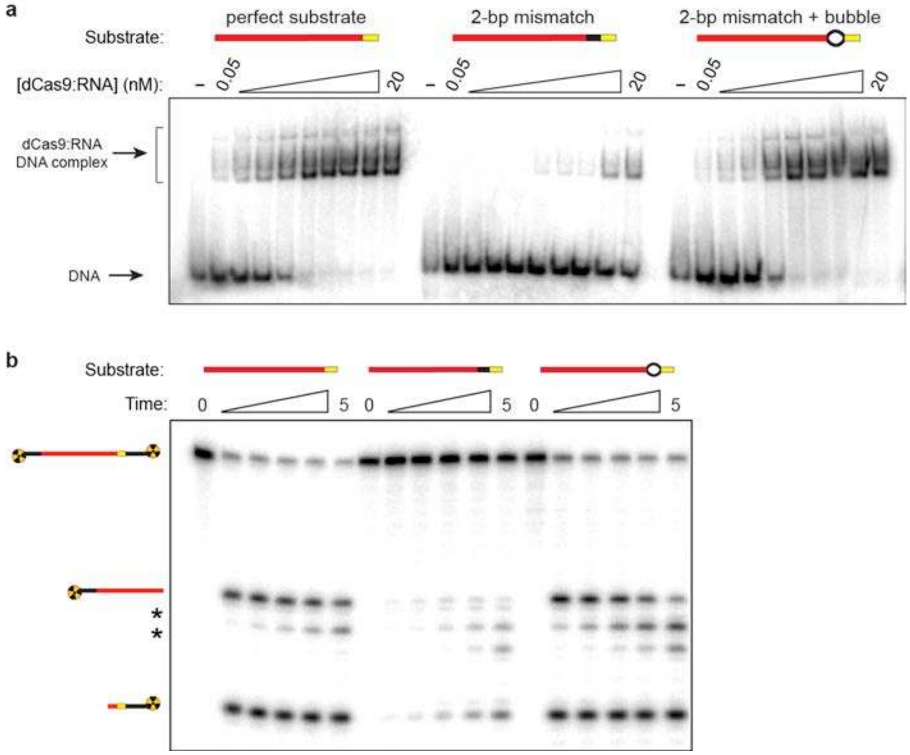


Figure 4.11 | Cas9:RNA binds and cleaves bubble-containing DNA substrates with mismatches to the crRNA that are otherwise discriminated against within the context of perfect duplexes. (a) dCas9:RNA has weak affinity for a substrate containing a 2-bp mismatch to the crRNA (middle), whereas a substrate presenting the same mismatches within a small 2-nt bubble (right) is bound with an affinity nearly indistinguishable from a perfect target substrate (left), in agreement with data presented in **Fig. 4.8e**. The indicated DNA substrates were incubated with increasing concentrations of dCas9:RNA for 60 minutes at 37 °C, and reactions were resolved by 5% native PAGE. **(b)** The same bubble-containing substrate in (a) is cleaved with similar kinetics as a perfect substrate (compare right and left time courses), whereas a perfectly base-paired substrate with the same pattern of complementarity to the crRNA is cleaved with substantially reduced kinetics (middle). Radiolabeled DNA substrates (1 nM) were incubated with 100 nM WT Cas9:RNA for the indicated time, and reaction products were resolved by 10% denaturing PAGE. * denotes further trimming of the non-target strand.

Importantly, Cas9:RNA recognizes the 5'-NGG-3' PAM on the non-target DNA strand (Jinek et al., 2012), so the ssDNA substrates did not contain a PAM but rather the complement to the PAM sequence. We hypothesized that the absence of the PAM on the ssDNA might explain why an otherwise fully complementary target is resistant to cleavage. To test this possibility, we prepared hybrid substrates with varying lengths of dsDNA at the 3' flanking sequence (**Fig. 4.12a**). Cleavage assays revealed that the ssDNA target strand could be activated for cleavage in the presence of flanking dsDNA that extended across the PAM sequence, but that this activating effect was lost when the dsDNA was truncated immediately before the PAM (**Fig. 4.12a,b** & **4.13**). Binding experiments confirmed these results were not a consequence of discrimination at the level of binding (**Fig. 4.12b**). Rather, the presence of the 5'-NGG-3' PAM on the non-target strand was critical for some step of the reaction that occurred after binding. These data suggest that the PAM acts as an allosteric regulator of Cas9:RNA nuclease activity.

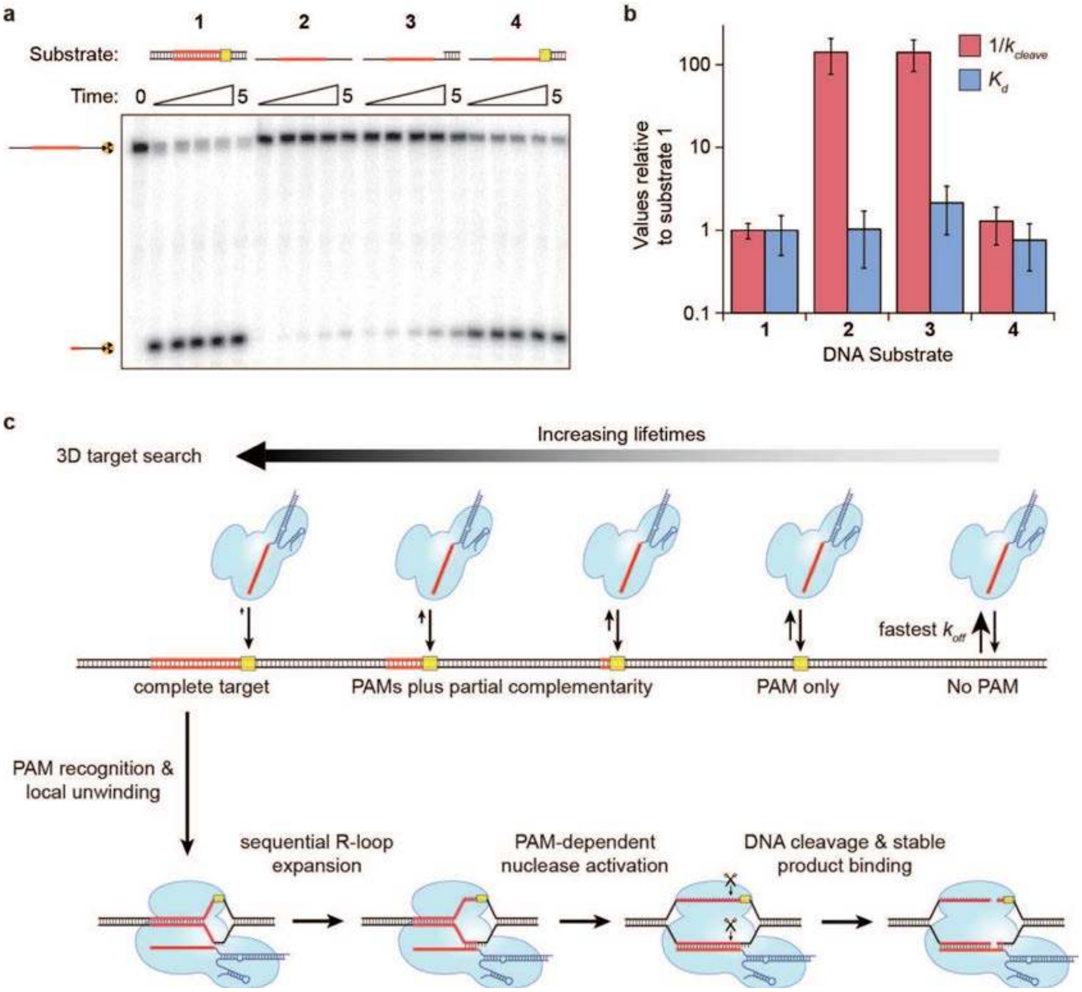


Figure 4.12 | PAM recognition regulates Cas9 nuclease activity. (a) Cleavage assay with single-stranded, double-stranded, and partially double-stranded substrates. **(b)** Relative affinities and cleavage rates; (mean ± s.d.). **(c)** Model for target search, recognition and cleavage by Cas9:RNA. The search initiates through random 3D collisions. Cas9:RNA rapidly dissociates from non-PAM DNA, but binds PAMs for longer times and samples adjacent DNA for guide RNA complementarity, giving rise to a heterogeneous population of intermediates. At correct targets, Cas9:RNA initiates formation of an RNA:DNA heteroduplex, and R-loop expansion propagates via sequential unwinding. The DNA is cleaved and Cas9:RNA remains bound to the cleaved products.

4.5 Discussion

Our results suggest a model for target binding and cleavage by Cas9:RNA involving an unanticipated level of importance for PAM sequences at each stage of the reaction (**Fig. 4.12c**). Although details may differ, we hypothesize that PAM interactions may function similarly for other CRISPR RNA-guided surveillance complexes (Esvelt et al., 2013; Hou et al., 2013; Marraffini and Sontheimer, 2010b; Mojica et al., 2009; Sashital et al., 2012; Semenova et al., 2011; Wiedenheft et al., 2011b). The Cas9:RNA target search begins with random collisions with DNA. However, rather than sampling all DNA equivalently, Cas9:RNA accelerates the search by rapidly dissociating from non-PAM sites, thereby reducing the amount of time spent at off-targets. Only upon binding to a PAM site does Cas9:RNA interrogate the flanking DNA for guide RNA complementarity, as was previously hypothesized for Cas9 (Jinek et al., 2012) and a distinct CRISPR RNA-guided complex (Cascade) (Sashital et al., 2012). A requirement for initial PAM recognition also eliminates the potential for suicidal self-targeting, since perfectly matching targets within the bacterial CRISPR locus are not flanked by PAMs. Our results suggest that PAM recognition coincides with initial destabilization of the adjacent sequence, as evidenced from experiments using a bubble-containing DNA substrate, followed by sequential extension of the RNA:DNA heteroduplex. This mechanism explains the emergence of seed sequences, because mismatches encountered early in a directional melting-in process would prematurely abort target interrogation. Moreover, the complex dissociation kinetics observed on non-target λ-DNA would arise from heterogeneity in the potential target sites as Cas9:RNA probes sequences adjacent to PAMs for guide RNA complementarity. Binding to a correct target then leads to activation of both nuclease domains. This step also requires PAM recognition, providing an unanticipated level of PAM-dependent regulation that may ensure further protection against self-cleavage of the CRISPR locus. Interestingly, Cas9:RNA does not dissociate from the cleaved DNA except under extremely harsh conditions, suggesting that Cas9:RNA may remain bound to the cleaved site in vivo (Garneau et al., 2010) and require other cellular factors to promote recycling. Finally, our data indicate that efforts to minimize off-target effects during genome engineering using Cas9:RNA complexes need only consider off-targets adjacent to a PAM, because potential targets lacking a PAM are unlikely to be interrogated (Fu et al., 2013; Hsu et al., 2013; Jiang et al., 2013; Pattanayak et al., 2013).

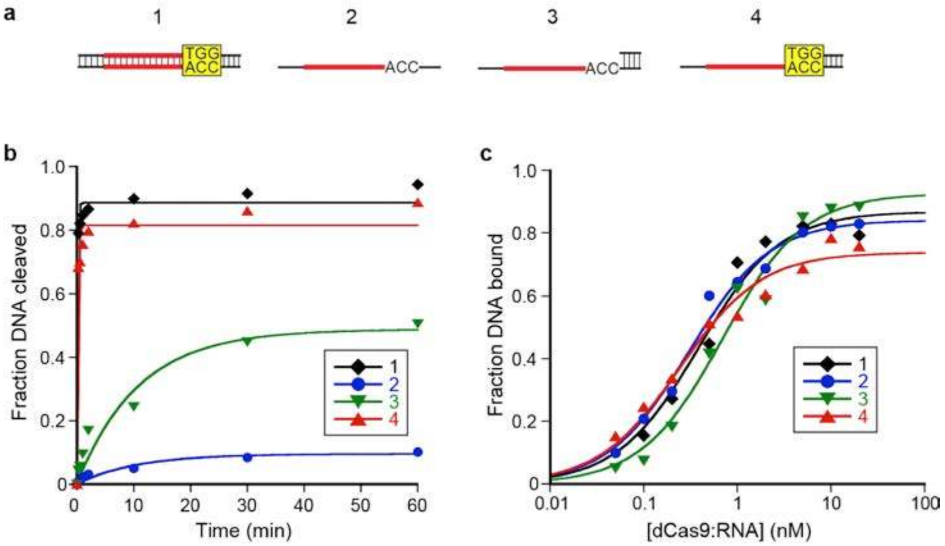


Figure 4.13 | PAM recognition activates the nuclease activity of Cas9. (a) DNA substrates were prepared using the λ2 target sequence as indicated (top), where the flanking region extending beyond the PAM was 16 bp (cleavage experiments) or 26 bp (binding experiments). **(b)** For cleavage experiments, substrates were prepared by annealing the radiolabeled target strand (i.e. substrate 2) to a 5X excess of cold complement, and 1 nM DNA was reacted with 50 nM Cas9:RNA at room temperature. Reaction products were resolved by 10% denaturing PAGE, and the quantified data were fit with single-exponential decays (solid lines). Results from three independent experiments

yielded apparent pseudo-first order cleavage rate constants of 9.0 ± 2.0 min^{-1} (substrate 1), 0.067 ± 0.027 min^{-1} (substrate 2), 0.066 ± 0.024 min^{-1} (substrate 3), and 7.3 ± 3.2 min^{-1} (substrate 4), and are presented as values relative to substrate 1 in **Fig. 4.12b**. Rate constants for substrates 2 and 3 are likely overestimates, since the reactions did not approach completion and the data were best fit with amplitudes well below 1. **(c)** For binding experiments, substrates were gel purified after annealing the radiolabeled target strand (i.e. substrate 2) to a 10X excess of cold complement. Binding reactions contained ~0.1 nM DNA and increasing concentrations of dCas9:RNA, and were incubated at 37 °C for one hour before being resolved by 5% native PAGE. The quantified data were fit with standard binding isotherms (solid lines). Results from three independent experiments yielded apparent K_d values of 0.27 ± 0.14 nM (substrate 1), 0.28 ± 0.12 nM (substrate 2), 0.59 ± 0.18 nM (substrate 3), and 0.21 ± 0.06 nM (substrate 4), and are presented as values relative to substrate 1 in **Fig. 4.12b**.

Chapter 5

Structures of Cas9 endonucleases reveal RNA-mediated conformational activation

† Part of the work presented in this chapter has previously been published in the following research article: Jinek, M.*, Jiang, F.*, Taylor, D.W*., Sternberg, S.H.*, Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., Kaplan, M., Iavarone, A.T., Charpentier, E., Nogales, E., Doudna, J.A. (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science *343*, 1247997-1–11. (*These authors contributed equally.)

‡ Martin Jinek, Enbo Ma, and Fuguo Jiang solved crystal structures of SpyCas9 and AnaCas9, with the assistance of Carolin Anders, Michael Hauer, Kaihong Zhou, Steven Lin, and Emmanuelle Charpentier. David Taylor performed all electron microscopy experiments. Samuel Sternberg prepared samples for electron microscopy experiments with the assistance of Matias Kaplan, conceived the UV-crosslinking experiments, and performed biochemical experiments. Emine Kaya performed biochemical and UV-crosslinking experiments. Anthony Iavarone performed mass spectrometry analysis. All authors were involved in designing various experiments. Eva Nogales and Jennifer Doudna supervised the project.

5.1 Abstract

Type II CRISPR-Cas systems use an RNA-guided DNA endonuclease, Cas9, to generate double-strand breaks in invasive DNA during an adaptive bacterial immune response. Cas9 has been harnessed as a powerful tool for genome editing and gene regulation in many eukaryotic organisms. Here we report 2.6 and 2.2 Å resolution crystal structures of two major Cas9 enzymes subtypes, revealing the structural core shared by all Cas9 family members. The architectures of Cas9 enzymes define nucleic acid binding clefts, and single-particle electron microscopy reconstructions show that the two structural lobes harboring these clefts undergo guide RNA-induced reorientation to form a central channel where DNA substrates are bound. The observation that extensive structural rearrangements occur before target DNA duplex binding implicates guide RNA loading as a key step in Cas9 activation.

5.2 Introduction

Bacteria and archaea target invasive DNA using RNA-guided adaptive immune systems encoded by CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)-Cas (CRISPR-associated) genomic loci (Al-Attar et al., 2011; Sorek et al., 2013; Terns and Terns, 2011; Wiedenheft et al., 2012). Following integration of short fragments of invader-derived DNA into a CRISPR array within the host chromosome (Barrangou et al., 2007), enzymatic processing of CRISPR transcripts produces mature CRISPR RNAs (crRNAs) that direct Cas protein-mediated targeting of DNA bearing complementary sequences (protospacers) to foreign nucleic acids (Brouns et al., 2008). While Type I and III CRISPR–Cas systems rely on multi-protein complexes for crRNA-guided DNA targeting (Makarova et al., 2011b; Sorek et al., 2013; Wiedenheft et al., 2012), Type II systems employ a single RNA-guided endonuclease, Cas9, that requires both a mature crRNA and a *trans*-activating crRNA (tracrRNA) for target DNA recognition and cleavage (Jinek et al., 2012; Karvelis et al., 2013). Both a seed sequence in the crRNA and conserved protospacer adjacent motif (PAM) sequence in the target are crucial for Cas9-mediated cleavage (Gasiunas et al., 2012; Jinek et al., 2012).

Cas9 proteins are abundant across the bacterial kingdom, but vary widely in both sequence and size. All known Cas9 enzymes contain an HNH domain that cleaves the DNA strand complementary to the guide RNA sequence (target strand), and a RuvC nuclease domain required for cleaving the non-complementary strand (non-target strand), yielding double-strand DNA breaks (DSBs) (Gasiunas et al., 2012; Jinek et al., 2012). In addition, Cas9 enzymes contain a highly conserved arginine-rich (Arg-rich) region previously suggested to mediate nucleic acid binding (Sampson et al., 2013). Based on CRISPR-Cas locus architecture and protein sequence phylogeny, Cas9 genes cluster into three subfamilies: Type II-A, II-B, and II-C (Chylinski et al., 2013; Makarova et al., 2011a). Cas9 proteins found in II-A and II-C subfamilies typically contain ~1400 or ~1100 amino acids, respectively.

The ability to program Cas9 for DNA cleavage at specific sites defined by guide RNAs has led to its adoption as a versatile platform for genome engineering (Mali et al., 2013b). When directed to target loci in eukaryotes by either dual crRNA:tracrRNA guides or chimeric single-guide RNAs, Cas9 generates site-specific DSBs that are repaired either by non-homologous end joining (NHEJ) or homologous recombination (HR) (Cong et al., 2013; Jinek et al., 2013; Mali et al., 2013c), which can be exploited to modify genomic sequences in the vicinity of the Cas9-generated DSBs. Furthermore, catalytically inactive Cas9 alone or fused to transcriptional activation or repression domains can be used to control transcription at sites defined by guide RNAs (Gilbert et al., 2013; Mali et al., 2013a; Qi et al., 2013). Both Type II-A and Type II-C Cas9 proteins have been used in eukaryotic genome editing (Esvelt et al., 2013; Hou et al., 2013). Smaller Cas9 proteins, encoded by more compact genes, are potentially advantageous for cellular delivery using vectors that have limited size such as adeno-associated virus (AAV) and lentivirus.

Here we present the crystal structures of Cas9 enzymes from the two major enzyme subclasses (Type II-A and Type II-C). Both structures reveal the fundamental RNA-guided DNA endonuclease architecture, the locations of both active sites, and the likely nucleic acid binding clefts. Biochemical experiments show that PAM recognition occurs through a composite binding site that is disordered in the absence of guide RNA and substrate interactions. Single-particle electron microscopy structures demonstrate that guide RNA loading triggers a conformational change in Cas9 for productive DNA surveillance. Together these data provide insights into the function, regulation and evolution of the Cas9 enzyme family.

5.3 Materials and Methods

5.3.1 SpyCas9 expression and purification

Streptococcus pyogenes Cas9 (SpyCas9) was cloned into a custom pET-based expression vector encoding an N-terminal His$_6$-tag followed by Maltose-Binding Protein (MBP) and a TEV protease cleavage site (Jinek et al., 2012). Point mutations were introduced into SpyCas9 using site-directed mutagenesis and verified by DNA sequencing.

For crystallization, wild-type (WT) and K848C mutant SpyCas9 proteins were expressed and purified essentially as described (Jinek et al., 2012). The protein was purified by a combination of Ni-NTA affinity, cation exchange (SP sepharose) and gel filtration (Superdex 200) chromatography steps. The final gel filtration step was carried out in elution buffer containing 20 mM HEPES-KOH pH 7.5, 250 mM KCl and 1 mM TCEP. The protein was concentrated to 4-6 mg ml^{-1} and flash frozen in liquid N$_2$. Selenomethionine (SeMet)-substituted SpyCas9 was expressed as described (Wiedenheft et al., 2009) and purified as for native SpyCas9, except that all chromatographic solutions were supplemented with 5 mM TCEP.

For crosslinking and biochemical assays, WT and mutant SpyCas9 proteins were expressed as His$_{10}$-MBP-TEV fusions and purified as described (Jinek et al., 2012), with the following modifications: All buffers contained 20 mM Tris-Cl pH 7.5, 5% glycerol, and 1 mM TCEP. The NaCl concentration was maintained at 500 mM during Ni-NTA chromatography and overnight dialysis with TEV protease. In order to remove TEV protease, His$_{10}$-MBP, and any uncleaved His$_{10}$-MBP-SpyCas9, the TEV-treated protein sample was run over Ni-NTA agarose resin again. SpyCas9 was dialyzed into Buffer A (20 mM Tris-Cl pH 7.5, 125 mM KCl, 5% glycerol, 1 mM TCEP) for 3 h at 4°C, and then applied onto a 5 ml HiTrap SP HP sepharose column (GE Healthcare). After washing with three column volumes of Buffer A, SpyCas9 was eluted using a linear gradient from 0-100% Buffer B (20 mM Tris-Cl pH 7.5, 1 M KCl, 5% glycerol, 1 mM TCEP) over 20 column volumes. The protein was further purified by gel filtration chromatography on a Superdex 200 16/60 column (GE Healthcare) in SpyCas9 Storage Buffer (20 mM Tris-Cl pH 7.5, 200 mM KCl, 5% glycerol, 1 mM TCEP).

5.3.2 SpyCas9 crystallization and structure determination

SpyCas9 crystals were grown using the hanging drop vapor diffusion method at 20 °C by mixing equal volumes (1.5 ml + 1.5 ml) of protein solution and crystallization buffer (0.1 M Tris-Cl pH 8.5, 0.2-0.3 M Li_2SO_4 and 14-15% (w/v) PEG 3350). Crystal nucleation and growth was gradually improved using iterative microseeding. For diffraction experiments, the crystals were cryoprotected in situ by stepwise exchange into a solution containing 0.1 M Tris-Cl pH 8.5, 0.1 M Li_2SO_4, 35% (w/v) PEG 3350, and 10% ethylene glycol in five steps executed at 5 min intervals. In each step, 0.5 ml of mother liquor was removed from the crystal drop and replaced with 0.5 ml cryoprotectant. After the final cryoprotectant addition, the crystals were incubated for an additional 5 min, transferred to a drop containing 100% cryoprotectant for 30 s, and then flash cooled in liquid N_2. Diffraction data were measured at beamlines 8.2.1 and 8.2.2 of the Advanced Light Source (Lawrence Berkeley National Laboratory), and beamlines PXI and PXIII of the Swiss Light Source (Paul Scherer Institute) and processed using XDS (Kabsch, 2010). Data collection statistics are shown in **Table 5.1**. The crystals belonged to space group $P2_12_12$ and contained two molecules of SpyCas9 in the asymmetric unit related by pseudotranslational, non-crystallographic symmetry. High-resolution native data to 2.62 Å resolution were measured from an unusually large crystal cryoprotected in the presence of 1 mM $MgCl_2$. A complete native data set was obtained by collecting four datasets (40° rotation per dataset) from different exposed parts of the crystal.

Phasing was performed as follows. A 4.2 Å resolution single-wavelength anomalous diffraction (SAD) dataset was measured at the selenium peak wavelength using a SeMet-substituted SpyCas9 crystal. However, due to small crystal size and low resolution, the anomalous signal in this dataset was too weak to locate the selenium sites. Additional phases were therefore obtained from SpyCas9 crystals soaked in sodium tungstate. The crystals were soaked by stepwise exchange of the lithium sulfate containing mother liquor with 0.1 M Tris-Cl pH 8.5, 0.1 M Na_2WO_4, 15% (w/v) PEG 3350, and then cryoprotected by stepwise exchange (as described above) of the soak solution with cryoprotectant solution supplemented with 10 mM Na_2WO_4. Using these crystals, a highly redundant SAD 3.9 Å dataset was measured at the tungsten L-III absorption edge (1.2149 Å), and 16 tungstate sites were located using SHELXD (Schneider and Sheldrick, 2002). Further phase information came from peak-wavelength SAD datasets obtained from a crystal of SpyCas9 K848C mutant soaked in 1 mM thimerosal for 6 hr prior to cryoprotection (thimerosal soak), a WT SpyCas9 crystal soaked with 10 mM $CoCl_2$ during the cryoprotection procedure (Co soak), and a WT SpyCas9 crystal grown in the presence of 1 mM Er(III)-acetate. Refinement of the substructures and phase calculations were performed using the MIRAS procedure in AutoSHARP (Vonrhein et al., 2007) by combining initial tungstate SAD phases with the additional SAD data sets (SeMet, Co, Er and thimerosal) and the high-resolution native data. Phases were improved by density modification and two-fold non-crystallographic symmetry averaging using the Resolve module of the Phenix suite (Adams et al., 2010; Terwilliger, 2004). The resulting electron density maps were of excellent quality and allowed manual model building in COOT (Emsley and Cowtan, 2004). Selenium positions aided in assigning the sequence register. The atomic model of SpyCas9 was completed by iterative model building in COOT and refinement using Phenix.refine (Afonine et al., 2012). Refinement and model statistics are provided in **Table 5.1**.

The final atomic model has R_{work} and R_{free} values of 0.253 and 0.286, respectively, and good stereochemistry, as assessed with MolProbity (Davis et al., 2007), with 96.6% of the residues in the most favored regions of the Ramachandran plot and no outliers. The model contains two SpyCas9 molecules that superimpose with an overall rmsd of 1.1 Å over 1060 Ca atoms, the major difference being a ~5° hinge-like rotation of the HNH domain. In the atomic model, molecule A contains residues 4-102, 115-307, 314-447, 503-527, 540-567, 587-672, 677-714, 718-764, 775-791, 799-859, 862-902, 908-1027, 1036-1102, 1137-1146, 1159-1186, 1192-1242, and 1259-1363. Molecule B contains residues 4-103, 116-308, 310-447, 502-527, 539-570, 587-673, 676-713, 718-764, 773-791, 800-859, 862-902, 908-1025, 1036-1102, 1137-1148, 1160-1185, 1188-1241, and 1256-1363. The remaining residues do not appear ordered in electron density maps and could not be built. In the manuscript, the discussion of the SpyCas9 structure is based on molecule B, which is better ordered.

An additional dataset (at 3.1 Å resolution) was measured using a SpyCas9 crystal soaked in 20 mM $MnCl_2$ during the cryoprotection procedure. F_o-F_c difference maps calculated using the high-resolution model revealed two Mn^{2+} ions bound in the RuvC domain active site and 4 additional Mn^{2+} ions bound to each of the two SpyCas9 molecules. The HNH domain active site remained poorly ordered in this structure, and no Mn^{2+} binding was observed. The model was refined to an R_{work} and R_{free} of 0.252 and 0.278, respectively.

5.3.3 Endonuclease cleavage assays with SpyCas9

A synthetic 42-nt crRNA targeting a protospacer from the bacteriophage λ genome was purchased from Integrated DNA Technologies (IDT) and purified via 10% denaturing PAGE. tracrRNA was *in vitro* transcribed from a synthetic DNA template (IDT) using T7 RNA polymerase and corresponds to nucleotides 15-87 as described previously (Jinek et al., 2012). crRNA:tracrRNA duplexes (10 µM) were prepared by mixing equimolar amounts of crRNA and tracrRNA in Hybridization Buffer (20 mM Tris-Cl pH 7.5, 100 mM KCl, 5 mM $MgCl_2$), heating at 95 °C for 30 sec, and slow-cooling on the benchtop. SpyCas9:RNA complexes were reconstituted by mixing SpyCas9 with a 2X molar excess of the crRNA:tracrRNA duplex in Reconstitution Buffer (20 mM Tris-Cl pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 1 mM DTT) and incubating at 37°C for 10 minutes.

A 55 base-pair (bp) DNA target derived from the bacteriophage λ genome was prepared by mixing equimolar amounts of individual synthetic oligonucleotides (IDT) in Hybridization Buffer supplemented with 5% glycerol, heating for 1-2 minutes, and slow-cooling on the benchtop. Duplexes were separated from single-stranded DNA by 6% native PAGE conducted at 4°C, with 5 mM $MgCl_2$ added to the gel and the running buffer. The DNA was excised, eluted into 10 mM Tris-Cl, pH 8 at 4°C overnight, ethanol precipitated, and resuspended in Hybridization Buffer. Br-dU containing ssDNAs used in analytical crosslinking reactions were radiolabeled and hybridized with a 5X molar excess of the unlabeled complementary strand. Cleavage reactions were performed at room temperature in Reaction Buffer (20 mM Tris-Cl pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 5% glycerol, 1 mM DTT) using 1 nM radiolabeled dsDNA substrates and 1 nM or 10 nM Cas9:RNA. Aliquots (10 µl) were removed at various time points and quenched by mixing with an equal volume of formamide gel loading buffer supplemented with 50 mM EDTA. Cleavage products were resolved by 10% denaturing PAGE and visualized by phosphorimaging (GE Healthcare). The sequences of DNA and RNA oligonucleotides used in this study are listed in **Table 5.4**.

5.3.4 Preparation of crosslinked peptide-DNA heteroconjugates for mass spectrometry

200 pmol of catalytically inactive (D10A/H840A) Cas9 was reconstituted with crRNA:tracrRNA and incubated with a 10X molar excess of Br-dU containing dsDNA substrate for 30 min at room temperature in Reaction Buffer. Reactions were transferred into the lid of open PCR tubes and irradiated with UV-light (308 nm) for 30 min at room temperature. Crosslinked samples were denatured with 6 M urea for 1 h at 65°C, diluted to 0.5 M urea with 25 mM ammonium bicarbonate, and digested with 1 ng trypsin overnight at room temperature. Samples were concentrated to a final volume of 50 µL and desalted with Illustra MicroSpin G-25 Columns (GE Healthcare). Samples were then treated with 1,000 Units of Nuclease S1 (Sigma Aldrich) for 1 h at 37 °C in 30 mM ammonium acetate pH 5.7, 10 mM $CaCl_2$ and 0.1 mM $ZnCl_2$ in a total volume of 60 µL. In order to remove remaining phosphate groups at the crosslink site, 7 µL of 10X Antarctic Phosphatase buffer and 5 Units of Antarctic Phosphatase (New England BioLabs) were added to the reactions, and samples were incubated for an additional hour at 37 °C.

5.3.5 Liquid chromatography-tandem mass spectrometry (LS-MS/MS)

Tryptic digests of crosslinked proteins were analyzed using a Dionex UltiMate3000 RSLCnano liquid chromatograph that was connected in-line with an LTQ Orbitrap XL mass spectrometer equipped with a nanoelectrospray ionization source (nanoESI; Thermo Fisher Scientific). The LC was equipped with a C18 analytical column (Acclaim® PepMap RSLC, 150 mm length × 0.075 mm inner diameter, 2 µm particles, 100 Å pores, Thermo) and a 1 µL sample loop. Solvent A was 99.9% water/0.1% formic acid and solvent B was 99.9% acetonitrile/0.1% formic acid (v/v). Samples were placed in polypropylene autosampler vials with septa caps (Wheaton) and loaded into the autosampler compartment (maintained at 4 °C) prior to analysis. The elution program consisted of isocratic flow at 5% B for 4 min, a linear gradient to 35% B over 98 min, isocratic flow at 95% B for 6 min, and isocratic flow at 5% B for 12 min, at a flow rate of 300 nL min^{-1}. The column exit was connected to the nanoESI emitter in the ion source of the mass spectrometer using polyimide-coated, fused-silica tubing (20 µm inner diameter × 280 µm outer diameter, Thermo).

Full-scan mass spectra were acquired in the positive ion mode over the range m/z = 350 to 1500 using the Orbitrap mass analyzer, in profile format, with a mass resolution setting of 60,000 (at m/z = 400, measured at full width at half-maximum peak height). Under these conditions, isotopic distributions of singly and multiply charged peptide ions were resolved in the full-scan mass spectra. Thus, a precursor ion's mass and charge were determined independently, i.e. the ion charge was determined from the reciprocal of the spacing between adjacent isotope peaks in the m/z spectrum. In the data-dependent mode, the six most intense ions exceeding an intensity threshold of 30,000 counts were selected from each full-scan mass spectrum for tandem mass spectrometry (MS/MS) analysis using collision-induced dissociation (CID). MS/MS spectra were acquired using the linear ion trap, in centroid format, with the following parameters: isolation width 3 m/z units, normalized collision energy 28%, default charge state 2+, activation Q 0.25, and activation time 30 ms. Real-time charge state screening was enabled to exclude singly charged ions and unassigned charge states from MS/MS analysis. To avoid the occurrence of redundant MS/MS measurements, real-time dynamic exclusion was enabled to preclude re-selection of previously analyzed precursor ions, with the following parameters: repeat count 2, repeat duration 10 s, exclusion list size 500, exclusion duration 60 s, and exclusion mass width 20 ppm (relative to mass). Data were analyzed using Xcalibur (version 2.0.7 SP1, Thermo) and Proteome Discoverer (version 1.3, Thermo, SEQUEST algorithm) software. Validation of identified cross-linked peptides was by manual inspection of the MS/MS spectra, i.e. to verify the occurrence of b- and y-type fragment ions (*67*) that identify the peptide sequences.

5.3.6 DNA binding experiments

SpyCas9:crRNA:tracrRNA complexes (containing wild-type SpyCas9 or PAM loop mutants $PWN_{475-477}$→AAA, $DWD_{1125-1127}$→AAA , and $PWN_{475-477}/DWD_{1125-1127}$→AAA/AAA) were reconstituted for 10 min at 37 °C in Reaction Buffer before being incubated with ~1 nM radiolabeled DNA target for 60 minutes at 37 °C. Reactions were resolved by 5% native PAGE and visualized by phosphorimaging (GE Healthcare).

5.3.7 AnaCas9 expression and purification

Full-length *Actinomyces naeslundii* Cas9 (AnaCas9; residues 1-1101) was subcloned into a custom pET-based expression vector with an N-terminal His_{10}-tag followed by Maltose-Binding Protein (MBP) and a TEV protease cleavage site. The protein was overexpressed in *Escherichia coli* strain Rosetta (DE3) and was purified to homogeneity by immobilized metal ion affinity chromatography and heparin affinity chromatography. An additional gel filtration chromatography step (HiLoad 16/60 Superdex200, GE Healthcare) was added to further purify AnaCas9 and remove trace nucleic acid contaminants prior to crystallization. Purified AnaCas9 protein in gel filtration buffer (50 mM HEPES 7.5, 300 mM KCl, 2 mM TCEP, 5% glycerol) was snap frozen in liquid nitrogen and stored at -80°C. Selenomethionine–labeled AnaCas9 protein was expressed in Rosetta (DE3) cells grown in M9 minimal medium supplemented with 50 mg ml^{-1} L-SeMet (Sigma) and specific amino acids to inhibit endogenous methionine synthesis. The SeMet-substituted protein was then purified using the same procedure as for the native AnaCas9 protein.

5.3.8 AnaCas9 crystallization and structure determination

Crystals of native and SeMet-substituted AnaCas9 were grown by the hanging drop vapor diffusion method at 20 °C. Aliquots (2.5 µl) of 4.5 mg ml^{-1} native AnaCas9 protein in 50 mM HEPES 7.5, 300 mM KCl, 2mM TCEP, 5% glycerol were mixed with 2.5 µl of reservoir solution containing 10% (w/v) PEG 8000, 0.25 M calcium acetate, 50 mM magnesium acetate and 5 mM spermidine. Crystals appeared after 1–2 days, and they grew to a maximum size of 0.15 × 0.20 × 0.35 mm over the course of 6 days. SeMet-substituted AnaCas9 crystals were grown and optimized under the same conditions. For cryogenic data collection, crystals were transferred into crystallization solutions containing 30% (v/v) glycerol as the cryoprotectant and then flash-cooled at 100 K. Native and SeMet single-wavelength anomalous diffraction (SAD) datasets were collected at beamline 8.3.1 of the Advanced Light Source, Lawrence Berkeley National Laboratory. Data from manganese-soaked AnaCas9 crystals were collected at the 8.2.2 beamline of the Advanced Light Source, Lawrence Berkeley National Laboratory. All diffraction data were integrated using Mosflm and scaled in SCALA (*54*, *55*).

The AnaCas9 structure was solved using the single anomalous dispersion phasing method. Using SeMet data between 79.0 and 3.2 Å resolution, both SHELXD/HKL2MAP (Schneider and Sheldrick, 2002) and HySS in Phenix (Zwart et al., 2008) detected a total of 13 out of 18 possible selenium sites in the asymmetric unit. Initial phases were calculated using SOLVE followed by solvent flattening with RESOLVE to produce an electron-density map into which most of the protein residues could be unambiguously built (Terwilliger, 2004). The initial model automatically generated from Phenix AutoBuild module was subjected to subsequent iterative rounds of manual building with COOT (Emsley and Cowtan, 2004) and refinement against the 2.2 Å native data in Refmac (Murshudov et al., 1997) and Phenix (Afonine et al., 2012). The final model contains one zinc ion, two magnesium ions, AnaCas9 residues 8-49, 65-98, 134-170, and 225-1101, and has R_{work} and R_{free} values of 0.19 and 0.23, respectively. The N terminus (residues 1–7), loop regions (residues 50-64), and a portion of the alpha-helical lobe (residues 99-133, 171-224) are completely disordered. Model validation showed 94% of the residues in the most favored and 5.8% in the allowed regions of the Ramachandran plot. The structure of Mn^{2+}-bound AnaCas9 was obtained by molecular replacement using the program Phaser (McCoy et al., 2007), which revealed two unambiguously refined Mn^{2+} ions present in the RuvC active site. All statistics of the data processing and structure refinement of AnaCas9 are summarized in **Table 5.2**.

5.3.9 Complex reconstitution for negative-stain EM

All samples for EM (10 µl volumes) were prepared in Reaction Buffer at a final Cas9 concentration of 1 µM. Cas9:RNA complexes contained 2 µM crRNA:tracrRNA duplex and were incubated at 37 °C for 10 minutes before storing on ice until grid preparation. Cas9:RNA:DNA complexes were prepared by first generating Cas9:RNA as before and then adding the DNA duplex at 5 µM (unlabeled) or 2 µM (biotin labeled) and incubating an additional 10 minutes at 37 °C. When present, streptavidin (New England Biolabs) was added after formation of Cas9:RNA or Cas9:RNA:DNA complexes at a 2X unit excess over the biotinylated species, according to the manufacturer's unit definition (~65 ng/µL in the final reaction volume), followed by an additional 10 minute incubation at 37 °C before storing on ice. Catalytically inactive Cas9 (D10A/H840A) was used to generate the following samples: unlabeled Cas9:RNA:DNA, Cas9:RNA:DNA containing biotin modifications on one or both ends of the duplex, and Cas9:RNA:DNA containing an N-terminal MBP. Wild-type Cas9 was used to generate apo-Cas9 and all Cas9:RNA complexes.

5.3.10 Negative-stain electron microscopy

We diluted Cas9 complexes for negative-stain EM to a concentration of ~25-60 nM in 20 mM Tris-HCl pH 7.5, 200 mM KCl, 1 mM DTT, and 5% glycerol immediately before applying the sample to glow-discharged 400 mesh continuous carbon grids. After adsorption for 1 min, we stained the samples consecutively with six droplets of 2% (w/v) uranyl acetate solution, gently blotted off the residual stain, and air-dried the sample in a fume hood. Data were acquired using a Tecnai F20 Twin transmission electron microscope operated at 120 keV at a nominal magnification of either 80,000X (1.45 Å at the specimen level) or 100,000X (1.08 Å at the specimen level) using low-dose exposures (~20 e^{-}Å^{-2}) with a randomly set defocus ranging from −0.5 to −1.3 µm. A total of 300–400 images of each Cas9 sample were automatically recorded on a Gatan 4k x 4k CCD camera using the MSI-Raster application within the automated macromolecular microscopy software LEGINON (Suloway et al., 2005).

5.3.11 Single-particle pre-processing

All image pre-processing and two-dimensional classification was performed in Appion as described previously (Wiedenheft et al., 2011a). The contrast transfer function (CTF) of each micrograph was estimated, and particles were selected concurrently with data collection using ACE2 (Mallick et al., 2005) and a template-based particle picker (Roseman, 2004), respectively. Micrograph phases were corrected using ACE2 (Mallick et al., 2005), and the negatively-stained Cas9 particles were extracted using a 288 × 288-pixel box size. The particle stacks were binned by a factor of 2 for processing, and particles were normalized to remove pixels whose values were above or below 4.5-σ of the mean pixel value using XMIPP (Scheres et al., 2008).

5.3.12 Random conical tilt reconstruction

Initial models for reconstructions of both apo-Cas9 and Cas9:RNA:DNA samples were determined using random conical tilt (RCT) methodology (Radermacher et al., 1987). Briefly, tilt-pairs of micrographs were recorded manually at 0° and 55°, and *ab initio* models were generated using the RCT module (Voss et al., 2010) in Appion (Lander et al., 2009). Particles were correlated between tilt-pairs using TiltPicker (Voss et al., 2009), binned by 2, and extracted from raw micrographs. Reference-free class averages were produced from untilted particle images by iterative 2D alignment and classification using MSA-MRA in IMAGIC (van Heel et al., 1996). These class averages served as references for SPIDER (Frank et al., 1996) reference-based alignment and classification, and RCT volumes were calculated for each class average using back-projection in SPIDER based on these angles and shifts. The RCT model from the most representative class (largest number of particles) was low-pass filtered to 60-Å resolution and used to assign Euler angles to the entire data set of reference-free class averages. The resulting low-resolution model was again low-pass filtered to 60-Å resolution and used as the initial model for refinement of the three-dimensional structure by iterative projection matching using the untilted particle images as previously described (Wiedenheft et al., 2011b), with libraries from EMAN2 and SPARX software packages (Hohn et al., 2007; Tang et al., 2007).

5.3.13 Domain mapping and localization of RNA- and DNA-ends

Particle stacks were binned by a factor of 2 and subjected to five rounds of iterative multivariate statistical analysis (MSA) and multi-reference alignment (MRA) using the IMAGIC (van Heel et al., 1996) software package, to generate two-dimensional class averages of each complex. The resulting set of class averages for each species was normalized using 'proc2d' in EMAN (Ludtke et al., 1999). The EMAN classification program 'classesbymra' was used to match the labeled class average to the best-matching unlabeled class average based on cross-correlation coefficients. The difference maps were calculating by subtracting the unlabeled class average from the labeled class averages using 'proc2d' in EMAN. This same strategy was used to match the unlabeled class average to the best-matching reprojection of the corresponding structure. The Euler angles used for creating the reprojection were applied to the 3D electron density using ‘proc3d,’ and the surface representation visualized in Chimera (Pettersen et al., 2004) is shown along with its corresponding reprojection.

5.3.14 3D reconstruction and analysis

Three-dimensional reconstructions were all performed using an iterative projection-matching refinement with libraries from the EMAN2 and SPARX software packages (Hohn et al., 2007; Tang et al., 2007). Refinement of the RCT starting models began using an angular increment of 25°, progressing down to 4° for all reconstructions. The resulting model was again low-pass filtered to 60-Å resolution and subjected to iterative projection-matching refinement to obtain the final structure. In an alternative approach for apo-Cas9 and Cas9:RNA:DNA, we used a low-pass filtered model of the other structure after initial refinement with untilted particles as an initial model for the above-mentioned projection matching refinement. This led to EM densities with similar structural features as the RCT models, and the structures converged to the final models presented. The resolution was estimated by splitting the particle stack into two equally sized data sets and calculating the Fourier shell correlation (FSC) between each of the back-projected volumes. The final reconstructions of Cas9, Cas9:RNA, and Cas9:RNA:DNA showed structural features to ~19-Å, ~21-Å, and ~19-Å resolution, respectively, based on the 0.5 Fourier shell correlation criterion. Reprojections of the final three-dimensional reconstruction showed excellent agreement with the reference-free class averages and displayed a large distribution of Euler angles, despite some preferential orientations of the particles on the carbon film.

The final reconstruction was segmented using Segger (Pintilie et al., 2010) in Chimera (Pettersen et al., 2004) based on inspection of the similarities between lobes in the apo-Cas9 and Cas9:RNA:DNA reconstructions. A modeled A-form duplex was manually docked into the map with Chimera, using information from the labeling experiments and map segmentation, and by accommodating the substrate within the channel in the EM reconstruction. While the absolute handedness of our apo-Cas9 reconstruction could be confirmed using the X-ray crystal structure, the relative handedness of our Cas9:RNA:DNA reconstruction is uncertain. Free hand tests performed on this sample failed, likely due to the small and/or dynamic nature of the enzyme. The model we present is based on the alpha-helical domain from the crystal structure having a more optimal CCC with the larger lobe of this reconstruction (0.83) than this lobe using the reconstruction of opposite handedness (0.74).

5.3.15 Enzymatic footprinting experiments

DNA targets (55 bp) were prepared by 5'-radiolabeling either the target or displaced non-target strand and then hybridizing it to a 5X molar excess of unlabeled complementary strand. After incubating catalytically inactive (D10A/H840A) SpyCas9:crRNA:tracrRNA complexes (100 nM) with ~1 nM DNA substrate for 30 minutes at 37 °C in Reaction Buffer, 100 units of exonuclease III (NEB) or 1.2 µg nuclease P1 (Sigma) was added and reactions were incubated an additional 10 minutes at 37 °C before quenching with formamide gel loading buffer supplemented with 50 mM EDTA. Reaction products were resolved by 15% denaturing (7M urea) PAGE and visualized by phosphorimaging (GE Healthcare). Control reactions contained a non-targeting crRNA that is not complementary to the 55-bp DNA substrate. To define the sequence register of enzymatic reaction products, a DNA ladder was generated by 5'-radiolabeling the synthetic target or non-target strand without prior gel purification and compared to DNA cleavage products using active SpyCas9:RNA or FokI and BglI restriction enzymes (NEB). Note that we observed SpyCas9:RNA cleaving the non-target strand between nucleotides 4 and 5 from the PAM end, in contrast to the cleavage site observed previously (Jinek et al., 2012).

5.4 Results

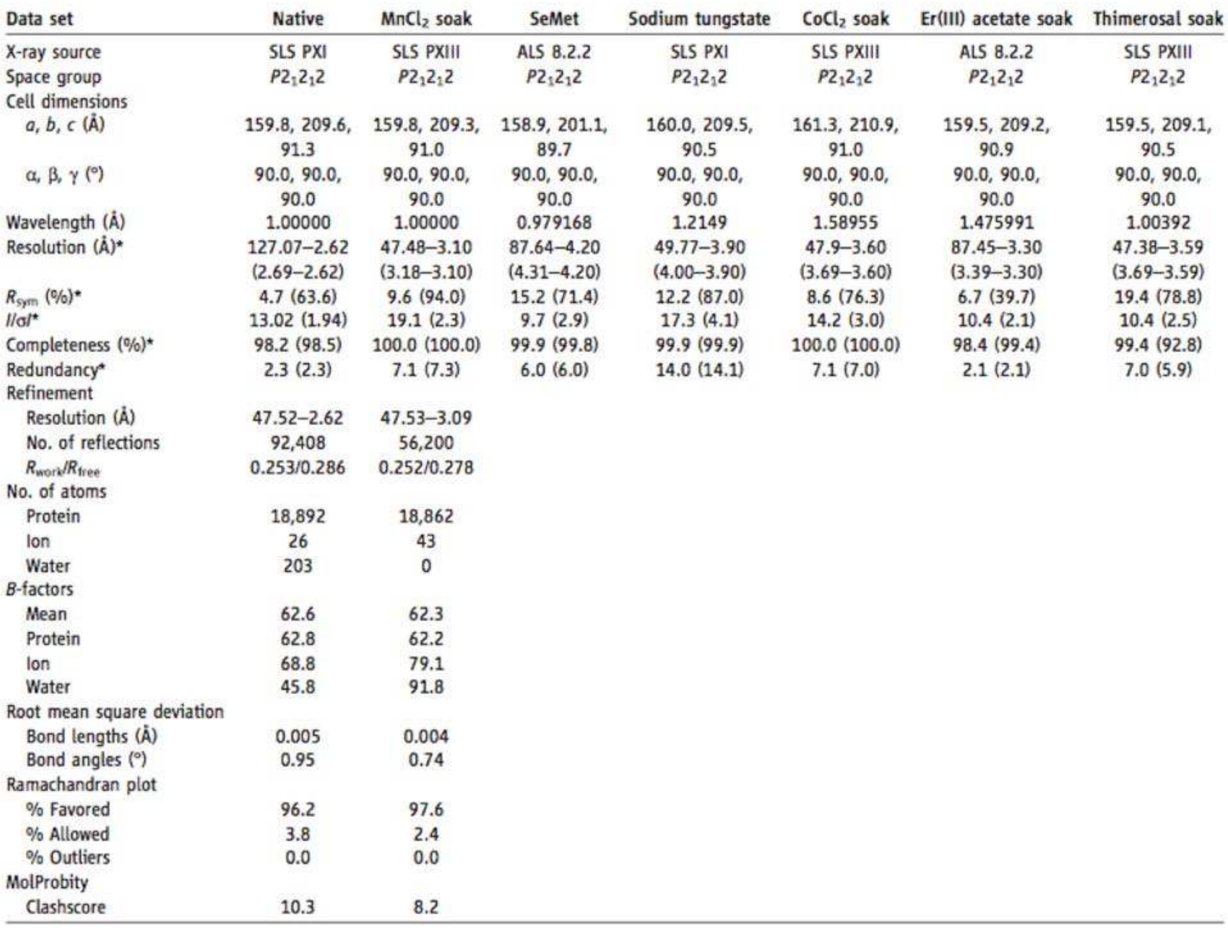
5.4.1 *S. pyogenes* Cas9 structure reveals a two-lobed architecture with adjacent active sites

Streptococcus pyogenes Cas9 (SpyCas9) is a prototypical Type II-A Cas9 protein consisting of well-conserved RuvC and HNH domains, and flanking regions lacking apparent sequence similarity to known protein structures (**Fig. 5.1a**). As the first biochemically characterized Cas9, SpyCas9 is used in the majority of current CRISPR-based genetic engineering methodologies (Mali et al., 2013b). To obtain structural insights into the architecture of SpyCas9, we determined the 2.6 Å resolution crystal structure of the enzyme (**Table 5.1**). The structure reveals that SpyCas9 is a crescent-shaped molecule with approximate dimensions of 100 Å x 100 Å x 50 Å (**Fig. 5.1b & 5.2**). The enzyme adopts a distinct bi-lobed architecture comprising the nuclease domains and C-terminal domain in one lobe (the nuclease lobe) and a large alpha-helical domain in the other. The RuvC domain forms the structural core of the nuclease lobe, a six-stranded beta sheet surrounded by four alpha helices, with all three conserved motifs contributing catalytic residues to the active site (**Fig. 5.2**). The HNH and RuvC domains are juxtaposed in the SpyCas9 structure, with their active sites located ~25 Å apart. The HNH domain active site is poorly ordered in apo-SpyCas9 crystals, suggesting that the active site may undergo conformational ordering upon nucleic acid binding. The C-terminal region of SpyCas9 contains a b-b-a-b Greek key domain that bears structural similarity to a domain found in topoisomerase II (Berger et al., 1996) (hereafter referred to as the Topo-homology domain, residues 1136^{Spy}-1200^{Spy}). A mixed a/b region (C-terminal domain, residues 1201^{Spy} -1363^{Spy}) forms a protrusion on the nuclease domain lobe. The structural halves of SpyCas9 are connected by two linking segments, one formed by the Arginine-rich region (residues 59^{Spy} -76^{Spy}) and the other by a disordered linker comprising residues 714^{Spy} -717^{Spy} (**Fig. 5.1b**). The total surface area buried between the two structural lobes in SpyCas9 is 1034 Å^2.

5.4.2 SpyCas9 contains two putative nucleic acid binding grooves

SpyCas9 contains two prominent clefts on one face of the molecule: a deep and narrow groove located within the nuclease lobe and a somewhat wider groove within the alpha-helical lobe (**Fig. 5.1c**). The nuclease lobe cleft is approximately 40 Å long, 15-20 Å wide and 15 Å deep, with the RuvC active site located at its bottom. The C-terminal domain forms one side of the cleft, while the HNH domain and a protrusion of the alpha-helical lobe forms the other. The concave surface of the alpha-helical lobe creates a wider, shallower groove that extends over almost its entire length (**Fig. 5.1c**). The groove is more than 25 Å across at its widest point, which would be sufficient to accommodate an RNA-RNA or DNA-RNA duplex. Its surface is highly positively charged (**Fig. 5.1d**), especially at the Arg-rich segment comprising $R69^{Spy}$, R70^{Spy}, R71^{Spy}, R75^{Spy} and K76^{Spy}. Multiple sulfate or tungstate ions are bound to the alpha-helical lobe in the SpyCas9 crystals (**Fig. 5.3**), hinting at a possible role in nucleic acid binding. Amino acid residues located in both the nuclease and alpha-helical lobe clefts are highly conserved within Type II-A Cas9 proteins (**Fig. 5.1e**), suggesting that both clefts play important functional roles. Since the RuvC domain mediates cleavage of the non-target DNA strand (Gasiunas et al., 2012; Jinek et al., 2012), the nuclease domain cleft likely binds the displaced non-target strand. Conversely, the alpha-helical lobe, which contains the Arg-rich segment, might be involved in binding the crRNA:tracrRNA guide RNA and/or the crRNA-target DNA heteroduplex. This would be consistent with the observation that a mutation in the Arg-rich region in *Francisella novicida* Cas9 leads to loss of RNA-guided targeting *in vivo* (Sampson et al., 2013).

Table 5.1 | X-ray data collection, refinement, and model statistics for SpyCas9.



Data set	Native	$MnCl_2$ soak	SeMet	Sodium tungstate	$CoCl_2$ soak	Er(III) acetate soak	Thimerosal soak
X-ray source	SLS PXI	SLS PXIII	ALS 8.2.2	SLS PXI	SLS PXIII	ALS 8.2.2	SLS PXIII
Space group	$P2_12_12$	$P2_12_12$	$P2_12_12$	$P2_12_12$	$P2_12_12$	$P2_12_12$	$P2_12_12$
Cell dimensions							
a, *b*, *c* (Å)	159.8, 209.6, 91.3	159.8, 209.3, 91.0	158.9, 201.1, 89.7	160.0, 209.5, 90.5	161.3, 210.9, 91.0	159.5, 209.2, 90.9	159.5, 209.1, 90.5
α, β, γ (°)	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0	90.0, 90.0, 90.0
Wavelength (Å)	1.00000	1.00000	0.979168	1.2149	1.58955	1.475991	1.00392
Resolution (Å)*	127.07–2.62 (2.69–2.62)	47.48–3.10 (3.18–3.10)	87.64–4.20 (4.31–4.20)	49.77–3.90 (4.00–3.90)	47.9–3.60 (3.69–3.60)	87.45–3.30 (3.39–3.30)	47.38–3.59 (3.69–3.59)
R_{sym} (%)*	4.7 (63.6)	9.6 (94.0)	15.2 (71.4)	12.2 (87.0)	8.6 (76.3)	6.7 (39.7)	19.4 (78.8)
$I/\sigma I$*	13.02 (1.94)	19.1 (2.3)	9.7 (2.9)	17.3 (4.1)	14.2 (3.0)	10.4 (2.1)	10.4 (2.5)
Completeness (%)*	98.2 (98.5)	100.0 (100.0)	99.9 (99.8)	99.9 (99.9)	100.0 (100.0)	98.4 (99.4)	99.4 (92.8)
Redundancy*	2.3 (2.3)	7.1 (7.3)	6.0 (6.0)	14.0 (14.1)	7.1 (7.0)	2.1 (2.1)	7.0 (5.9)
Refinement							
Resolution (Å)	47.52–2.62	47.53–3.09					
No. of reflections	92,408	56,200					
R_{work}/R_{free}	0.253/0.286	0.252/0.278					
No. of atoms							
Protein	18,892	18,862					
Ion	26	43					
Water	203	0					
B-factors							
Mean	62.6	62.3					
Protein	62.8	62.2					
Ion	68.8	79.1					
Water	45.8	91.8					
Root mean square deviation							
Bond lengths (Å)	0.005	0.004					
Bond angles (°)	0.95	0.74					
Ramachandran plot							
% Favored	96.2	97.6					
% Allowed	3.8	2.4					
% Outliers	0.0	0.0					
MolProbity							
Clashscore	10.3	8.2					

*Values in parentheses denote highest-resolution shell.

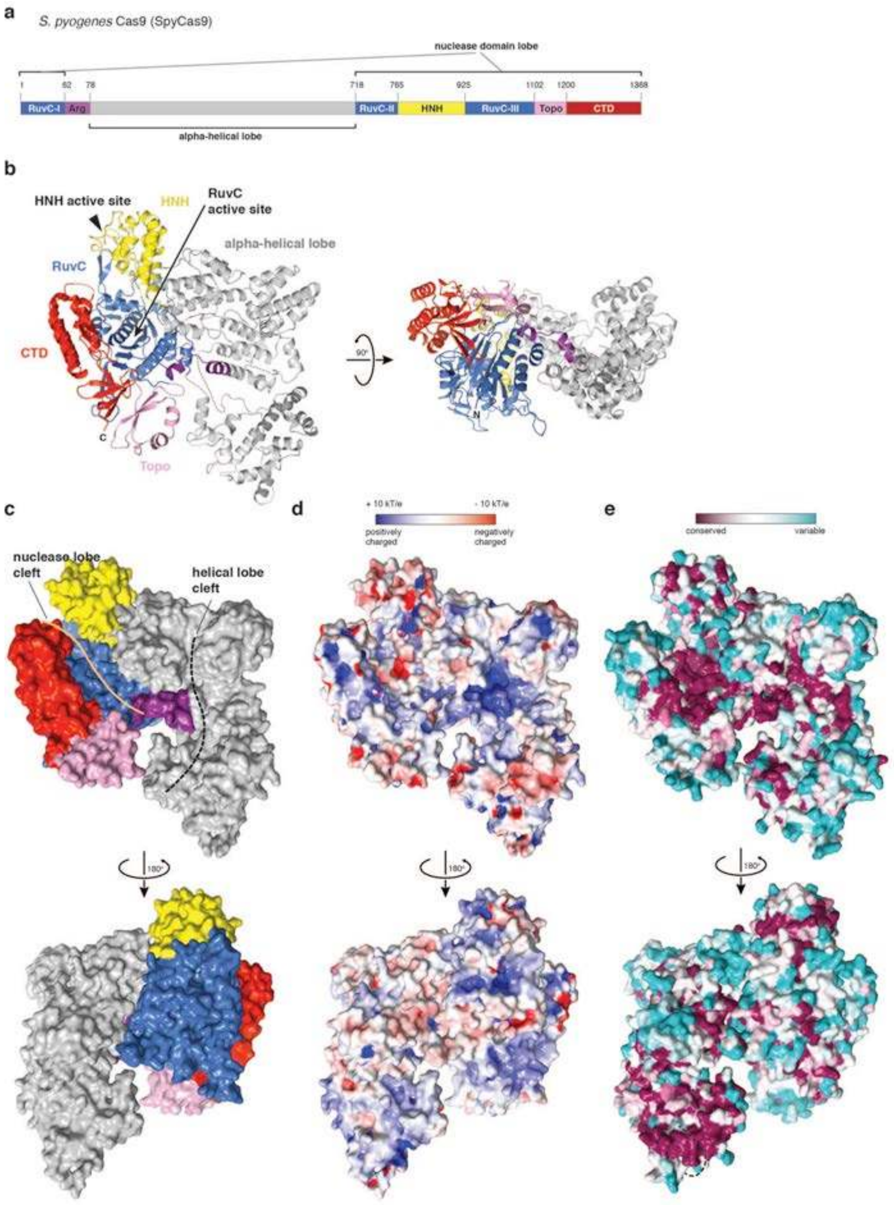
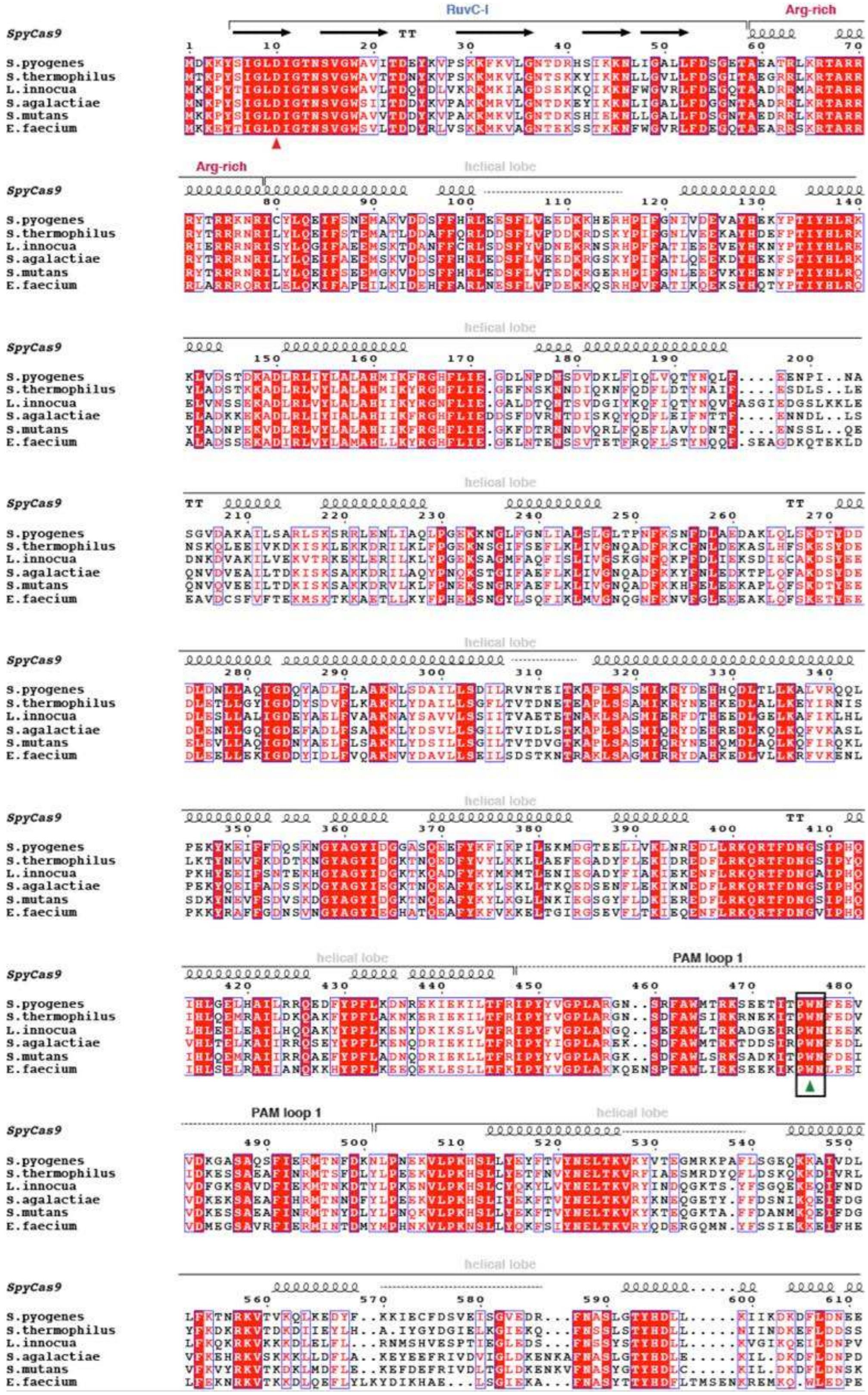
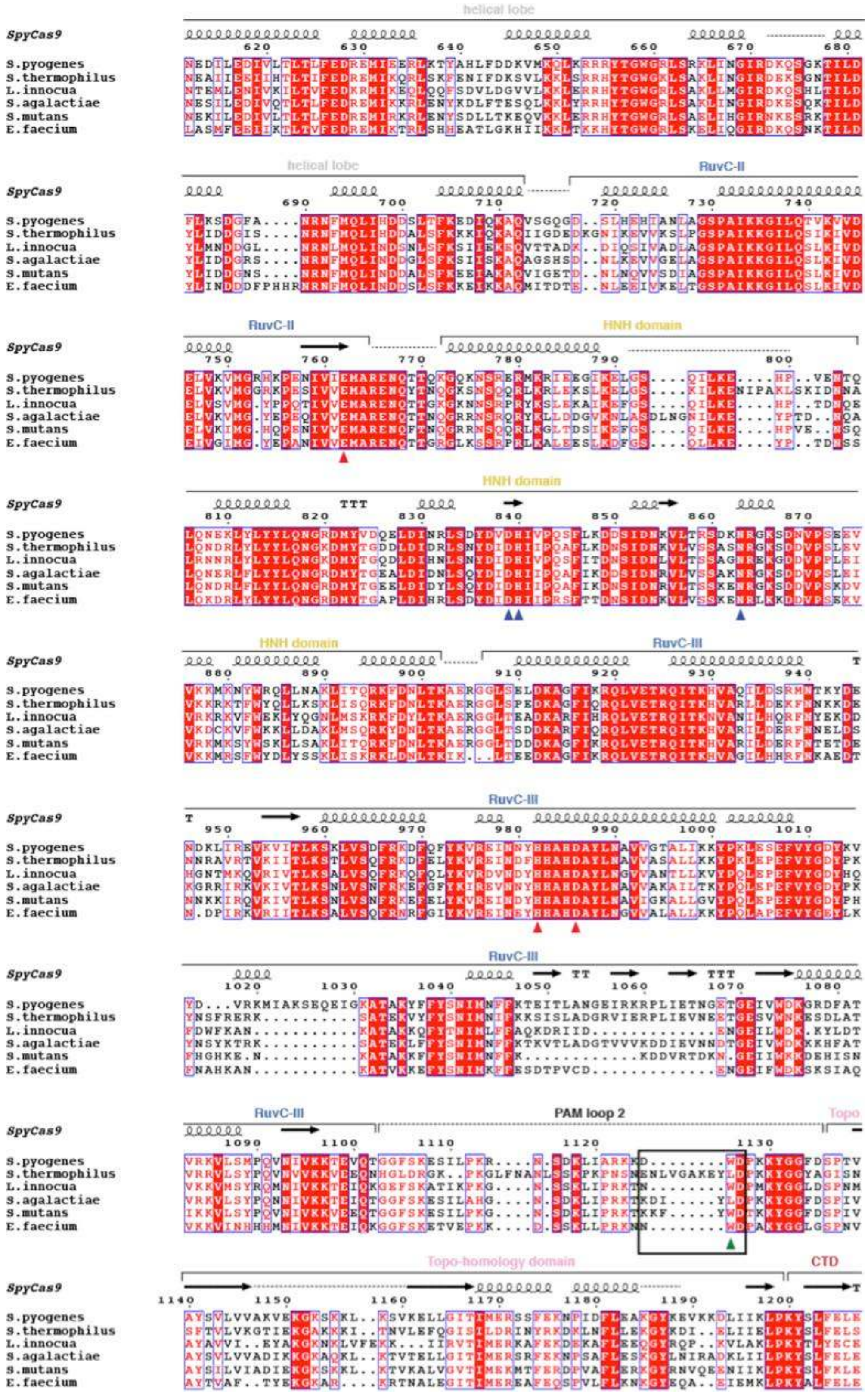


Figure 5.1 | Crystal structure of SpyCas9 reveals an open bi-lobed architecture and nucleic acid binding clefts. (a) Cartoon schematic of the polypeptide sequence and domain organization for the Type II-A Cas9 protein from *S. pyogenes* (SpyCas9). Cas9 is predicted to contain a single HNH nuclease domain and a single RuvC nuclease domain. The RuvC domain is made up of three discontinuous segments (RuvC-I-III), with the alpha-helical lobe inserted between the first and the second segments, and the HNH domain inserted between second and the third segments. Arg, arginine-rich region. Topo, Topo-homology domain. CTD, C-terminal domain. **(b)** Orthogonal views of the overall structure of SpyCas9 shown in ribbon and surface representations. Individual Cas9 domains are colored according to the scheme in (a). SpyCas9 consists of a nuclease domain lobe and an alpha-helical lobe. Disordered segments of the polypeptide chain are denoted with dotted lines. **(c)** Surface representation of SpyCas9 depicting the two nucleic acid binding clefts on the molecular surface. **(d)** Surface electrostatic potential map of SpyCas9 colored from -10 kT/e (red) to +10 kT/e (blue) (Baker et al., 2001). **(e)** Surface representation of SpyCas9 colored according to evolutionary conservation. The representation was generated using the Consurf server (Ashkenazy et al., 2010) based on the multiple sequence alignment of Type II-A Cas9 proteins shown in fig. S1. A disordered segment (residues 571^{Spy}-586^{Spy}, indicated with a black dashed line) covers the apparently conserved patch on the reverse convex surface of SpyCas9.


SpyCas9
RuvC-I
Arg-rich
helical lobe
PAM loop 1
S.pyogenes
S.thermophilus
L.innocua
S.agalactiae
S.mutans
E.faecium



helical lobe
RuvC-II
HNH domain
RuvC-III
PAM loop 2
Topo
Topo-homology domain
CTD
SpyCas9
S.pyogenes
S.thermophilus
L.innocua
S.agalactiae
S.mutans
E.faecium
620
630
640
650
660
670
680
690
700
710
720
730
740
750
760
770
780
790
800
810
820
830
840
850
860
870
880
890
900
910
920
930
940
950
960
970
980
990
1000
1010
1020
1030
1040
1050
1060
1070
1080
1090
1100
1110
1120
1130
1140
1150
1160
1170
1180
1190
1200
TT
TTT
T

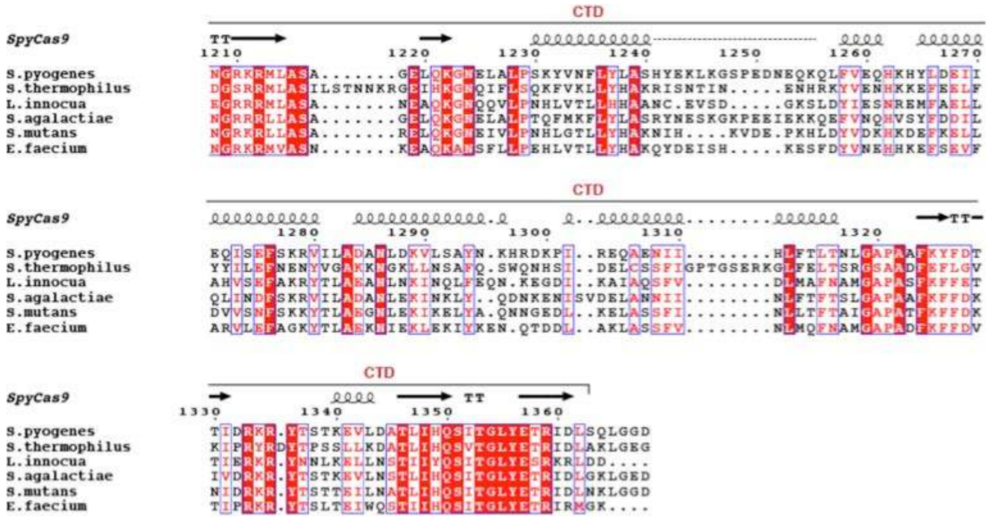

CTD
SpyCas9
TT
1210
1220
1230
1240
1250
1260
1270
S.pyogenes NGRKRMLASA.......GELQKGNELALPSKYVNFLYLASHYEKLKGSPEDNEQKQLFVEQHKHYLDEII
S.thermophilus DGSRRMLASILSTNNKRGEIHKGNQIFLSQKFVKLLYHAKRISNTIN.....ENHRKYVENHKKEFEELF
L.innocua EGRRRMLASA.......NEAQKGNQQVLPNHLVTLLHHAANC.EVSD.....GKSLDYIESNREMFAELL
S.agalactiae NGRRRLLASA.......GELQKGNELALPTQFMKFLYLASRYNESKGKPEEIEKKQEFVNQHVSYFDDIL
S.mutans NGRKRLLASA.......RELQKGNEIVLPNHLGTLLYHAKNIH....KVDE.PKHLDYVDKHKDEFKELL
E.faecium NGRKRMVASN.......KEAQKANSFLLPEHLVTLLYHAKQYDEISH.....KESFDYVNEHHKEFSEVF
CTD
SpyCas9
TT
1280
1290
1300
1310
1320
S.pyogenes EQISEFSKRVILADANLDKVLSAYN.KHRDKPI..REQAENII........HLFTLTNLGAPAAFKYFDT
S.thermophilus YYILEFNENYVGAKKNGKLLNSAFQ.SWQNHSI..DELCSSFIGPTGSERKGLFELTSRGSAADFEFLGV
L.innocua AHVSEFAKRYTLAEAHLNKINQLFEQN.KEGDI..KAIAQSFV........DLMAFNAMGAPASFKFFET
S.agalactiae QLINDFSKRVILADANLEKINKLY..QDNKENISVDELANNII........NLFTFTSLGAPAAFKFFDK
S.mutans DVVSNFSKKYTLAEGNLEKIKELYA.QNNGEDL..KELASSFI........NLLTFTAIGAPATFKFFDK
E.faecium ARVLEFAGKYTLAEKNIEKLEKIYKEN.QTDDL..AKLASSFV........NLMQFNAMGAPADFKFFDV
CTD
SpyCas9
TT
1330
1340
1350
1360
S.pyogenes TIDRKR.YTSTKEVLDATLIHQSITGLYETRIDLSQLGGD
S.thermophilus KIPRYRDYTPSSLLKDATLIHQSVTGLYETRIDLAKLGEG
L.innocua TIERKR.YNNLKELLNSTIIYQSITGLYESRKRLDD....
S.agalactiae IVDRKR.YTSTKEVLNSTLIHQSITGLYETRIDLGKLGED
S.mutans NIDRKR.YTSTTEILNATLIHQSITGLYETRIDLNKLGGD
E.faecium TIPRKR.YTSLTEIWQSTIIHQSITGLYETRIRMGK....

Figure 5.2 | Multiple sequence alignment of Cas9 proteins associated with Type II-A CRISPR loci. Primary sequences of Cas9 proteins from *Streptococcus pyogenes* (GI 15675041), *Streptococcus thermophilus* LMD-9 (GI 11662823), *Listeria innocua* Clip 11262 (GI 16801805), *Streptococcus agalactiae* A909 (GI 76788458), *Streptococcus mutans* UA159 (GI 24379809), and *Enterococcus faecium* 1,231,408 (GI 257893735) were aligned using MAFFT (*82*). The alignment was generated in ESPript (*83*) using default settings. Strictly conserved residues are shown with white letters on red background. Residues with >70% similarity are shown in red and boxed in blue. The domain organization of SpyCas9 (as in **Fig. 5.1a**) and secondary structure are shown above the sequences. Disordered segments of the polypeptide chain are indicated with dashed lines. RuvC domain catalytic residues are denoted with red arrowheads. HNH domain active site residues are denoted with blue arrowheads. Tryptophan residues that crosslinked to nucleotides flanking the PAM are denoted with green arrowheads, and tryptophan-containing motifs mutated in **Fig. 5.5d** are boxed in black.

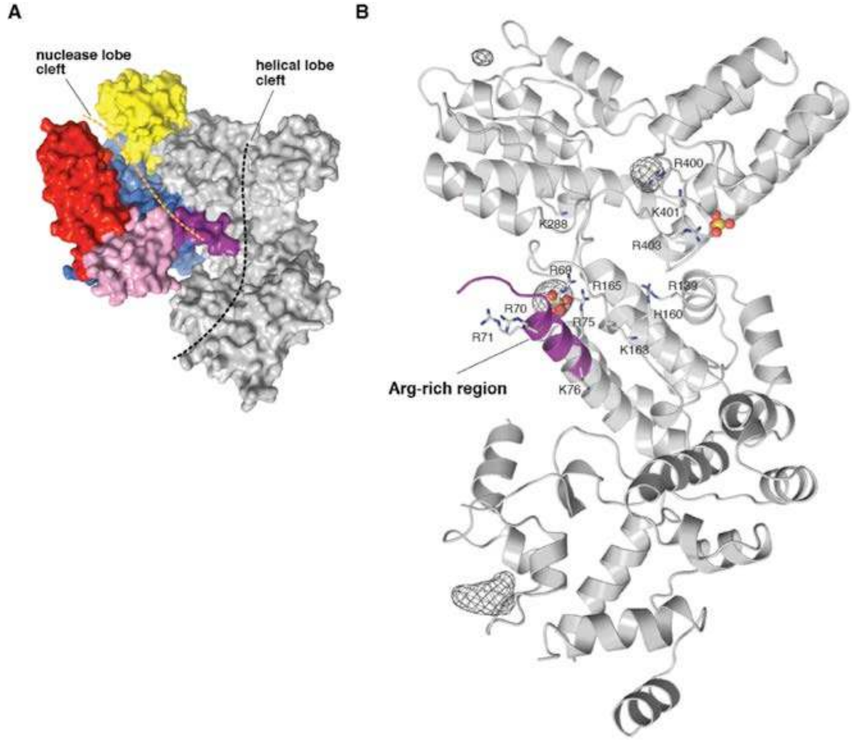


Figure 5.3 | The helical lobe of SpyCas9 features a putative nucleic acid binding cleft. (a) Surface representation of SpyCas9, colored according to the scheme in **Fig. 5.1a**. The surface clefts located on the nuclease and alpha-helical lobes of the protein are indicated with orange and black dashed lines, respectively. (**b**) Close-up view of the helical lobe of SpyCas9. Arg-rich region is depicted in purple. Conserved basic (Arg, Lys) residues lining the cleft are shown in stick format. Sulfate ions bound to the cleft are shown in ball-and-stick format. Anomalous difference electron density map (black mesh, contoured at 5.0 s) indicates positions of tungstate ions bound to SpyCas9 in crystals soaked with 10 mM Na_2WO_4.

5.4.3 PAM recognition by SpyCas9 involves two tryptophan-containing flexible loops

SpyCas9 recognizes a 5'-NGG-3' PAM sequence located three base pairs to the 3' side of the cleavage site on the non-complementary DNA strand, whereas other Cas9 orthologs have different PAM requirements (Esvelt et al., 2013; Fonfara et al., 2013; Garneau et al., 2010; Gasiunas et al., 2012; Jinek et al., 2012; Sapranauskas et al., 2011; Zhang et al., 2013). To gain insight into PAM binding by SpyCas9, we superimposed the SpyCas9 RuvC nuclease domain structure with that of the RuvC Holliday junction resolvase-substrate complex (Górecka et al., 2013) (**Fig. 5.4a**), which enabled us to model the likely trajectory of the non-target DNA strand in the SpyCas9 holoenzyme (**Fig. 5.4b,c & 5.5a**). The DNA strand is located along the length of the nuclease lobe cleft in an orientation that would position the 3' end of the DNA, and hence the PAM, at the junction of the two lobes, in the vicinity of the Arg-rich segment and the Topo-homology domain (**Fig. 5.5b**).

To directly identify regions of Cas9 involved in PAM binding, we reconstituted catalytically inactive SpyCas9 (D10A/H840A) with a crRNA:tracrRNA guide RNA and bound it to DNA targets carrying a photoactivatable 5-bromodeoxyuridine (Br-dU) nucleotide adjacent to either end of the GG PAM motif on the non-target strand (**Fig. 5.5c**). Following UV irradiation and trypsin digestion, covalent peptide-DNA crosslinks were detected (**Fig. 5.5c & 5.6**). DNA substrate containing Br-dU on the target strand opposite the PAM failed to produce a crosslink (**Fig. 5.6**). Following nuclease and phosphatase digestion of cross-linked DNA, nano-HPLC MS/MS was performed to identify tryptic peptides containing covalent dU or p-dU adducts (**Fig. 5.5c, 5.7 & 5.8**). The nucleotide immediately 5' to the GG motif cross-linked to residue W476^{Spy}, whereas the residue immediately 3' to the motif cross-linked to residue W1126^{Spy} (**Fig. 5.7 & 5.8**). Both tryptophans are located in disordered regions of SpyCas9 located ~30 Å apart. W476^{Spy} resides in a 53-aa loop at the edge of the alpha helical lobe underneath the Arg-rich region, whereas W1126^{Spy} is located in a 33-aa loop connecting the RuvC and Topo-homology domains (**Fig. 5.5b**). These tryptophan residues are conserved among Type II-A Cas9 proteins that utilize the same NGG PAM to cleave target DNA *in vitro* (Fonfara et al., 2013; Jinek et al., 2012), but are absent from Type II-C Cas9 proteins, which are known to recognize different PAMs (Esvelt et al., 2013; Fonfara et al., 2013; Garneau et al., 2010; Zhang et al., 2013) (**Fig. 5.2 & 5.9**). Interestingly, the Type II-B Cas9 protein from *Francisella novicida* , whose PAM was recently shown to be 5'-NG-3', contains a tryptophan (W501^{Fno}) at the position corresponding to W476^{Spy}, but lacks an aromatic residue equivalent to W1126^{Spy} (Fonfara et al., 2013).

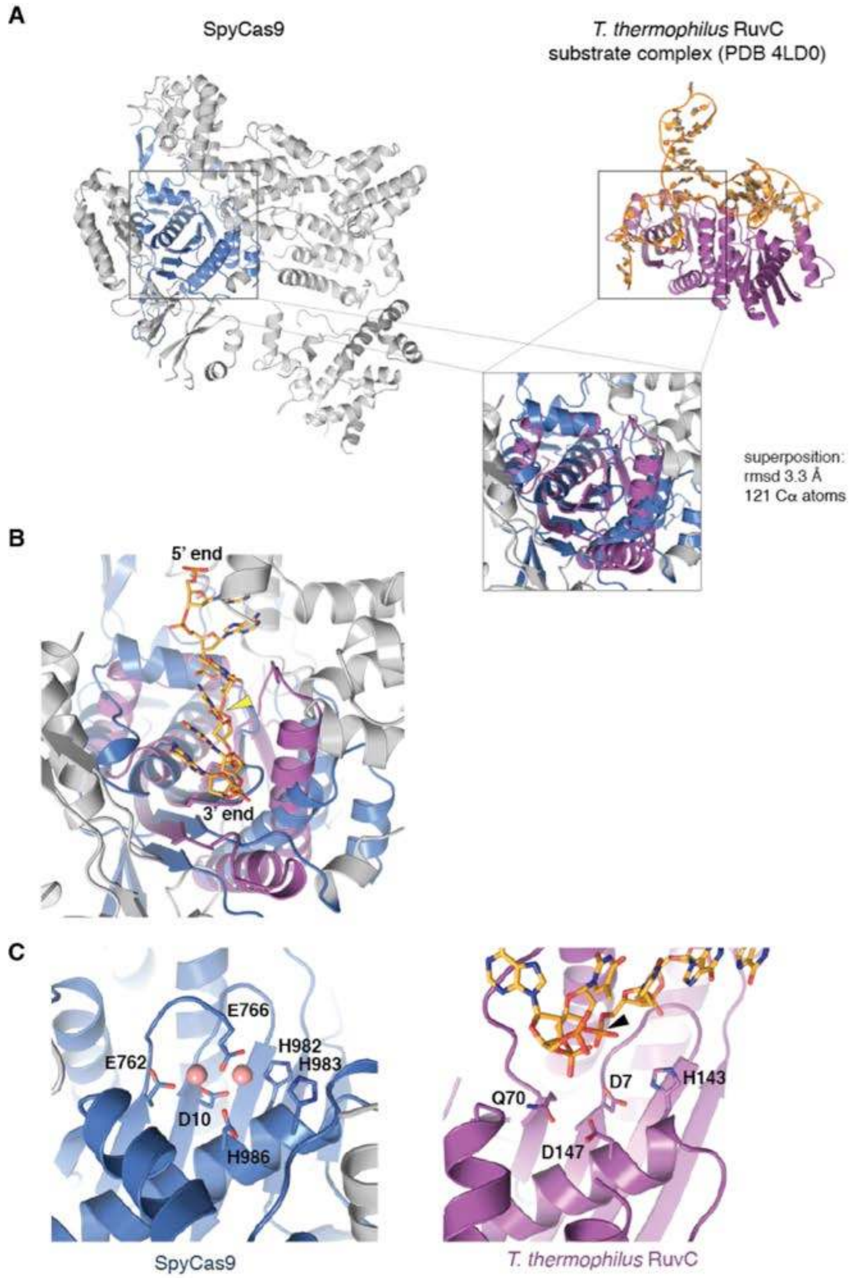


Figure 5.4 | Structural superposition of SpyCas9 with RuvC resolvase defines the directionality of non-target DNA strand in DNA-bound SpyCas9 holoenzyme. (a) Structural superposition of SpyCas9 with *Thermus thermophilus* RuvC resolvase bound to a Holliday junction substrate (PDB entry 4LD0) (*28*). The structures were superimposed using DALI (*84*) and are shown in the same orientation. The SpyCas9 RuvC domain is depicted in blue, and the RuvC resolvase is colored purple. Inset shows the superposition of the two structures. The proteins superimpose with an rmsd of 3.3 Å over 121 Ca atoms. **(b)** Close-up view of the SpyCas9 nuclease lobe cleft harbouring the RuvC active site. Six nucleotides of single stranded DNA are modeled in the cleft (stick format, colored orange) based on the superposition in (a). The position of the scissile phosphate is indicated with a yellow arrowhead. **(c)** Close-up views of the catalytic sites in SpyCas9 (left) and *T. thermophilus* RuvC (right). Active site residues are shown in stick format. Pink spheres represent two Mn^{2+} ions bound to the SpyCas9 RuvC domain in crystals soaked with 20 mM $MnCl_2$. The DNA substrate is show in stick format, and the position of the scissile phosphate is indicated with a black arrowhead.

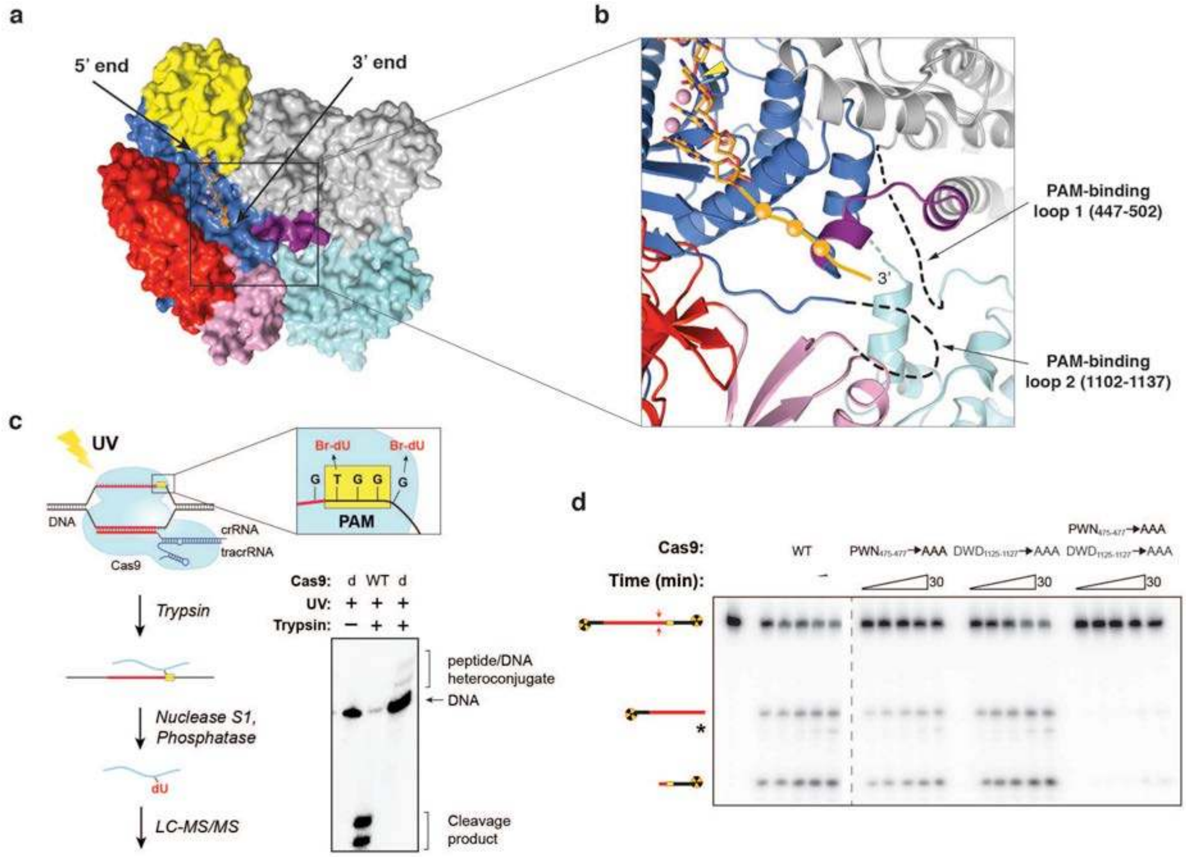


Figure 5.5 | Crosslinking identifies a PAM binding region adjacent to the active-site cleft. (a) Model of non-complementary DNA strand bound to the RuvC domain based on a superposition with the DNA-bound complex of *Thermus thermophilus* RuvC Holliday junction resolvase (PDB entry 4LD0). The modeled DNA strand contains three nucleotides upstream and three nucleotides downstream of the scissile phosphate. Divalent ions in the RuvC active site are depicted as pink spheres. **(b)** Zoomed-in view of the modeled DNA binding site showing the modeled non-target DNA strand (stick format, scissile phosphate indicated with yellow arrowhead) and the predicted path of the downstream (3') sequence containing the PAM (orange ball and string). Disordered loops identified by crosslinking are denoted with dashed lines. **(c)** Cartoon (left) showing the design and workflow of crosslinking experiments with DNA substrates containing 5-bromodeoxyuridine (Br-dU) nucleotides for LC-MS/MS analysis. The guide/target sequence is depicted in red and the PAM is highlighted in yellow. The denaturing polyacrylamide gel (right) demonstrates the generation of covalent peptide-DNA adducts with Br-dU$_1$ and catalytically inactive SpyCas9 (dCas9) following UV irradiation and trypsin digestion. **(d)** DNA cleavage activity assays with SpyCas9 constructs containing mutations in residues identified by crosslinking and LC-MS/MS experiments.

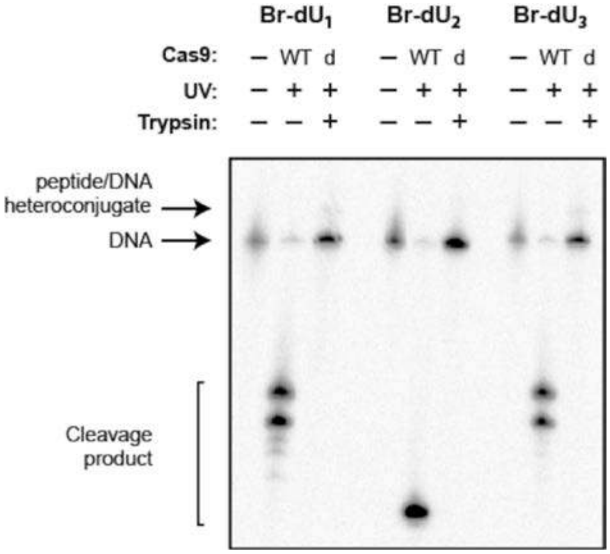
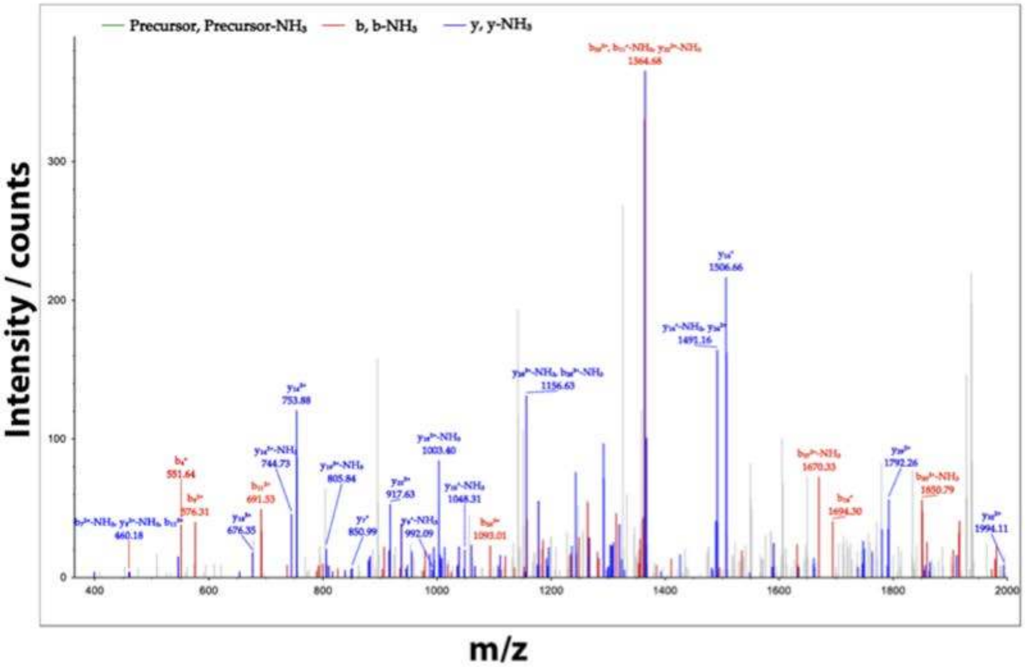
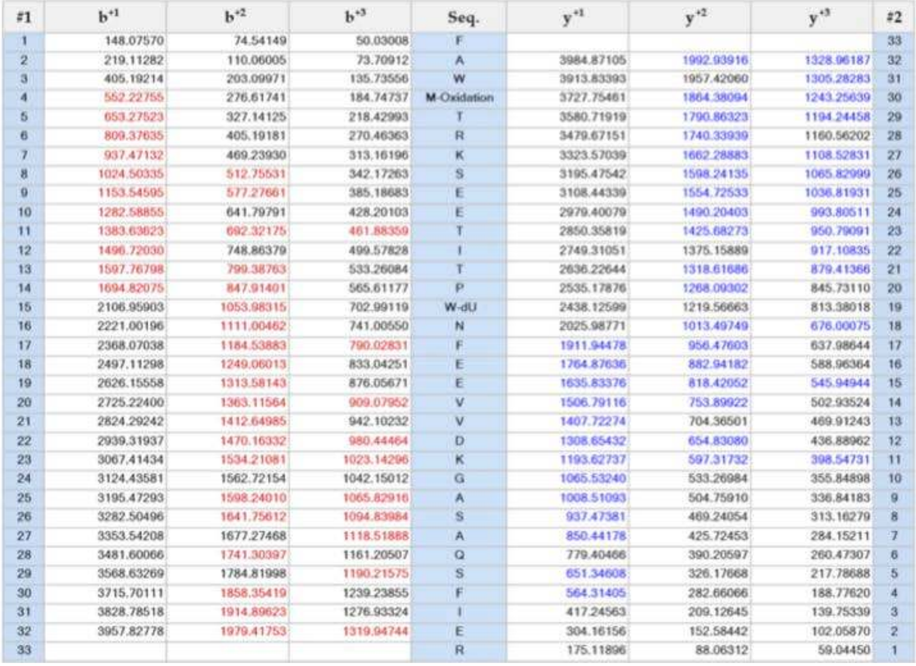

Br-dU$_1$
Br-dU$_2$
Br-dU$_3$
Cas9: – WT d – WT d – WT d
UV: – + + – + + – + +
Trypsin: – – + – – + – – +
peptide/DNA heteroconjugate
DNA
Cleavage product

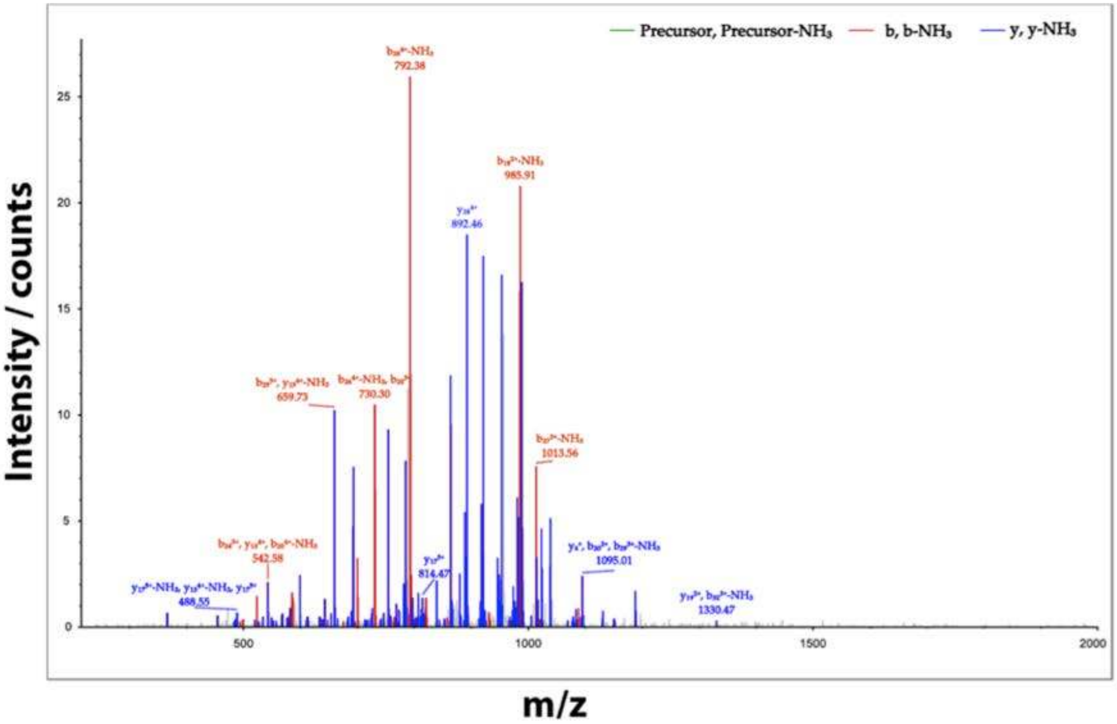
Figure 5.6 | Br-dU containing dsDNA substrates are cleaved by WT SpyCas9 and crosslink to catalytically inactive dCas9. DNA cleavage assays were performed and analysed by denaturing PAGE to verify that modified dsDNA substrates do not impair cleavage by WT SpyCas9. Sequences for each substrate (Br-dU$_1$, Br-dU$_2$, and Br-dU$_3$) can be found in **Table 5.4**. Reactions with catalytically inactive (D10A/H840A) dCas9 that can bind but not cleave DNA showed an additional band of higher molecular weight following UV irradiation and trypsin digestion, providing evidence for the generation of a peptide-DNA heteroconjugate. Crosslinking reactions with Br-dU$_1$, Br-dU$_2$, and Br-dU$_3$ were analyzed by LC-MS/MS, but only reactions with Br-dU$_1$ and Br-dU$_3$ dsDNA substrates resulted in the identification of crosslinked peptides.

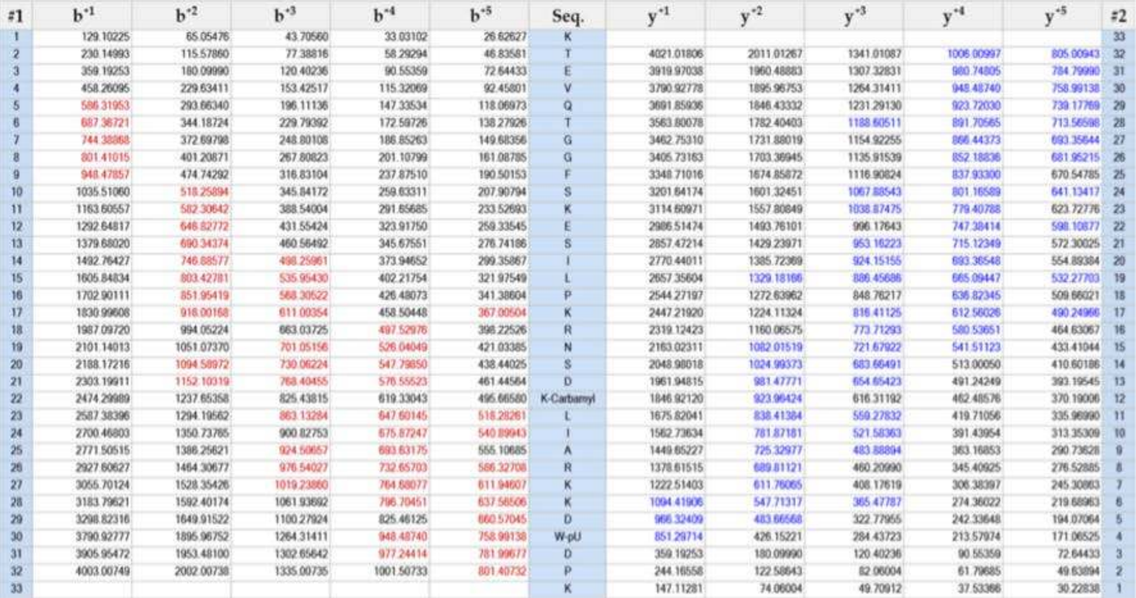




#1	b^{+1}	b^{+2}	b^{+3}	Seq.	y^{+1}	y^{+2}	y^{+3}	#2
1	148.07570	74.54149	50.03008	F				33
2	219.11282	110.06005	73.70912	A	3984.87105	1992.93916	1328.96187	32
3	405.19214	203.09971	135.73556	W	3913.83393	1957.42060	1305.28283	31
4	552.22755	276.61741	184.74737	M-Oxidation	3727.75461	1864.38094	1243.25639	30
5	653.27523	327.14125	218.42993	T	3580.71919	1790.86323	1194.24458	29
6	809.37635	405.19181	270.46363	R	3479.67151	1740.33939	1160.56202	28
7	937.47132	469.23930	313.16196	K	3323.57039	1662.28883	1108.52831	27
8	1024.50335	512.75531	342.17263	S	3195.47542	1598.24135	1065.82999	26
9	1153.54595	577.27661	385.18683	E	3108.44339	1554.72533	1036.81931	25
10	1282.58855	641.79791	428.20103	E	2979.40079	1490.20403	993.80511	24
11	1383.63623	692.32175	461.88359	T	2850.35819	1425.68273	950.79091	23
12	1496.72030	748.86379	499.57828	I	2749.31051	1375.15889	917.10835	22
13	1597.76798	799.38763	533.26084	T	2636.22644	1318.61686	879.41366	21
14	1694.82075	847.91401	565.61177	P	2535.17876	1268.09302	845.73110	20
15	2106.95903	1053.98315	702.99119	W-dU	2438.12599	1219.56663	813.38018	19
16	2221.00196	1111.00462	741.00550	N	2025.98771	1013.49749	676.00075	18
17	2368.07038	1184.53883	790.02831	F	1911.94478	956.47603	637.98644	17
18	2497.11298	1249.06013	833.04251	E	1764.87636	882.94182	588.96364	16
19	2626.15558	1313.58143	876.05671	E	1635.83376	818.42052	545.94944	15
20	2725.22400	1363.11564	909.07952	V	1506.79116	753.89922	502.93524	14
21	2824.29242	1412.64985	942.10232	V	1407.72274	704.36501	469.91243	13
22	2939.31937	1470.16332	980.44464	D	1308.65432	654.83080	436.88962	12
23	3067.41434	1534.21081	1023.14296	K	1193.62737	597.31732	398.54731	11
24	3124.43581	1562.72154	1042.15012	G	1065.53240	533.26984	355.84898	10
25	3195.47293	1598.24010	1065.82916	A	1008.51093	504.75910	336.84183	9
26	3282.50496	1641.75612	1094.83984	S	937.47381	469.24054	313.16279	8
27	3353.54208	1677.27468	1118.51888	A	850.44178	425.72453	284.15211	7
28	3481.60066	1741.30397	1161.20507	Q	779.40466	390.20597	260.47307	6
29	3568.63269	1784.81998	1190.21575	S	651.34608	326.17668	217.78688	5
30	3715.70111	1858.35419	1239.23855	F	564.31405	282.66066	188.77620	4
31	3828.78518	1914.89623	1276.93324	I	417.24563	209.12645	139.75339	3
32	3957.82778	1979.41753	1319.94744	E	304.16156	152.58442	102.05870	2
33				R	175.11896	88.06312	59.04450	1

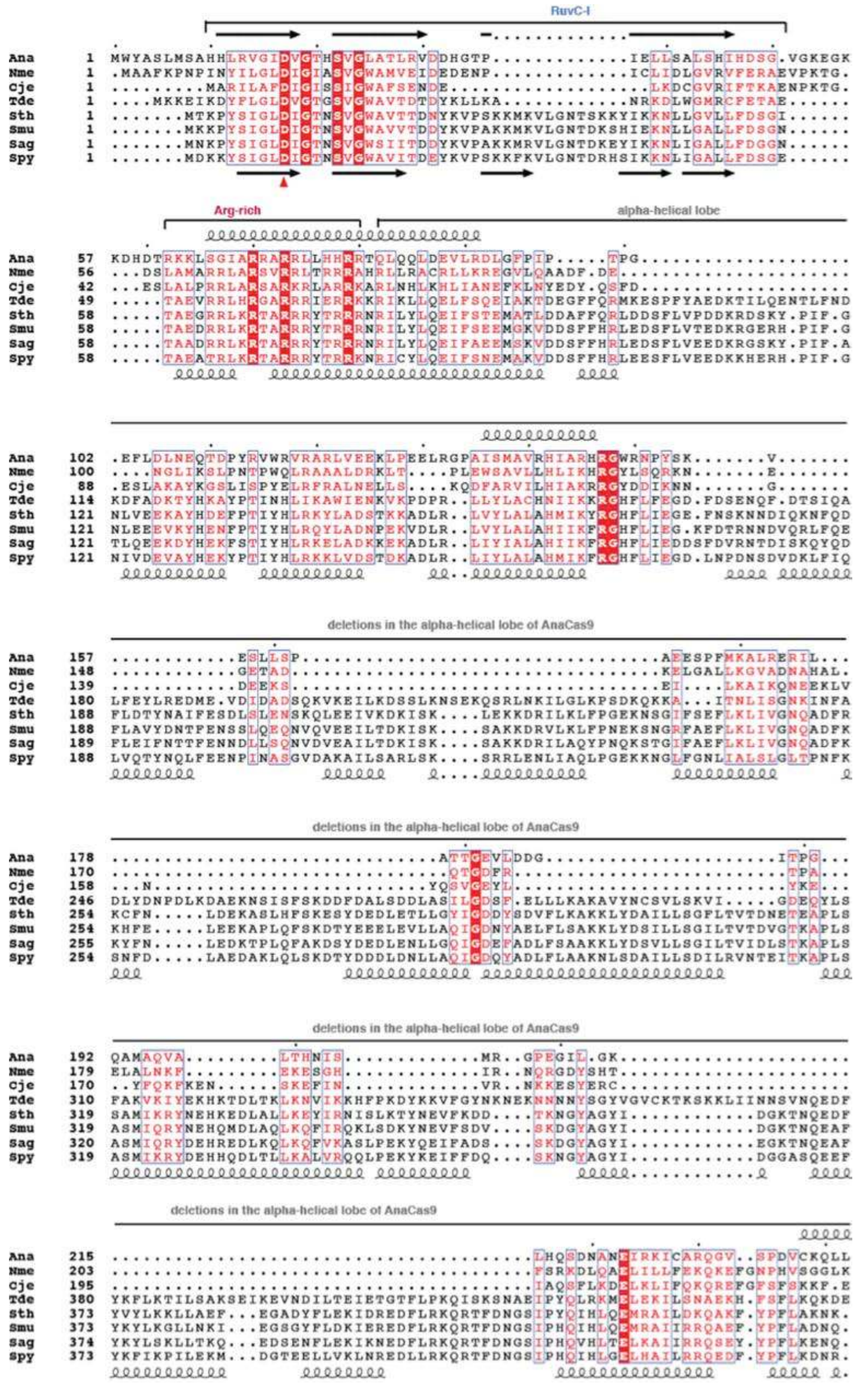
Figure 5.7 | Trp476^{Spy} crosslinks to Br-dU$_1$ dsDNA target. Tandem mass spectrum (MS/MS) and fragment ion list resulting from collision-induced dissociation (CID) of the 3+ ion occurring at mass-to-charge ratio *m*/*z* = 1377.9835. This corresponds to the $[M + 3H]^{3+}$ ion of the peptide, FAWMTRKSEETITP(**W-dU**)NFEEVVDKGASAQSFIER, which corresponds to residues 462-494 of SpyCas9, in which Trp476^{Spy} is crosslinked to deoxyuridine (dU) and Met468^{Spy} is oxidized (i.e. methionine sulfoxide). (Crosslinking to deoxyuridine and oxidation result in exact, monoisotopic mass additions of 226.05896 Da and 15.994915 Da, respectively.) Fragment ions b15 through b32 and y19 through y32 exhibit the deoxyuridine mass addition. Detected b-ions are shown in red and y-ions are shown in blue.

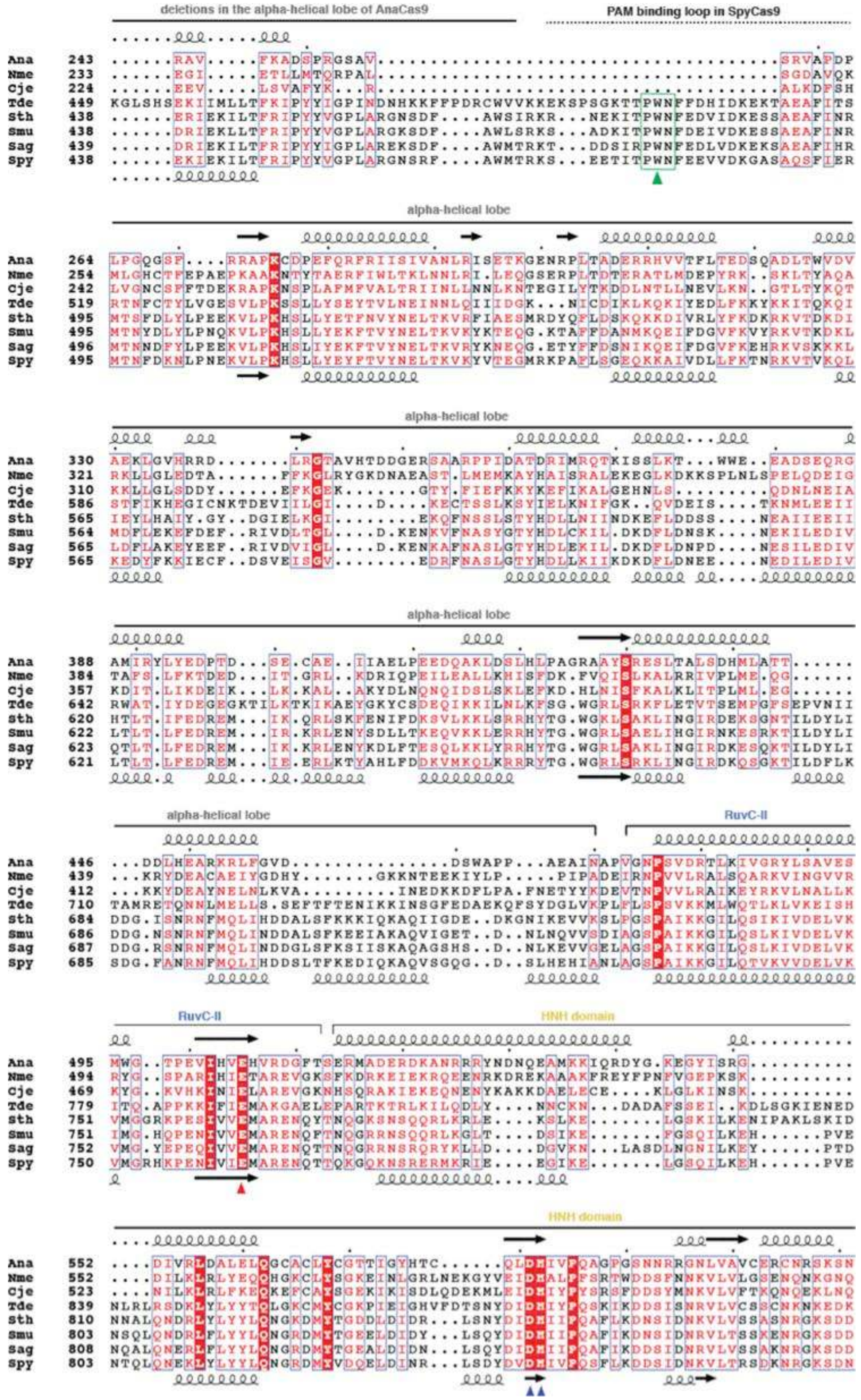




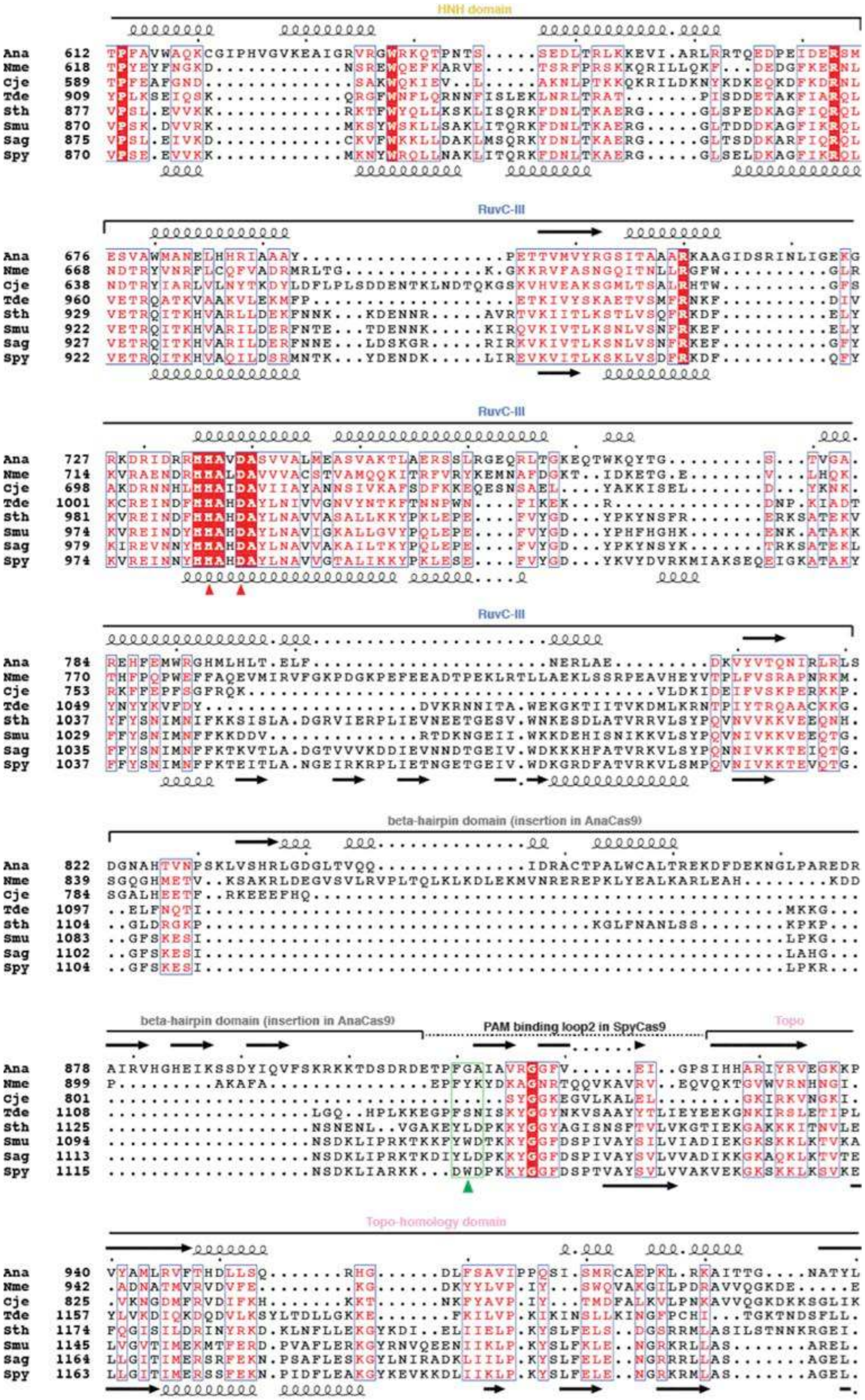
#1	b^{+1}	b^{+2}	b^{+3}	b^{+4}	b^{+5}	Seq.	y^{+1}	y^{+2}	y^{+3}	y^{+4}	y^{+5}	#2
1	129.10225	65.05476	43.70560	33.03102	26.62627	K						33
2	230.14993	115.57860	77.38816	58.29294	46.83581	T	4021.01806	2011.01267	1341.01087	1006.00997	805.00943	32
3	359.19253	180.09990	120.40236	90.55359	72.64433	E	3919.97038	1960.48883	1307.32831	980.74805	784.79990	31
4	458.26095	229.63411	153.42517	115.32069	92.45801	V	3790.92778	1895.96753	1264.31411	948.48740	758.99138	30
5	586.31953	293.66340	196.11136	147.33534	118.06973	Q	3691.85936	1846.43332	1231.29130	923.72030	739.17769	29
6	687.36721	344.18724	229.79392	172.59726	138.27926	T	3563.80078	1782.40403	1188.60511	891.70565	713.56598	28
7	744.38868	372.69798	248.80108	186.85263	149.68356	G	3462.75310	1731.88019	1154.92255	866.44373	693.35644	27
8	801.41015	401.20871	267.80823	201.10799	161.08785	G	3405.73163	1703.36945	1135.91539	852.18836	681.95215	26
9	948.47857	474.74292	316.83104	237.87510	190.50153	F	3348.71016	1674.85872	1116.90824	837.93300	670.54785	25
10	1035.51060	518.25894	345.84172	259.63311	207.90794	S	3201.64174	1601.32451	1067.88543	801.16589	641.13417	24
11	1163.60557	582.30642	388.54004	291.65685	233.52693	K	3114.60971	1557.80849	1038.87475	779.40788	623.72776	23
12	1292.64817	646.82772	431.55424	323.91750	259.33545	E	2986.51474	1493.76101	996.17643	747.38414	598.10877	22
13	1379.68020	690.34374	460.56492	345.67551	276.74186	S	2857.47214	1429.23971	953.16223	715.12349	572.30025	21
14	1492.76427	746.88577	498.25961	373.94652	299.35867	I	2770.44011	1385.72369	924.15155	693.36548	554.89384	20
15	1605.84834	803.42781	535.95430	402.21754	321.97549	L	2657.35604	1329.18166	886.45686	665.09447	532.27703	19
16	1702.90111	851.95419	568.30522	426.48073	341.38604	P	2544.27197	1272.63962	848.76217	636.82345	509.66021	18
17	1830.99608	916.00168	611.00354	458.50448	367.00504	K	2447.21920	1224.11324	816.41125	612.56026	490.24966	17
18	1987.09720	994.05224	663.03725	497.52976	398.22526	R	2319.12423	1160.06575	773.71293	580.53651	464.63067	16
19	2101.14013	1051.07370	701.05156	526.04049	421.03385	N	2163.02311	1082.01519	721.67922	541.51123	433.41044	15
20	2188.17216	1094.58972	730.06224	547.79650	438.44025	S	2048.98018	1024.99373	683.66491	513.00050	410.60186	14
21	2303.19911	1152.10319	768.40455	576.55523	461.44564	D	1961.94815	981.47771	654.65423	491.24249	393.19545	13
22	2474.29989	1237.65358	825.43815	619.33043	495.66580	K-Carbamyl	1846.92120	923.96424	616.31192	462.48576	370.19006	12
23	2587.38396	1294.19562	863.13284	647.60145	518.28261	L	1675.82041	838.41384	559.27832	419.71056	335.96990	11
24	2700.46803	1350.73765	900.82753	675.87247	540.89943	I	1562.73634	781.87181	521.58363	391.43954	313.35309	10
25	2771.50515	1386.25621	924.50657	693.63175	555.10685	A	1449.65227	725.32977	483.88894	363.16853	290.73628	9
26	2927.60627	1464.30677	976.54027	732.65703	586.32708	R	1378.61515	689.81121	460.20990	345.40925	276.52885	8
27	3055.70124	1528.35426	1019.23860	764.68077	611.94607	K	1222.51403	611.76065	408.17619	306.38397	245.30863	7
28	3183.79621	1592.40174	1061.93692	796.70451	637.56506	K	1094.41906	547.71317	365.47787	274.36022	219.68963	6
29	3298.82316	1649.91522	1100.27924	825.46125	660.57045	D	966.32409	483.66568	322.77955	242.33648	194.07064	5
30	3790.92777	1895.96752	1264.31411	948.48740	758.99138	W-pU	851.29714	426.15221	284.43723	213.57974	171.06525	4
31	3905.95472	1953.48100	1302.65642	977.24414	781.99677	D	359.19253	180.09990	120.40236	90.55359	72.64433	3
32	4003.00749	2002.00738	1335.00735	1001.50733	801.40732	P	244.16558	122.58643	82.06004	61.79685	49.63894	2
33						K	147.11281	74.06004	49.70912	37.53366	30.22838	1

Figure 5.8 | Trp1126^{Spy} crosslinks to Br-dU$_3$ dsDNA target. MS/MS spectrum and fragment ion list resulting from CID of the 5+ ion occurring at m/z = 830.6273. This corresponds to the [M + 5H]$^{5+}$ ion of the peptide, KTEVQTGGFSKESILPKRNSDKLIARKKD(**W-pdU**)DPK, which corresponds to residues 1097-1129 of SpyCas9, in which Trp1126^{Spy} is crosslinked to deoxyuridine monophosphate (pdU) and Lys1121^{Spy} is carbamylated. (Cross-linking to deoxyuridine monophosphate and carbamylation result in exact, monoisotopic mass additions of 306.02529 Da and 43.005814 Da, respectively.) Fragment ions b30 through b32 and y4 through y32 exhibit the deoxyuridine monophosphate mass addition. Detected b-ions are shown in red and y-ions are shown in blue.


RuvC-I
Ana 1 MWYASLMSAHHLRVGIDVGTHSVGLATLRVDDHGTP.............IELLSALSHIHDSG.VGKEGK
Nme 1 .MAAFKPNPINYILGLDIGIASVGWAMVEIDEDENP.............ICLIDLGVRVFERAEVPKTG.
Cje 1 .........MARILAFDIGISSIGWAFSENDE....................LKDCGVRIFTKAENPKTG.
Tde 1 ....MKKEIKDYFLGLDVGTGSVGWAVTDTDYKLLKA............NRKDLWGMRCFETAE......
Sth 1 .......MTKPYSIGLDIGTNSVGWAVTTDNYKVPSKKMKVLGNTSKKYIKKNLLGVLLFDSGI......
Smu 1 .......MKKPYSIGLDIGTNSVGWAVVTDDYKVPAKKMKVLGNTDKSHIEKNLLGALLFDSGN......
Sag 1 .......MNKPYSIGLDIGTNSVGWSIITDDYKVPAKKMRVLGNTDKEYIKKNLIGALLFDGGN......
Spy 1 .......MDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGE......
Arg-rich
alpha-helical lobe
Ana 57 KDHDTRKKLSGIARRARRLLHHRRTQLQQLDEVLRDLGFPIP.....TPG....................
Nme 56 ...DSLAMARRLARSVRRLTRRRAHRLLRACRLLKREGVLQAADF.DE......................
Cje 42 ...ESLALPRRLARSARKRLARRKARLNHLKHLIANEFKLNYEDY.QSFD...................
Tde 49 .....TAEVRRLHRGARRRIERRKKRIKLLQELFSQEIAKTDEGFFQRMKESPFYAEDKTILQENTLFND
Sth 58 .....TAEGRRLKRTARRRYTRRRNRILYLQEIFSTEMATLDDAFFQRLDDSFLVPDDKRDSKY.PIF.G
Smu 58 .....TAEDRRLKRTARRRYTRRRNRILYLQEIFSEEMGKVDDSFFHRLEDSFLVTEDKRGERH.PIF.G
Sag 58 .....TAADRRLKRTARRRYTRRRNRILYLQEIFAEEMSKVDDSFFHRLEDSFLVEEDKRGSKY.PIF.A
Spy 58 .....TAEATRLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSFFHRLEESFLVEEDKKHERH.PIF.G
Ana 102 .EFLDLNEQTDPYRVWRVRARLVEEKLPEELRGPAISMAVRHIARHRGWRNPYSK.......V.......
Nme 100 ....NGLIKSLPNTPWQLRAAALDRKLT....PLEWSAVLLHLIKHRGYLSQRKN.......E.......
Cje 88 .ESLAKAYKGSLISPYELRFRALNELLS....KQDFARVILHIAKRRGYDDIKNN.......G.......
Tde 114 KDFADKTYHKAYPTINHLIKAWIENKVKPDPR..LLYLACHNIIKKRGHFLFEGD.FDSENQF.DTSIQA
Sth 121 NLVEEKAYHDEFPTIYHLRKYLADSTKKADLR..LVYLALAHMIKYRGHFLIEGE.FNSKNNDIQKNFQD
Smu 121 NLEEEVKYHENFPTIYHLRQYLADNPEKVDLR..LVYLALAHIIKFRGHFLIEG.KFDTRNNDVQRLFQE
Sag 121 TLQEEKDYHEKFSTIYHLRKELADKKEKADLR..LIYIALAHIIKFRGHFLIEDDSFDVRNTDISKQYQD
Spy 121 NIVDEVAYHEKYPTIYHLRKKLVDSTDKADLR..LIYLALAHMIKFRGHFLIEGD.LNPDNSDVDKLFIQ
deletions in the alpha-helical lobe of AnaCas9
Ana 157 ............ESLLSP...................................AEESPFMKALRERIL...
Nme 148 ............GETAD....................................KELGALLKGVADNAHAL.
Cje 139 ............DEEKS....................................EI....LKAIKQNEEKLV
Tde 180 LFEYLREDME.VDIDADSQKVKEILKDSSLKNSEKQSRLNKILGLKPSDKQKKA...ITNLISGNKINFA
Sth 188 FLDTYNAIFESDLSLENSKQLEEIVKDKISK....LEKKDRILKLFPGEKNSGIFSEFLKLIVGNQADFR
Smu 188 FLAVYDNTFENSSLQEQNVQVEEILTDKISK....SAKKDRVLKLFPNEKSNGRFAEFLKLIVGNQADFK
Sag 189 FLEIFNTTFENNDLLSQNVDVEAILTDKISK....SAKKDRILAQYPNQKSTGIFAEFLKLIVGNQADFK
Spy 188 LVQTYNQLFEENPINASGVDAKAILSARLSK....SRRLENLIAQLPGEKKNGLFGNLIALSLGLTPNFK
deletions in the alpha-helical lobe of AnaCas9
Ana 178 .............................ATTGEVLDDG.......................ITPG...
Nme 170 ................................QTGDFR.........................TPA...
Cje 158 ...N............................YQSVGEYL........................YKE...
Tde 246 DLYDNPDLKDAEKNSISFSKDDFDALSDDLASILGDSF.ELLLKAKAVYNCSVLSKVI.....GDEQYLS
Sth 254 KCFN.....LDEKASLHFSKESYDEDLETLLGYIGDDYSDVFLKAKKLYDAILLSGFLTVTDNETEAPLS
Smu 254 KHFE.....LEEKAPLQFSKDTYEEELEVLLAQIGDNYAELFLSAKKLYDSILLSGILTVTDVGTKAPLS
Sag 255 KYFN.....LEDKTPLQFAKDSYDEDLENLLGQIGDEFADLFSAAKKLYDSVLLSGILTVIDLSTKAPLS
Spy 254 SNFD.....LAEDAKLQLSKDTYDDDLDNLLAQIGDQYADLFLAAKNLSDAILLSDILRVNTEITKAPLS
deletions in the alpha-helical lobe of AnaCas9
Ana 192 QAMAQVA.........LTHNIS.............MR..GPEGIL.GK......................
Nme 179 ELALNKF.........EKESGH.............IR..NQRGDYSHT......................
Cje 170 ..YFQKFKEN......SKEFIN.............VR..NKKESYERC......................
Tde 310 FAKVKIYEKHKTDLTKLKNVIKKHFPKDYKKVFGYNKNEKNNNNYSGYVGVCKTKSKKLIINNSVNQEDF
Sth 319 SAMIKRYNEHKEDLALLKEYIRNISLKTYNEVFKDD....TKNGYAGYI............DGKTNQEDF
Smu 319 ASMIQRYNEHQMDLAQLKQFIRQKLSDKYNEVFSDV....SKDGYAGYI............DGKTNQEAF
Sag 320 ASMIQRYDEHREDLKQLKQFVKASLPEKYQEIFADS....SKDGYAGYI............EGKTNQEAF
Spy 319 ASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQ....SKNGYAGYI............DGGASQEEF
deletions in the alpha-helical lobe of AnaCas9
Ana 215 ......................................LHQSDNANLIRKICARQGV..SPDVCKQLL
Nme 203 ......................................FSRKDLQAELILLFEKQKEFGNPHVSGGLK
Cje 195 ......................................IAQSFLKDELKLIFQKQREFGFSFSKKF.E
Tde 380 YKFLKTILSAKSEIKEVNDILTEIETGTFLPKQISKSNAEIPYQLRKMELEKILSNAEKH.FSFLKQKDE
Sth 373 YVYLKKLLAEF...EGADYFLEKIDREDFLRKQRTFDNGSIPYQIHLQEMRAILDKQAKF.YPFLAKNK.
Smu 373 YKYLKGLLNKI...EGSGYFLDKIEREDFLRKQRTFDNGSIPHQIHLQEMRAILRRQAEF.YPFLADNQ.
Sag 374 YKYLSKLLTKQ...EDSENFLEKIKNEDFLRKQRTFDNGSIPHQVHLTELKAIIRRQSEY.YPFLKENQ.
Spy 373 YKFIKPILEKM...DGTEELLVKLNREDLLRKQRTFDNGSIPHQIHLGELHAILRRQEDF.YPFLKDNR.



```
Ana    243  ......RAV.....FKADSPRGSAV......................................SRVAPDP
Nme    233  ......EGI.....ETLLMTQRPAL......................................SGDAVQK
Cje    224  ......EEV.....LSVAFYK...R......................................ALKDFSH
Tde    449  KGLSHSEKIIMLLTFKIPYYIGPINDNHKKFFPDRCWVVKKEKSPSGKTTPWNFFDHIDKEKTAEAFITS
Sth    438  ......ERIEKILTFRIPYYVGPLARGNSDF....AWSIRKR...NEKITPWNFEDVIDKESSAEAFINR
Smu    438  ......DRIEKLLTFRIPYYVGPLARGKSDF....AWLSRKS...ADKITPWNFDEIVDKESSAEAFINR
Sag    439  ......DRIEKILTFRIPYYIGPLAREKSDF....AWMTRKT...DDSIRPWNFEDLVDKEKSAEAFIHR
Spy    438  ......EKIEKILTFRIPYYVGPLARGNSRF....AWMTRKS...EETITPWNFEEVVDKGASAQSFIER

Ana    264  LPGQGSF....RRAPKCDPEFQRFRIISIVANLRISETKGENRPLTADERRHVVTFLTEDSQADLTWVDV
Nme    254  MLGHCTFEPAEPKAAKNTYTAERFIWLTKLNNLRI.LEQGSERPLTDTERATLMDEPYRK..SKLTYAQA
Cje    242  LVGNCSFFTDEKRAPKNSPLAFMFVALTRIINLLNNLKNTEGILYTKDDLNTLLNEVLKN..GTLTYKQT
Tde    519  RTNFCTYLVGESVLPKSSLLYSEYTVLNEINNLQIIIDGK...NICDIKLKQKIYEDLFKKYKKITQKQI
Sth    495  MTSFDLYLPEEKVLPKHSLLYETFNVYNELTKVRFIAESMRDYQFLDSKQKKDIVRLYFKDKRKVTDKDI
Smu    495  MTNYDLYLPNQKVLPKHSLLYEKFTVYNELTKVKYKTEQG.KTAFFDANMKQEIFDGVFKVYRKVTKDKL
Sag    496  MTNNDFYLPEEKVLPKHSLIYEKFTVYNELTKVRYKNEQG.ETYFFDSNIKQEIFDGVFKEHRKVSKKKL
Spy    495  MTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVTEGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQL

Ana    330  AEKLGVHRRD.......LRGTAVHTDDGERSAARPPIDATDRIMRQTKISSLKT...WWE..EADSEQRG
Nme    321  RKLLGLEDTA......FFKGLRYGKDNAEAST.LMEMKAYHAISRALEKEGLKDKKSPLNLSPELQDEIG
Cje    310  KKLLGLSDDY......EFKGEK.......GTY.FIEFKKYKEFIKALGEHNLS.........QDNLNEIA
Tde    586  STFIKHEGICNKTDEVIILGI....D....KECTSSLKSYIELKNIFGK..QVDEIS....TKNMLEEII
Sth    565  IEYLHAIY.GY..DGIELKGI........EKQFNSSLSTYHDLLNIINDKEFLDDSS....NEAIIEEII
Smu    564  MDFLEKEFDEF..RIVDLTGL....D.KENKVFNASYGTYHDLCKIL.DKDFLDNSK....NEKILEDIV
Sag    565  LDFLAKEYEEF..RIVDVIGL....D.KENKAFNASLGTYHDLEKIL.DKDFLDNPD....NESILEDIV
Spy    565  KEDYFKKIECF..DSVEISGV........EDRFNASLGTYHDLLKIIKDKDFLDNEE....NEDILEDIV

Ana    388  AMIRYLYEDPTD...SE.CAE..IIAELPEEDQAKLDSLHLPAGRAAYSRESLTALSDHMLATT......
Nme    384  TAFS.LFKTDED...IT.GRL..KDRIQPEILEALLKHISFDK.FVQISLKALRRIVPLME.QG......
Cje    357  KDIT.LIKDEIK...LK.KAL..AKYDLNQNQIDSLSKLEFKD.HLNISFKALKLITPLML.EG......
Tde    642  RWAT.IYDEGEGKTILKTKIKAEYGKYCSDEQIKKILNLKFSG.WGRLSRKFLETVTSEMPGFSEPVNII
Sth    620  HTLT.IFEDREM...IK.QRLSKFENIFDKSVLKKLSRRHYTG.WGKLSAKLINGIRDEKSGNTILDYLI
Smu    622  LTLT.LFEDREM...IR.KRLENYSDLLTKEQVKKLERRHYTG.WGRLSAELIHGIRNKESRKTILDYLI
Sag    623  QTLT.LFEDREM...IK.KRLENYKDLFTESQLKKLYRRHYTG.WGRLSAKLINGIRDKESQKTILDYLI
Spy    621  LTLT.LFEDREM...IE.ERLKTYAHLFDDKVMKQLKRRRYTG.WGRLSRKLINGIRDKQSGKTILDFLK

Ana    446  ...DDLHEARKRLFGVD...............DSWAPP...AEAINAPVGNPSVDRTLKIVGRYLSAVES
Nme    439  ...KRYDEACAEIYGDHY.......GKKNTEEKIYLP.....PIPADEIRNPVVLRALSQARKVINGVVR
Cje    412  ...KKYDEAYNELNLKVA.........INEDKKDFLPA.FNETYYKDEVTNPVVLRAIKEYRKVLNALLK
Tde    710  TAMRETQNNLMELLS.SEFTFTENIKKINSGFEDAEKQFSYDGLVKPLFLSPSVKKMLWQTLKLVKEISH
Sth    684  DDG.ISNRNFMQLIHDDALSFKKKIQKAQIIGDE..DKGNIKEVVKSLPGSPAIKKGILQSIKIVDELVK
Smu    686  DDG.NSNRNFMQLINDDALSFKEEIAKAQVIGET..D..NLNQVVSDIAGSPAIKKGILQSLKIVDELVK
Sag    687  DDG.RSNRNFMQLINDDGLSFKSIISKAQAGSHS..D..NLKEVVGELAGSPAIKKGILQSLKIVDELVK
Spy    685  SDG.FANRNFMQLIHDDSLTFKEDIQKAQVSGQG..D..SLHEHIANLAGSPAIKKGILQTVKVVDELVK

Ana    495  MWG..TPEVIHVEHVRDGFTSERMADERDKANRRRYNDNQEAMKKIQRDYG.KEGYISRG..........
Nme    494  RYG..SPARIHIETAREVGKSFKDRKEIEKRQEENRKDREKAAAKFREYFPNFVGEPKSK..........
Cje    469  KYG..KVHKINIELAREVGKNHSQRAKIEKEQNENYKAKKDAELECE....KLGLKINSK..........
Tde    779  ITQ.APPKKIFIEMAKGAELEPARTKTRLKILQDLY....NNCKN...DADAFSSEI..KDLSGKIENED
Sth    751  VMGGRKPESIVVEMARENQYTNQGKSNSQQRLKRLE....KSLKE.......LGSKILKENIPAKLSKID
Smu    751  IMG.HQPENIVVEMARENQFTNQGRRNSQQRLKGLT....DSIKE.......FGSQILKEH......PVE
Sag    752  VMG.YEPEQIVVEMARENQTTNQGRRNSRQRYKLLD....DGVKN...LASDLNGNILKEY......PTD
Spy    750  VMGRHKPENIVIEMARENQTTQKGQKNSRERMKRIE....EGIKE.......LGSQILKEH......PVE

Ana    552  ....DIVRLDALELQGCACLYCGTTIGYHTC......QLDHIVPQAGPGSNNRRGNLVAVCERCNRSKSN
Nme    552  ....DILKLRLYEQQHGKCLYSGKEINLGRLNEKGYVEIDHALPFSRTWDDSFNNKVLVLGSENQNKGNQ
Cje    523  ....NILKLRLFKEQKEFCAYSGEKIKISDLQDEKMLEIDHIYPYSRSFDDSYMNKVLVFTKQNQEKLNQ
Tde    839  NLRLRSDKLYLYYTQLGKCMYCGKPIEIGHVFDTSNYDIDHIYPQSKIKDDSISNRVLVCSSCNKNKEDK
Sth    810  NNALQNDRLYLYYLQNGKDMYTGDDLDIDR...LSNYDIDHIIPQAFLKDNSIDNKVLVSSASNRGKSDD
Smu    803  NSQLQNDRLFLYYLQNGRDMYTGEELDIDY...LSQYDIDHIIPQAFIKDNSIDNRVLTSSKENRGKSDD
Sag    808  NQALQNERLFLYYLQNGRDMYTGEALDIDN...LSQYDIDHIIPQAFIKDDSIDNRVLVSSAKNRGKSDD
Spy    803  NTQLQNEKLYLYYLQNGRDMYVDQELDINR...LSDYDVDHIVPQSFLKDDSIDNKVLTRSDKNRGKSDN
```


HNH domain
Ana 612 TPFAVWAQKCGIPHVGVKEAIGRVRGWRKQTPNTS.....SEDLTRLKKEVI.ARLRRTQEDPEIDERSM
Nme 618 TPYEYFNGKD............NSREWQEFKARVE.....TSRFPRSKKQRILLQKF...DEDGFKERNL
Cje 589 TPFEAFGND..............SAKWQKIEV..L.....AKNLPTKKQKRILDKNYKDKEQKDFKDRNL
Tde 909 YPLKSEIQSK............QRGFWNFLQRNNFISLEKLNRLTRAT.......PISDDETAKFIARQL
Sth 877 VPSL.EVVKK............RKTFWYQLLKSKLISQRKFDNLTKAERG.....GLSPEDKAGFIQRQL
Smu 870 VPSK.DVVRK............MKSYWSKLLSAKLITQRKFDNLTKAERG.....GLTDDDKAGFIKRQL
Sag 875 VPSL.EIVKD............CKVFWKKLLDAKLMSQRKYDNLTKAERG.....GLTSDDKARFIQRQL
Spy 870 VPSE.EVVKK............MKNYWRQLLNAKLITQRKFDNLTKAERG.....GLSELDKAGFIKRQL
RuvC-III
Ana 676 ESVAWMANELHHRIAAAY...................PETTVMVYRGSITAAARKAAGIDSRINLIGEKG
Nme 668 NDTRYVNRFLCQFVADRMRLTG.............K.GKKRVFASNGQITNLLRGFW..........GLR
Cje 638 NDTRYIARLVLNYTKDYLDFLPLSDDENTKLNDTQKGSKVHVEAKSGMLTSALRHTW..........GFS
Tde 960 VETRQATKVAAKVLEKMFP...................ETKIVYSKAETVSMFRNKF..........DIV
Sth 929 VETRQITKHVARLLDEKFNNK...KDENNR.....AVRTVKIITLKSTLVSQFRKDF..........ELY
Smu 922 VETRQITKHVARILDERFNTE...TDENNK.....KIRQVKIVTLKSNLVSNFRKEF..........ELY
Sag 927 VETRQITKHVARILDERFNNE...LDSKGR.....RIRKVKIVTLKSNLVSNFRKEF..........GFY
Spy 922 VETRQITKHVAQILDSRMNTK...YDENDK.....LIREVKVITLKSKLVSDFRKDF..........QFY
RuvC-III
Ana 727 RKDRIDRRHHAVDASVVALMEASVAKTLAERSSLRGEQRLTGKEQTWKQYTG.........S...TVGA.
Nme 714 KVRAENDRHHALDAVVVACSTVAMQQKITRFVRYKEMNAFDGKT..IDKETG.E.......V...LHQK.
Cje 698 AKDRNNHLHHAIDAVIIAYANNSIVKAFSDFKKEQESNSAEL....YAKKISEL.......D...YKNK.
Tde 1001 KCREINDFHHAHDAYLNIVVGNVYNTKFTNNPWN....FIKEK...R..............DNP.KIADT
Sth 981 KVREINDFHHAHDAYLNAVVASALLKKYPKLEPE....FVYGD...YPKYNSFR.......ERKSATEKV
Smu 974 KVREINDYHHAHDAYLNAVIGKALLGVYPQLEPE....FVYGD...YPHFHGHK.......ENK.ATAKK
Sag 979 KIREVNNYHHAHDAYLNAVVAKAILTKYPQLEPE....FVYGD...YPKYNSYK.......TRKSATEKL
Spy 974 KVREINNYHHAHDAYLNAVVGTALIKKYPKLESE....FVYGD...YKVYDVRKMIAKSEQEIGKATAKY
RuvC-III
Ana 784 REHFEMWRGHMLHLT.ELF......................NERLAE.........DKVYVTQNIRLRLS
Nme 770 THFPQPWEFFAQEVMIRVFGKPDGKPEFEEADTPEKLRTLLAEKLSSRPEAVHEYVTPLFVSRAPNRKM.
Cje 753 RKFFEPFSGFRQK......................................VLDKIDEIFVSKPERKKP.
Tde 1049 YNYYKVFDY....................DVKRNNITA.WEKGKTIITVKDMLKRNTPIYTRQAACKKG.
Sth 1037 YFYSNIMNIFKKSISLA.DGRVIERPLIEVNEETGESV.WNKESDLATVRRVLSYPQVNVVKKVEEQNH.
Smu 1029 FFYSNIMNFFKKDDV..............RTDKNGEII.WKKDEHISNIKKVLSYPQVNIVKKVEEQTG.
Sag 1035 FFYSNIMNFFKTKVTLA.DGTVVVKDDIEVNNDTGEIV.WDKKKHFATVRKVLSYPQNNIVKKTEIQTG.
Spy 1037 FFYSNIMNFFKTEITLA.NGEIRKRPLIETNGETGEIV.WDKGRDFATVRKVLSMPQVNIVKKTEVQTG.
beta-hairpin domain (insertion in AnaCas9)
Ana 822 DGNAHTVNPSKLVSHRLGDGLTVQQ..............IDRACTPALWCALTREKDFDEKNGLPAREDR
Nme 839 SGQGHMETV..KSAKRLDEGVSVLRVPLTQLKLKDLEKMVNREREPKLYEALKARLEAH........KDD
Cje 784 SGALHEETF..RKEEEFHQ...
Tde 1097 ..ELFNQTI..MKKG...
Sth 1104 ..GLDRGKP....................................KGLFNANLSS........KPKP...
Smu 1083 ..GFSKESI..LPKG...
Sag 1102 ..GFSKESI..LAHG...
Spy 1104 ..GFSKESI..LPKR...
beta-hairpin domain (insertion in AnaCas9)
PAM binding loop2 in SpyCas9
Topo
Ana 878 AIRVHGHEIKSSDYIQVFSKRKKTDSDRDETPFGAIAVRGGFV......EI..GPSIHHARIYRVEGKKP
Nme 899 P.........AKAFA...............EPFYKYDKAGNRTQQVKAVRV..EQVQKTGVWVRNHNGI.
Cje 801 .....................................SYGGKEGVLKALEL........GKIRKVNGKI.
Tde 1108 ....LGQ..HPLKKEGPFSNISKYGGYNKVSAAYYTLIEYEEKGNKIRSLETIPL
Sth 1125 ....NSNENL..VGAKEYLDPKKYGGYAGISNSFTVLVKGTIEKGAKKKITNVLE
Smu 1094 ...................NSDKLIPRKTKKFYWDTKKYGGFDSPIVAYSILVIADIEKGKSKKLKTVKA
Sag 1113 ...................NSDKLIPRKTKDIYLDPKKYGGFDSPIVAYSVLVVADIKKGKAQKLKTVTE
Spy 1115 ....NSDKLIARKK...DWDPKKYGGFDSPTVAYSVLVVAKVEKGKSKKLKSVKE
Topo-homology domain
Ana 940 VYAMLRVFTHDLLSQ........RHG.....DLFSAVIPPQSI.SMRCAEPKL.RKAITTG....NATYL
Nme 942 .ADNATMVRVDVFE.........KG......DKYYLVP.IY...SWQVAKGILPDRAVVQGKDE.....E
Cje 825 .VKNGDMFRVDIFKH.......KKT......NKFYAVP.IY...TMDFALKVLPNKAVVQGKDKKSGLIK
Tde 1157 YLVKDIQKDQDVLKSYLTDLLGKKE.......FKILVP.KIKINSLLKINGFPCHI...TGKTNDSFLL.
Sth 1174 FQGISILDRINYRKD.KLNFLLEKGYKDI..ELIIELP.KYSLFELS..DGSRRMLASILSTNNKRGEI.
Smu 1145 LVGVTIMEKMTFERD.PVAFLERKGYRNVQEENIIKLP.KYSLFKLE..NGRKRLLAS.......AREL.
Sag 1164 LLGITIMERSRFEKN.PSAFLESKGYLNIRADKLIILP.KYSLFELE..NGRRRLLAS.......AGEL.
Spy 1163 LLGITIMERSSFEKN.PIDFLEAKGYKEVKKDLIIKLP.KYSLFELE..NGRKRMLAS.......AGEL.

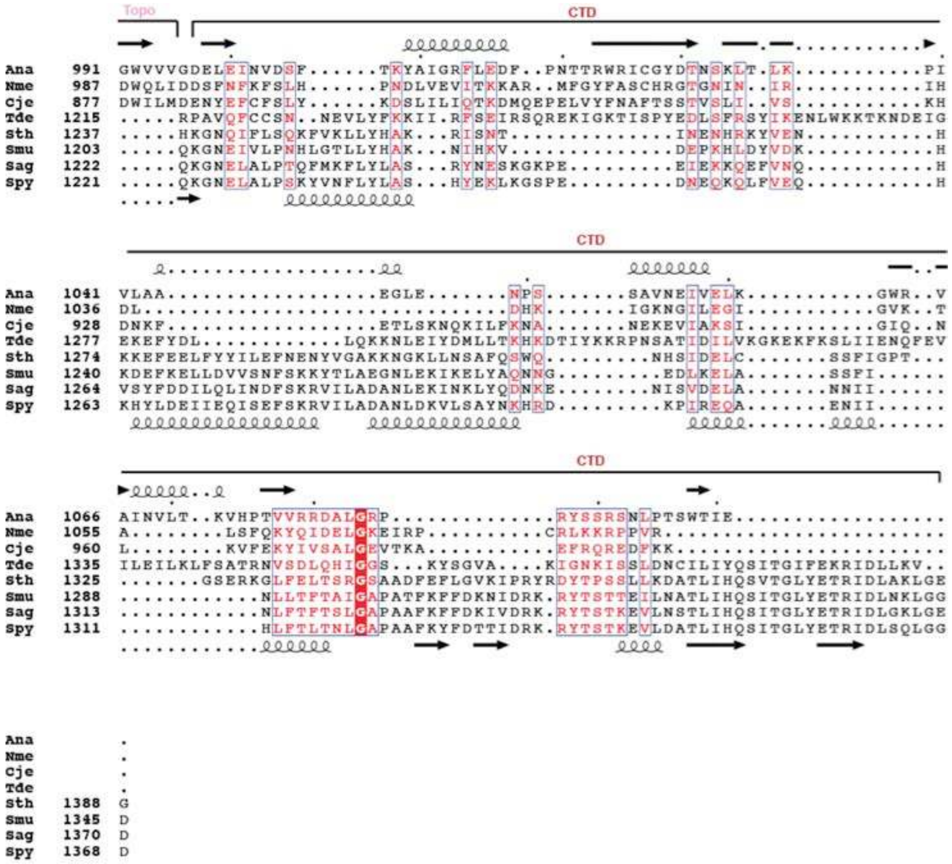

Topo
CTD
Ana 991 GWVVVGDELEINVDSF......TKYAIGRFLEDF..PNTTRWRICGYDTNSKLT.LK...........PI
Nme 987 DWQLIDDSFNFKFSLH......PNDLVEVITKKAR..MFGYFASCHRGTGNIN..IR...........IH
Cje 877 DWILMDENYEFCFSLY......KDSLILIQTKDMQEPELVYFNAFTSSTVSLI..VS...........KH
Tde 1215 .....RPAVQFCCSN..NEVLYFKKII.RFSEIRSQREKIGKTISPYEDLSFRSYIKENLWKKTKNDEIG
Sth 1237 .....HKGNQIFLSQKFVKLLYHAK...RISNT..............INENHRKYVEN...........H
Smu 1203 .....QKGNEIVLPNHLGTLLYHAK...NIHKV..............DEPKHLDYVDK...........H
Sag 1222 .....QKGNELALPTQFMKFLYLAS...RYNESKGKPE.........EIEKKQEFVNQ...........H
Spy 1221 .....QKGNELALPSKYVNFLYLAS...HYEKLKGSPE.........DNEQKQLFVEQ...........H
CTD
Ana 1041 VLAA..................EGLE.......NPS.......SAVNEIVELK...........GWR..V
Nme 1036 DL.............................DHK.......IGKNGILEGI...........GVK..T
Cje 928 DNKF..................ETLSKNQKILFKNA.......NEKEVIAKSI...........GIQ..N
Tde 1277 EKEFYDL............LQKKNLEIYDMLLTKHKDTIYKKRPNSATIDILVKGKEKFKSLIIENQFEV
Sth 1274 KKEFEELFYYILEFNENYVGAKKNGKLLNSAFQSWQ.........NHSIDELC.......SSFIGPT...
Smu 1240 KDEFKELLDVVSNFSKKYTLAEGNLEKIKELYAQNNG.........EDLKELA.......SSFI......
Sag 1264 VSYFDDILQLINDFSKRVILADANLEKINKLYQDNKE.........NISVDELA.......NNII......
Spy 1263 KHYLDEIIEQISEFSKRVILADANLDKVLSAYNKHRD.........KPIREQA.......ENII......
CTD
Ana 1066 AINVLT..KVHPTVVRRDALGRP...............RYSSRSNLPTSWTIE..................
Nme 1055 A........LSFQKYQIDELGKEIRP..........CRLKKRPPVR........................
Cje 960 L........KVFEKYIVSALGEVTKA............EFRQREDPKK........................
Tde 1335 ILEILKLFSATRNVSDLQHIGGS...KYSGVA....KIGNKISSLDNCILIYQSITGIPEKRIDLLKV..
Sth 1325 .......GSERKGLFELTSRGSAADFEFLGVKIPRYRDYTPSSLLKDATLIHQSVTGLYETRIDLAKLGE
Smu 1288 ............NLLTFTAIGAPATFKFFDKNIDRK.RYTSTTEILNATLIHQSITGLYETRIDLNKLGG
Sag 1313 ............NLFTFTSLGAPAAFKFFDKIVDRK.RYTSTKEVLNSTLIHQSITGLYETRIDLGKLGE
Spy 1311 ............HLFTLTNLGAPAAFKYFDTTIDRK.RYTSTKEVLDATLIHQSITGLYETRIDLSQLGG
Ana .
Nme .
Cje .
Tde .
Sth 1388 G
Smu 1345 D
Sag 1370 D
Spy 1368 D

Figure 5.9 | Multiple sequence alignment of Type II-A and II-C Cas9 orthologs. The primary sequences of Type II-C Cas9 orthologs from *Actinomyces naeslundii* (Ana), *Neisseria meningitidis* (Nme) and *Campylobacter jejuni* (Cje), together with type II-A Cas9 orthologs from *Treponema denticola* (Tde), *Streptococcus thermophilus* (Sth), *Streptococcus mutans* (Smu), *Streptococcus agalactiae* (Sag) and *Streptococcus pyogenes* (Spy) were aligned using CLUSTALW (*85*). The alignment was generated in ESPript (*83*) using default settings**.** Absolutely conserved residues are shown as white text on a red background, while similar residues are shown as red text with a white background. Red triangles indicate conserved residues in the RuvC active site, whereas conserved residues located in the HNH active site are denoted with a blue triangle. Green triangles indicates the tryptophan residues involved in PAM binding based on SpyCas9 crosslinking assay. The secondary structure of AnaCas9 derived from the crystal structure is marked on the top of the sequence alignment, whereas the secondary structure of SpyCas9 is shown at the bottom. Accession numbers for each Cas9 ortholog are as follows: Ana (*Actinomyces naeslundii* str. Howell 279, EJN84392.1), Nme (*Neisseria meningitidis*, WP_019742773.1), Cje (*Campylobacter jejuni*, WP_002876341.1), Tde (*Treponema denticola*, WP_002676671.1), Sth (*Streptococcus thermophilus* LMD-9, YP_820832.1), Smu (*Streptococcus mutans*, WP_019803776.1), Sag (*Streptococcus agalactiae*, WP_001040088.1), and Spy (*Streptococcus pyogenes*, YP_282132.1).

To test the roles of both loops in DNA target recognition and cleavage, we made triple alanine substitutions of residues 475^{Spy}-477^{Spy} (P-W-N) and 1125^{Spy}-1127^{Spy} (D-W-D) and performed cleavage assays with double-stranded DNA targets (**Fig. 5.5d** & **5.10**). SpyCas9 mutated in residues 1125^{Spy} -1127^{Spy} showed wild-type cleavage activity, whereas mutations in residues 475^{Spy}-477^{Spy} caused a subtle but reproducible decrease in activity (**Fig. 5.11**). Remarkably, mutating both loops simultaneously almost completely abolished SpyCas9 activity, indicating that at least one tryptophan-containing segment is necessary to promote DNA cleavage (**Fig. 5.5d** & **5.12**). The distance of both tryptophan residues from either nuclease domain argue against their direct catalytic role in DNA cleavage, instead suggesting that the residues are involved in PAM recognition. Consistent with this, DNA binding assays showed that each triple-mutant protein is moderately defective in DNA binding, whereas the dual triple-mutant protein has markedly reduced DNA binding affinity (**Fig. 5.13**).

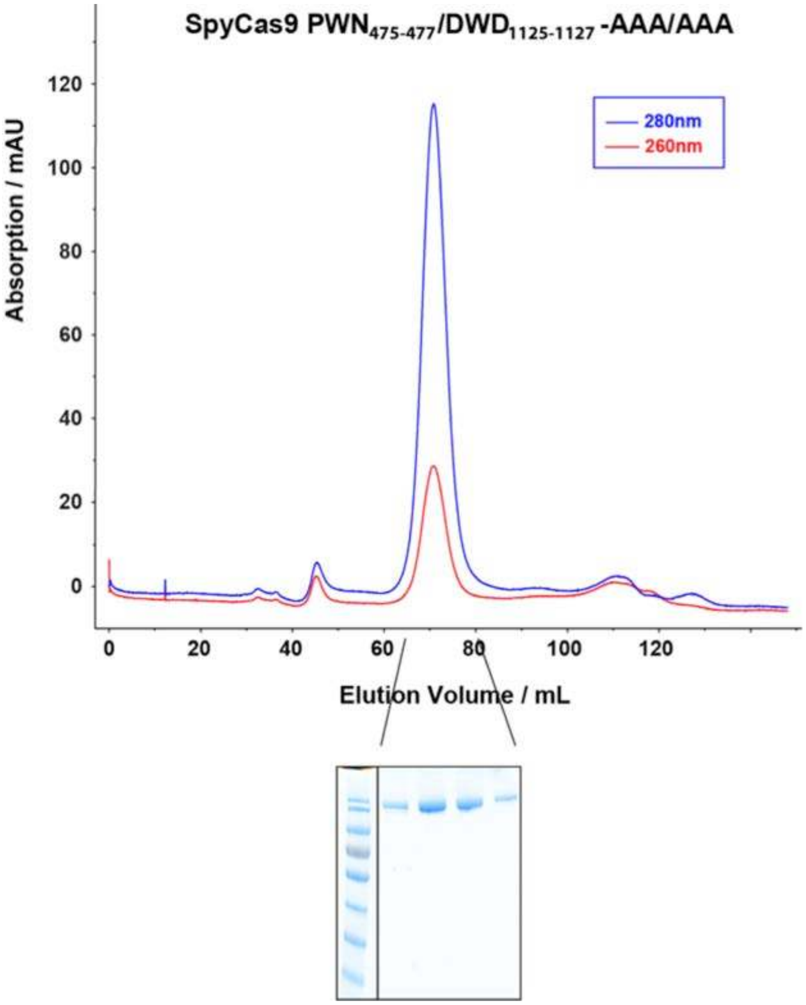


Figure 5.10 | Size exclusion chromatogram of SpyCas9 $PWN_{475\text{-}477}/DWD_{1125\text{-}1127}$→AAA/AAA mutant. All SpyCas9 mutants in this study showed the same properties during purification as observed for the wild-type SpyCas9. The retention time during gel filtration chromatography on a Superdex 200 16/60 column (GE Healthcare) is comparable to WT SpyCas9 (*8*).

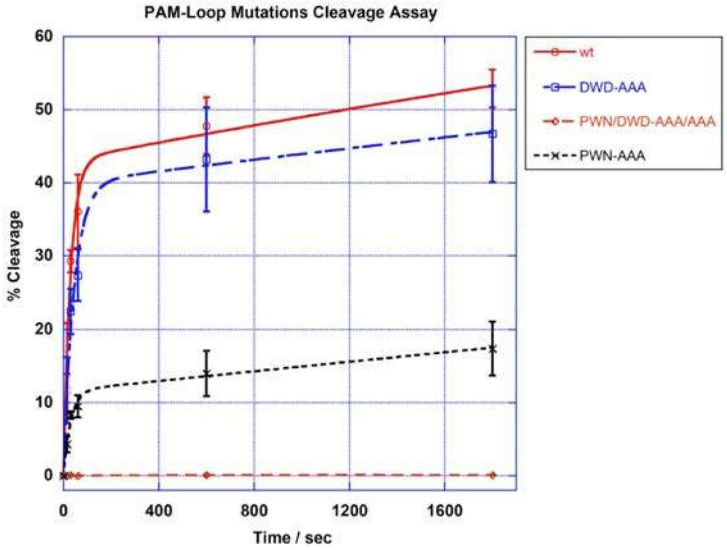


Figure 5.11 | Quantification of DNA cleavage experiments with PAM-binding mutants. For cleavage experiments, 1 nM radiolabeled 55-bp dsDNA substrate was incubated with equimolar Cas9:RNA variants (wildtype, $PWN_{475-477}$→AAA and/or $DWD_{1125-1127}$→AAA/AAA) at room temperature. The reactions were quenched at various time points and resolved by 10% denaturing PAGE. DNA was visualized by phosphorimaging, quantified with ImageQuant (GE Healthcare), and analyzed with Kaleidagraph (Synergy Software). The results presented here show a decreased cleavage activity for the $PWN_{475-477}$→AAA mutant, whereas SpyCas9 mutated in both regions leads to a severe defect in dsDNA cleavage.

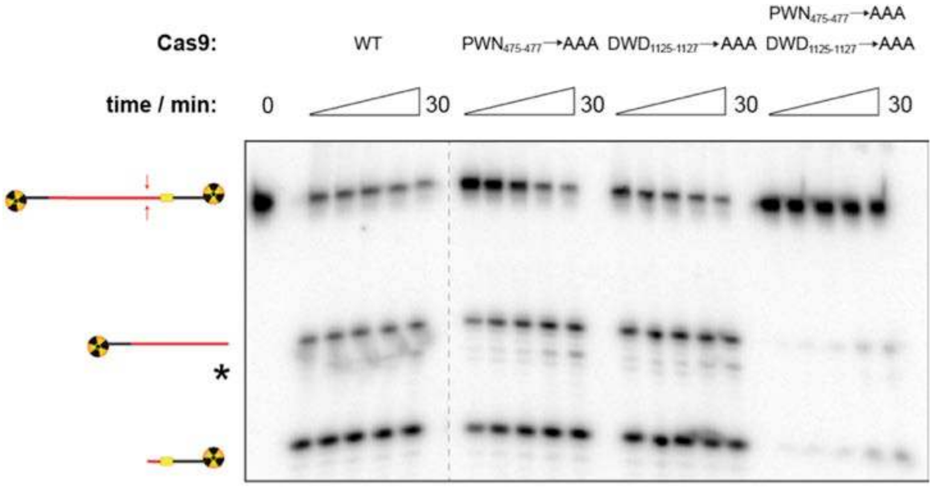


Figure 5.12 | SpyCas9 $PWN_{475-477}/DWD_{1125-1127}$→AAA/AAA mutant is impaired in dsDNA substrate cleavage. In addition to equimolar cleavage conditions (**Fig. 5.5d**), reconstituted SpyCas9 variants were also tested at a 10-fold molar excess over dsDNA substrate concentration. Reactions contained 1 nM radiolabeled DNA substrate and 10 nM Cas9:RNA complex, and were conducted at room temperature. Aliquots were removed at 0.25, 0.5, 1, 10, and 30 minutes, quenched by mixing with formamide gel loading buffer containing 50 mM EDTA, and resolved by 10% denaturing PAGE. Reaction products were visualized by phosphorimaging.

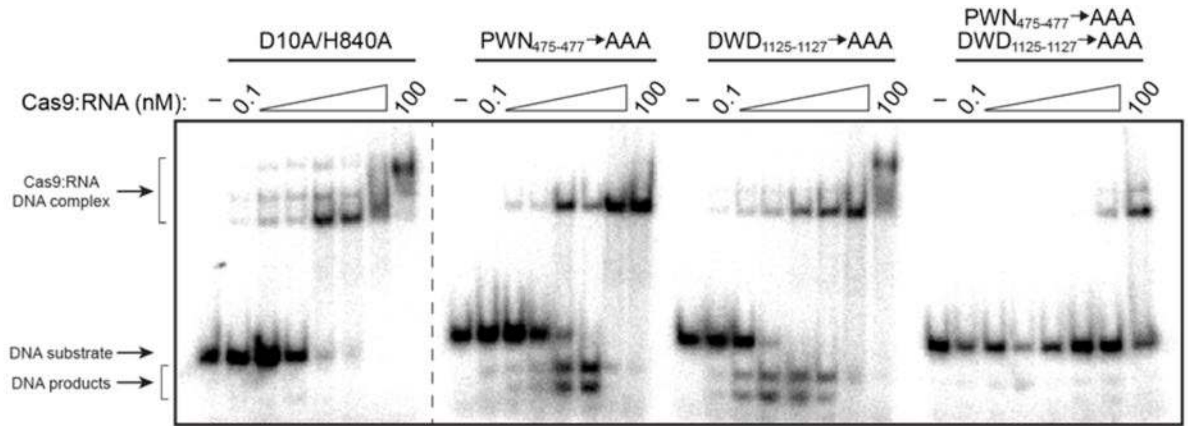
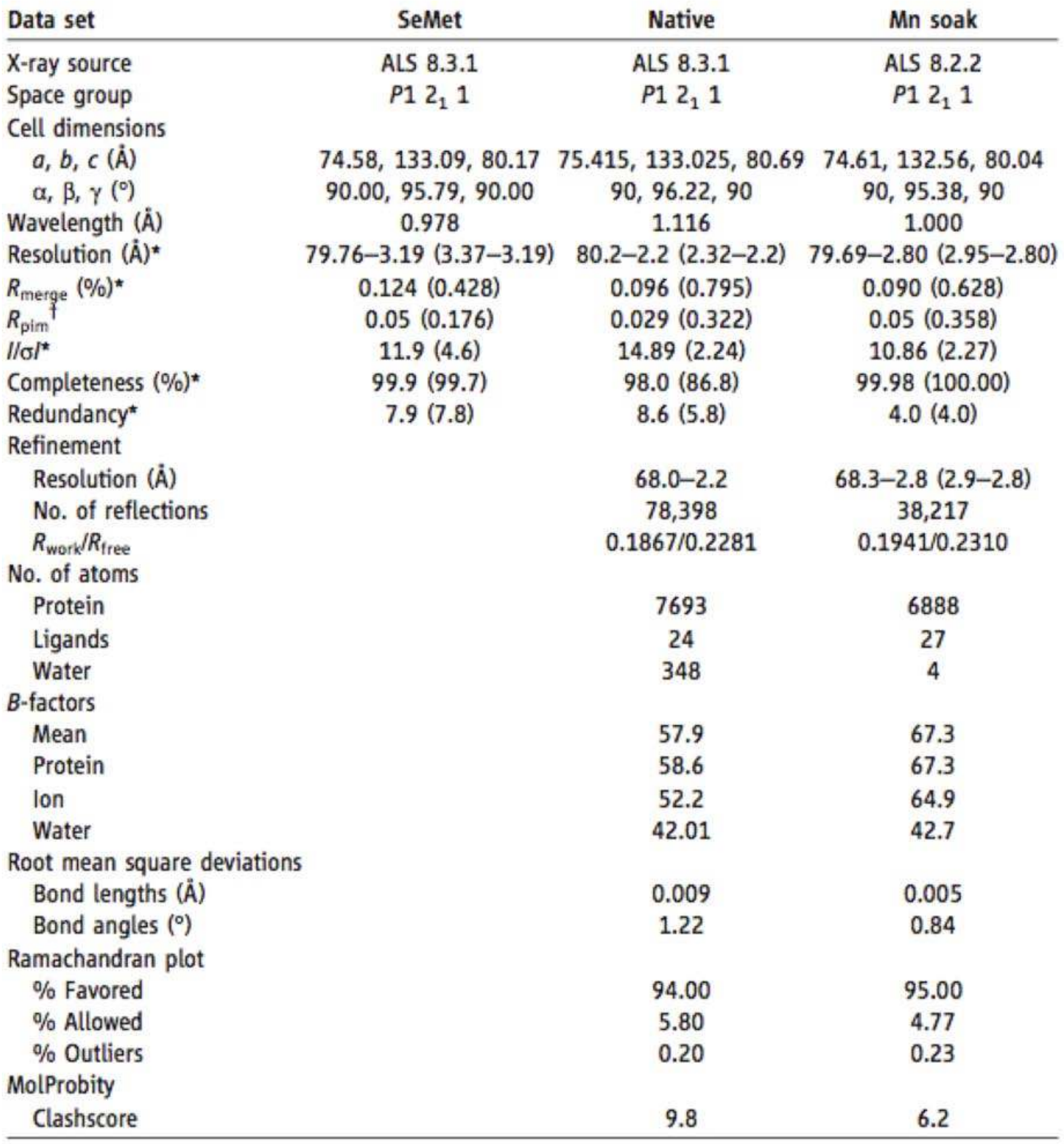


Figure 5.13 | SpyCas9 PWN$_{475-477}$/DWD$_{1125-1127}$→AAA/AAA mutant is impaired in dsDNA binding. Target 55-bp dsDNA was incubated with increasing concentrations of the indicated Cas9:RNA mutants for 60 min before being resolved by 5% native PAGE. SpyCas9 mutated individually at PWN$_{475-477}$→AAA or DWD$_{1125-1127}$→AAA/AAA binds dsDNA with an affinity similar to catalytically inactive dCas9 (D10A/H840A), whereas SpyCas9 mutated in both regions is defective in dsDNA binding. Note that unbound DNA cleavage products exhibit a distinct mobility from intact substrate DNA.

5.4.4 *A. naeslundii* Cas9 structure reveals the architecture of a smaller Cas9 variant

Although most genome engineering methodologies currently utilize SpyCas9, there is considerable interest in exploiting more compact Cas9 enzymes for such applications (Esvelt et al., 2013; Hou et al., 2013). To understand how the large and small Cas9 variants are related and how they carry out similar catalytic functions, we determined the 2.2 Å resolution crystal structure of the Type II-C Cas9 enzyme from *Actinomyces naeslundii* (AnaCas9) (**Table 5.2**). AnaCas9 also folds into a bi-lobed structure with approximate dimensions of 105 Å x 80 Å x 55 Å. The RuvC and HNH nuclease domains, a Topo-homology domain, and the C-terminal domain form an extended nuclease lobe with the RuvC domain located at its center (**Fig. 5.14a,b**). Similar to SpyCas9, the RuvC and HNH domains comprise a compact catalytic core, with the two active sites positioned ~30 Å apart. In contrast to SpyCas9, an additional domain (residues 822^{Ana}-924^{Ana}, hereafter referred to as the beta-hairpin domain) is found between the RuvC and Topo-homology domains, and adopts a novel fold composed primarily of three anti-parallel beta-hairpins. As in SpyCas9, the polypeptide sequence found between the RuvC-I and RuvC-II motifs forms an alpha-helical lobe. However, the AnaCas9 alpha-helical lobe is much smaller in size, and its orientation relative to the nuclease lobe is different (**Fig. 5.14c** & **5.15a-c**). Comparison of the helical lobes in AnaCas9 and SpyCas9 reveals that regions 95^{Ana}-251^{Ana} and 77^{Spy}-447^{Spy} are highly divergent and do not align in sequence and structure (**Fig. 5.9**). Moreover, the 95^{Ana}-251^{Ana} region is poorly ordered (**Fig. 5.16a**), and only parts of it could be modeled. By contrast, residues 252^{Ana}-468^{Ana} and 502^{Spy}-713^{Spy}, which share ~32% sequence identity, superimpose well with a root mean square deviation (rmsd) of ~3.6 Å over 149 Cα atoms (**Fig. 5.14c** & **5.15a-h**). Intriguingly, the position and orientation of this portion of the alpha-helical domain with respect to the RuvC domain in the AnaCas9 and SpyCas9 structures are substantially different, with a large displacement of ~70 Å towards the RuvC domain and an approximately 35° rotation about the junction between two domains in AnaCas9 (**Fig. 5.15c**).

Table 5.2 | X-ray data collection, refinement, and model statistics for AnaCas9.



Data set	SeMet	Native	Mn soak
X-ray source	ALS 8.3.1	ALS 8.3.1	ALS 8.2.2
Space group	$P1\ 2_1\ 1$	$P1\ 2_1\ 1$	$P1\ 2_1\ 1$
Cell dimensions			
a, b, c (Å)	74.58, 133.09, 80.17	75.415, 133.025, 80.69	74.61, 132.56, 80.04
α, β, γ (°)	90.00, 95.79, 90.00	90, 96.22, 90	90, 95.38, 90
Wavelength (Å)	0.978	1.116	1.000
Resolution (Å)*	79.76–3.19 (3.37–3.19)	80.2–2.2 (2.32–2.2)	79.69–2.80 (2.95–2.80)
R_{merge} (%)*	0.124 (0.428)	0.096 (0.795)	0.090 (0.628)
R_{pim}†	0.05 (0.176)	0.029 (0.322)	0.05 (0.358)
$I/\sigma I$*	11.9 (4.6)	14.89 (2.24)	10.86 (2.27)
Completeness (%)*	99.9 (99.7)	98.0 (86.8)	99.98 (100.00)
Redundancy*	7.9 (7.8)	8.6 (5.8)	4.0 (4.0)
Refinement			
Resolution (Å)		68.0–2.2	68.3–2.8 (2.9–2.8)
No. of reflections		78,398	38,217
R_{work}/R_{free}		0.1867/0.2281	0.1941/0.2310
No. of atoms			
Protein		7693	6888
Ligands		24	27
Water		348	4
B-factors			
Mean		57.9	67.3
Protein		58.6	67.3
Ion		52.2	64.9
Water		42.01	42.7
Root mean square deviations			
Bond lengths (Å)		0.009	0.005
Bond angles (°)		1.22	0.84
Ramachandran plot			
% Favored		94.00	95.00
% Allowed		5.80	4.77
% Outliers		0.20	0.23
MolProbity			
Clashscore		9.8	6.2

*Values in parentheses denote highest-resolution shell.
†R_{pim} = precision-indicating (multiplicity-weighted) R_{merge}.

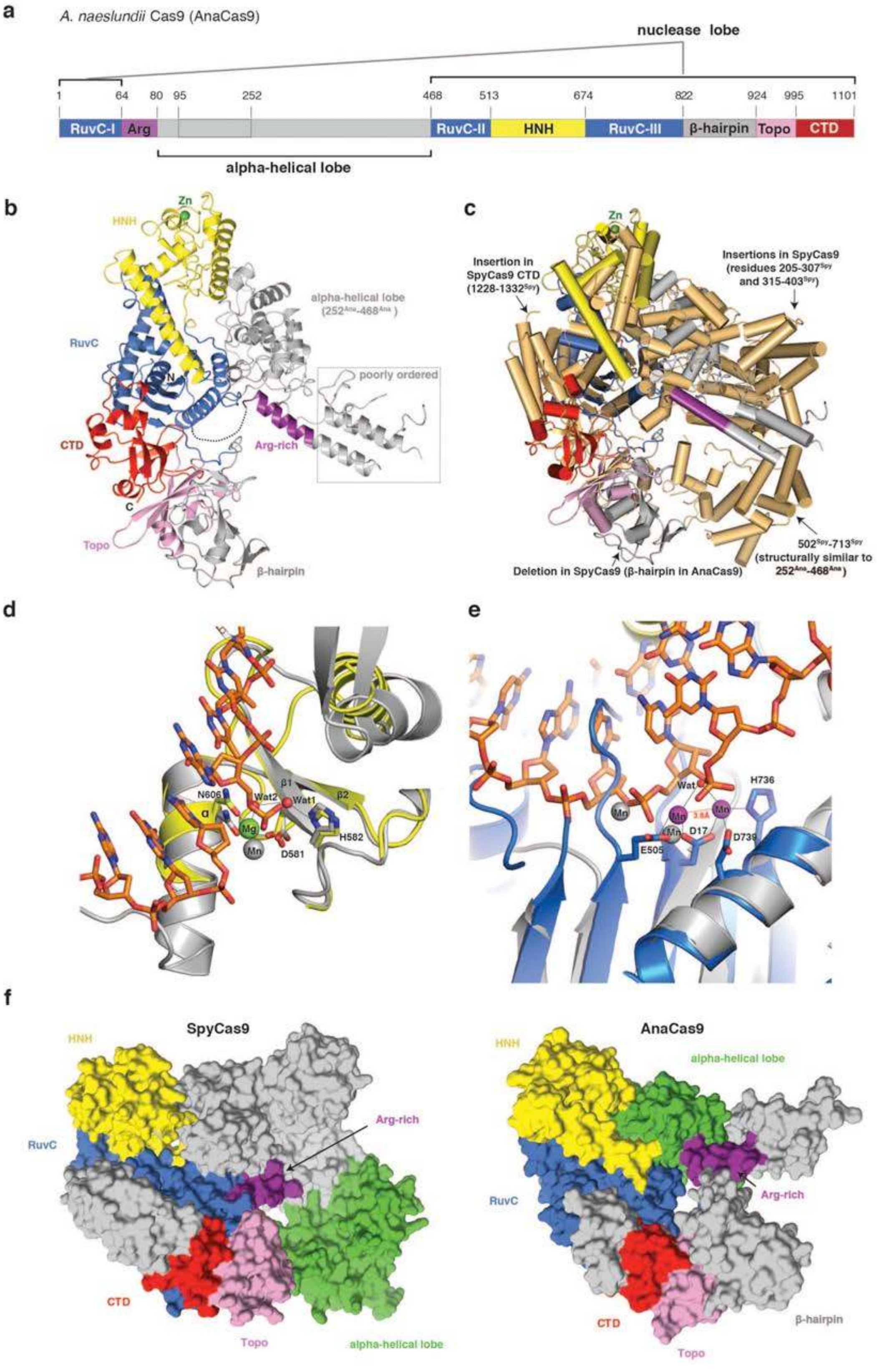

a
A. naeslundii Cas9 (AnaCas9)
nuclease lobe
1
64
80
95
252
468
513
674
822
924
995
1101
RuvC-I
Arg
RuvC-II
HNH
RuvC-III
β-hairpin
Topo
CTD
alpha-helical lobe
b
Zn
HNH
alpha-helical lobe
(252^{Ana}-468^{Ana})
RuvC
N
poorly ordered
CTD
Arg-rich
C
Topo
β-hairpin
c
Zn
Insertion in
SpyCas9 CTD
(1228-1332^{Spy})
Insertions in SpyCas9
(residues 205-307^{Spy}
and 315-403^{Spy})
502^{Spy}-713^{Spy}
(structurally similar to
252^{Ana}-468^{Ana})
Deletion in SpyCas9 (β-hairpin in AnaCas9)
d
N606
Wat2
Wat1
β1
β2
α
Mg
Mn
D581
H582
e
H736
Wat
Mn
Mn
3.8Å
Mn
Mn
D17
D739
E505
f
SpyCas9
HNH
RuvC
Arg-rich
CTD
Topo
alpha-helical lobe
AnaCas9
HNH
alpha-helical lobe
RuvC
Arg-rich
CTD
Topo
β-hairpin

Figure 5.14 | Crystal structure of AnaCas9 defines the conserved structural core of Cas9 enzymes. (a) Cartoon schematic of the polypeptide sequence and domain organization for the Type II-C Cas9 protein from *A. naeslundii* (AnaCas9). The dotted-line box represents the disordered region in the alpha-helical lobe. **(b)** Orthogonal views of the overall structure of AnaCas9 shown in ribbon representation. Individual Cas9 domains are colored according to the scheme in (a). A disordered segment connecting a RuvC motif and Arg-rich region is denoted with a dashed line. A green sphere denotes a bound zinc ion in the HNH domain. **(c)** Superposition of AnaCas9 (colored as in panel (a)) with SpyCas9 (colored in light orange). **(d)** Close-up view of the active site of AnaCas9 HNH domain (yellow) superimposed with the structure of I-HmuI-DNA complex (PDB entry 1U3E). The DNA cleavage product in I-HmuI-DNA complex is colored in orange, and I-HmuI and its bound Mn^{2+} ion are colored gray. **(e)** Close-up view of the AnaCas9 RuvC active site (marine, bound Mn^{2+} ions shown as purple spheres) overlaid with the structure of RNase H and its bound Mn^{2+} ions (gray) complexed with a DNA/RNA duplex (orange) (PDB entry 3O3H). **(f)** Surface representations of SpyCas9 (left panel) and AnaCas9 (right panel) with conserved RuvC, HNH, Arg-rich, Topo-homology, and the conserved cores of the C-terminal domains, colored as in **Fig. 5.1a**. The structurally preserved portion of the alpha-helical lobe is colored green. The non-conserved regions of each protein are colored in gray.

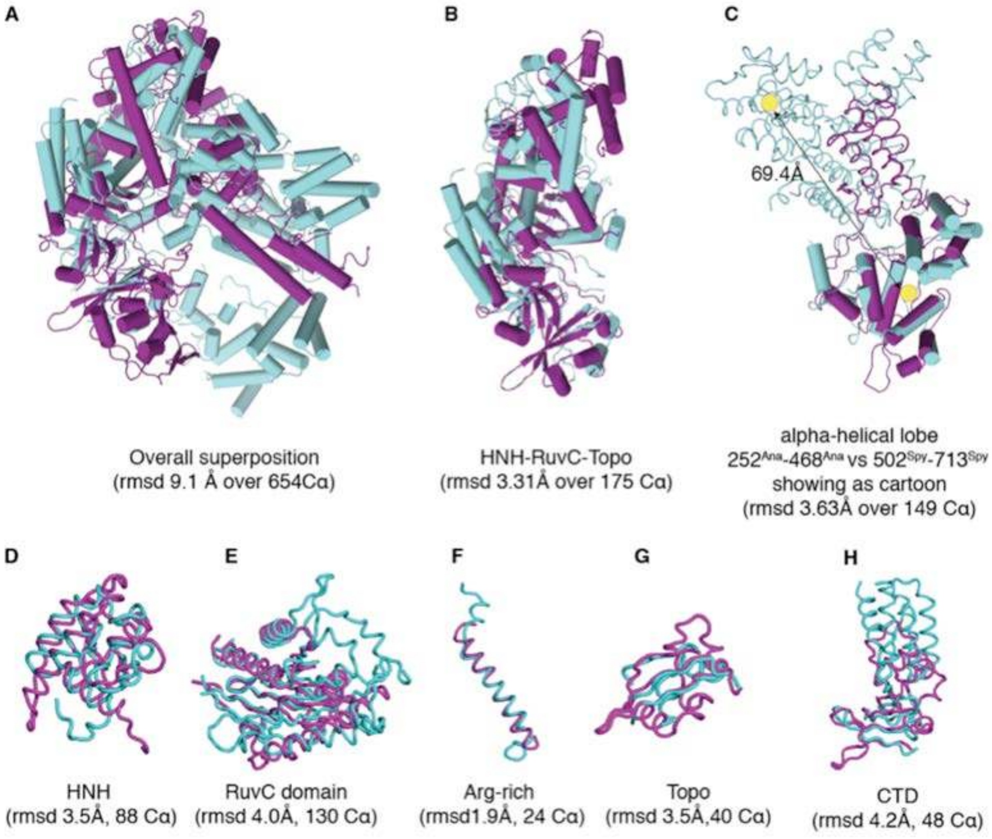


Figure 5.15 | Pairwise structural comparisons of SpyCas9 and AnaCas9. (a) Overall structural alignment of AnaCas9 (purple) and SpyCas9 (cyan) showing a good alignment of the nuclease lobe but distinct structural features in the alpha-helical lobe. The superpositions were generated using the jCE algorithm (http://source.rcsb.org/jfatcatserver/). **(b)** Superposition of the catalytic core. For clarity, the alpha-helical lobe is not shown. **(c)** Superposition of the alpha-helical lobe, revealing structural similarity between 252^{Ana}-468^{Ana} and 502^{Spy}-713^{Spy}, with a large displacement of 69.4 Å towards the RuvC domain and an approximately 35° rotation about the junction between two domains in AnaCas9. The putative domain centers are labeled with yellow circles. **(d-h)** Individual domains of AnaCas9 superimposed onto the corresponding domains in SpyCas9 with root mean square deviation (rmsd) values for the equivalent alpha-carbons indicated.

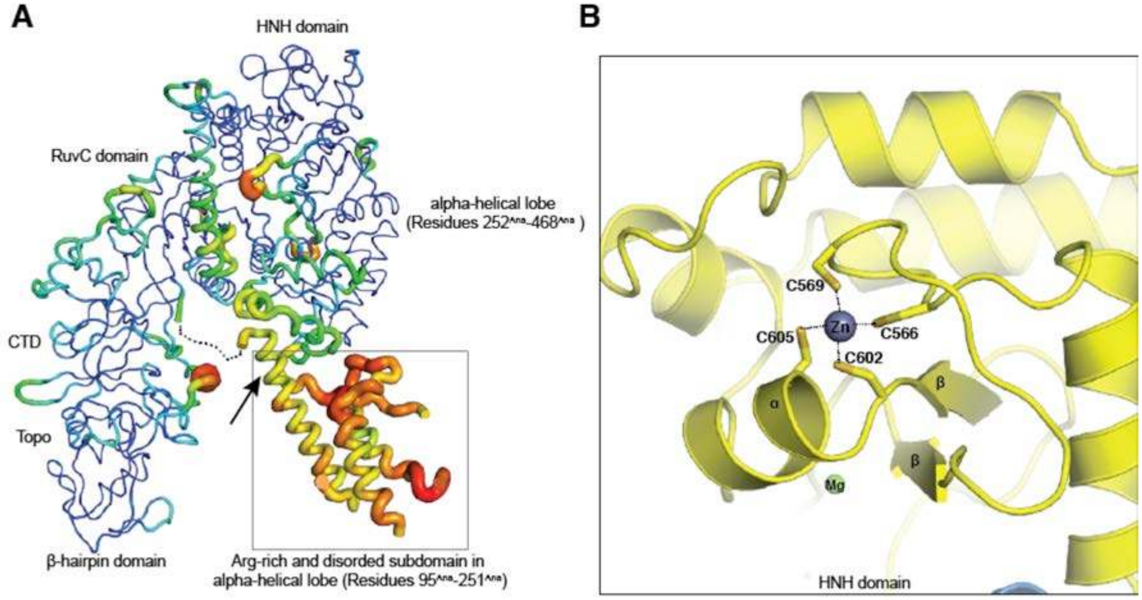

A
HNH domain
RuvC domain
alpha-helical lobe (Residues 252^Ana-468^Ana)
CTD
Topo
β-hairpin domain
Arg-rich and disorded subdomain in alpha-helical lobe (Residues 95^Ana-251^Ana)
B
C569
Zn
C566
C605
C602
α
β
β
Mg
HNH domain

Figure 5.16 | Analysis of disordered regions and HNH domain of AnaCas9. (a) AnaCas9 displayed by B-factor putty. Thin blue loops represent low B-values, while broad red tubes represent high B-values. The Arg-rich region and the neighboring alpha-helical part (box) have the highest B-factors in the structure, suggesting high flexibility in these regions. The hinge connecting the RuvC domain and the Arg-rich region is drawn as a dotted line. **(b)** Close-up view of the zinc-binding site in the HNH domain of AnaCas9. The zinc site is coordinated by residues C566^{Ana}, C569 ^{Ana}, C602^{Ana} and C605^{Ana}, and may serve to stabilize the AnaCas9 HNH domain architecture (ββα-Me fold).

The higher resolution of the AnaCas9 structure provides insights into active-site chemistries for both nuclease domains. The well-defined AnaCas9 HNH domain contains a two-stranded antiparallel β-sheet flanked by two α-helices on each side, as well as a non-conserved non-catalytic zinc binding site (**Fig. 5.14c & 5.16b**). The HNH active site reveals D581^{Ana} and N606^{Ana} coordinating a hydrated magnesium ion that would be involved in binding the scissile phosphate in the target DNA strand (**Fig. 5.14d**), and the general base residue H582^{Ana} (corresponding to H840^{Spy}) involved in deprotonating the attacking water nucleophile, in agreement with a one-metal-ion catalytic mechanism for the endonucleases containing the ββα-Metal motif (Yang, 2008). In the RuvC domain, two Mn^{2+} ions, spaced 3.8 Å apart and coordinated by the invariant residues D17^{Ana}, E505^{Ana}, H736^{Ana} and D739^{Ana}, are consistent with a two-metal ion mechanism, as observed in other nucleases containing the RNase H fold (**Fig. 5.14e**) (Yang, 2011; Yang et al., 2006).

5.4.5 A common Cas9 functional core suggests structural plasticity that supports RNA-guided DNA cleavage

Comparison of the SpyCas9 and AnaCas9 structures reveals a conserved functional core consisting of the RuvC and HNH domains, the Arg-rich region, and the Topo-homology domain, with divergent C-terminal and alpha-helical domains (**Fig. 5.14f**). In both SpyCas9 and AnaCas9, the Arg-rich region connects the nuclease and helical lobes of the proteins. The central position of the Arg-rich segment and its proximity to the PAM-binding loops in SpyCas9 suggests that this region may be involved in guide RNA and/or target DNA binding and could function as a hinge to enable conformational rearrangements in the enzyme.

Although the helical lobes of SpyCas9 and AnaCas9 share a common region (residues 252^{Ana}-468^{Ana} versus 502^{Spy}-713^{Spy}), the orientation of this region relative to the nuclease lobe varies in the two structures (**Fig. 5.14f**). Differences between SpyCas9 and AnaCas9 thus illustrate the structural divergence likely responsible for the diversity of guide RNA structures and PAM specificities within the Cas9 superfamily. The PAM-interacting regions identified in SpyCas9 are located in loops that are highly variable within Cas9 enzymes (Chylinski et al., 2013; Makarova et al., 2011a). In AnaCas9, the beta-hairpin domain (residues 822^{Ana}-924^{Ana}) is inserted at a position corresponding to one of the SpyCas9 PAM loops (1102^{Spy}-1136^{Spy}), suggesting that AnaCas9 employs a distinct mechanism of PAM recognition (**Fig. 5.14c** & **5.9**). The beta-hairpin domain is not conserved in all Type II-C Cas9 proteins (**Fig. 5.9** & **5.17)**, further underscoring the notion that the sequence- and structurally-divergent regions of Cas9 proteins may have co-evolved with specific guide RNA structures and PAM sequences (Chylinski et al., 2013; Fonfara et al., 2013).

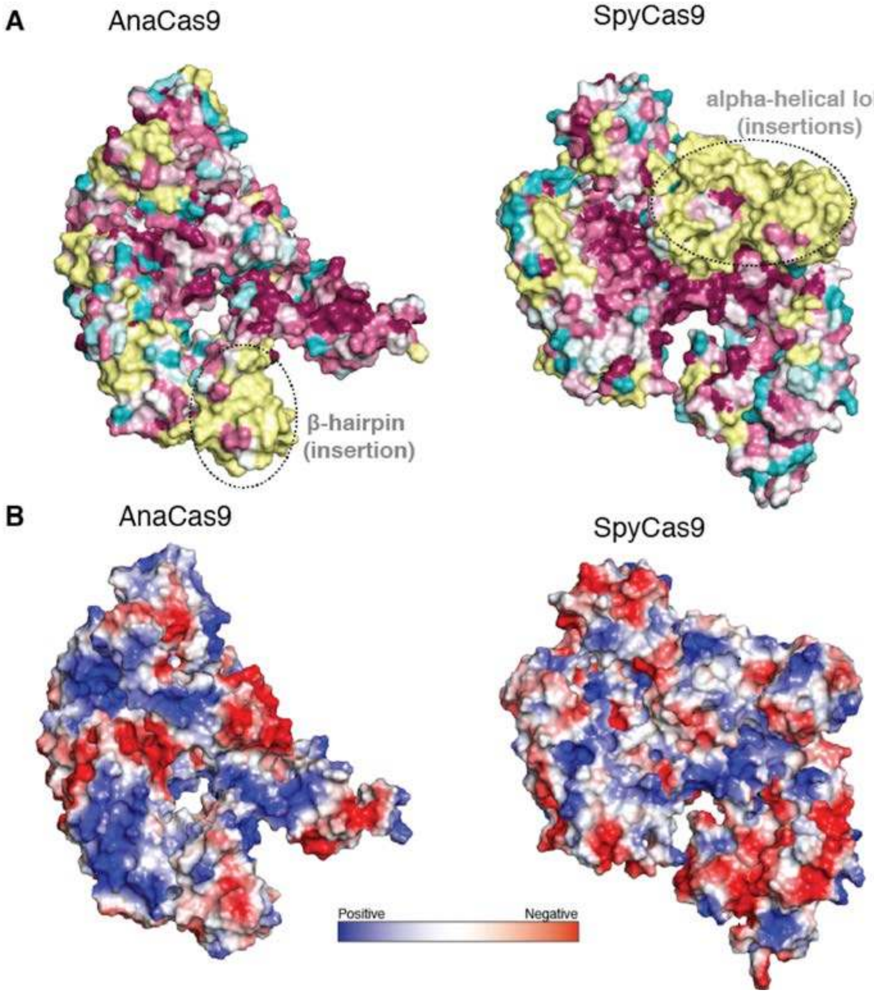


Figure 5.17 | Surface features of SpyCas9 and AnaCas9 based on sequence conservation and electrostatic potential. (a) Surface conservation of AnaCas9 (left) and SpyCas9 (right), with the same orientation as in **Fig. 5.14**. The surface is colored according to amino acid conservation among the Type-II Cas9 proteins shown in **Fig. 5.9** by the Consurf Server (*61*), where purple/red represents highly conserved residues, while yellow/light green denotes the most variant residues in Type-II Cas9 orthologs. Notably, AnaCas9 harbors a β-hairpin domain insertion, whereas SpyCas9 has a large insertion in the alpha-helical lobe. **(b)** The same molecular surface representations of AnaCas9 (left) and SpyCas9 (right) are color-coded by electrostatic potential, as calculated by APBS (*60*) electrostatics in PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC).

5.4.6 SpyCas9 and AnaCas9 adopt auto-inhibited conformations in the apo state

Target DNA cleavage by Cas9 RuvC and HNH domains is thought to occur upon base-pairing between the crRNA guide and the target DNA (Gasiunas et al., 2012; Ivancic-Bace et al., 2012; Jinek et al., 2012). Although SpyCas9 and AnaCas9 adopt distinct conformations in their helical lobes, the relative orientations of the RuvC and HNH active sites within the nuclease lobes are very similar (**Fig. 5.14c & 5.15**). In both structures, the HNH active site faces outwards, away from the putative nucleic acid binding clefts (**Fig. 5.1b & 5.3b**). Structural superpositions with the DNA-bound complex of the HNH homing endonuclease I-HmuI (Shen et al., 2004) suggest that this orientation is unlikely to be compatible with target DNA binding and cleavage (**Fig. 5.18a**). In SpyCas9, the HNH domain active site is blocked by a beta-hairpin formed by residues 1049^{Spy}-1059^{Spy} of the RuvC domain. The RNA-DNA heteroduplex would additionally clash sterically with the C-terminal domain (**Fig. 5.18a,b**). In AnaCas9, the bound crRNA-target DNA heteroduplex would conversely make few contacts with the protein outside of the HNH domain in the absence of HNH domain reorientation (**Fig. 5.18a**, right panel). The finding that two highly divergent Cas9 orthologs exhibit similar inactive states suggests that this may be a general property of Cas9 enzymes and not a consequence of crystallization. It is also consistent with the observation that Cas9 enzymes are inactive as nucleases in the absence of bound guide RNAs (Jinek et al., 2012; Karvelis et al., 2013). Taken together, these observations suggest that the enzymes undergo a conformational rearrangement upon guide RNA and/or target DNA binding.

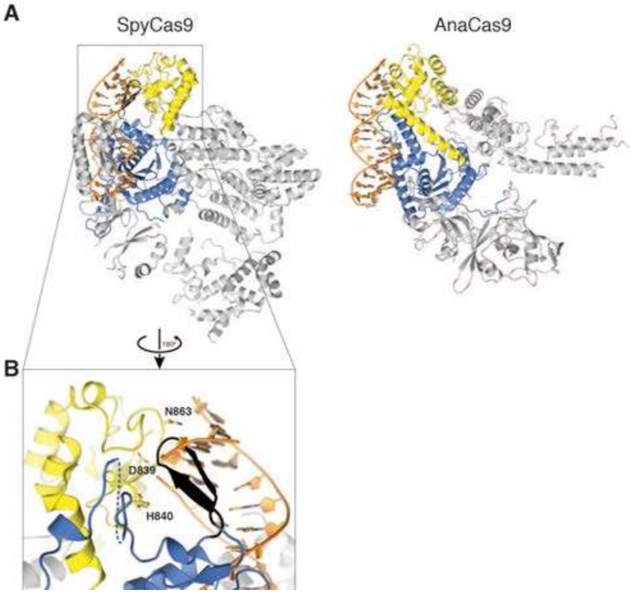


Figure 5.18 | Both SpyCas9 and AnaCas9 adopt auto-inhibited conformations in the apo state. (a) Models of substrate binding by the HNH domains in SpyCas9 (left) and AnaCas9 (right), based on the superposition of the Cas9 structures with the product-bound complex of the homing endonuclease I-HmuI (PDB entry 1U3E). Shown is a 17-base pair B-form DNA duplex that covers three base pairs 5' and 14 base pairs 3' of the scissile phosphate, respectively. The Cas9 proteins are shown in the same orientation, based on superposition of the respective HNH domains. The HNH domains are depicted in yellow, the RuvC domains are depicted in blue, and residues 1049^{Spy}-1059^{Spy} of the RuvC domain are shown in black. **(b)** Zoomed-in view of the HNH domain (yellow) active site in SpyCas9 occluded by the 1049^{Spy}-1059^{Spy} beta-hairpin (black).

5.4.7 RNA loading rearranges the two lobes of SpyCas9 to form a central channel

To visualize conformational states adopted by Cas9 upon guide RNA and target DNA binding, we determined the molecular architectures of SpyCas9 without guide RNA (apo-SpyCas9), in complex with crRNA:tracrRNA (SpyCas9:RNA), and bound to target DNA (SpyCas9:RNA:DNA) using negative-stain electron microscopy. Raw micrographs of the ~160-kDa apo-SpyCas9 enzyme show mono-disperse, globular particles with approximate dimensions of 120 Å x 140 Å, and two-dimensional (2D) reference-free class averages reveal that the enzyme has a two-lobed morphology in agreement with the crystal structure (**Fig. 5.19**). We used the random conical tilt (RCT) method (Radermacher et al., 1987) to obtain an initial, *ab initio* three-dimensional (3D) model of the complex (**Fig. 5.19**). Using multiple refinement procedures (see Materials and Methods), we arrived at a final reconstruction of apo-SpyCas9 at ~19-Å resolution (using the 0.5 Fourier Shell Correlation (FSC) criterion) that reveals a clam-shaped morphology with one large, globular lobe connected to a smaller lobe (**Fig. 5.20a**). Using SITUS (Chacón and Wriggers, 2002), we were able to computationally dock the alpha-helical and nuclease domain lobes of the SpyCas9 crystal structure as rigid bodies into the larger and smaller lobes, with cross-correlation coefficients (CCC) of 0.74 and 0.66, respectively (**Table 5.3** & **Fig. 5.21**). To further support our lobe assignment, we generated a 3D reconstruction of Cas9 containing an N-terminal maltose-binding protein (N-MBP) fusion directly upstream of the RuvC-I motif; N-MBP-Cas9 retains full DNA cleavage activity (**Fig. 5.21**). By 3D difference mapping, the additional density observed in this reconstruction was found to localize to the smaller lobe containing the RuvC nuclease domain (**Fig. 5.21**).

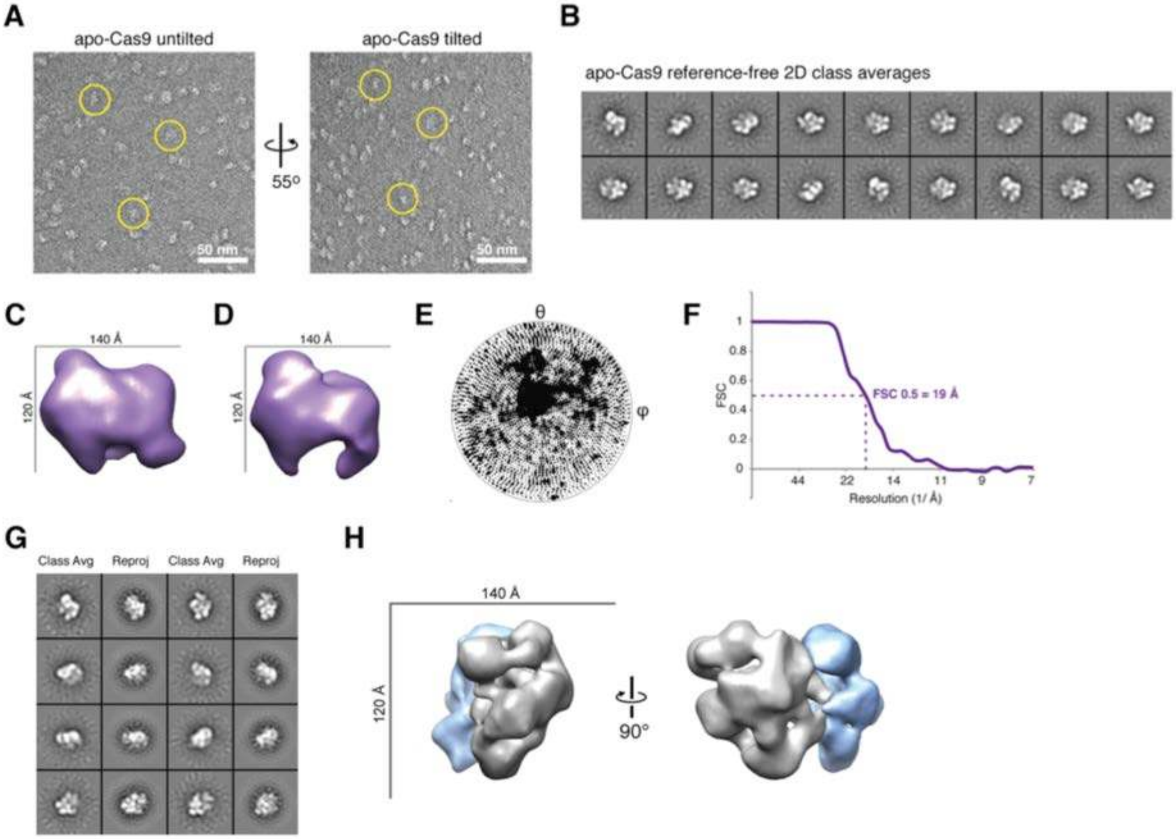


Figure 5.19 | Molecular architecture of apo-SpyCas9. (a) Representative untilted (left) and tilted (right) micrographs of negatively stained apo-SpyCas9. Scale bar indicates 50 nm. **(b)** Reference-free 2D class averages of apo-SpyCas9. The width of the boxes is ~316 Å. **(c)** Random conical tilt (RCT) class volume showing the *ab initio* structure of apo-SpyCas9. **(d)** Initial model generated by assigning Euler angles of the reference-free class averages with respect to the RCT volume. This initial model was used for refinement of the raw particle images of apo-SpyCas9. **(e)** Euler angle distribution for the final reconstruction. **(f)** Fourier shell correlation (FSC) curve for the final reconstruction, showing the resolution to be ~19 Å using the 0.5 FSC criterion. **(g)** Reference-free 2D class averages of apo-SpyCas9 (first and third columns) matched to reprojections of the final reconstruction (second and fourth columns). The width of the boxes is ~316 Å. **(h)** Final reconstruction of apo-SpyCas9 using the map in (d) as the initial model for refinement. The final map is segmented and colored as in **Fig. 5.20a**.

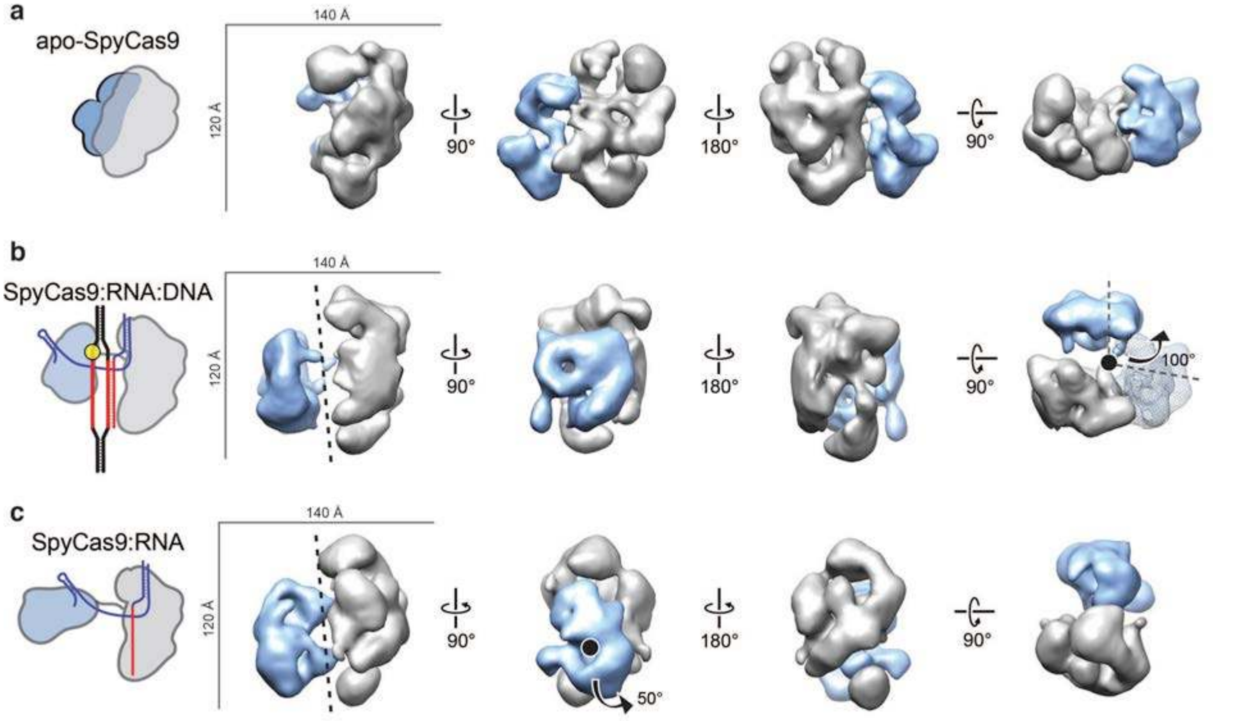


Figure 5.20 | RNA loading positions the two major lobes of SpyCas9 around a central channel. (a-c) Single particle EM reconstructions of negatively stained apo-SpyCas9 (a), SpyCas9:RNA:DNA (b), and SpyCas9:RNA (c) at 19-, 19-, and 21-Å resolution (using the 0.5 FSC criterion), respectively. Cartoon representations of the structures are shown (left). The structures are aligned based on the optimal cross-correlation coefficients between the independent alpha-helical lobes (grey). The smaller RuvC lobe (blue) in SpyCas9:RNA:DNA and SpyCas9:RNA rotates by ~100° (arrow in (b)) with respect to this lobe in the apo-Cas9 structure (transparent mesh) to form a central channel (black dashed line). There is a ~50° rotation (arrow in (c)) of the smaller lobe of SpyCas9:RNA along an axis perpendicular to this channel relative to SpyCas9:RNA:DNA.

Table 5.3 | Cross-correlation coefficient (CCC) analysis of docking results using SITUS.

	apo-SpyCas9 EM density		**SpyCas9:RNA:DNA EM density**	
	α-helical lobe	**α-helical lobe opposite hand**	**α-helical lobe**	**α-helical lobe opposite hand**
α-helical lobe crystal structure	0.74	0.70	0.83	0.74

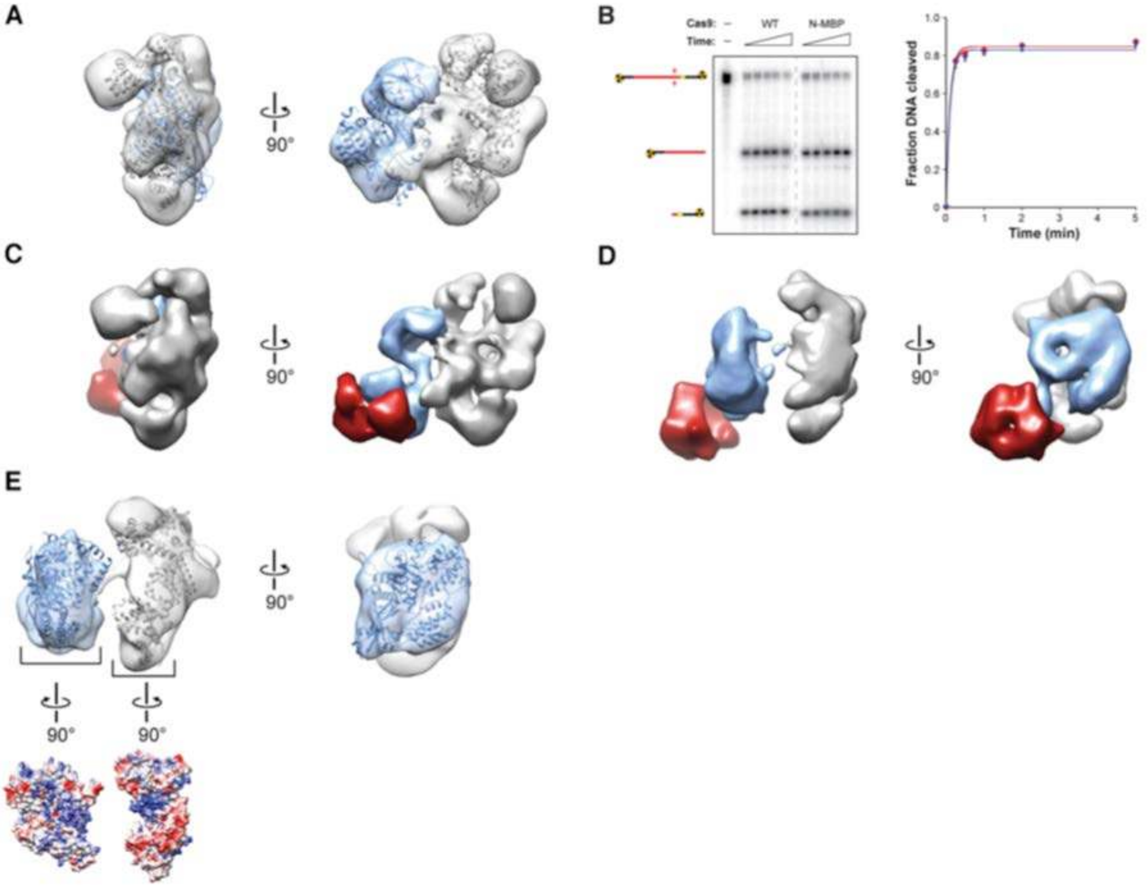


Figure 5.21 | Structural similarities between the apo-SpyCas9 EM structure and X-ray crystal structure. (a) The X-ray crystal structure of SpyCas9 was split into the alpha-helical lobe (residues 66-713) and the RuvC nuclease-containing lobe (residues 1-65 and 744-1363). Both lobes were computationally docked into the apo-SpyCas9 EM density as separate rigid bodies using SITUS (*86*), due to flexibility in the RuvC nuclease-containing lobe (blue) in the absence of bound nucleic acids (see Fig. S1B, blurry, smaller lobe in class averages). The HNH domain was excluded from docking for the same reason. **(b)** Activity assay with WT and N-MBP SpyCas9. DNA cleavage experiments were performed and resolved by 10% denaturing polyacrylamide gel electrophoresis (left). The data were plotted (right) and fit with single-exponentials (solid lines); error bars represent the standard deviation from three independent experiments and are not always visible. **(c,d)** 3D difference maps (> 7-σ) (red density) between the N-terminal MBP-labeled and unlabeled reconstructions of apo-SpyCas9 (c) and SpyCas9:RNA:DNA (d) were mapped onto the corresponding unlabeled reconstructions. **(e)** The X-ray crystal structure of SpyCas9 was again split into the alpha-helical lobe and nuclease-containing lobe and both lobes were computationally docked into the SpyCas9:RNA:DNA EM density as separate rigid bodies using SITUS (*86*) (top). This docking result is consistent with a 100° rigid body rotation of the nuclease lobe toward the alpha-helical lobe and places the two nucleic acid binding clefts across from one another (electrostatic surface potential, below). Further experiments and/or higher-resolution structures will be required to verify this working model.

We next prepared ribonucleoprotein complexes containing catalytically inactive D10A/H840A-SpyCas9 mutant, full-length crRNA and tracrRNA (SpyCas9:RNA) and bound these complexes to a 55 base-pair (bp) double-stranded DNA substrate at substrate concentrations expected to saturate Cas9, given an equilibrium dissociation constant of ~4 nM (**Fig. 5.22**). Reference-free 2D class averages of the DNA-bound complex (SpyCas9:RNA:DNA) hinted at a large-scale conformational change, with both lobes separating from one another into discrete structural units (**Fig. 5.22**). Using the apo-SpyCas9 structure low-pass filtered to 60 Å as a starting model, we obtained a 3D reconstruction of SpyCas9:RNA:DNA at ~19 Å resolution (using the 0.5 FSC criterion) that further reveals a substantial reorganization of the major lobes (**Fig. 5.20b**). An independently determined *ab initio* 3D structure using the RCT method (Radermacher et al., 1987) yielded similar results (**Fig. 5.22**). The shape of the larger lobe remains relatively unchanged from that in apo-Cas9 (CCC of 0.78), but the smaller lobe rotates by ~100° with respect to its position in the apo structure (**Fig. 5.20b**). An alternative model, assuming opposite handedness, also shows a large conformational change relative to the apo-Cas9 structure (**Fig. 5.23**). A reconstruction of SpyCas9:RNA:DNA using the N-terminal MBP fusion (**Fig. 5.22**) confirmed that the nuclease domain-containing lobe is rearranged with respect to the alpha-helical lobe in this complex. In this rearrangement, the nuclease domain lobe closes over the putative nucleic acid binding cleft on the alpha-helical lobe, forming a central channel with a width of ~25-Å that spans the length of both lobes (**Fig. 5.21**). Because nucleic acids cannot be visualized directly in EM structures of negatively stained complexes, this channel could be occupied by RNA and/or DNA.

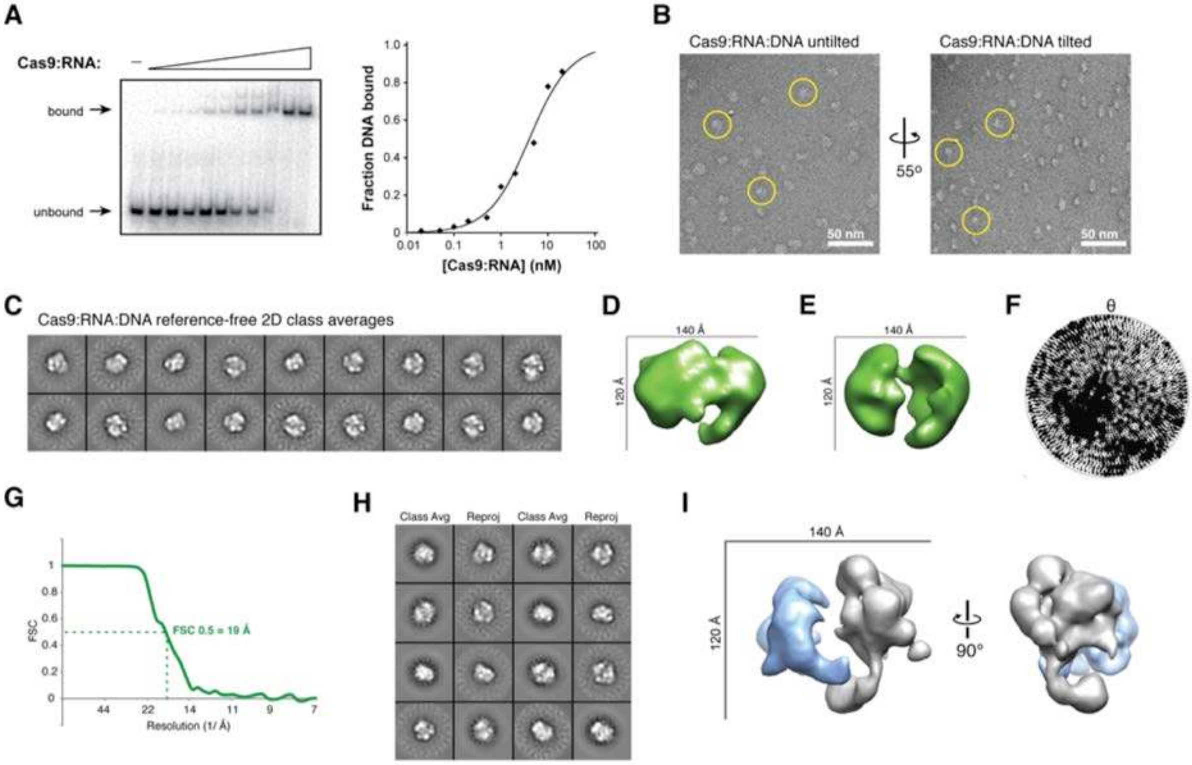


Figure 5.22 | Molecular architecture of SpyCas9:RNA:DNA. (a) Electrophoretic mobility gel shift assay (left) with radiolabeled target DNA and increasing concentrations of catalytically inactive (D10A/H840A) SpyCas9:RNA complex. Fitting of the quantified data with a standard binding isotherm (solid line, right) yields an equilibrium dissociation constant (K_d) of 4.0 ± 0.4 nM. **(b)** Representative untilted (left) and tilted (right) micrographs of negatively stained SpyCas9:RNA:DNA. Scale bar indicates 50 nm. **(c)** Reference-free 2D class averages of SpyCas9:RNA:DNA. The width of the boxes is ~316 Å. **(d)** Random conical tilt (RCT) class volume showing the *ab initio* structure of SpyCas9:RNA:DNA. **(e)** Initial model generated by assigning Euler angles of the reference-

free class averages with respect to the RCT volume. This initial model was used for refinement of the raw particle images of SpyCas9:RNA:DNA. **(f)** Euler angle distribution for the final reconstruction. **(g)** Fourier shell correlation (FSC) curve for the final reconstruction, showing the resolution to be ~19 Å using the 0.5 FSC criterion. **(h)** Reference-free 2D class averages of SpyCas9:RNA:DNA (first and third columns) matched to reprojections of the final reconstruction (second and fourth columns). The width of the boxes is ~316 Å. **(i)** Final reconstruction of SpyCas9:RNA:DNA using the map in (e) as the initial model for refinement. The final map is segmented and colored as in **Fig. 5.20b**.

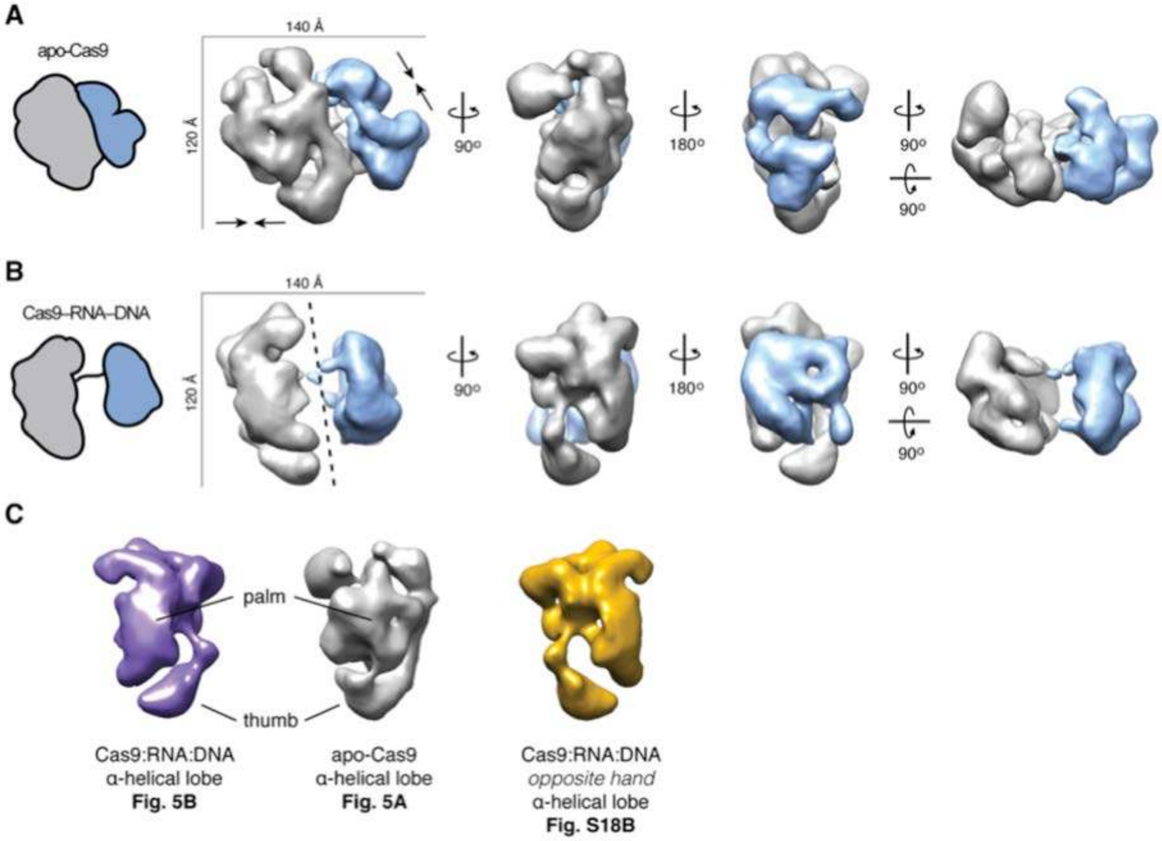


Figure 5.23 | Alternative model for the conformational change in SpyCas9:RNA:DNA complex considering the opposite handedness. (a,b) Single particle EM reconstructions of negatively stained apo-SpyCas9 **(a) (**as in the main text) and SpyCas9:RNA:**DNA (b) with** opposite handedness as the structure presented in Fig. 5.20b. Cartoon representations are shown on the left. In this alternative model, the movement of the smaller lobe with respect to the larger one is subtler. The blue lobe rotates in towards the larger lobe and reorganizes to form the central channel spanning the length of the enzyme (black dashed line). Note that the grey lobes of the two structures are aligned differently than the alignment in the main text, to maintain the blue lobe in a similar relative position for the two structures. (c) From left to right: the α-helical lobe of SpyCas9:RNA:DNA from Fig. 5.20b (purple), apo-Cas9 (grey), and Cas9:RNA:DNA with opposite handedness **from (b) (**gold), aligned to one another based on optimal cross correlation coefficient (CCC). The favored model presented in Fig. 5.20b of the main text is based on the more obvious, direct correspondence between the features of the apo-SpyCas9 α-helical lobe (grey) and the SpyCas9:RNA:DNA α-helical lobe (purple). Additionally, the α-helical domain from the crystal structure exhibits a higher CCC with the α-helical lobe from the model presented in Fig. 5.20b of the main text (purple) than the α-helical lobe of opposite handedness (gold) (0.83 versus 0.74).

We next wondered whether guide RNA alone induces the observed conformational rearrangements in SpyCas9, or whether both RNA and DNA are required for this structural change. To distinguish between these possibilities, we examined the architecture of SpyCas9:RNA in the absence of a bound target DNA molecule. Strikingly, reference-free 2D class averages of the SpyCas9:RNA showed a clear central channel similar to SpyCas9:RNA:DNA (**Fig. 5.24**). Using the 3D reconstruction of SpyCas9:RNA:DNA low-pass filtered to 60 Å as a starting model, we obtained a reconstruction of SpyCas9:RNA at ~21 Å resolution (using the 0.5 FSC criterion), which reveals a conformation similar to that of the DNA-bound complex (CCC of 0.89 with DNA-bound versus 0.81 with apo), with a central channel extending between the two lobes (**Fig. 5.20c** & **5.24**). Both the SpyCas9:RNA and SpyCas9:RNA:DNA complexes were more resistant to limited proteolysis by trypsin than apo-SpyCas9 and displayed similar digestion patterns, in agreement with these nucleic acid-bound complexes occupying a similar structural state (**Fig. 5.25**). While the smaller lobe appears to undergo an additional ~50° rotation along an axis perpendicular to the channel in the DNA-bound complex compared to SpyCas9:RNA, the same ~100° rotation around the channel is observed in both structures. Thus, loading of crRNA and tracrRNA alone is sufficient to convert the endonuclease into an active conformation for target surveillance.

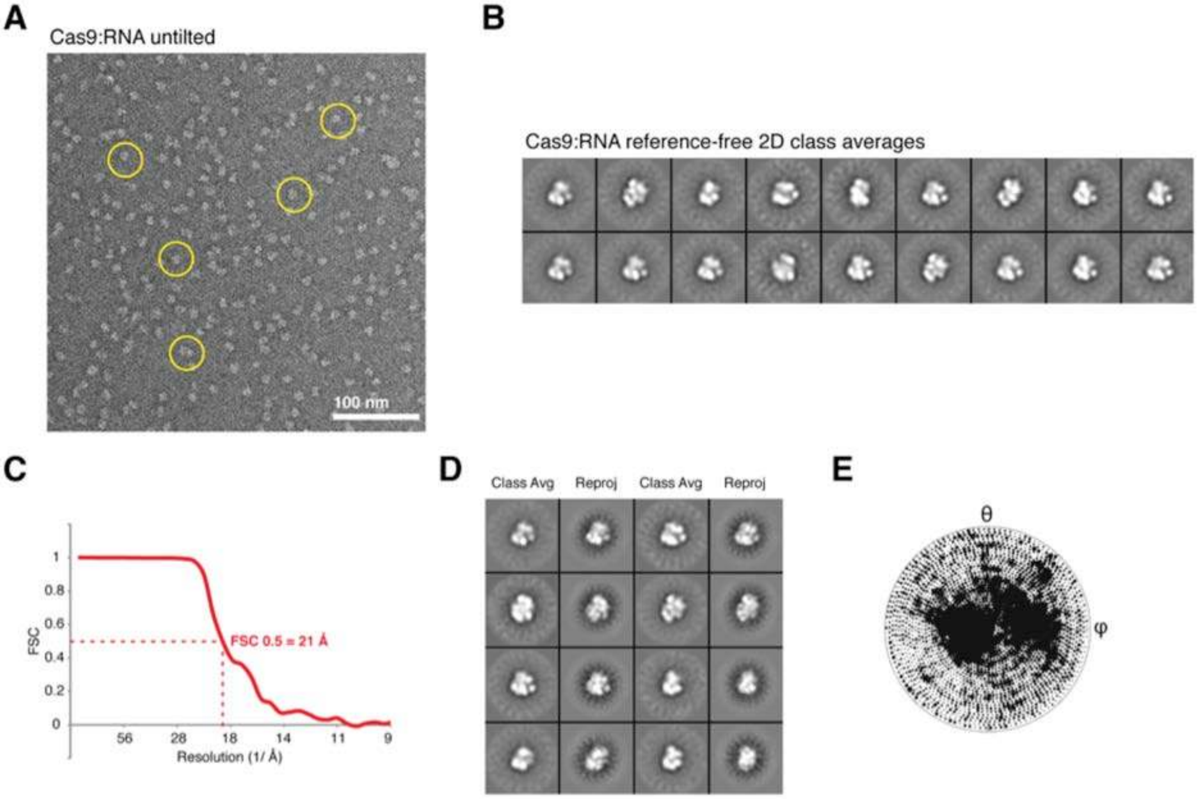


Figure 5.24 | Molecular architecture of SpyCas9:RNA. (a) Representative untilted micrograph of negatively stained SpyCas9:RNA. Scale bar indicates 100 nm. **(b)** Reference-free 2D class averages of SpyCas9:RNA. The width of the boxes is ~316 Å. **(c)** Fourier shell correlation (FSC) curve for the final reconstruction, showing the resolution to be ~21 Å using the 0.5 FSC criterion. **(d)** Reference-free 2D class averages of SpyCas9:RNA (first and third columns) matched to reprojections of the final reconstruction (second and fourth columns). The width of the boxes is ~316 Å. **(e)** Euler angle distribution for the final reconstruction.

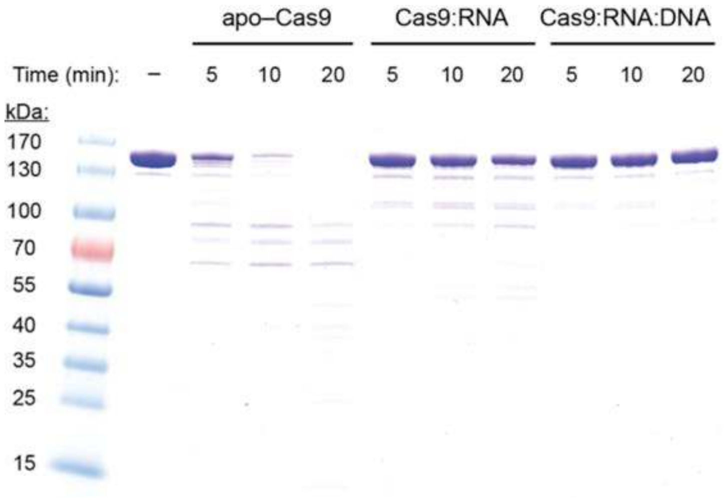


Figure 5.25 | Limited proteolysis of SpyCas9 with and without nucleic acid substrates suggests that nucleic acid-bound complexes adopt similar structural states. Apo-SpyCas9, SpyCas9 bound to full-length crRNA and tracrRNA (SpyCas9:RNA), or RNA-programmed SpyCas9 in complex with target DNA (SpyCas9:RNA:DNA) were prepared at a concentration of 2.5 µM and incubated with 2 ng/µl trypsin at 37 °C for the indicated time before quenching with 2X SDS gel-loading buffer. Samples were resolved by SDS-PAGE on a 4-20% gradient polyacrylamide gel (Bio-Rad). Apo-SpyCas9 is rapidly proteolyzed, whereas both SpyCas9:RNA and SpyCas9:RNA:DNA complexes are resistant to digestion by trypsin, suggesting that SpyCas9 undergoes similar structural rearrangements in both cases that mitigate proteolysis. Complexes were prepared with catalytically inactive D10A/H840A-SpyCas9 under the same conditions used to prepare samples for electron microscopy imaging.

5.4.8 The central channel in SpyCas9 accommodates bound target DNA and guide-RNAs

Based on the dimensions of the channel and the requirement for SpyCas9:RNA to recognize ~23 bps of its DNA substrate, we hypothesized that target DNA spans the central channel. To test this, we reconstituted SpyCas9:RNA:DNA complexes using DNA substrates containing 3'-biotin modifications (**Table 5.4**) to visualize the duplex ends via streptavidin labeling. Negative-stain EM analysis of samples labeled at either the PAM-distal (non-PAM) end, or both ends, showed additional circular density below, or both above and below the complex, respectively, along the central channel positioned between the two structural lobes (**Fig 5.26a,b**). These data support the conclusion that the major lobes of SpyCas9 enclose the target DNA, positioning the RNA:DNA heteroduplex along the central channel with the PAM oriented near the top. Interestingly, the additional streptavidin densities in the double-end labeled class averages are not completely parallel with the channel and instead wrap around the nuclease lobe (**Fig. 5.26b**), consistent with some degree of bending of the target DNA. Finally, we determined the orientation of RNA within SpyCas9:RNA complexes using streptavidin labeling of crRNA and tracrRNA containing biotin at their 3' termini, after ensuring that SpyCas9 retains full activity with these modified RNAs (**Fig. 5.27**). Using the same 2D and 3D difference mapping approach, we pinpointed the 3' end of the crRNA to the top of the channel (**Fig. 5.26c**), while the 3' end of the tracrRNA is extended roughly perpendicular to the central channel from the side of the nuclease domain lobe (**Fig. 5.26d**). The finding that the 3' end of the crRNA localizes to a similar position above the channel as the PAM-proximal side of the target suggests that the crRNA:DNA heteroduplex may be oriented roughly in parallel with the crRNA:tracrRNA duplex.

The channel between the lobes of SpyCas9:RNA:DNA can easily accommodate ~25 bps of a modeled A-form helix (**Fig. 5.28a**). Corroborating this, exonuclease III footprinting experiments indicated that Cas9 protects a ~26-bp segment of the target DNA (**Fig. 5.28b**). Additionally, P1 nuclease mapping experiments reveal that the displaced non-target strand is susceptible to degradation towards the 5' end of the protospacer, while the target strand that hybridizes to crRNA is protected along nearly its entire length. These results are consistent with the formation of an R-loop structure (**Fig. 5.28b**), as observed for other CRISPR-Cas targeting complexes (Jore et al., 2011b).

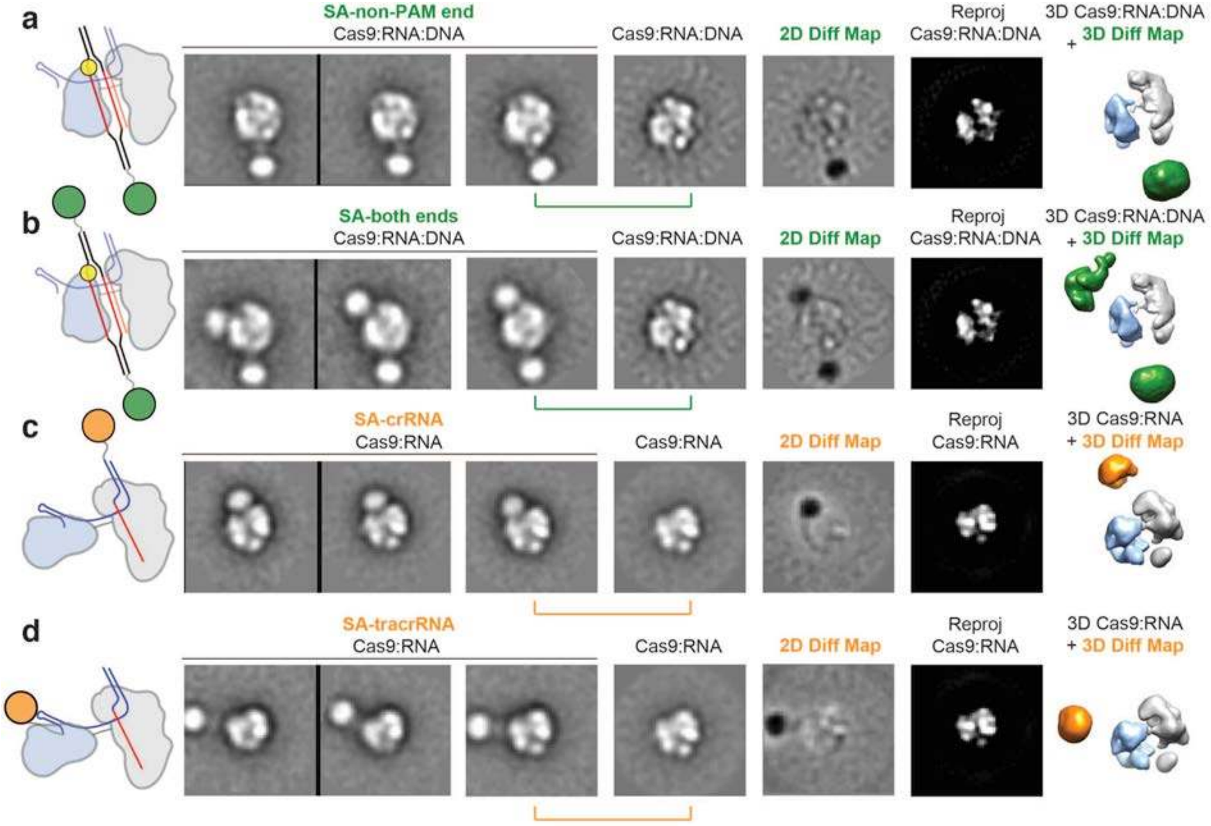


Figure 5.26 | Bound target DNA and guide RNAs span the central channel. (a,b) Biotinylated DNA duplexes were labeled with streptavidin (SA) at either the end distal to the PAM ((a), non-PAM) or at both ends (b). From left to right: schematic of structures and labels, five representative, reference-free 2D class averages, the corresponding reference-free 2D class average of unlabeled SpyCas9:RNA:DNA, a 2D difference map between the unlabeled and labeled structures, and the corresponding reprojection of the SpyCas9:RNA:DNA structure. The SpyCas9:RNA:DNA reconstruction is shown on the right with superimposed 3D difference density at ≥ 5-σ (green) between the SpyCas9:RNA:DNA reconstruction and the SA-labeled reconstruction. **(c,d)** Single particle EM analyses of SpyCas9:RNA labeled with SA at the 3' end of the crRNA (c) or tracrRNA (d). Data are shown as in (a), with the 3D difference density at ≥ 6-σ depicted in orange. The width of the boxes is ~316 Å.

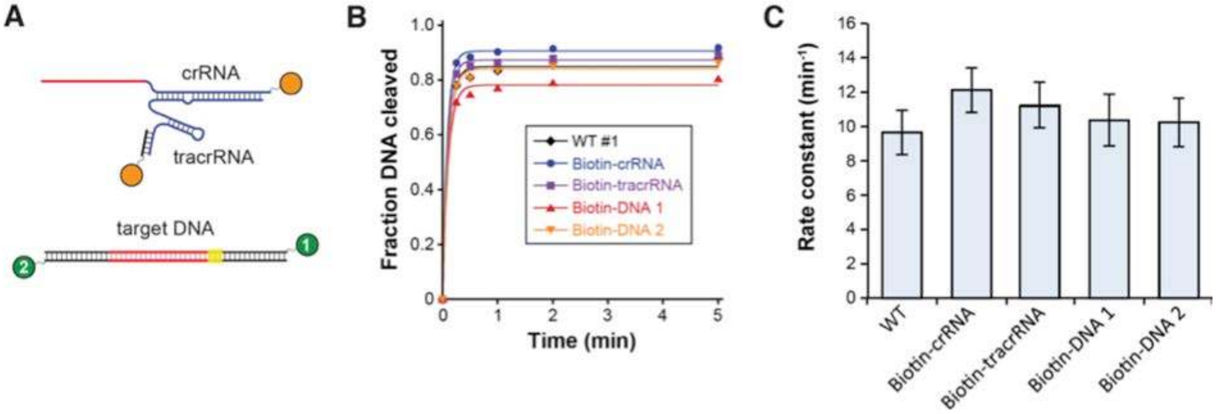


Figure 5.27 | Activity assays with biotin-RNA and biotin-DNA substrates used in streptavidin labeling experiments. **(a)** Schematic depicting the attachment of biotin (orange and green circles) to each nucleic acid substrate. Note that the crRNA and each strand of the DNA target are covalently linked to biotin at their 3' ends, whereas tracrRNA is hybridized to a short biotinylated DNA oligonucleotide at its 3' end. **(b)** DNA cleavage assays were conducted with biotin-labeled nucleic acids to verify that the modification does not perturb DNA recognition and cleavage. Data from representative time courses were plotted and fit with single-exponentials (solid line) to yield first-order rate constants for the DNA cleavage reaction. Note that the steep part of the curve (<15 seconds) could not be well defined due to the rapid reaction rate, limiting the accuracy of these measurements. **(c)** Three independent DNA cleavage time courses were conducted for each SpyCas9 construct, and the averaged rate constants are shown in the bar graph. The fitting error for individual single-exponential fits was greater than the standard deviation in rate constants between independent replicates, and so error bars represent the fitting error averaged from three independent experiments.

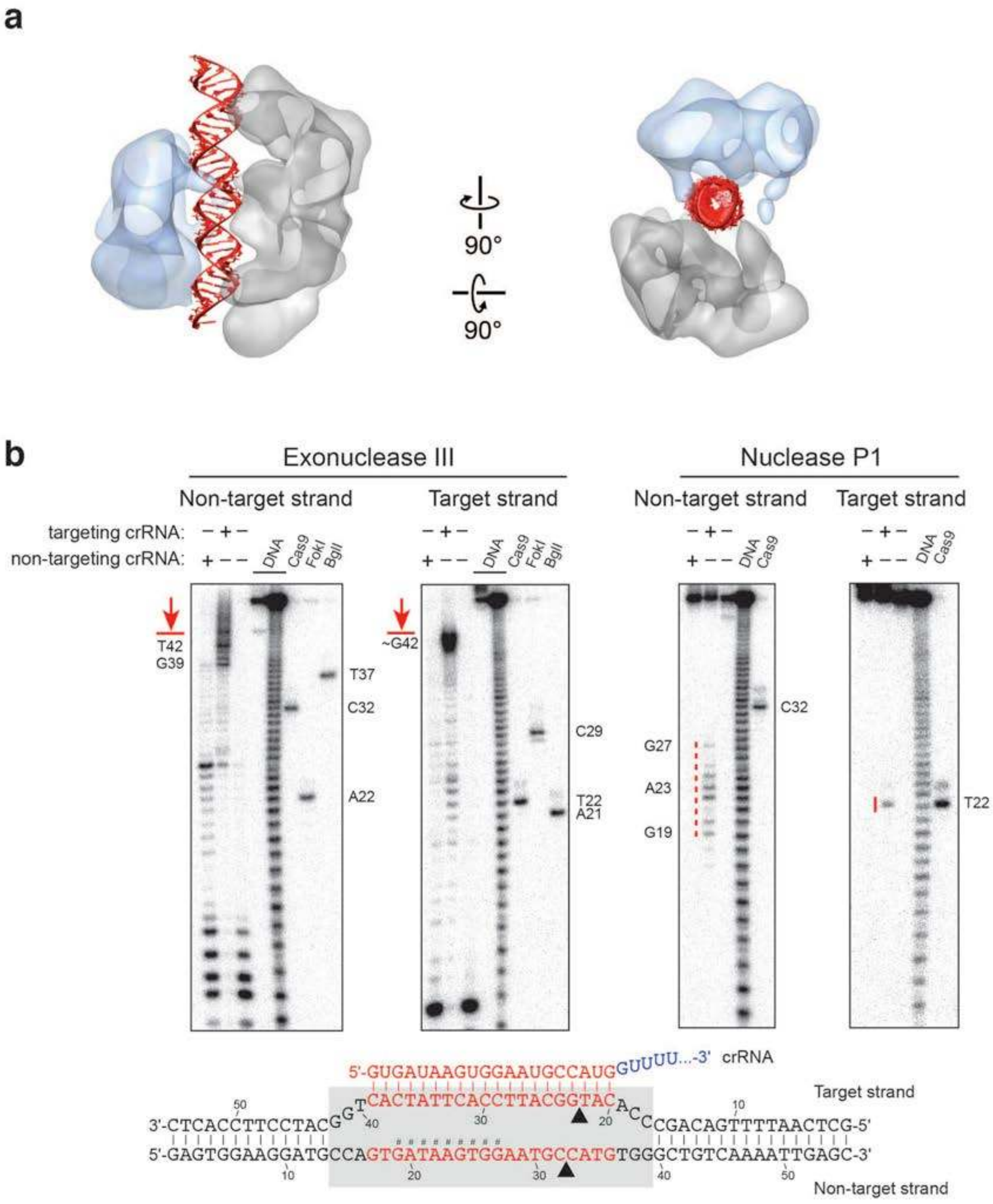


Figure 5.28 | SpyCas9 wraps around target DNA. (a) The central channel of the SpyCas9:RNA:DNA reconstruction (transparent surface) can easily accommodate ~25 bp of an A-form duplex (red). **(b)** Footprinting experiment with target DNA bound by SpyCas9:RNA. A 55-bp DNA substrate was 5'-radiolabeled on either the target or non-target strand and incubated with catalytically inactive SpyCas9:RNA containing a complementary crRNA (targeting) or a mismatched control crRNA (non-targeting), before being subjected to exonuclease III (left) or nuclease P1 (right) treatment. Reaction products were resolved by denaturing polyacrylamide gel electrophoresis; markers generated via digestion with BglI and FokI restriction enzymes and WT SpyCas9:RNA are labeled. The borders of the DNA target protected by SpyCas9:RNA are indicated in red next to the gel and with a grey box (bottom), and nucleotides susceptible to P1 digestion are indicated in red next to the gel and with hash tags in the schematic at the bottom.

Table 5.4 | List of nucleic acid reagents used in this study.

#	Description	Sequence (5'→3')
1	tracrRNA (nts 15-87)	GGACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUU
2	Targeting crRNA	GUGAUAAGUGGAAUGCCAUGGUUUUAGAGCUAUGCUGUUUUG
3	55-bp DNA substrate, non-target strand[a]	GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATGTGGGCTGTCAAAATTGAGC
4	55-bp DNA substrate, target strand[a]	GCTCAATTTTGACAGCCCACATGGCATTCCACTTATCACTGGCATCCTTCCACTC
5	Br-dU$_1$ containing 55 nt DNA substrate, non-target strand[a]	GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATG(Br-dU$_1$)GGGCTGTCAAAATTGAGC
6	Br-dU$_2$ containing 55 nt DNA substrate, target strand[a]	GCTCAATTTTGACAGCCC(Br-dU$_2$)CATGGCATTCCACTTATCACTGGCATCCTTCCACTC
7	reverse complement for # 6[a]	GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATGAGGGCTGTCAAAATTGAGC
8	Br-dU$_3$ containing 55 nt DNA substrate, non-target strand[a]	GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATGTGG(Br-dU$_3$)CTGTCAAAATTGAGC
9	reverse complement for #8[a]	GCTCAATTTTGACAGACCACATGGCATTCCACTTATCACTGGCATCCTTCCACTC
10	tracrRNA_ext[b]	GGACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUUGCUCGUGCGC
11	Biotinylated DNA oligo to hybridize to tracrRNA_ext[b]	Biotin-TTGCGCACGAGCAAA
12	Non-targeting crRNA (control, Fig. 7b)	GACGCAUAAAGAUGAGACGCGUUUUAGAGCUAUGCUGUUUUG
13	3'-Biotinylated DNA, non-target strand[c]	GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATG**TGG**GCTGTCAAAATTGAGC-Biotin
14	3'-Biotinylated DNA, target strand[c]	GCTCAATTTTGACAGCCCACATGGCATTCCACTTATCACTGGCATCCTTCCACTC-Biotin
15	ssDNA template for transcribing tracrRNA[c]	AAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATGCTGT**CCTATAGTGAGTCGTATTA**
16	ssDNA template for transcribing tracrRNA_ext[c]	GCGCACGAGCAAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATGCTGT**CCTATAGTGAGTCGTATTA**
17	Oligo for preparing double-stranded T7 promoters for *in vitro* transcription	TAATACGACTCACTATA

[a] The protospacer is depicted in red. The PAM is underlined.
[b] Nucleotides hybridizing between the tracrRNA_ext and biotin-DNA oligo are in blue.
[c] The reverse complement of the T7 promoter is indicated in **bold**.

5.5 Discussion

The crystal structures of Type II-A and II-C Cas9 proteins described here highlight the features in Cas9 enzymes that support their function as RNA-guided endonucleases. Cas9 enzymes adopt a bi-lobed architecture composed of a nuclease lobe containing juxtaposed RuvC and HNH nuclease domains and a variable alpha-helical lobe likely to be involved in nucleic acid binding. The identification of variable regions appended to a conserved Cas9 structural core provides a rationale for the diversity of crRNA:tracrRNA guide structures recognized by Cas9 enzymes and outlines a framework for protein engineering approaches aimed at altering catalytic function, guide RNA specificity, or PAM requirements.

Crosslinking experiments conducted in this study suggest that two unstructured tryptophan-containing loops in SpyCas9 contact the PAM in the target-bound complex. The location of the PAM-binding loops suggests that Cas9 interrogates DNA using flexible regions that may form an ordered binding site upon engaging target DNA. It is tempting to speculate that the two tryptophan residues in SpyCas9 mediate base-stacking interactions with the GG dinucleotide PAM. Alternatively, the tryptophan residues could be involved in extrahelical base extrusion of the PAM motif, in a manner similar to the mechanisms of numerous DNA modification and repair enzymes (Crenshaw et al., 2012; Qi et al., 2009; Song et al., 2011; Yang et al., 2008). It is also possible that the tryptophan residues are not directly involved in PAM binding, but instead reside near the PAM-binding pocket in the enzyme. We note, however, that the involvement of aromatic loop regions in SpyCas9 is highly reminiscent of the mechanism employed by the Cascade complex in Type I CRISPR-Cas systems (Sashital et al., 2012). The lack of conservation of the PAM binding region in Type II-C Cas9 enzymes is consistent with different PAM specificities observed for these endonucleases and points to distinct mechanisms of PAM recognition across the Cas9 enzyme family.

Single-molecule and biochemical experiments underscore the singular importance of PAM binding in both DNA interrogation and cleavage by Cas9 (Sternberg et al., 2014). The observed prevalence of PAM mutation as a mechanism of viral escape from CRISPR/Cas9 targeting (Semenova et al., 2011; 2009) has presumably spurred the evolution of Cas9 proteins with a variety of PAM specificities. It will be interesting to elucidate how PAM binding couples to Cas9 activation across the Cas9 superfamily, which has important implications for the use of these enzymes in genome engineering applications.

Both SpyCas9 and AnaCas9 structures support the conclusion that Cas9 enzymes maintain an autoinhibited conformation in the absence of nucleic acid ligands that requires restructuring upon guide RNA and target DNA binding. Consistent with this finding, electron microscopic reconstructions of SpyCas9-nucleic acid complexes show that the two lobes of the protein reorient substantially upon guide RNA association. Based on these observations, we propose a model for Cas9 function in which RNA loading drives structural rearrangements of the enzyme to enable productive encounters with target DNA (**Fig. 5.29**). Binding of crRNA:tracrRNA to Cas9 causes a substantial rotation of the small nuclease lobe relative to the larger lobe. This RNA-induced conformational change could occur either through direct interactions between the RNA and both lobes, or indirectly through allosteric effects. This reorganization may position the two major catalytic centers of the enzyme on opposite sides of the central channel, where the two separated strands are threaded into either active site.

While Type I and III CRISPR-Cas RNA-guided surveillance complexes form helical architectures that wrap around the crRNA (Rouillon et al., 2013; Spilman et al., 2013; Staals et al., 2013; Wiedenheft et al., 2011a), Cas9 instead forms a central channel. The helical morphology in these other systems may have evolved to accommodate the topological requirements of a longer crRNA:DNA heteroduplex, and the open helical arrangement exhibited by the Type I multi-subunit Cascade complex likely facilitates recruitment of the trans-acting Cas3 nuclease for target cleavage (Sinkunas et al., 2013). In contrast, Cas9 functions alone to both bind and cleave the DNA target, which could be facilitated by sequestering the substrate within the interior surface of the channel formed by both lobes. Although we do not observe extensive connecting density between the two lobes, we hypothesize that only one face will enable dsDNA to enter the central channel during 3D target search. While further experiments will be necessary to elucidate the precise search and recognition mechanisms used by Cas9, our structural analysis shows that RNA-loading serves as a key conformational switch in the activation and regulation of Cas9 enzymes.

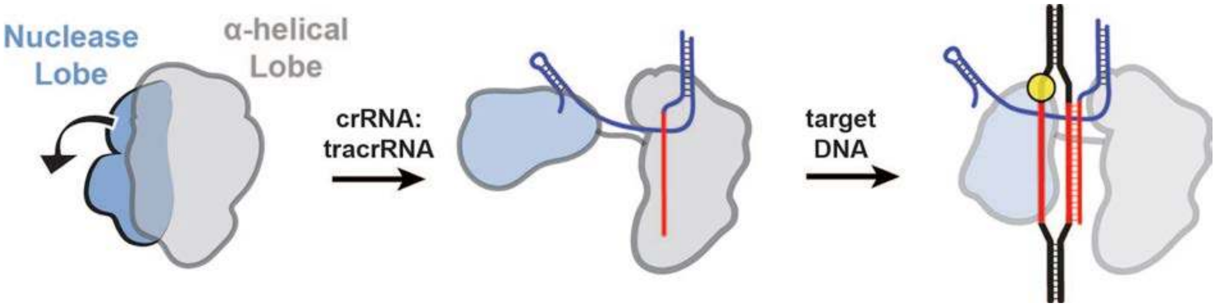


Figure 5.29 | Model for RNA-induced conversion of Cas9 into a structurally activated DNA surveillance complex. Upon binding the crRNA:tracrRNA guide, the two structural lobes of Cas9 reorient such that the two nucleic acid binding clefts face each other. This generates a central DNA binding channel, which allows access to double stranded DNA. Target DNA binding in the central channel and PAM-dependent R-loop formation result in a further structural rearrangement. Here, the nuclease domain lobe undergoes further rotation relative to the alpha-helical lobe, fully enclosing the DNA target, and the two nuclease domains engage both DNA strands for cleavage.

Chapter 6

Programmable RNA recognition and cleavage by CRISPR/Cas9

† Part of the work presented in this chapter has previously been published in the following research article: O'Connell, M.R., Oakes, B.L., Sternberg, S.H., East-Seletsky, A., Kaplan, M., Doudna, J.A. (2014). Programmable RNA recognition and cleavage by CRISPR/Cas9. Nature, doi:10.1038/nature13769.

‡ Mitchell O'Connell and Samuel Sternberg conceived the project. Mitchell O'Connell and Benjamin Oakes designed experiments, performed Cas9 binding and cleavage assays, and conducted pull-down experiments. Samuel Sternberg helped design experiments and performed biochemical assays. Alexandra East-Seletsky assisted with experimental design and performed cell-based experiments. Matias Kaplan generated biotin-Cas9 with the assistance of Samuel Sternberg. Jennifer Doudna supervised the project.

6.1 Abstract

The CRISPR-associated protein Cas9 is an RNA-guided DNA endonuclease that uses RNA:DNA complementarity to identify target sites for sequence-specific double-stranded DNA (dsDNA) cleavage. In its native context, Cas9 acts on DNA substrates exclusively because both binding and catalysis require recognition of a short DNA sequence, the protospacer adjacent motif (PAM), next to and on the strand opposite the 20-nucleotide target site in dsDNA. Cas9 has proven to be a versatile tool for genome engineering and gene regulation in many cell types and organisms, but it has been thought to be incapable of targeting RNA. Here we show that Cas9 binds with high affinity to single-stranded RNA (ssRNA) targets matching the Cas9-associated guide RNA sequence when the PAM is presented *in trans* as a separate DNA oligonucleotide. Furthermore, PAM-presenting oligonucleotides (PAMmers) stimulate site-specific endonucleolytic cleavage of ssRNA targets, similar to PAM-mediated stimulation of Cas9-catalyzed DNA cleavage. Using specially designed PAMmers, Cas9 can be specifically directed to bind or cut RNA targets while avoiding corresponding DNA sequences, and we demonstrate that this strategy enables the isolation of a specific endogenous mRNA from cells. These results reveal a fundamental connection between PAM binding and substrate selection by Cas9, and highlight the utility of Cas9 for programmable and tagless transcript recognition.

6.2 Materials and Methods

6.2.1 Cas9 and nucleic acid preparation

Wild-type Cas9 and catalytically inactive dCas9 (D10A/H840A) from *S. pyogenes* were purified as previously described (Jinek et al., 2012). crRNAs (42 nt) were either ordered synthetically (Integrated DNA Technologies) or transcribed *in vitro* with T7 polymerase using single-stranded DNA templates, as described (Sternberg et al., 2012). tracrRNA was transcribed *in vitro* and contained nucleotides 15–87 following the numbering scheme used previously (Jinek et al., 2012). λ-targeting sgRNAs were *in vitro* transcribed from linearized plasmids and contain full-length crRNA and tracrRNA connected via a GAAA tetraloop insertion. *GAPDH* mRNA-targeting sgRNAs were *in vitro* transcribed from dsDNA PCR products based on an optimized sgRNA design (Chen et al., 2013). Target ssRNAs (55–56 nt) were *in vitro* transcribed using single-stranded DNA templates. Sequences of all nucleic acid substrates used in this study can be found in **Table 6.1**.

All RNAs were purified using 10–15% denaturing polyacrylamide gel electrophoresis (PAGE). crRNA–tracrRNA duplexes were prepared by mixing equimolar concentrations of each RNA in hybridization buffer (20 mM Tris-HCl pH 7.5, 100 mM KCl, 5 mM $MgCl_2$), heating to 95 °C for 30 s and slow cooling. Fully double-stranded DNA/RNA substrates were prepared by mixing equimolar concentrations of each nucleic acid strand in hybridization buffer, heating to 95 °C for 30 s, and slow cooling. RNA, DNA, and chemically modified PAMmers were synthesized commercially (Intergrated DNA Technologies). DNA and RNA substrates were 5'-radiolabeled using [γ-^{32}P]-ATP (PerkinElmer) and T4 polynucleotide kinase (New England Biolabs). dsDNA and dsRNA substrates were 5'-radiolabeled on both strands, whereas only the target ssRNA was 5'-radiolabeled in other experiments.

6.2.2 Cleavage assays

Cas9–gRNA complexes were reconstituted before cleavage experiments by incubating Cas9 and the crRNA–tracrRNA duplex for 10 min at 37 °C in reaction buffer (20 mM Tris-HCl pH 7.5, 75 mM KCl, 5 mM $MgCl_2$, 1 mM dithiothreitol (DTT), 5% glycerol). Cleavage reactions were conducted at 37 °C and contained ~1 nM 5′-radiolabeled target substrate, 100 nM Cas9–RNA, and 100 nM PAMmer, where indicated. Aliquots were removed at each time point and quenched by the addition of RNA gel loading buffer (95% deionized formamide, 0.025% (w/v) bromophenol blue, 0.025% (w/v) xylene cyanol, 50 mM EDTA (pH 8.0), 0.025% (w/v) SDS). Samples were boiled for 10 min at 95 °C prior to being resolved by 12% denaturing PAGE. Reaction products were visualized by phosphorimaging and quantified with ImageQuant (GE Healthcare).

6.2.3 RNA cleavage site mapping

A hydrolysis ladder (OH^-) was obtained by incubating ~25 nM 5′-radiolabeled λ2 target ssRNA in hydrolysis buffer (25 mM CAPS (*N*-cyclohexyl-3-aminopropanesulfonic acid), pH 10.0, 0.25 mM EDTA) at 95 °C for 10 min, before quenching on ice. An RNase T1 ladder was obtained by incubating ~25 nM 5′-radiolabeled λ2 target ssRNA with 1 Unit RNase T1 (NEB) for 5 min at 37 °C in RNase T1 buffer (20 mM sodium citrate, pH 5.0, 1 mM EDTA, 2 M urea, 0.1 mg mL^{-1} yeast tRNA). The reaction was quenched by phenol/chloroform extraction before adding RNA gel loading buffer. All products were resolved by 15% denaturing PAGE.

6.2.4 Electrophoretic mobility shift assays

In order to avoid dissociation of the Cas9–gRNA complex at low concentrations during target ssRNA binding experiments, binding reactions contained a constant excess of dCas9 (300 nM), increasing concentrations of sgRNA, and 0.1–1 nM of target ssRNA. The reaction buffer was supplemented with 10 μg ml^{-1} heparin in order to avoid non-specific association of apo-dCas9 with target substrates (Sternberg et al., 2014). Reactions were incubated at 37 °C for 45 min before being resolved by 8% native PAGE at 4 °C (0.5× TBE buffer with 5 mM $MgCl_2$). RNA and DNA were visualized by phosphorimaging, quantified with ImageQuant (GE Healthcare), and analyzed with Kaleidagraph (Synergy Software).

6.2.5 Cas9 biotin labeling

To ensure specific labeling at a single residue on Cas9, two naturally occurring cysteine residues were mutated to serine (C80S and C574S) and a cysteine point mutant was introduced at residue M1. To attach the biotin moiety, 10 μM WT Cas9 or dCas9 was reacted with a 50-fold molar excess of EZ-Link Maleimide-PEG2-Biotin (Thermo Scientific) at 25 °C for 2 h. The reaction was quenched by the addition of 10 mM DTT, and unreacted Maleimide-PEG2-Biotin was removed using a Bio-Gel P-6 column (Bio-Rad). Labeling was verified using a streptavidin bead binding assay, where 8.5 pmol of biotinylated Cas9 or non-biotinylated Cas9 was mixed with either 25 μL streptavidin-agarose (Pierce Avidin Agarose; Thermo Scientific) or 25 μL streptavidin magnetic beads (Dynabeads MyOne Streptavidin C1; Life Technologies). Samples were incubated in Cas9 reaction buffer at RT for 30 minutes, followed by three washes with

Cas9 reaction buffer and elution in boiling SDS-PAGE loading buffer. Elutions were analysed using SDS-PAGE. Cas9 M1C biotinylation was also confirmed using mass spectroscopy performed in the QB3/Chemistry Mass Spectrometry Facility at UC Berkeley. Samples of intact Cas9 proteins were analyzed using an Agilent 1200 liquid chromatograph equipped with a Viva C8 (100 mm × 1.0 mm, 5 µm particles, Restek) analytical column and connected in-line with an LTQ Orbitrap XL mass spectrometer (Thermo Fisher Scientific). Mass spectra were recorded in the positive ion mode. Mass spectral deconvolution was performed using ProMass software (Novatia).

6.2.6 GAPDH mRNA pull-down

HeLa-S3 cell lysates were prepared as previously described (Lee et al., 2013). Total RNA was isolated from HeLa-S3 cells using Trizol reagent according to the manufacturer's instructions (Life Technologies). Cas9–sgRNA complexes were reconstituted before pull-down experiments by incubating a two-fold molar excess of Cas9 with sgRNA for 10 min at 37 °C in reaction buffer. HeLa total RNA (40 µg) or HeLa lysate (~5×10^6 cells) was added to reaction buffer with 40U RNasin (Promega), PAMmer (5 µM) and the biotin-dCas9 (50 nM):sgRNA (25 nM) in a total volume of 100 µL and incubated at 37 °C for 1 h. This mixture was then added to 25 µL magnetic streptavidin beads (Dynabeads MyOne Streptavidin C1; Life Technologies) pre-equilibrated in reaction buffer and agitated at 4 °C for 2 h. Beads were then washed six times with 300 µL wash buffer (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 5mM $MgCl_{2,}$ 0.1% Triton X-100, 5% glycerol, 1mM DTT, 10 µg ml^{-1} heparin). Immobilized RNA was eluted by heating beads at 70 °C in the presence of DEPC-treated water and a phenol/chloroform mixture. Eluates were then treated with an equal volume of glyoxal loading dye (Life Technologies) and heated at 50 °C for 1 h before separation via 1% BPTE agarose gel (30 mM Bis-Tris, 10 mM PIPES, 10 mM EDTA, pH 6.5). Northern blot transfers were carried out according to Chomczynski *et al.* (Chomczynski, 1992). Following transfer, membranes were crosslinked using UV radiation and incubated in pre-hybridization buffer (UltraHYB Ultrasensitive Hybridization Buffer; Life Technologies) for 1 h at 46 °C prior to hybridization. Radioactive Northern probes were synthesized using random priming of *GAPDH* and *β-actin* partial cDNAs (for cDNA primers, see **Table 6.1**) in the presence of [α-^{32}P]-dATP (PerkinElmer), using a Prime-It II Random Primer Labeling kit (Agilent Technologies). Hybridization was carried out for 3 h in pre-hybridization buffer at 46 °C followed by two washes with 2×SSC (300mM NaCl, 30 mM trisodium citrate, pH 7, 0.5% (w/v) SDS) for 15 min at 46 °C. Membranes were imaged using a phosphorscreen.

6.3. Results and Discussion

CRISPR–Cas immune systems must discriminate between self and non-self to avoid an autoimmune response (Marraffini and Sontheimer, 2010b). In type I and II systems, foreign DNA targets which contain adjacent PAM sequences are targeted for degradation, whereas potential targets in CRISPR loci of the host do not contain PAMs and are avoided by RNA-guided interference complexes (Garneau et al., 2010; Gasiunas et al., 2012; Mojica et al., 2009; Sashital et al., 2012). Single-molecule and bulk biochemical experiments showed that PAMs act both to recruit Cas9–guide RNA complexes (Cas9–gRNA) to potential target sites and to trigger nuclease domain activation (Sternberg et al., 2014). Cas9 from *Streptococcus pyogenes* recognizes a 5’-NGG-3’ PAM on the non-target (displaced) DNA strand (Jinek et al., 2012; Mojica et al., 2009), suggesting that PAM recognition may stimulate catalysis through allosteric regulation. Moreover, the HNH nuclease domain of Cas9, which is responsible for target strand cleavage (Gasiunas et al., 2012; Jinek et al., 2012), is homologous to other HNH domains shown previously to cleave RNA substrates (Hsia et al., 2004; Pommer et al., 2001). Based on the observation that single-stranded DNA (ssDNA) targets can be activated for cleavage by a separate PAMmer oligonucleotide (Sternberg et al., 2014), and that similar HNH domains can cleave RNA, we wondered whether a similar strategy would enable Cas9 to bind and cleave ssRNA targets in a programmable fashion (**Fig. 6.1a**).

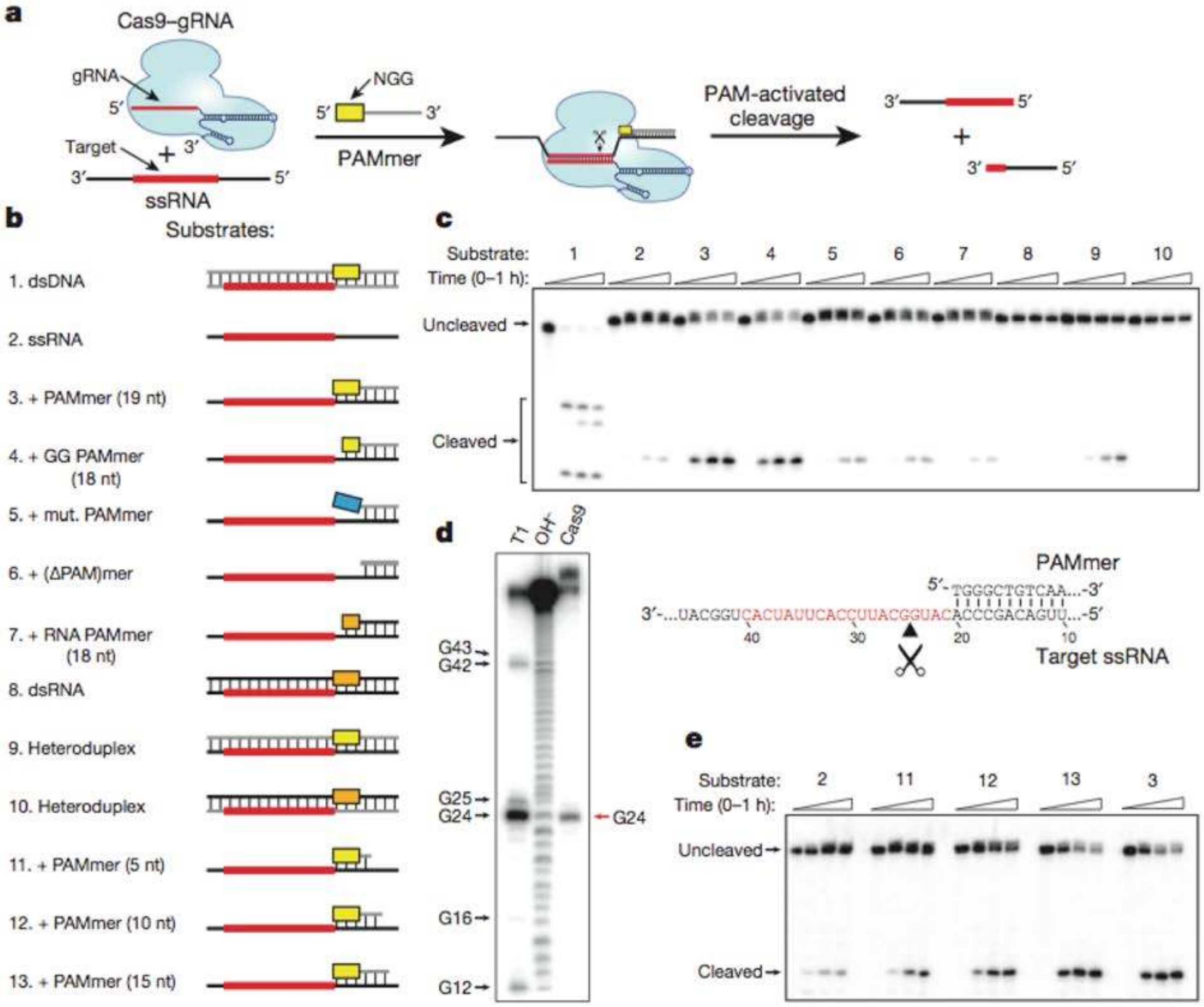


Figure 6.1 | RNA-guided Cas9 cleaves ssRNA targets in the presence of a short PAM-presenting DNA oligonucleotide (PAMmer). (a) Schematic depicting the approach used to target ssRNA for programmable, sequence-specific cleavage. (b) The panel of nucleic acid substrates examined in this study. Substrate elements are colored as follows: DNA (grey), RNA (black), guide RNA target sequence (red), DNA PAM (yellow), mutated DNA PAM (blue), RNA PAM (orange). **(c)** Representative cleavage assay for 5’-radiolabeled nucleic acid substrates using Cas9–gRNA, numbered as in (b). **(d)** Cas9–gRNA cleavage site mapping assay for substrate 3. T1 and OH$^-$ denote RNase T1 and hydrolysis ladders, respectively; the sequence of the target ssRNA is shown at right. **(e)** Representative ssRNA cleavage assay in the presence of PAMmers of increasing length, numbered as in (b).

Using *S. pyogenes* Cas9 and dual-guide RNAs (Methods), we performed *in vitro* cleavage experiments using a panel of different RNA and DNA targets (**Fig. 6.1b**). Deoxyribonucleotide-comprised PAMmers specifically activated Cas9 to cleave ssRNA (**Fig. 6.1c**), an effect that required a 5'-NGG-3' or 5'-GG-3' PAM. RNA cleavage was not observed using ribonucleotide-based PAMmers, suggesting that Cas9 may recognize the local helical geometry and/or deoxyribose moieties within the PAM. Consistent with this idea, dsRNA targets were not cleavable, and RNA–DNA heteroduplexes could only be cleaved when the non-target strand was composed of deoxyribonucleotides. Interestingly, we found that Cas9 cleaved the ssRNA target strand between positions 4 and 5 of the base-paired guide RNA-target RNA hybrid (**Fig. 6.1d**), in contrast to the cleavage between positions 3 and 4 observed for dsDNA (Garneau et al., 2010; Gasiunas et al., 2012; Jinek et al., 2012) likely due to subtle differences in substrate positioning. However, we did observe a significant reduction in the pseudo-first order cleavage rate constant of PAMmer-activated ssRNA as compared to ssDNA (Sternberg et al., 2014) (**Fig. 6.2**).

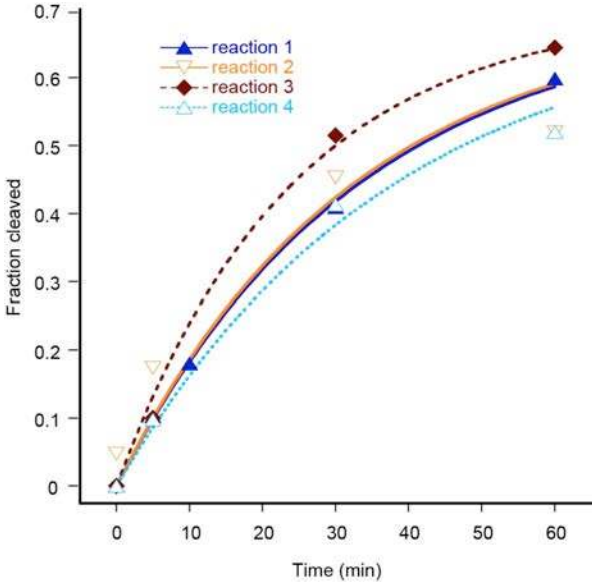


Figure 6.2 | Quantified data for cleavage of ssRNA by Cas9–gRNA in the presence of a 19-nt PAMmer. Cleavage assays were conducted as described in the **Materials and Methods**, and the quantified data were fit with single-exponential decays. Results from four independent experiments yielded an average apparent pseudo-first order cleavage rate constant of 0.032 ± 0.007 min^{-1}. This is slower than the rate constant determined previously for ssDNA in the presence of the same 19-nt PAMmer (7.3 ± 3.2 min^{-1}) (Sternberg et al., 2014).

We hypothesized that PAMmer nuclease activation would depend on the stability of the hybridized PAMmer–ssRNA duplex and tested this by varying the PAMmer length. As expected, ssRNA cleavage was lost when the predicted melting temperature for the duplex decreased below the temperature used in our experiments (**Fig. 6.1e).** In addition, large molar excesses of di- or tri-deoxyribonucleotides in solution were poor activators of Cas9 cleavage (**Fig. 6.3**). Collectively, these data demonstrate that hybrid substrate structures composed of ssRNA and deoxyribonucleotide-based PAMmers that anneal upstream of the RNA target sequence can be efficiently cleaved by RNA-guided Cas9.

We next investigated the binding affinity of catalytically inactive (dCas9; D10A/H840A) dCas9–gRNA for ssRNA targets with and without PAMmers using native gel mobility shift experiments. Intriguingly, while our previous results showed that ssDNA and PAMmer-activated ssDNA targets are bound with indistinguishable affinity (Sternberg et al., 2014), PAMmer-activated ssRNA targets were bound >500-fold tighter than ssRNA alone (**Fig. 6.4a,b**). A recent crystal structure of Cas9 bound to a ssDNA target revealed deoxyribose -specific van der Waals interactions between the protein and the DNA backbone (Nishimasu et al., 2014), suggesting that energetic penalties associated with ssRNA binding must be attenuated by favorable compensatory binding interactions with the provided PAM. The equilibrium dissociation constant measured for a PAMmer–ssRNA substrate was within 5-fold of that for dsDNA (**Fig. 6.4b**), and this high-affinity interaction again required a cognate deoxyribonucleotide-comprised 5'-GG-3' PAM (**Fig. 6.4a**). Tight binding also scaled with the PAMmer length (**Fig. 6.4c**), consistent with the cleavage data presented above.

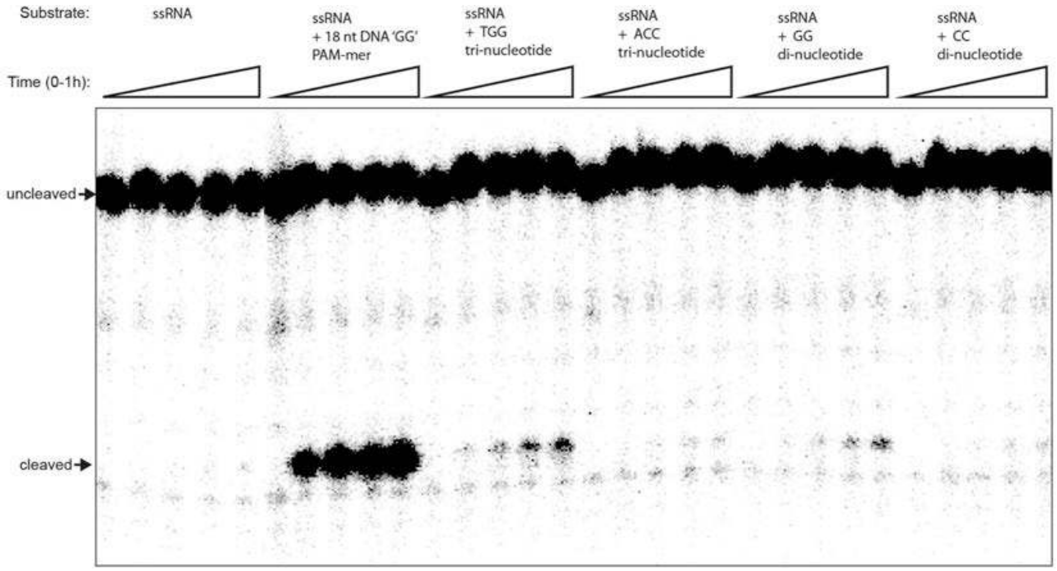


Figure 6.3 | RNA cleavage is marginally stimulated by di- and tri-deoxyribonucleotide PAMmers. Cleavage reactions contained ~1 nM 5′- radiolabelled target ssRNA and no PAMmer (left), 100 nM 18-nt PAMmer (second from left), or 1 mM of the indicated di- or tri-nucleotide (remaining lanes). Reaction products were resolved by 12% denaturing polyacrylamide gel electrophoresis (PAGE) and visualized by phosphorimaging.

To verify the programmable nature of PAMmer-mediated ssRNA cleavage by Cas9–gRNA, we prepared three distinct guide RNAs (λ2, λ3, and λ4) and showed that their corresponding ssRNA targets could be efficiently cleaved using complementary PAMmers without any detectable cross-reactivity (**Fig. 6.5a**). This result indicates that complementary RNA–RNA base-pairing is critical in these reactions. Surprisingly, though, dCas9 programmed with the λ2 guide RNA bound all three PAMmer–ssRNA substrates with similar affinity (**Fig. 6.5b**). This observation suggests that high-affinity binding in this case may not require correct base-pairing between the guide RNA and the ssRNA target, particularly given the compensatory role of the PAMmer.

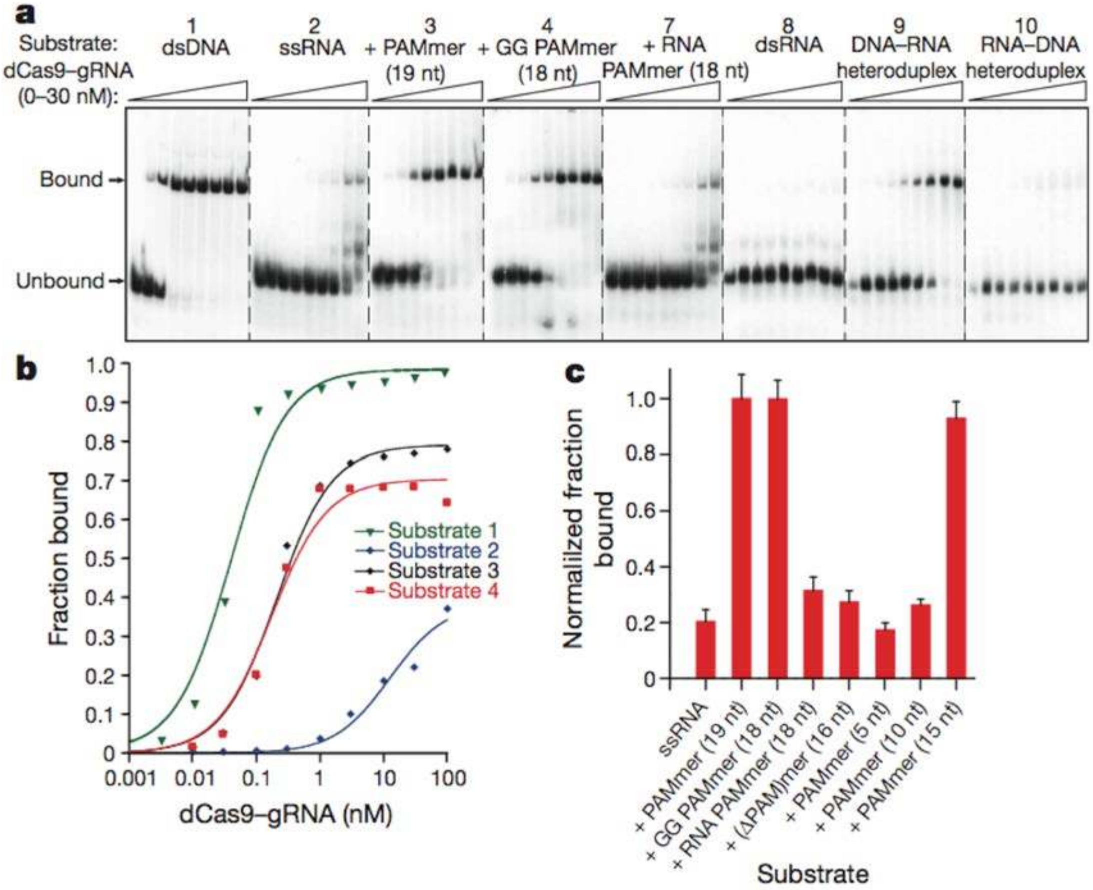


Figure 6.4 | dCas9–gRNA binds ssRNA targets with high affinity in the presence of PAMmers. (a) Representative electrophoretic mobility shift assay for binding reactions with dCas9–gRNA and a panel of 5'-radiolabeled nucleic acid substrates, numbered as in **Fig. 6.1b**. **(b)** Quantified binding data for substrates 1–4 from (a) fit with standard binding isotherms. Measured dissociation constants from three independent experiments (mean ± s.d.) were 0.036 ± 0.003 nM (**1**), >100 nM (**2**), 0.20 ± 0.09 nM (**3**), and 0.18 ± 0.07 nM (**4**). **(c)** Relative binding data for 1nM dCas9–gRNA and 5'-radiolabeled ssRNA with a panel of different PAMmers. The data are normalized to the amount of binding observed at 1 nM dCas9–gRNA with a 19 nt PAMmer; error bars represent the standard deviation from three independent experiments.

During dsDNA targeting by Cas9–gRNA, duplex melting proceeds directionally from the PAM and strictly requires formation of complementary RNA–DNA base-pairs to offset the energetic costs associated with dsDNA unwinding (Sternberg et al., 2014). We therefore wondered whether binding specificity for ssRNA substrates would be recovered using PAMmers containing 5'-extensions that create a partially double-stranded target region requiring unwinding (**Fig. 6.5c**). Indeed, we found that use of a 5'-extended PAMmer enabled dCas9 bearing the λ2 guide sequence to bind sequence-selectively to the λ2 PAMmer–ssRNA target. The λ3 and λ4 PAMmer–ssRNA targets were not recognized under these conditions (**Fig. 6.5d** & **6.6**), although we did observe a 10-fold reduction in overall ssRNA substrate binding affinity. By systematically varying the length of the 5' extension, we found that PAMmers containing 2–8 additional nucleotides upstream of the 5'-NGG-3' offer an optimal compromise between gains in binding specificity and concomitant losses in binding affinity and cleavage efficiency (**Fig. 6.7**).

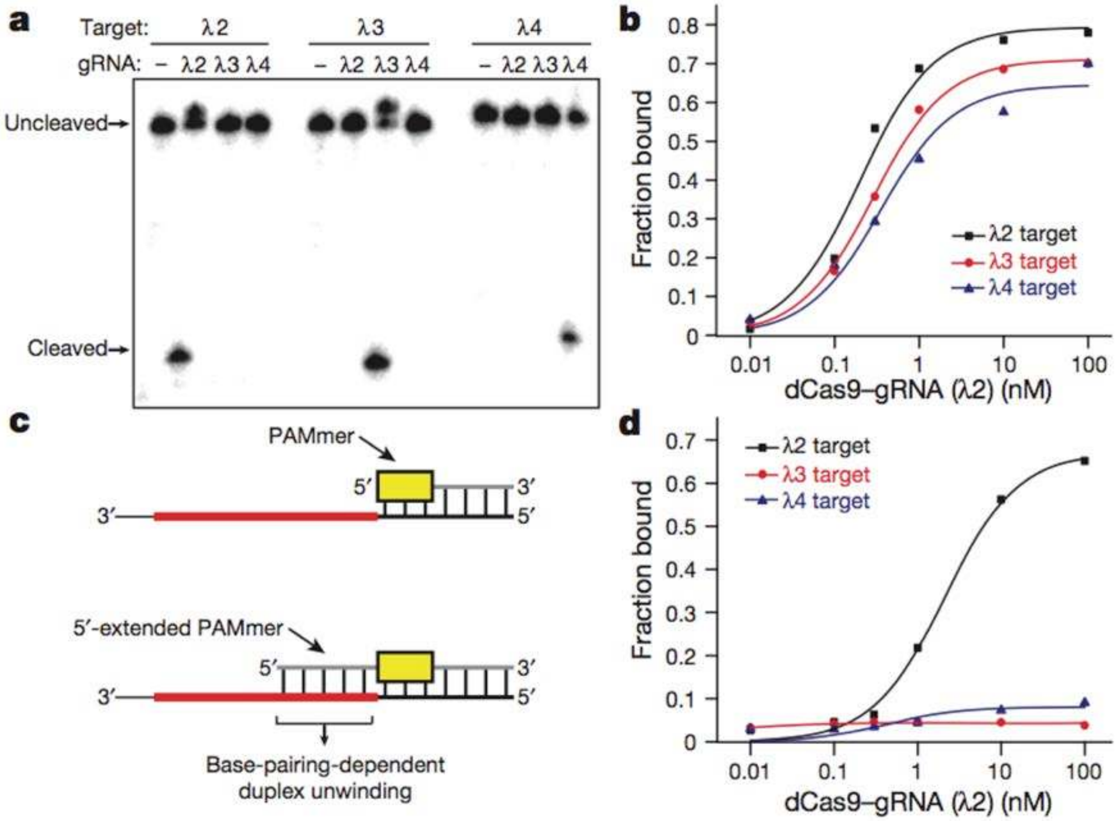


Figure 6.5 | 5’-extended PAMmers are required for specific target ssRNA binding. **(a)** Cas9 programmed with either λ2, λ3, or λ4-targeting gRNAs exhibits sequence-specific cleavage of 5’-radiolabeled λ2, λ3, and λ4 target ssRNAs, respectively, in the presence of cognate PAMmers. **(b)** dCas9 programmed with a λ2-targeting gRNA exhibits similar binding affinity to λ2, λ3, and λ4 target ssRNAs in the presence of cognate PAMmers. Dissociation constants from three independent experiments (mean ± s.d.) were 0.20 ± 0.09 nM (**λ2**), 0.33 ± 0.14 nM (**λ3**), and 0.53 ± 0.21 nM (**λ4**). **(c)** Schematic depicting the approach used to restore guide RNA-mediated ssRNA binding specificity, which involves 5’-extensions to the PAMmer that cover part of the target sequence. **(d)** dCas9 programmed with a λ2-targeting gRNA specifically binds the λ2 ssRNA but not λ3 and λ4 ssRNAs in the presence of 5’-extended PAMmers. Dissociation constants from three independent experiments (mean ± s.d.) were 3.3 ± 1.2 nM (**λ2**) and >100 nM (**λ3** and **λ4**).

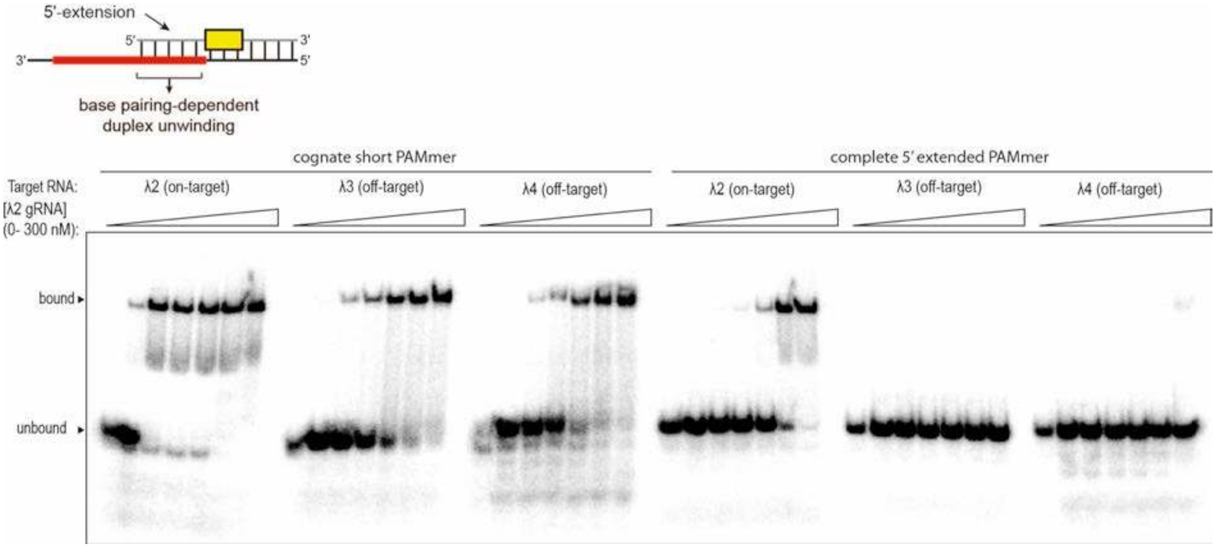


Figure 6.6 | Representative binding experiment demonstrating guide-specific ssRNA binding with 5'-extended PAMmers. Gel shift assays were conducted as described in the **Materials and Methods**. Binding reactions contained Cas9 programmed with λ2-gRNA and either λ2 (on-target), λ3 (off-target) or λ4 (off-target) ssRNA in the presence of short cognate PAMmers or cognate PAMmers with complete 5'-extensions, as indicated. The presence of a cognate 5'-extended PAM- mer abrogates off-target binding. Three independent experiments were conducted to produce the data shown in **Fig. 6.5b,d**.

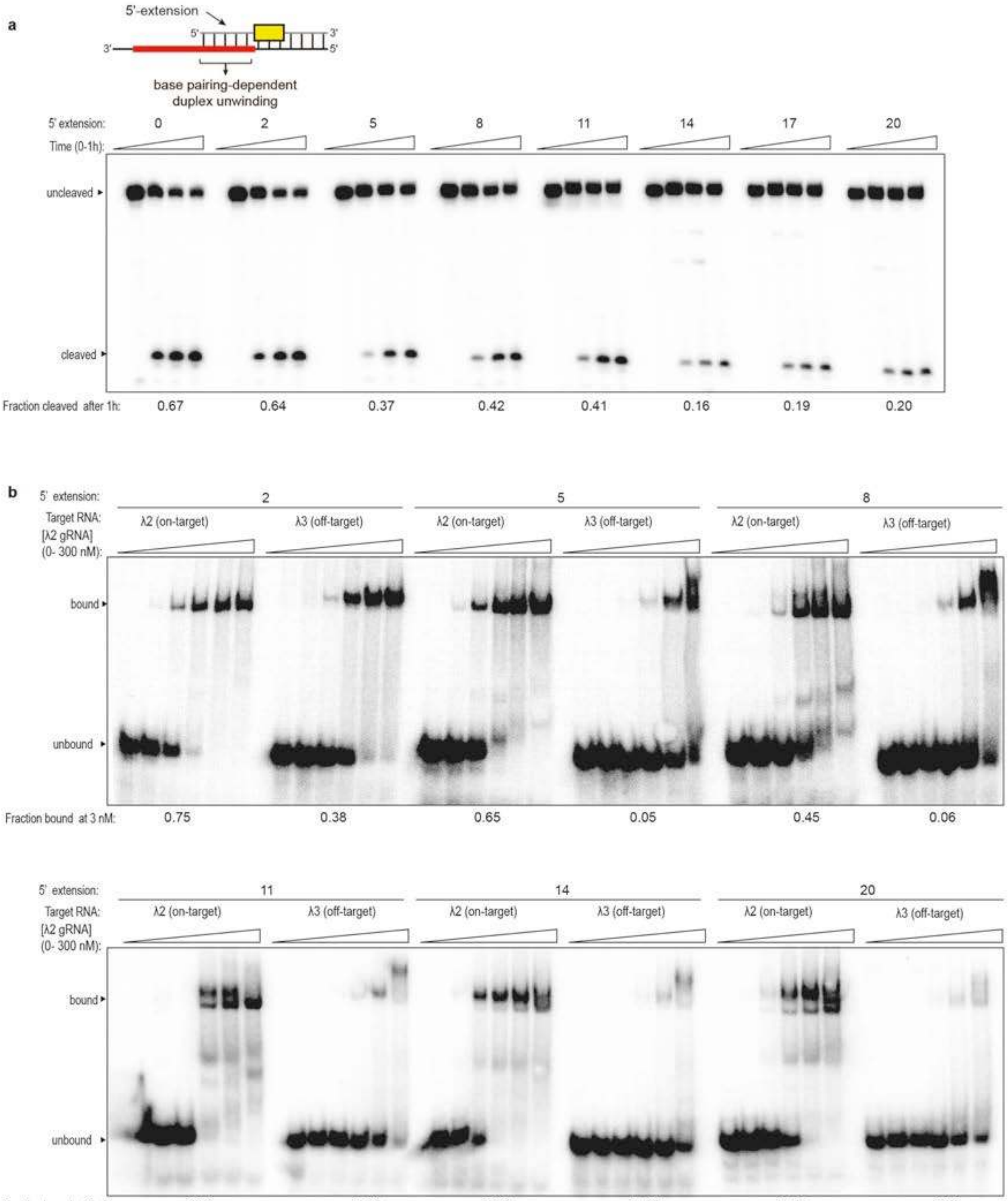


Figure 6.7 | Exploration of RNA cleavage efficiencies and binding specificity using PAMmers with variable 5'-extensions. (a) Cleavage assays were conducted as described in the **Materials and Methods**. Reactions contained Cas9 programmed with λ2-gRNA and λ2 ssRNA target in the presence of PAMmers with 5'-extensions of variable length. The ssRNA cleavage efficiency decreases as the PAMmer extends further into the target region, as indicated by the fraction RNA cleaved after 1 h. **(b)** Binding assays were conducted as described in the **Materials and Methods**, using mostly the same panel of 5'-extended PAMmers as in (a). Binding reactions contained Cas9 programmed with λ2-gRNA and either λ2 (on-target) or λ3 (off-target) ssRNA in the presence of cognate PAMmers with 5'-extensions of variable length. The binding specificity increases as the PAMmer extends further into the target region, as indicated by the fraction of λ3 (off-target) ssRNA bound at 3 nM Cas9-gRNA. PAMmers with 5' extensions also cause a slight reduction in the relative binding affinity of λ2 (on-target) ssRNA.

Next we investigated whether nuclease activation by PAMmers requires base-pairing between the 5’-NGG-3’ and corresponding nucleotides on the ssRNA. Prior studies showed that DNA substrates containing a cognate PAM that is mismatched with the corresponding nucleotides on the target strand are cleaved as efficiently, under some conditions, as a fully base-paired PAM (Jinek et al., 2012). Importantly, this could enable targeting of RNA while precluding binding or cleavage of corresponding genomic DNA sites lacking PAMs (**Fig. 6.8a**). To test this possibility, we first demonstrated that Cas9–gRNA cleaves PAMmer–ssRNA substrates regardless of whether or not the PAM is base-paired (**Fig. 6.8b,c**). When Cas9–RNA was incubated with both a PAMmer–ssRNA substrate and the corresponding dsDNA template containing a cognate PAM, both targets were cleaved. In contrast, when a dsDNA target lacking a PAM was incubated together with a PAMmer-ssRNA substrate bearing a mismatched 5’-NGG-3’ PAM, Cas9–gRNA selectively targeted the ssRNA for cleavage (**Fig. 6.8c**). The same result was obtained using a mismatched PAMmer with a 5’ extension (**Fig. 6.8c**), demonstrating that this general strategy enables the specific targeting of RNA transcripts while effectively eliminating any targeting of their corresponding dsDNA template loci.

We next explored whether Cas9-mediated RNA targeting could be applied for tagless transcript isolation from HeLa cells (**Fig. 6.8d**). To immobilize Cas9 on a solid-phase resin, we mutagenized Cas9 to remove both wild-type cysteine residues and introduced a unique cysteine at the N-terminus distal from any nucleic acid binding surfaces. Chemical labeling of purified Cas9 at this position with a biotin moiety was specific and robust, and the resulting biotin-Cas9 protein was fully active and could be quantitatively retained by magnetic streptavidin beads (**Fig. 6.9**).

As a proof of concept, we first isolated *GAPDH* mRNA from HeLa total RNA using biotinylated dCas9, gRNAs and PAMmers that target four non-PAM-adjacent sequences within exons 5–7 (**Fig. 6.8e**). We observed a substantial enrichment of *GAPDH* mRNA relative to a control *β-actin* mRNA by Northern blot analysis, but saw no enrichment using a non-targeting gRNA or dCas9 alone (**Fig. 6.8f**).

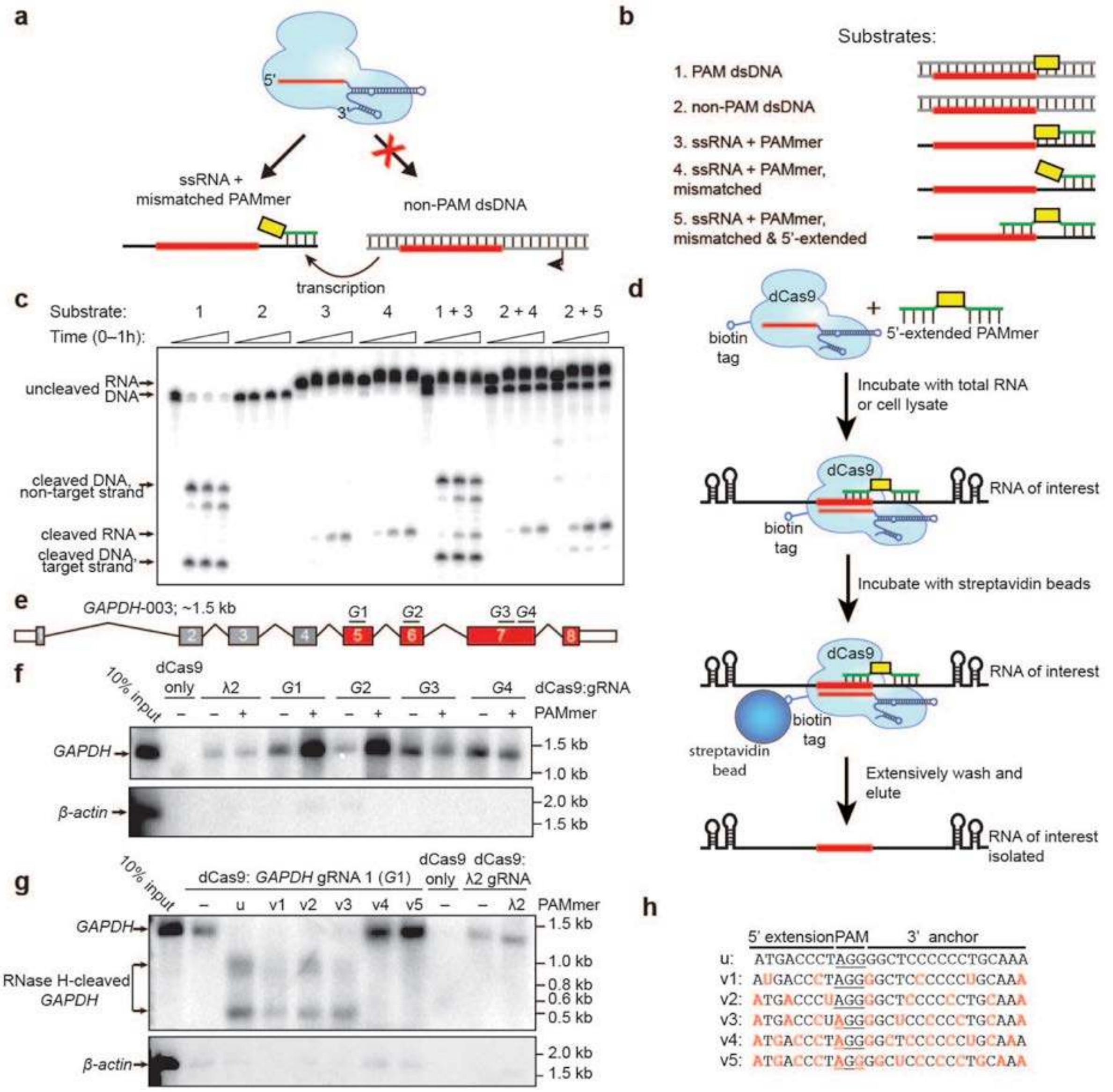
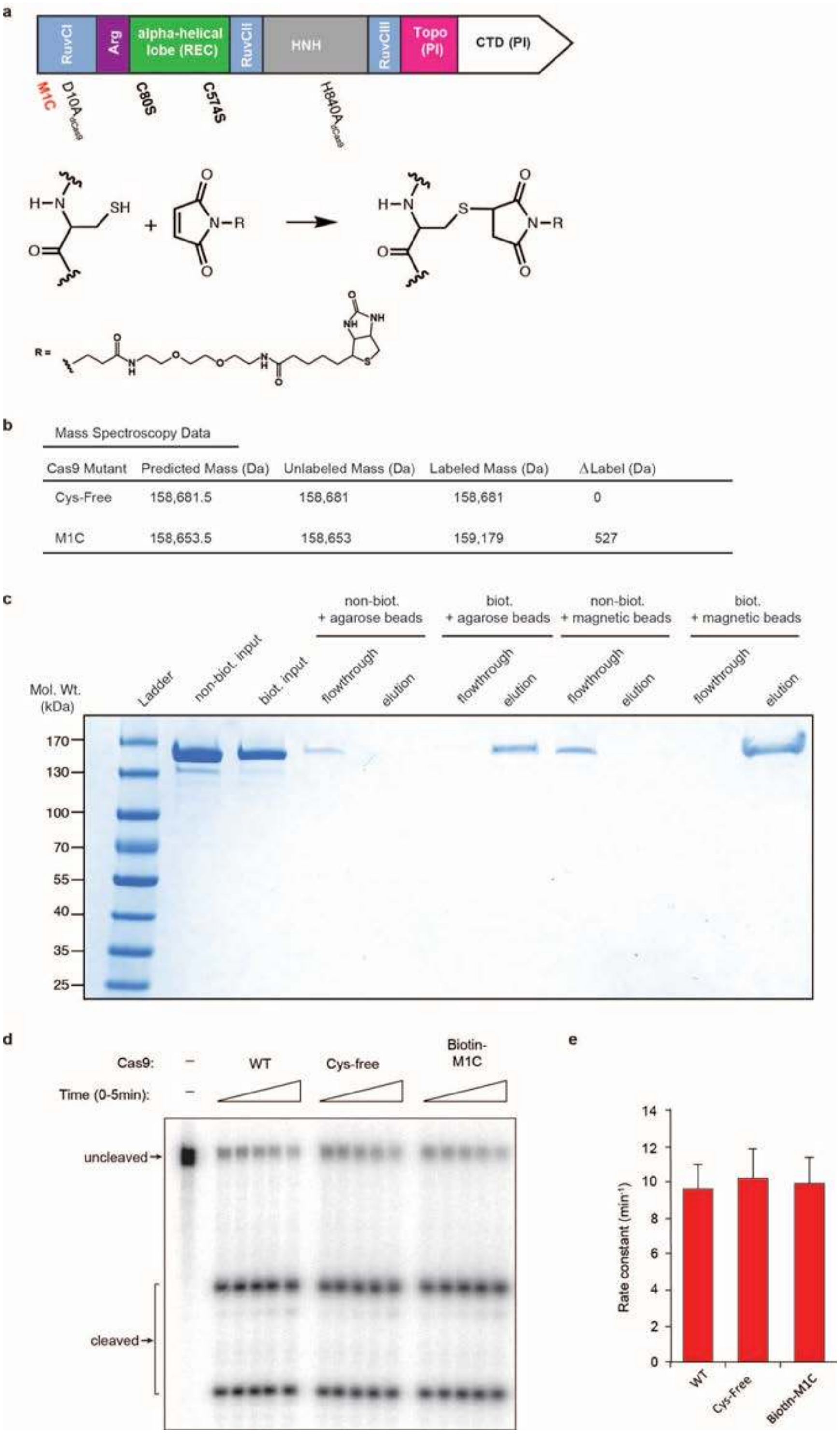

a
5'
3'
ssRNA +
mismatched PAMmer
non-PAM dsDNA
transcription
b
Substrates:
1. PAM dsDNA
2. non-PAM dsDNA
3. ssRNA + PAMmer
4. ssRNA + PAMmer, mismatched
5. ssRNA + PAMmer, mismatched & 5'-extended
c
Substrate: 1 2 3 4 1 + 3 2 + 4 2 + 5
Time (0–1h):
uncleaved RNA
uncleaved DNA
cleaved DNA, non-target strand
cleaved RNA
cleaved DNA, target strand
d
dCas9
biotin tag
5'-extended PAMmer
Incubate with total RNA or cell lysate
RNA of interest
Incubate with streptavidin beads
streptavidin bead
Extensively wash and elute
RNA of interest isolated
e
GAPDH-003; ~1.5 kb
G1 G2 G3 G4
2 3 4 5 6 7 8
f
10% input
dCas9 only
λ2 G1 G2 G3 G4
dCas9:gRNA
– – + – + – + – + – + PAMmer
GAPDH
β-actin
1.5 kb
1.0 kb
2.0 kb
1.5 kb
g
10% input
dCas9: GAPDH gRNA 1 (G1)
dCas9 only
dCas9: λ2 gRNA
– u v1 v2 v3 v4 v5 – – λ2 PAMmer
GAPDH
RNase H-cleaved GAPDH
β-actin
1.5 kb
1.0 kb
0.8 kb
0.6 kb
0.5 kb
2.0 kb
1.5 kb
h
5' extension PAM 3' anchor
u: ATGACCCTAGGGGCTCCCCCCTGCAAA
v1: AUGACCCTAGGGGCTCCCCCCUGCAAA
v2: ATGACCCUAGGGGCTCCCCCCTGCAAA
v3: ATGACCCUAGGGGCUCCCCCCTGCAAA
v4: ATGACCCTAGGGGCTCCCCCCUGCAAA
v5: ATGACCCTAGGGGCUCCCCCCTGCAAA

Figure 6.8 | RNA-guided Cas9 can target non-PAM sites on ssRNA and isolate *GAPDH* mRNA from HeLa cells in a tagless manner. **(a)** Schematic of the approach designed to avoid cleavage of template DNA by targeting non-PAM sites in the ssRNA target. **(b)** The panel of nucleic acid substrates tested in (c). **(c)** Cas9–gRNA cleaves ssRNA targets with equal efficiency when the 5'-NGG-3' of the PAMmer is mismatched with the ssRNA. This strategy enables selective cleavage of ssRNA in the presence of non-PAM target dsDNA; at cognate PAM sites, Cas9–gRNA cleaves both ssRNA and dsDNA. **(d)** Schematic of the dCas9 RNA pull-down expriment. **(e),** *GAPDH* mRNA transcript isoform 3 shown schematically, with exons common to all *GAPDH* protein-coding transcripts in red and gRNA/PAMmer targets 1-4 indicated. **(f)** Northern blot showing that gRNAs and 5'-extended PAMmers enable tagless isolation of *GAPDH* mRNA from HeLa total RNA; *β-actin* mRNA is shown as a control. **(g)** Northern blot showing tagless isolation of *GAPDH* mRNA from HeLa cell lysate with varying 2'-OMe-modified PAMmers. RNase H cleavage is abrogated with v4 and v5 PAMmers; *β-actin* mRNA is shown as a control. **(h)** Sequences of unmodified and modified *GAPDH* PAMmers used in (g); 2'-OMe-modified nucleotides are shown in red.



Mass Spectroscopy Data

Cas9 Mutant	Predicted Mass (Da)	Unlabeled Mass (Da)	Labeled Mass (Da)	ΔLabel (Da)
Cys-Free	158,681.5	158,681	158,681	0
M1C	158,653.5	158,653	159,179	527

Figure 6.9 | Site-specific biotin labeling of Cas9. (a) In order to introduce a single biotin moiety on Cas9, the solvent accessible, non-conserved N- terminal methionine was mutated to a cysteine (M1C; red text) and the naturally occurring cysteine residues were mutated to serine (C80S and C57S; bold text). This enabled cysteine-specific labeling with EZ-link Maleimide-PEG2-biotin through an irreversible reaction between the reduced sulfhydryl group of the cysteine and the maleimide group present on the biotin label. dCas9 mutations are also indicated in the domain schematic. **(b)** Mass spectrometry analysis of the Cas9 biotin labeling reaction confirmed

that successful biotin labeling only occurs when the M1C mutation is present in the Cys-Free background (C80S,C574S). The mass of the Maleimide- PEG2-biotin reagent is 525.6 Da. **(c)** Streptavidin bead binding assay with biotinylated (biot.) or non-biotinylated (non-biot.) Cas9 and streptavidin agarose or streptavidin magnetic beads. Cas9 only remains specifically bound to the beads after biotin labeling. **(d)** Cleavage assays were conducted as described in the **Materials and Methods** and resolved by denaturing PAGE. Reactions contained 100 nM Cas9 programmed with λ2-gRNA and ~1 nM 5′-radiolabelled λ2 dsDNA target. **(e)** Quantified cleavage data from triplicate experiments were fit with single-exponential decays to calculate the apparent pseudo-first order cleavage rate constants (average ± standard deviation). Both Cys-Free and Biotin-M1C Cas9 retain WT activity.

We then used this approach to isolate endogenous *GAPDH* transcripts from HeLa cell lysate under physiological conditions. In initial experiments, we found that Cas9–gRNA captured two *GAPDH*-specific RNA fragments rather than the full-length mRNA (**Fig. 6.8g**). Based on the sizes of these bands, we hypothesized that RNA:DNA heteroduplexes formed between the mRNA and PAMmer were cleaved by cellular RNase H. Previous studies have shown that modified DNA oligonucleotides can abrogate RNase H activity (Wu et al., 1999), and therefore we investigated whether Cas9 would tolerate chemical modifications to the PAMmer. We found that a wide range of modifications still enabled PAMmer-mediated nuclease activation, including locked nucleic acids, 2'-OMe and 2'-F ribose moieties (**Fig. 6.10**). Importantly, by varying the pattern of 2'-OMe modifications in the PAMmer, we could completely eliminate RNase H-mediated cleavage during the pull-down experiment and successfully isolate intact *GAPDH* mRNA (**Fig. 6.8g,h**). Interestingly, we consistently observed specific isolation of *GAPDH* mRNA in the absence of any PAMmer, albeit with lower efficiency, suggesting that Cas9–gRNA can bind to *GAPDH* mRNA through direct RNA:RNA hybridization (**Fig. 6.8f,g & 6.11**). Taken together, these experiments demonstrate that RNA-guided Cas9 can be used to purify endogenous untagged RNA transcripts. In contrast to current oligonucleotide-mediated RNA-capture methods, this approach works well under physiological salt conditions and doesn't require crosslinking or large sets of biotinylated probes (Chu et al., 2011; Engreitz et al., 2013; Simon et al., 2011).

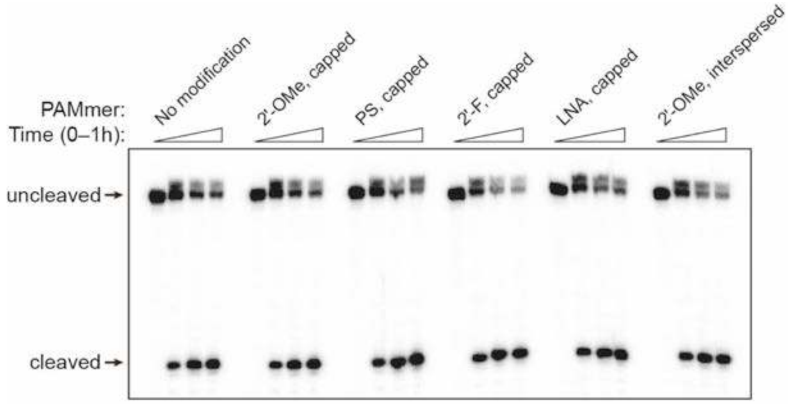


Figure 6.10 | RNA-guided Cas9 can utilize chemically modified PAMmers. 19-nt PAMmer derivatives containing various chemical modifications on the 5' and 3' ends (capped) or interspersed still activate Cas9 for cleavage of ssRNA targets. These types of modification are often used to increase the in vivo half-life of short oligonucleotides by preventing exo- and endonuclease-mediated degradation. Cleavage assays were conducted as described in the Methods. PS, phosphorothioate bonds; LNA, locked nucleic acid.

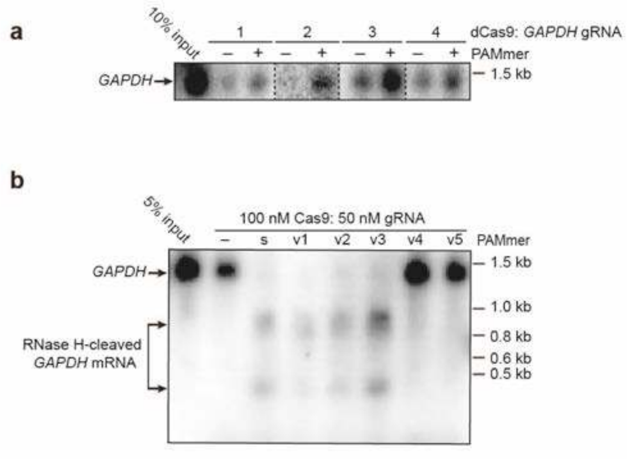


Figure 6.11 | Cas9 programmed with GAPDH-specific gRNAs can pull-down GAPDH mRNA in the absence of PAMmer. (a) Northern blot showing that, in some cases, Cas9-gRNA is able to pull down detectable amounts of GAPDH mRNA from total RNA without requiring a PAMmer. **(b)** Northern blot showing that Cas9-gRNA 1 is also able to pull-down quantitative amounts of GAPDH mRNA from HeLa cell lysate without requiring a PAMmer. s: standard; v: 2'-OMe-modified PAMmers.

Table 6.1 | RNA and DNA substrates used in this study

Description	**Sequence[a]**
Oligo for preparing dsDNA T7 promoter, in vitro transcription	**5′-TAATACGACTCACTATA-3′**
λ2-targeting crRNA	**5′-GUGAUAAGUGGAAUGCCAUG**GUUUUAGAGCUAUGCUGUUUUG-3′
λ3-targeting crRNA	**5′-CUGGUGAACUUCCGAUAGUG**GUUUUAGAGCUAUGCUGUUUUG-3′
λ4-targeting crRNA	**5′-CAGATATAGCCTGGTGGTTC**GUUUUAGAGCUAUGCUGUUUUG-3′
ssDNA T7 template[b]: tracrRNA	5′-AAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATGCTGT**CCTATAGTGAGTCGTATTA**
tracrRNA (nt 15-87)	**GG**ACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUU
λ2-targeting sgRNA T7 template[c]	5′**TAATACGACTCACTATAGG**TGATAAGTGGAATGCCATGGTTTTAGAGCTATGCTGTTTTGGAAACAAAACAGCATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3′
λ2-targeting sgRNA	5′-**GG**UGAUAAGUGGAAUGCCAUGGUUUUAGAGCUAUGCUGUUUUGGAAACAAAACAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUU-3′
λ2 target dsDNA duplex	5'-GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATG**TGG**GCTGTCAAAATTGAGC-3' 3'-CTCACCTTCCTACGGT**CACTATTCACCTTACGGTAC**ACCCGACAGTTTTAACTCG-5'
λ2 ssDNA target strand (used to make heteroduplex DNA:RNA)	3'-CTCACCTTCCTACGGT**CACTATTCACCTTACGGTAC**ACCCGACAGTTTTAACTCG-5'

λ2 ssDNA non-target strand (used to make heteroduplex DNA:RNA)	5'-GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATG**TGG**GCTGTCAAAATTGAGC-3'
λ2 ssRNA target strand T7 template	5′-GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATGTGGGCTGTCAAAATTGAG**CCTATAGTGAGTCGTATTA**-3′
λ2 ssRNA target strand	3'-CUCACCUUCCUACGGU**CACUAUUCACCUUACGGUAC**ACCCGACAGUUUUAACUCG**G**-5'
λ2 ssRNA non-target strand T7 template	5′-GCTCAATTTTGACAGCCCACATGGCATTCCACTTATCACTGGCATCCTTCCACTC**CTATAGTGAGTCGTATTA**-3′
λ2 ssRNA non-target strand (used to make dsRNA)	**5′-G**GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATG**TGG**GCTGTCAAAATTGAGC-3′
19 nt λ2 DNA PAMmer	5'-**TGG**GCTGTCAAAATTGAGC-3'
18 nt λ2 "GG" PAMmer	5'-**GG**GCTGTCAAAATTGAGC-3'
19 nt λ2 DNA mutated PAMmer	5'-ACCGCTGTCAAAATTGAGC-3'
16 nt λ2 DNA "PAM-less" PAMmer	5'-GCTGTCAAAATTGAGC-3'
18 nt λ2 RNA PAMmer	5'-**GG**GCUGUCAAAAUUGAGC-3'
5 nt λ2 DNA PAMmer	5'-**TGG**GC-3'
10 nt λ2 DNA PAMmer	5'-**TGG**GCTGTCA-3'
15 nt λ2 DNA PAMmer	5'-**TGG**GCTGTCAAAATT-3'
λ3 ssRNA target strand T7 template	5′-AACGTGCTGCGGCTGGCTGGTGAACTTCCGATAGTGCGGGTGTTGAATGATTTCC**TATAGTGAGTCGTATTA**-3′
λ3 ssRNA target strand	3'-UUGCACGACGCCGACC**GACCACUUGAAGGCUAUCAC**GCCCACAACUUACUAAAGG-5'
λ4 ssRNA target strand T7 template	5′-TCACAACAATGAGTGGCAGATATAGCCTGGTGGTTCAGGCGGCGCATTTTTATTG**CCTATAGTGAGTCGTATTA**-3′
λ4 ssRNA target strand	3'-AGUGUUGUUACUCACC**GUCUAUAUCGGACCACCAAG**UCCGCCGCGUAAAAAUAAC**GG**-5'

λ3 ssDNA non-target strand	5'-AACGTGCTGCGGCTGGCTGGTGAACTTCCGATAGTG**CGG**GTGTTGAATGATTTCC-3'
λ4 ssDNA non-target strand	5'-TCACAACAATGAGTGGCAGATATAGCCTGGTGGTTC**AGG**CGGCGCATTTTTATTG-3'
19 nt λ3 DNA PAMmer	5′-**CGG**GTGTTGAATGATTTCC-3′
19 nt λ4 DNA PAMmer	5′-**AGG**CGGCGCATTTTTATTG-3′
21 nt λ2 5'-extended DNA PAMmer	5′-TG**TGG**GCTGTCAAAATTGAGC-3′
21 nt λ3 5'-extended DNA PAMmer	5′-TG**CGG**GTGTTGAATGATTTCC-3′
24 nt λ2 5'-extended DNA PAMmer	5′-CCATG**TGG**GCTGTCAAAATTGAGC-3′
24 nt λ3 5'-extended DNA PAMmer	5′-TAGTG**CGG**GTGTTGAATGATTTCC-3′
27 nt λ2 5'-extended DNA PAMmer	5′-ATGCCATG**TGG**GCTGTCAAAATTGAGC-3′
27 nt λ3 5'-extended DNA PAMmer	5′-CGATAGTG**CGG**GTGTTGAATGATTTCC-3′
30 nt λ2 5'-extended DNA PAMmer	5′-GGAATGCCATG**TGG**GCTGTCAAAATTGAGC-3′
30 nt λ3 5'-extended DNA PAMmer	5′-TTCCGATAGTG**CGG**GTGTTGAATGATTTCC-3′
33 nt λ2 5'-extended DNA PAMmer	5′-AGTGGAATGCCATG**TGG**GCTGTCAAAATTGAGC-3′
33 nt λ3 5'-extended DNA PAMmer	5′-AACTTCCGATAGTG**CGG**GTGTTGAATGATTTCC-3′
36 nt λ2 5'-extended DNA PAMmer	5′-ATAAGTGGAATGCCATG**TGG**GCTGTCAAAATTGAGC-3′
39 nt λ2 5'-extended DNA PAMmer	5′-GTGATAAGTGGAATGCCATG**TGG**GCTGTCAAAATTGAGC-3′

39 nt λ3 5′-extended DNA PAMmer	5′-CTGGTGAACTTCCGATAGTG**CGG**GTGTTGAATGATTTCC-3′
non-PAM λ2 dsDNA	5'-GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATGACCGCTGTCAAAATTGAGC-3' 3'-CTCACCTTCCTACGGT**CACTATTCACCTTACGGTAC**TGGCGACAGTTTTAACTCG-5'
non-PAM λ2 ssRNA target strand T7 template	5′-GAGTGGAAGGATGCCAGTGATAAGTGGAATGCCATGACCGCTGTCAAAATTGAG**CCTATAGTGAGTCGTATTA**-3′
non-PAM λ2 ssRNA target strand	3'-CUCACCUUCCUACGGU**CACUAUUCACCUUACGGUAC**TGGCGACAGUUUUAACUC**GG**-5'
λ2 2′OMe capped PAMmer[d]	5′-***UGG**GCTGTCAAAATTGAG*C-3′
λ2 PS capped PAMmer[d]	5′-**T*GG**GCTGTCAAAATTGAG*C-3′
λ2 2′F capped PAMmer[d]	5′-***UGG**GCTGTCAAAATTGAG*C-3′
λ2 LNA capped PAMmer[d]	5′-***TGG**GCTGTCAAAATTGAG*C-3′
λ2 19 nt 2′OMe interspersed PAMmer[d]	5′-***UGG**GC*UGTCA*AAATT*GAG*C-3′
GAPDH-targeting sgRNA 1 T7 template[e]	5′-**TAATACGACTCACTATA**GGGGCAGAGATGATGACCCTGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3′
GAPDH-targeting sgRNA 1	5′-GGGGCAGAGAUGAUGACCCUGUUUAAGAGCUAUGCUGGAAACAGCAUAGCAAGUUUAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUU-3′
GAPDH-targeting sgRNA 2 T7 template[e]	5′-**TAATACGACTCACTATAGG**CCAAAGTTGTCATGGATGACGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3′
GAPDH-targeting sgRNA 2	**5′-GG**CCAAAGUUGUCAUGGAUGACGUUUAAGAGCUAUGCUGGAAACAGCAUAGCAAGUUUAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUU-3′
GAPDH-targeting sgRNA 3 T7 template[e]	5′-**TAATACGACTCACTATAGG**CCAAAGTTGTCATGGATGACGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3′
GAPDH-targeting sgRNA 3	**5′-GG**AUGUCAUCAUAUUUGGCAGGGUUUAAGAGCUAUGCUGGAAACAGCAUAGCAAGUUUAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUU-3′
GAPDH-targeting sgRNA 4 T7 template[e]	5′-**TAATACGACTCACTATAGG**ATGTCATCATATTTGGCAGGGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3′
GAPDH-targeting sgRNA 4	5′-**GG**ATGTCATCATATTTGGCAGGGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT-3′
GAPDH PAMmer 1	5′-ATGACCCTTGGGGCTCCCCCCTGCAAA-3′
GAPDH PAMmer 2	5′-TGGATGACCGGGGCCAGGGGTGCTAAG-3′
GAPDH PAMmer 3	5′-TTGGCAGGTGGTTCTAGACGGCAGGTC-3′

GAPDH PAMmer 4	5'-CCCCAGCGTGGAAGGTGGAGGAGTGGG-3'
GAPDH PAMmer 1 2'OMe v1	5'-A*UGACC*CTAGG*GGCTC*CCCCC*UGCAA*A-3'
GAPDH PAMmer 1 2'OMe v2	5'-*ATG*ACCC*UAGG*GGCT*CCCC*CCTG*CAA*A-3'
GAPDH PAMmer 1 2'OMe v3	5'-*ATG*ACC*CU*AGG*GGC*UCC*CCC*CTG*CAA*A-3'
GAPDH PAMmer 1 2'OMe v4	5'-*AT*GA*CC*CT*AGG*GG*CT*CC*CC*CC*UG*CA*AA-3'
GAPDH PAMmer 1 2'OMe v5	5'-*AT*GA*CC*CT*AG*GG*GC*TC*CC*CC*CU*GC*AA*A-3'
GAPDH cDNA primer Fwd	5'-CTCACTGTTCTCTCCCTCCGC-3'
GAPDH cDNA primer Rev	5'-AGGGGTCTACATGGCAACTG-3'
β-actin cDNA primer Fwd	5'-AGAAAATCTGGCACCACACC-3'
β-actin cDNA primer Rev	5'-GGAGTACTTGCGCTCAGGAG-3'

[a] Guide crRNA sequences and complementary DNA target strand sequences are shown in red. PAM sites (5'-NGG-3') are highlighted in yellow on the non-target strand when adjacent to the target sequence or in the PAMmer oligonucleotides.

[b] The T7 promoter is indicated in **bold** (or reverse complement of), as well as 5' G or GG included in the ssRNA product by T7 polymerase.

NA, not applicable.

[c] sgRNA template obtained in pIDT, subsequently linearised by AflII for run-off transcription.

[d] Positions of modifications depicted with asterisks preceding each modified nucleotide in each case (except for PS linkages which are depicted between bases)

PS: phosphorothioate bond

LNA: locked nucleic acid

[e] sgRNAs for *GAPDH* were designed according to (Chen et al., 2013).

Here we have demonstrated the ability to re-direct the dsDNA targeting capability of CRISPR/Cas9 for RNA-guided ssRNA binding and/or cleavage (RCas9). Programmable RNA recognition and cleavage has the potential to transform the study of RNA function much as site-specific DNA targeting is changing the landscape of genetic and genomic research (Mali et al., 2013b) (**Fig. 6.12**). Although certain engineered proteins such as PPR proteins and Pumilio/FBF (PUF) repeats show promise as platforms for sequence-specific RNA targeting (Filipovska and Rackham, 2011; Mackay et al., 2011; Wang et al., 2013c; Yagi et al., 2014; Yin et al., 2013), these strategies suffer from the need to re-design the protein for every new RNA sequence of interest. While RNA interference has proven useful for manipulating gene regulation in certain organisms (Kim and Rossi, 2008), there has been a strong motivation to develop orthogonal nucleic acid-based RNA recognition systems, such as the CRISPR/Cas Type III-B Cmr complex (Hale et al., 2009; 2012; Spilman et al., 2013; Staals et al., 2013; Terns and Terns, 2014) and the atypical Cas9 from *Francisella novicida* (Sampson and Weiss, 2014; Sampson et al., 2013). In contrast to these systems, the molecular basis for RNA recognition by RCas9 is now clear and requires only the design and synthesis of a matching gRNA and complementary PAMmer. The ability to recognize endogenous RNAs within complex mixtures with high affinity and in a programmable manner paves the way for direct transcript detection, analysis and manipulation without the need for genetically encoded affinity tags.

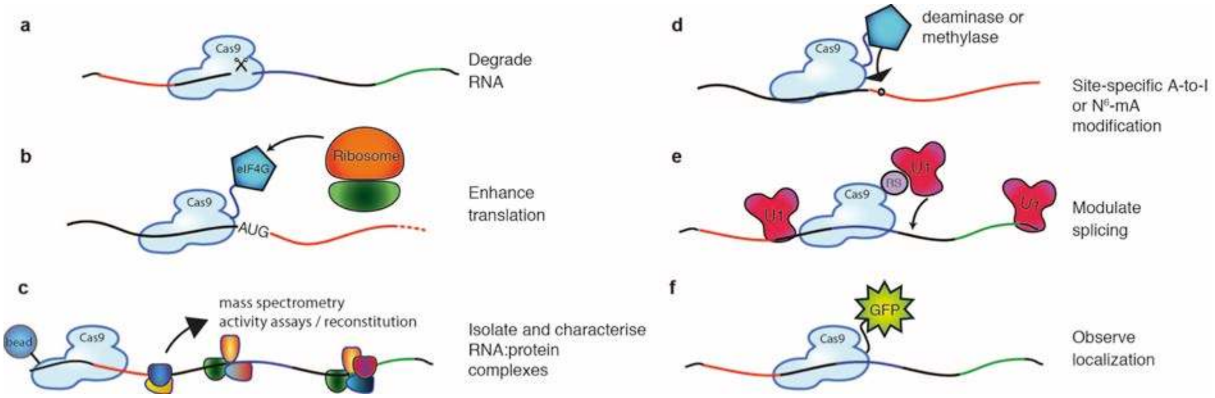


Figure 6.12 | Potential applications of RCas9 for untagged transcript analysis, detection, and manipulation. **(a)** Catalytically-active RCas9 could be used to target and cleave RNA, particularly those for which RNAi-mediated repression/degradation is not possible. **(b)** Tethering the eukaryotic initiation factor eIF4G to a catalytically inactive dRCas9 targeted to the 5' untranslated region of an mRNA could drive translation. **(c)** dRCas9 tethered to beads could be used to specifically isolate RNA or native RNA:protein complexes of interest from cells for downstream analysis or assays including identification of bound protein complexes, probing of RNA structure under native protein-bound conditions, and enrichment of rare transcripts for sequencing analysis. **(d)** dRCas9 tethered to RNA deaminase or N^6-mA methylase domains could direct site-specific A-to-I editing or methylation or RNA, respectively. **(e)** dRCas9 fused to a U1 recruitment domain (arginine- and serine- rich (RS) domain) could be programmed to recognize a splicing enhancer site and thereby promote the inclusion of a targeted exon. **(f)** dRCas9 tethered to a fluorescent protein such as GFP could be used to observe RNA localization and transport in living cells.

Chapter 7

Expanding the biologist's toolkit with CRISPR-Cas9

† Part of the work presented in this chapter has recently been submitted for publication in the following review article: Sternberg, S.H., Doudna, J.A. Expanding the biologist's toolkit with CRISPR-Cas9.

7.1 Abstract

Few discoveries transform a discipline overnight, but biologists today can manipulate cells in ways never possible before, thanks to a peculiar form of prokaryotic adaptive immunity mediated by CRISPRs (clustered regularly interspaced short palindromic repeats). From elegant studies that deciphered how these immune systems function in bacteria, researchers quickly uncovered the technological potential of Cas9, an RNA-guided DNA cleaving enzyme, for genome engineering. Here we highlight the recent explosion in visionary applications of CRISPR-Cas9 that promises to usher in a new era of biological understanding and control.

7.2 Introduction

It was only six years ago that a fledgling group of international scientists met at the University of California, Berkeley, for the first annual meeting on CRISPRs. A diverse range of expertise was represented—microbiology, biochemistry, metagenomics, food science—allowing the mystery of CRISPR immune system function to be unraveled collectively from multiple lines of experimentation. Each subsequent conference boasted more breakthrough discoveries, and the increasing rate of CRISPR-related publications reflected an intensifying interest in the topic. A description of the molecular function of Cas9 and suggestion of its use for genome engineering, presented at the 2012 meeting, foreshadowed an explosion of research using CRISPR-Cas9 that was soon to come.

Beginning in January 2013, a flurry of studies demonstrated that site-specific DNA editing in eukaryotic cells could be achieved through the heterologous expression of Cas9 together with a guide RNA. Two years and >500 publications later (**Fig. 7.1a**), the technology has gone viral. The genomes of virtually all model plants and animals have been modified with CRISPR-Cas9, and creative new tools continue to expand the capabilities of this system. While CRISPR biology remains an active area of study, the memorable acronym is now more commonly associated with genome engineering than it is with prokaryotic adaptive immunity.

In this perspective, we provide a concise summary of how the CRISPR-Cas9 technology emerged and is enabling remarkable innovations in the biological sciences. We encourage the reader to consult the recent literature for more comprehensive reviews (Doudna and Charpentier, 2014; Hsu et al., 2014; Mali et al., 2013b). Detailed protocols for the numerous applications involving CRISPR-Cas9 can be found in a recent volume of *Methods in Enzymology* (Doudna and Sontheimer, 2014). Finally, we apologize to our many colleagues whose work we could not mention or discuss due to length constraints.

7.3 "The biological significance of these sequences is not known"

So concluded a study published in 1987, in which the authors inadvertently discovered the first genomic CRISPR locus in *Escherichia coli* while sequencing the *iap* gene (Ishino et al., 1987). CRISPRs have since been found in roughly 40% and 90% of all bacterial and archaeal species, respectively (Grissa et al., 2007b), and are characterized by short direct repeats interrupted at regular intervals by unique spacer sequences. Yet it wasn't until 2005 that a potential connection between CRISPRs and antiviral immune defense was established, when multiple laboratories reported that spacers derive from foreign genetic elements (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). A landmark study in 2007 from Barrangou and colleagues provided the definitive link: working in *Steptococcus thermophilus*, the authors demonstrated that CRISPR spacers confer potent resistance to bacterial viruses (bacteriophage) bearing matching DNA sequences, and that bacteria could actively vaccinate themselves against bacteriophage by integrating new spacers into the pre-existing CRISPR locus (Barrangou et al., 2007). The central role of noncoding CRISPR RNA (crRNA) in this pathway was revealed shortly thereafter by pioneering work from the van der Oost laboratory (Brouns et al., 2008).

CRISPRs function together with CRISPR-associated (*cas*) genes that typically flank CRISPR loci in the genome, and the entire pathway is consequently referred to as CRISPR-Cas. Adaptive immunity proceeds in three stages: acquisition (or adaptation), CRISPR RNA biogenesis, and interference (for recent reviews see (Barrangou and Marraffini, 2014; van der Oost et al., 2014)). During acquisition, new spacers are selected from foreign nucleic acids and integrated at one end of the CRISPR locus. RNA precursors are then transcribed from the CRISPR locus and enzymatically processed into mature crRNAs, which are bound by one or more Cas proteins to form ribonucleoprotein targeting complexes that each contain a single spacer (guide) sequence. Finally, Cas nucleases cleave target nucleic acids that are marked for degradation via complementary base pairing to the crRNA.

CRISPR-Cas immune systems have been classified into three types and numerous subtypes based primarily on *cas* gene phylogeny (Makarova et al., 2011b), and while many mechanistic features are widely conserved, significant differences exist. For example, Type I and III systems require only crRNA for targeting, while Type II systems also use *trans*-activating CRISPR RNA (tracrRNA) (Deltcheva et al., 2011). In addition, the protein composition of crRNA-Cas targeting complexes is highly variable, with complexes in Type I and III systems typically comprising >8 subunits (Brouns et al., 2008; Hale et al., 2009). In contrast, Type II systems require only a single polypeptide: Cas9 (Sapranauskas et al., 2011).

7.4 RNA-guided DNA targeting by Cas9

Cas9 is a DNA endonuclease that generates double-strand breaks (DSBs) in DNA target sequences identified through base pairing to the guide RNA (**Fig. 7.1b**) (Gasiunas et al., 2012; Jinek et al., 2012). Cas9 functions naturally with a dual-guide RNA composed of crRNA and tracrRNA (Jinek et al., 2012); the 5' end of the crRNA base-pairs with target DNA, whereas the 3' end forms a double-stranded stem with the tracrRNA to facilitate Cas9 recruitment. Because DNA targets are recognized via RNA-DNA base pairing, changing the sequence of the guide RNA easily alters DNA specificity.

Efficient targeting also requires the presence of a short sequence motif proximal to the DNA target sequence, known as the protospacer adjacent motif (PAM) (Mojica et al., 2009). PAM recognition enables CRISPR-Cas immune systems to discriminate between self and non-self sequences, since only targets found in foreign DNA contain a PAM; matching targets in the CRISPR locus itself, from which the crRNA is transcribed, do not contain a PAM and are thereby avoided. Cas9 from *Streptococcus pyogenes*, which has been the focus of most studies to date, recognizes a 5'-NGG-3' PAM sequence (Jinek et al., 2012; Mojica et al., 2009).

The ability to edit genomic DNA inside cells has been limited by the dearth of effective tools to introduce site-specific DSBs. Earlier methods relied on protein-only systems such as zinc finger nucleases (ZFNs), and transcription activator-like effector nucleases (TALENs), but the feasibility of engineering these designer enzymes to recognize new sequences was limited. In contrast, the CRISPR-Cas system has the distinct advantage of relying on RNA for specificity. And while Cas9 shares molecular capabilities with other CRISPR-Cas systems, its compositional simplicity has been paramount to its successful application. Not only does it encompass only a single polypeptide, but remarkably, it retains full activity with a chimeric single-guide RNA (sgRNA), generated by connecting the 3' end of the crRNA to the 5' end of the tracrRNA (Jinek et al., 2012).

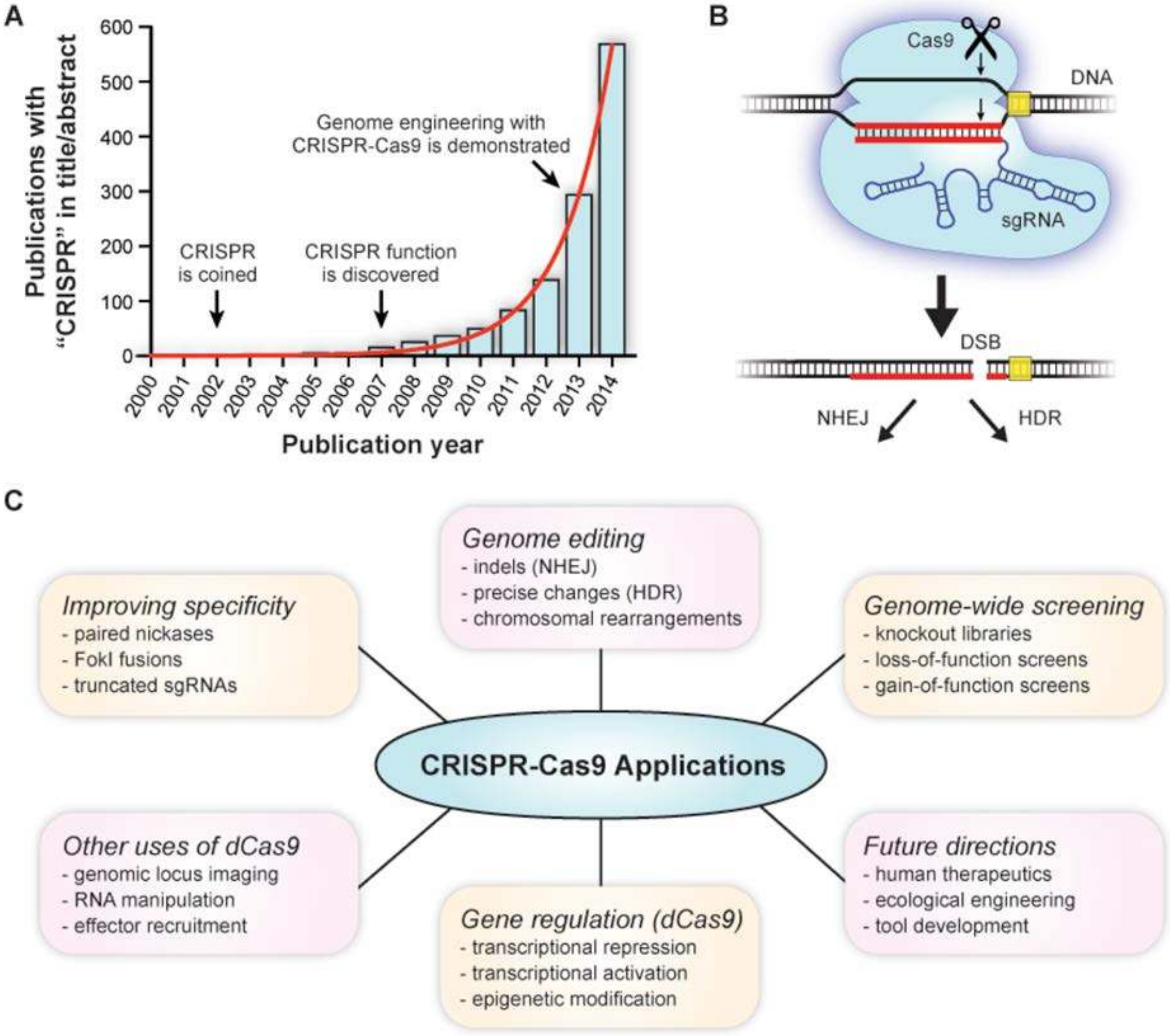


Figure 7.1 | Development of CRISPR-Cas9 for genome engineering. (a) The surge in CRISPR-Cas9 applications is highlighted by the exponential growth (red line) of publications with "CRISPR" in the title or abstract. Data were taken from PubMed. **(b)** Cas9 functions together with a single-guide RNA (sgRNA) to identify DNA target sequences adjacent to a PAM (yellow box) using RNA-DNA base pairing (red). Both strands of the target DNA are cleaved, generating a double-strand break (DSB) that is repaired by either non-homologous end joining (NHEJ) or homology-directed repair (HDR). **(c)** Programmable RNA-guided DNA targeting by Cas9 has been exploited in numerous diverse applications, some of which are shown.

7.5 Genome editing with CRISPR-Cas9

Six months after the first description of the molecular function of Cas9 (Jinek et al., 2012), six studies demonstrated that Cas9 together with sgRNA can be used to specifically edit the genomes of human cells (Cho et al., 2013; Cong et al., 2013; Jinek et al., 2013; Mali et al., 2013c), zebrafish embryos (Hwang et al., 2013), and bacteria (Jiang et al., 2013). The technology was rapidly extended to other model organisms, and by May of 2013, the Jaenisch laboratory reported the one-step generation of mice mutated at multiple alleles via zygotic injection of Cas9 mRNA and sgRNAs (Wang et al., 2013b). What had once been laborious and time-consuming was now facile and rapidly achievable.

In the simplest embodiment of the CRISPR-Cas9 technology (**Fig. 7.1c**), the sgRNA guide sequence is designed to target a complementary 20-base pair (bp) site flanked by NGG within the gene of interest. Heterologous expression of Cas9 together with sgRNA may be accomplished by stable lentiviral transfection, transient plasmid transfection, direct DNA or RNA injection, or transfection with purified ribonucleprotein complex, with the optimal strategy depending on the desired application and the cell or organism being edited. After trafficking to the nucleus through an appended nuclear localization signal, Cas9 targets the locus of interest and induces a DSB that is repaired by the cell's endogenous machinery via non-homologous end joining (NHEJ). Because NHEJ is an error-prone repair pathway that results in small insertions or deletions (indels), the open reading frame is disrupted and the gene becomes inactivated. While the editing efficiency with CRISPR-Cas9 can be as high 80% (Kim et al., 2014; Zuris et al., 2014), it is cell- and site-specific and depends on the delivery method. In addition, the resulting cell population will be inherently heterogeneous, both in the percentage of cells that were edited and in the specific genotype of the edited cells.

The genome can also be edited in a more precise manner using homology-directed repair (HDR). By combining Cas9–sgRNA delivery with a donor DNA that bears homology to sequences flanking the targeted site, DSBs are repaired using the donor DNA as a template. Importantly, this strategy enables new sequences to be introduced into the gene of interest, such as epitope tags, and specifically defined mutations to be installed, such as those that might mimic or correct disease-causing alleles. However, efficiencies of HDR are significantly lower than NHEJ, and more work is needed to develop strategies that bias the cell's natural DNA repair machinery towards the desired outcome.

Genome editing with CRISPR-Cas9 is now a routine procedure for virtually all model plants and animals, and recent progress has pushed the technology into ever more interesting directions. A unique advantage of CRISPR-Cas9 over earlier genome editing methods is that multiplexable targeting is easily achieved by co-expressing Cas9 with multiple sgRNAs simultaneously. In addition to editing multiple chromosomal loci in a single experiment with this approach, entire chromosomal deletions can be achieved by using two sgRNAs to induce DSBs at sites that flank the region of interest (Xiao et al., 2013). Furthermore, large-scale chromosomal rearrangements resembling those found in specific tumors can be introduced (Choi and Meyerson, 2014; Torres et al., 2014). Indeed, CRISPR-Cas9 offers great promise in transforming the tools available to recreate, model, and treat human cancers (Maddalo et al., 2014; Platt et al., 2014; Sánchez-Rivera et al., 2014).

7.6 Leveraging CRISPR-Cas9 as a versatile DNA-binding system

Cas9 generates DSBs using two conserved nuclease domains (HNH and RuvC) that cleave both strands of DNA target sequences. Inactivating both catalytic active sites via point mutations results in catalytically inactive Cas9 (dCas9), which remains fully active for programmable, RNA-guided DNA binding (Jinek et al., 2012). Numerous studies have taken advantage of this discovery to develop powerful new tools for regulating gene expression.

When expressed in bacteria, dCas9 together with sgRNA can sterically occlude RNA polymerase from binding promoter sequences and thereby down-regulate the expression of specific transcripts (Bikard et al., 2013; Qi et al., 2013). More robust gene expression control in eukaryotes becomes possible by fusing Cas9 to specific effector domains, such as transcriptional repressors and activators, which are recruited to specific genomic loci via the sgRNA (Gilbert et al., 2013; Maeder et al., 2013; Perez-Pinera et al., 2013). Recent developments continue to increase the efficiency and dynamic range of gene regulation by CRISPR-Cas9. In addition to improvements in the sgRNA design (Chen and Huang, 2014), newer methods involve recruiting multiple effector proteins through engineered molecular scaffolds fused to Cas9 (Gilbert et al., 2014; Tanenbaum et al., 2014) or RNA aptamers fused to sgRNA (Konermann et al., 2014; Mali et al., 2013a).

dCas9 can also be used to probe and manipulate the genome in other ways that ultimately rely on specific nucleic acid targeting. For example, DNA loci can be imaged in live cells using dCas9-GFP fusions (Chen et al., 2013), offering new insights into the dynamics and conformation of genomic loci. dCas9 fusions to effector domains that install epigenetic markers may enable specific perturbation of epigenetic regulation. Finally, a recent report demonstrated that dCas9 can bind single-stranded RNA targets (O'Connell et al., 2014), suggesting that programmable manipulation of cellular RNA transcripts may become possible in the near future.

7.7 High-throughput screening using CRISPR-Cas9

CRISPR-Cas9 enables facile, targeted perturbation of specific genes in the cell, either through permanent genome editing or temporary gene regulation. A systematic investigation of gene function, however, requires this level of control to be extended across the genome. A number of recent studies have harnessed the programmable nature of CRISPR-Cas9 to conduct powerful genome-scale screens. Importantly, this approach enables the hypothesis-free discovery of novel pathways that underlie a given biological process.

Using lentiviral sgRNA libraries and catalytically active Cas9, loss-of-function gene knockout screens were performed in both human and mouse cells (Koike-Yusa et al., 2014; Shalem et al., 2014; Wang et al., 2014a; Zhou et al., 2014). Deep sequencing of the sgRNA pool after either positive or negative selection revealed genes essential for cell viability, as well as genes involved in resistance to specific small-molecule drugs. While focused libraries will prove useful for targeted screens in which the candidate genes are selected by the researcher, genome-wide libraries that query all protein-coding genes will have a greater likelihood of discovering novel hits that were not previously identified (Koike-Yusa et al., 2014; Shalem et al., 2014).

Catalytically inactive dCas9 has also been co-opted for genome-scale screening by directly up- or down-regulating gene expression (Gilbert et al., 2014; Konermann et al., 2014). In comparison to the indels generated by active Cas9, which may be insufficient to inactivate non-coding RNAs or disrupt a given open reading frame, transcriptional silencing by dCas9 can more effectively block gene expression in some contexts. However, the ability to perform gain-of-function screens using dCas9-mediated recruitment of transcriptional activators is arguably the most significant advantage of this approach over active Cas9, and has not been possible with earlier technologies.

7.8 Off-target effects of CRISPR-Cas9

The success of any genome engineering technique, either as a basic research tool or in therapeutic applications, will ultimately be limited by its specificity. Early reports warned that CRISPR-Cas9 causes frequent off-target editing events (Fu et al., 2013), leading many laboratories to analyze cleavage specificity more thoroughly using different approaches (Hsu et al., 2013; Mali et al., 2013a; Pattanayak et al., 2013). While positions throughout the 20-bp target sequence affect specificity, mismatches encountered proximal to the PAM, within a seed sequence of ~8-12 nucleotides, have the largest impact on cleavage accuracy. These findings are consistent with the mechanism of DNA interrogation by Cas9, in which the duplex is unwound beginning at the PAM, in a directional manner that depends on RNA-DNA complementarity (Sternberg et al., 2014; Szczelkun et al., 2014).

A number of strategies have been developed that reduce off-target effects. Requiring two independent Cas9 binding events for genome editing effectively increases the length of DNA being recognized and has been accomplished in two ways. First, a nickase variant of Cas9 can be used, in which only one active site is mutated (Mali et al., 2013a; Ran et al., 2013). Pairs of sgRNAs then direct Cas9 to nick two closely spaced target sites to mimic a DSB; off-target nicking events with just a single sgRNA are precisely repaired without indel formation. Second, dCas9-FokI fusions can be used, similarly to ZFNs and TALENs (Guilinger et al., 2014; Tsai et al., 2014). Pairs of sgRNAs direct dCas9-FokI to two adjacent target sites, and a DSB is generated upon FokI dimerization; off-target binding events with a Cas9-FokI monomer do not result in cleavage. Finally, truncated sgRNAs have been shown to reduce off-target cleavage events without largely affecting on-target editing efficiencies (Fu et al., 2014). While most specificity studies to date have restricted their analysis to predicted off-target sites, two recent reports applied unbiased, whole-genome sequencing to carefully assess the incidence of off-target mutations (Smith et al., 2014; Veres et al., 2014).

Recent ChIP-seq studies have revealed that DNA binding by Cas9 is far more promiscuous than DNA cleavage (Cencic et al., 2014; Duan et al., 2014; Kuscu et al., 2014; Wu et al., 2014). The relevance of these findings for gene regulation applications involving dCas9 remains unclear, since off-target binding events may be too transient to affect transcription. Notably, a careful analysis of dCas9-mediated transcriptional repression found minimal off-target activity from properly designed sgRNAs (Gilbert et al., 2014). Nevertheless, *in vitro* experiments confirm that off-target DNAs with mismatches distal from the cleavage site can be tightly bound but not cleaved (Sternberg et al., 2014). The molecular cues that regulate catalytic activity have yet to be fully determined.

7.9 Future directions of CRISPR-Cas9 technologies

Cas9 holds great promise as a therapeutic strategy to treat human genetic diseases, as evidenced by the recent emergence of numerous companies dedicated to this cause. In proof-of-concept experiments, a disease-causing *Fah* mutation was successfully corrected in adult mice by hydrodynamic injection of a donor DNA template and plasmid DNA encoding Cas9 and sgRNA (Yin et al., 2014), and Duchenne muscular dystrophy was prevented by direct injection of Cas9 mRNA, sgRNA, and donor DNA template into the mouse germline (Long et al., 2014). Significant hurdles exist before similar experiments can be performed on human patients, but successes in ZFN-based human clinical trials demonstrate the exciting potential of general approaches using programmable nucleases (Tebas et al., 2014).

A number of recent studies have highlighted the ability of CRISPR-Cas9 to specifically alter ecological populations. Within microbial communities, Cas9 was packaged in bacteriophage and programmed to selectively kill virulent bacteria by targeting virulence genes, while leaving other bacteria unaffected (Bikard et al., 2014; Citorik et al., 2014). In animal populations that undergo sexual reproduction, Cas9-based gene drives could be used to rapidly spread altered traits and control invasive species (Esvelt et al., 2014). Finally, CRISPR-Cas9 can be used to genetically improve major staple crops such as bread wheat (Wang et al., 2014b). Many of these applications will require renewed attention to existing and future regulatory challenges (Oye et al., 2014; Voytas and Gao, 2014).

Finally, there is still significant room for basic tool development in the CRISPR-Cas9 technology space. Recent high-resolution structures of Cas9 in both unbound and DNA-bound states have been particularly insightful for the rational design of new Cas9–sgRNA variants and will surely inform future protein engineering efforts (Anders et al., 2014; Jinek et al., 2014; Nishimasu et al., 2014). Advances in our biochemical understanding of DNA recognition by CRISPR-Cas9 have inspired strategies to target new nucleic acids substrates such as single-stranded RNA (O'Connell et al., 2014). And an exciting avenue of future research will be the characterization and application of naturally occurring Cas9 homologs for genome engineering beyond those already described (Esvelt et al., 2013; Hou et al., 2013). For example, smaller variants may be more easily delivered with viral vectors, and orthogonal sgRNA and PAM specificities will enable a wider range of multiplexable outputs, including the simultaneous up- and down-regulation of gene expression.

7.10 Conclusions

The remarkable speed at which the CRISPR-Cas9 technology has spread throughout the biological community attests to its substantial impact in transforming our ability to manipulate cells (**Fig. 7.1c**). Genome engineering with Cas9 and sgRNA has become so routine that soon, the CRISPR-Cas9 method for editing chromosomal sites in model organisms will require no more attention in research articles than that accorded to PCR and molecular cloning. Indeed, the ease with which this technology can be practiced, and its tremendous utility, suggests that CRISPR-Cas9 will increasingly become a tool of choice for the next generation of biologists.

BIBLIOGRAPHY

Abbondanzieri, E.A., Greenleaf, W.J., Shaevitz, J.W., Landick, R., and Block, S.M. (2005). Direct observation of base-pair stepping by RNA polymerase. Nature *438*, 460–465.

Adams, P.D., Afonine, P.V., Bunkóczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.-W., Kapral, G.J., Grosse-Kunstleve, R.W., et al. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr *66*, 213–221.

Afonine, P.V., Grosse-Kunstleve, R.W., Echols, N., Headd, J.J., Moriarty, N.W., Mustyakimov, M., Terwilliger, T.C., Urzhumtsev, A., Zwart, P.H., and Adams, P.D. (2012). Towards automated crystallographic structure refinement with phenix.refine. Acta Crystallogr D Biol Crystallogr *68*, 352–367.

Aguilera, A. (2002). The connection between transcription and genomic instability. Embo J *21*, 195–201.

Al-Attar, S., Westra, E.R., van der Oost, J., and Brouns, S.J.J. (2011). Clustered regularly interspaced short palindromic repeats (CRISPRs): the hallmark of an ingenious antiviral defense mechanism in prokaryotes. Biological Chemistry *392*, 277–289.

Anders, C., Niewoehner, O., Duerst, A., and Jinek, M. (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature.

Andersson, A.F., and Banfield, J.F. (2008). Virus Population Dynamics and Acquired Virus Resistance in Natural Microbial Communities. Science *320*, 1047–1050.

Aravin, A.A., Naumova, N.M., Tulin, A.V., Vagin, V.V., Rozovsky, Y.M., and Gvozdev, V.A. (2001). Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. Curr Biol *11*, 1017–1027.

Aravin, A.A., Hannon, G.J., and Brennecke, J. (2007). The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science *318*, 761–764.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. (2010). ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res *38*, W529–W533.

Auweter, S.D., Oberstrass, F.C., and Allain, F.H.-T. (2006). Sequence-specific binding of single-stranded RNA: is there a code for recognition? Nucleic Acids Res *34*, 4943–4959.

Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B., Gagarinova, A., Pogoutse, O., Brown, G., Binkowski, A., et al. (2011). A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. Mol Microbiol *79*, 484–502.

Baker, N.A., Sept, D., Joseph, S., Holst, M.J., and McCammon, J.A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci USA *98*, 10037–10041.

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

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

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281–297.

Bassett, A.R., Tibbit, C., Ponting, C.P., and Liu, J.-L. (2013). Highly Efficient Targeted Mutagenesis of Drosophila with the CRISPR/Cas9 System. Cell Rep *4*, 220–228.

Beloglazova, N., Petit, P., Flick, R., Brown, G., Savchenko, A., and Yakunin, A.F. (2011). Structure and activity of the Cas3 HD nuclease MJ0384, an effector enzyme of the CRISPR interference. Embo J *30*, 4616–4627.

Berger, J.M., Gamblin, S.J., Harrison, S.C., and Wang, J.C. (1996). Structure and mechanism of DNA topoisomerase II. Nature *379*, 225–232.

Bikard, D., Euler, C.W., Jiang, W., Nussenzweig, P.M., Goldberg, G.W., Duportet, X., Fischetti, V.A., and Marraffini, L.A. (2014). Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. Nat Biotechnol *32*, 1146–1150.

Bikard, D., Jiang, W., Samai, P., Hochschild, A., Zhang, F., and Marraffini, L.A. (2013). Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. Nucleic Acids Res *41*, 7429–7437.

Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., and Hugenholtz, P. (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics *8*, 209.

Bolotin, A., Quinquis, B., Sorokin, A., and Ehrlich, S.D. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology (Reading, Engl) *151*, 2551–2561.

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.

Busso, D., Delagoutte-Busso, B., and Moras, D. (2005). Construction of a set Gateway-based destination vectors for high-throughput cloning and expression screening in Escherichia coli. Anal Biochem *343*, 313–321.

Cady, K.C., and O'Toole, G.A. (2011). Non-identity-mediated CRISPR-bacteriophage interaction mediated via the Csy and Cas3 proteins. J Bacteriol *193*, 3433–3445.

Calnan, B., Tidor, B., Biancalana, S., Hudson, D., and Frankel, A. (1991). Arginine-mediated RNA recognition: the arginine fork. Science *252*, 1167–1171.

Carte, J., Pfister, N.T., Compton, M.M., Terns, R.M., and Terns, M.P. (2010). Binding and cleavage of CRISPR RNA by Cas6. Rna *16*, 2181–2188.

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.

Cencic, R., Miura, H., Malina, A., Robert, F., Ethier, S., Schmeing, T.M., Dostie, J., and Pelletier, J. (2014). Protospacer adjacent motif (PAM)-distal sequences engage CRISPR Cas9 DNA target cleavage. PLoS ONE *9*, e109213.

Chacón, P., and Wriggers, W. (2002). Multi-resolution contour-based fitting of macromolecular structures. J Mol Biol *317*, 375–384.

Chen, B., and Huang, B. (2014). Imaging Genomic Elements in Living Cells Using CRISPR/Cas9. Meth Enzymol *546*, 337–354.

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, C.-S., Korobkova, E., Chen, H., Zhu, J., Jian, X., Tao, S.-C., He, C., and Zhu, H. (2008). A proteome chip approach reveals new DNA damage recognition activities in Escherichia coli. Nat Meth *5*, 69–74.

Chen, V.B., Arendall, W.B., Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr *66*, 12–21.

Chen, V.B., Davis, I.W., and Richardson, D.C. (2009). KING (Kinemage, Next Generation): a versatile interactive molecular and scientific visualization program. Protein Sci *18*, 2403–2409.

Cho, S.W., Kim, S., Kim, J.M., and Kim, J.-S. (2013). Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. Nat Biotechnol *31*, 230–232.

Choi, P.S., and Meyerson, M. (2014). Targeted genomic rearrangements using CRISPR/Cas technology. Nat Commun *5*, 3728.

Chomczynski, P. (1992). One-hour downward alkaline capillary transfer for blotting of DNA and RNA. Anal Biochem *201*, 134–139.

Chu, C., Qu, K., Zhong, F.L., Artandi, S.E., and Chang, H.Y. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Mol Cell *44*, 667–678.

Chylinski, K., Le Rhun, A., and Charpentier, E. (2013). The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. RNA Biol *10*, 726–737.

Cilley, C.D., and Williamson, J.R. (1997). Analysis of bacteriophage N protein and peptide binding to boxB RNA using polyacrylamide gel coelectrophoresis (PACE). Rna *3*, 57–67.

Citorik, R.J., Mimee, M., and Lu, T.K. (2014). Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. Nat Biotechnol *32*, 1141–1145.

Cochrane, J.C., and Strobel, S.A. (2008). Catalytic strategies of self-cleaving ribozymes. Acc Chem Res *41*, 1027–1035.

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.

Crenshaw, C.M., Nam, K., Oo, K., Kutchukian, P.S., Bowman, B.R., Karplus, M., and Verdine, G.L. (2012). Enforced presentation of an extrahelical guanine to the lesion recognition pocket of human 8-oxoguanine glycosylase, hOGG1. J Biol Chem *287*, 24916–24928.

Davis, I.W., Leaver-Fay, A., Chen, V.B., Block, J.N., Kapral, G.J., Wang, X., Murray, L.W., Arendall, W.B., Snoeyink, J., Richardson, J.S., et al. (2007). MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res *35*, W375–W383.

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

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

Deveau, H., Garneau, J.E., and Moineau, S. (2010). CRISPR/Cas System and Its Role in Phage-Bacteria Interactions. Annu Rev Microbiol.

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

Doudna, J.A., and Sontheimer, E.J. (2014). Preface. Meth Enzymol *546*, xix–xx.

Draper, D.E. (1995). Protein-RNA recognition. Annu Rev Biochem *64*, 593–620.

Dsouza, M., Larsen, N., and Overbeek, R. (1997). Searching for patterns in genomic data. Trends Genet *13*, 497–498.

Duan, J., Lu, G., Xie, Z., Lou, M., Luo, J., Guo, L., and Zhang, Y. (2014). Genome-wide identification of CRISPR/Cas9 off-targets in human genome. Cell Res *24*, 1009–1012.

Ebihara, A., Yao, M., Masui, R., Tanaka, I., Yokoyama, S., and Kuramitsu, S. (2006). Crystal structure of hypothetical protein TTHB192 from Thermus thermophilus HB8 reveals a new protein family with an RNA recognition motif-like domain. Protein Sci *15*, 1494–1499.

Edgar, R.C. (2007). PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics *8*, 18.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr *60*, 2126–2132.

Engreitz, J.M., Pandya-Jones, A., McDonel, P., Shishkin, A., Sirokman, K., Surka, C., Kadri, S., Xing, J., Goren, A., Lander, E.S., et al. (2013). The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. Science *341*, 1237973.

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.

Esvelt, K.M., Smidler, A.L., Catteruccia, F., and Church, G.M. (2014). Concerning RNA-guided gene drives for the alteration of wild populations. Elife *3*, e03401.

Fazio, T., Visnapuu, M.-L., Wind, S., and Greene, E.C. (2008). DNA curtains and nanoscale curtain rods: high-throughput tools for single molecule imaging. Langmuir *24*, 10524–10531.

Fersht, A.R. (1987). The hydrogen bond in molecular recognition. Trends Biochem Sci *12*, 301–304.

Filipovska, A., and Rackham, O. (2011). Designer RNA-binding proteins: New tools for manipulating the transcriptome. RNA Biol *8*, 978–983.

Fonfara, I., Le Rhun, A., Chylinski, K., Makarova, K.S., Lécrivain, A.-L., Bzdrenga, J., Koonin, E.V., and Charpentier, E. (2013). Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. Nucleic Acids Res.

Frank, J., Radermacher, M., Penczek, P., Zhu, J., Li, Y., Ladjadj, M., and Leith, A. (1996). SPIDER and WEB: processing and visualization of images in 3D electron microscopy and related fields. J. Struct. Biol. *116*, 190–199.

Friedland, A.E., Tzur, Y.B., Esvelt, K.M., Colaiácovo, M.P., Church, G.M., and Calarco, J.A. (2013). Heritable genome editing in C. elegans via a CRISPR-Cas9 system. Nat Meth *10*, 741–743.

Fu, Y., Foden, J.A., Khayter, C., Maeder, M.L., Reyon, D., Joung, J.K., and Sander, J.D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol *31*, 822–826.

Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M., and Joung, J.K. (2014). Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nat Biotechnol *32*, 279–284.

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.

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

Gesner, E.M., Schellenberg, M.J., Garside, E.L., George, M.M., and MacMillan, A.M. (2011). Recognition and maturation of effector RNAs in a CRISPR interference pathway. Nat Struct Mol Biol *18*, 688–692.

Gherghe, C.M., Mortimer, S.A., Krahn, J.M., Thompson, N.L., and Weeks, K.M. (2008). Slow conformational dynamics at C2'-endo nucleotides in RNA. J Am Chem Soc *130*, 8884–8885.

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.

Gorman, J., Fazio, T., Wang, F., Wind, S., and Greene, E.C. (2010a). Nanofabricated racks of aligned and anchored DNA substrates for single-molecule imaging. Langmuir *26*, 1372–1379.

Gorman, J., Plys, A.J., Visnapuu, M.-L., Alani, E., and Greene, E.C. (2010b). Visualizing one-dimensional diffusion of eukaryotic DNA repair factors along a chromatin lattice. Nat Struct Mol Biol *17*, 932–938.

Gorman, J., Wang, F., Redding, S., Plys, A.J., Fazio, T., Wind, S., Alani, E.E., and Greene, E.C. (2012). Single-molecule imaging reveals target-search mechanisms during DNA mismatch repair. Proc Natl Acad Sci USA *109*, E3074–E3083.

Górecka, K.M., Komorowska, W., and Nowotny, M. (2013). Crystal structure of RuvC resolvase in complex with Holliday junction substrate. Nucleic Acids Res *41*, 9945–9955.

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

Grissa, I., Vergnaud, G., and Pourcel, C. (2007a). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res *35*, W52–W57.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007b). The CRISPRdb database and tools to display

CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics *8*, 172.

Groenen, P.M., Bunschoten, A.E., van Soolingen, D., and van Embden, J.D. (1993). Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method. Mol Microbiol *10*, 1057–1065.

Gudbergsdottir, S., Deng, L., Chen, Z., Jensen, J.V.K., Jensen, L.R., She, Q., and Garrett, R.A. (2011). Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. Mol Microbiol *79*, 35–49.

Guilinger, J.P., Thompson, D.B., and Liu, D.R. (2014). Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. Nat Biotechnol *32*, 577–582.

Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835–840.

Haft, D.H., Selengut, J., Mongodin, E.F., and Nelson, K.E. (2005). A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput Biol *1*, e60.

Hale, C.R., Majumdar, S., Elmore, J., Pfister, N., Compton, M., Olson, S., Resch, A.M., Glover, C.V.C., Graveley, B.R., Terns, R.M., et al. (2012). Essential Features and Rational Design of CRISPR RNAs that Function with the Cas RAMP Module Complex to Cleave RNAs. Mol Cell.

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.

Hale, C., Kleppe, K., Terns, R.M., and Terns, M.P. (2008). Prokaryotic silencing (psi)RNAs in Pyrococcus furiosus. Rna *14*, 2572–2579.

Han, D., and Krauss, G. (2009). Characterization of the endonuclease SSO2001 from Sulfolobus solfataricus P2. FEBS Lett *583*, 771–776.

Han, D., Lehmann, K., and Krauss, G. (2009). SSO1450--a CAS1 protein from Sulfolobus solfataricus P2 with high affinity for RNA and DNA. FEBS Lett *583*, 1928–1932.

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. Embo J *31*, 2824–2832.

Hippel, von, P.H., and Berg, O.G. (1986). On the specificity of DNA-protein interactions. Proc Natl Acad Sci USA *83*, 1608–1612.

Hippel, von, P.H., and Berg, O.G. (1989). Facilitated target location in biological systems. J Biol

Chem *264*, 675–678.

Hohn, M., Tang, G., Goodyear, G., Baldwin, P.R., Huang, Z., Penczek, P.A., Yang, C., Glaeser, R.M., Adams, P.D., and Ludtke, S.J. (2007). SPARX, a new environment for Cryo-EM image processing. J. Struct. Biol. *157*, 47–55.

Horvath, P., and Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. Science *327*, 167–170.

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

Hoskisson, P.A., and Smith, M.C.M. (2007). Hypervariation and phase variation in the bacteriophage 'resistome'. Curr Opin Microbiol *10*, 396–400.

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

Hsia, K.-C., Chak, K.-F., Liang, P.-H., Cheng, Y.-S., Ku, W.-Y., and Yuan, H.S. (2004). DNA binding and degradation by the HNH protein ColE7. Structure *12*, 205–214.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014). 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. Nat Biotechnol *31*, 827–832.

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. Nat Biotechnol *31*, 227–229.

Ilya J Finkelstein, M.-L.V.E.C.G. (2010). Single-molecule imaging reveals mechanisms of protein disruption by a DNA translocase. Nature *468*, 983–987.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakata, A. (1987). Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. J Bacteriol *169*, 5429–5433.

Ivancic-Bace, I., Howard, Al, J., and Bolt, E.L. (2012). Tuning in to Interference: R-Loops and Cascade Complexes in CRISPR Immunity. J Mol Biol.

Jansen, R., Embden, J.D.A.V., Gaastra, W., and Schouls, L.M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol *43*, 1565–1575.

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

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.

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.

Johnson, J.E., and Hoogstraten, C.G. (2008). Extensive backbone dynamics in the GCAA RNA tetraloop analyzed using 13C NMR spin relaxation and specific isotope labeling. J Am Chem Soc *130*, 16757–16769.

Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, U., Wurm, R., Wagner, R., et al. (2011a). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat Struct Mol Biol *18*, 529–536.

Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, U., Wurm, R., Wagner, R., et al. (2011b). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat Struct Mol Biol *18*, 529–536.

Kabsch, W. (2010). XDS. Acta Crystallogr D Biol Crystallogr *66*, 125–132.

Karginov, F.V., and Hannon, G.J. (2010). The CRISPR system: small RNA-guided defense in bacteria and archaea. Mol Cell *37*, 7–19.

Karvelis, T., Gasiunas, G., Miksys, A., Barrangou, R., Horvath, P., and Siksnys, V. (2013). crRNA and tracrRNA guide Cas9-mediated DNA interference in Streptococcus thermophilus. RNA Biol *10*, 841–851.

Katoh, H., Yoshinaga, M., Yanagita, T., Ohgi, K., Irie, M., Beintema, J.J., and Meinsma, D. (1986). Kinetic studies on turtle pancreatic ribonuclease: a comparative study of the base specificities of the B2 and P0 sites of bovine pancreatic ribonuclease A and turtle pancreatic ribonuclease. Biochim Biophys Acta *873*, 367–371.

Kim, D., and Rossi, J. (2008). RNAi mechanisms and applications. BioTechniques *44*, 613–616.

Kim, S., Kim, D., Cho, S.W., Kim, J., and Kim, J.-S. (2014). Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. Genome Res.

Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M.D.C., and Yusa, K. (2014). Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat Biotechnol *32*, 267–273.

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. (2014). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. Nature.

Kunin, V., Sorek, R., and Hugenholtz, P. (2007). Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol *8*, R61.

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. Nat Biotechnol *32*, 677–683.

Labrie, S.J., Samson, J.E., and Moineau, S. (2010). Bacteriophage resistance mechanisms. Nat Rev Microbiol *8*, 317–327.

Lander, G.C., Stagg, S.M., Voss, N.R., Cheng, A., Fellmann, D., Pulokas, J., Yoshioka, C., Irving, C., Mulder, A., Lau, P.-W., et al. (2009). Appion: an integrated, database-driven pipeline to facilitate EM image processing. J. Struct. Biol. *166*, 95–102.

LeCuyer, K.A., Behlen, L.S., and Uhlenbeck, O.C. (1995). Mutants of the bacteriophage MS2 coat protein that alter its cooperative binding to RNA. Biochemistry *34*, 10600–10606.

Lee, H.Y., Haurwitz, R.E., Apffel, A., Zhou, K., Smart, B., Wenger, C.D., Laderman, S., Bruhn, L., and Doudna, J.A. (2013). RNA-protein analysis using a conditional CRISPR nuclease. Proc Natl Acad Sci USA *110*, 5416–5421.

Legault, P., Li, J., Mogridge, J., Kay, L.E., and Greenblatt, J. (1998). NMR structure of the bacteriophage lambda N peptide/boxB RNA complex: recognition of a GNRA fold by an arginine-rich motif. Cell *93*, 289–299.

Li, J.-F., Norville, J.E., Aach, J., McCormack, M., Zhang, D., Bush, J., Church, G.M., and Sheen, J. (2013). Multiplex and homologous recombination-mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9. Nat Biotechnol *31*, 688–691.

Lintner, N.G., Kerou, M., Brumfield, S.K., Graham, S., Liu, H., Naismith, J.H., Sdano, M., Peng, N., She, Q., Copié, V., et al. (2011). Structural and functional characterization of an archaeal clustered regularly interspaced short palindromic repeat (CRISPR)-associated complex for antiviral defense (CASCADE). J Biol Chem *286*, 21643–21656.

Liu, F., Barrangou, R., Gerner-Smidt, P., Ribot, E.M., Knabel, S.J., and Dudley, E.G. (2011). Novel virulence gene and clustered regularly interspaced short palindromic repeat (CRISPR) multilocus sequence typing scheme for subtyping of the major serovars of Salmonella enterica subsp. enterica. Appl Environ Microbiol *77*, 1946–1956.

Long, C., McAnally, J.R., Shelton, J.M., Mireault, A.A., Bassel-Duby, R., and Olson, E.N. (2014). Prevention of muscular dystrophy in mice by CRISPR/Cas9-mediated editing of germline DNA. Science *345*, 1184–1188.

Ludtke, S.J., Baldwin, P.R., and Chiu, W. (1999). EMAN: semiautomated software for high-resolution single-particle reconstructions. J. Struct. Biol. *128*, 82–97.

Maag, D., and Lorsch, J.R. (2003). Communication between eukaryotic translation initiation factors 1 and 1A on the yeast small ribosomal subunit. J Mol Biol *330*, 917–924.

Mackay, J.P., Font, J., and Segal, D.J. (2011). The prospects for designer single-stranded RNA-binding proteins. Nat Struct Mol Biol *18*, 256–261.

Maddalo, D., Manchado, E., Concepcion, C.P., Bonetti, C., Vidigal, J.A., Han, Y.-C., Ogrodowski, P., Crippa, A., Rekhtman, N., de Stanchina, E., et al. (2014). In vivo engineering of oncogenic chromosomal rearrangements with the CRISPR/Cas9 system. Nature.

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

Makarova, K.S., Aravind, L., Wolf, Y.I., and Koonin, E.V. (2011a). Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biol Direct *6*, 38.

Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I., and Koonin, E.V. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol Direct *1*, 7.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J.M., Wolf, Y.I., Yakunin, A.F., et al. (2011b). Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol *9*, 467–477.

Mali, P., Aach, J., Stranges, P.B., Esvelt, K.M., Moosburner, M., Kosuri, S., Yang, L., and Church, G.M. (2013a). CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. Nat Biotechnol *31*, 833–838.

Mali, P., Esvelt, K.M., and Church, G.M. (2013b). Cas9 as a versatile tool for engineering biology. Nat Meth *10*, 957–963.

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

Mallick, S.P., Carragher, B., Potter, C.S., and Kriegman, D.J. (2005). ACE: automated CTF estimation. Ultramicroscopy *104*, 8–29.

Manica, A., Zebec, Z., Teichmann, D., and Schleper, C. (2011). In vivo activity of CRISPR-mediated virus defence in a hyperthermophilic archaeon. Mol Microbiol.

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

Marraffini, L.A., and Sontheimer, E.J. (2010a). CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nature Reviews Genetics *11*, 181–190.

Marraffini, L.A., and Sontheimer, E.J. (2010b). Self versus non-self discrimination during CRISPR RNA-directed immunity. Nature *463*, 568–571.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., and Read, R.J. (2007). Phaser crystallographic software. J Appl Crystallogr *40*, 658–674.

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

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

Mokrousov, I., Limeschenko, E., Vyazovaya, A., and Narvskaya, O. (2007). Corynebacterium diphtheriae spoligotyping based on combined use of two CRISPR loci. Biotechnol J *2*, 901–906.

Mortimer, S.A., and Weeks, K.M. (2009). C2'-endo nucleotides as molecular timers suggested by the folding of an RNA domain. Proc Natl Acad Sci USA *106*, 15622–15627.

Mulepati, S., and Bailey, S. (2011). Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). J Biol Chem *286*, 31896–31903.

Murshudov, G.N., Vagin, A.A., and Dodson, E.J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr *53*, 240–255.

Nekrasov, V., Staskawicz, B., Weigel, D., Jones, J.D.G., and Kamoun, S. (2013). Targeted mutagenesis in the model plant Nicotiana benthamiana using Cas9 RNA-guided endonuclease. Nat Biotechnol *31*, 691–693.

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.

Nolan, S., Shiels, J., Tuite, J., Cecere, K., and Baranger, A. (1999). Recognition of an essential adenine at a protein-RNA interface: Comparison of the contributions of hydrogen bonds and a stacking interaction. J Am Chem Soc *121*, 8951–8952.

O'Connell, M.R., Oakes, B.L., Sternberg, S.H., East-Seletsky, A., Kaplan, M., and Doudna, J.A. (2014). Programmable RNA recognition and cleavage by CRISPR/Cas9. Nature *516*, 263–266.

Obbard, D.J., Gordon, K.H.J., Buck, A.H., and Jiggins, F.M. (2009). The evolution of RNAi as a defence against viruses and transposable elements. Philos. Trans. R. Soc. Lond., B, Biol. Sci. *364*, 99–115.

Ogawa, T., Inoue, S., Yajima, S., Hidaka, M., and Masaki, H. (2006). Sequence-specific recognition of colicin E5, a tRNA-targeting ribonuclease. Nucleic Acids Res *34*, 6065–6073.

Otwinowski, Z., Schevitz, R.W., Zhang, R.G., Lawson, C.L., Joachimiak, A., Marmorstein, R.Q., Luisi, B.F., and Sigler, P.B. (1988). Crystal structure of trp repressor/operator complex at atomic resolution. Nature *335*, 321–329.

Oye, K.A., Esvelt, K., Appleton, E., Catteruccia, F., Church, G., Kuiken, T., Lightfoot, S.B.-Y., McNamara, J., Smidler, A., and Collins, J.P. (2014). Biotechnology. Regulating gene drives. Science *345*, 626–628.

Parker, J.S., Parizotto, E.A., Wang, M., Roe, S.M., and Barford, D. (2009). Enhancement of the seed-target recognition step in RNA silencing by a PIWI/MID domain protein. Mol Cell *33*, 204–214.

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. Nat Biotechnol *31*, 839–843.

Perez-Pinera, P., Kocak, D.D., Vockley, C.M., Adler, A.F., Kabadi, A.M., Polstein, L.R., Thakore, P.I., Glass, K.A., Ousterout, D.G., Leong, K.W., et al. (2013). RNA-guided gene activation by CRISPR-Cas9-based transcription factors. Nat Meth *10*, 973–976.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem *25*, 1605–1612.

Pintilie, G.D., Zhang, J., Goddard, T.D., Chiu, W., and Gossard, D.C. (2010). Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. J. Struct. Biol. *170*, 427–438.

Platt, R.J., Chen, S., Zhou, Y., Yim, M.J., Swiech, L., Kempton, H.R., Dahlman, J.E., Parnas, O., Eisenhaure, T.M., Jovanovic, M., et al. (2014). CRISPR-Cas9 knockin mice for genome editing and cancer modeling. Cell *159*, 440–455.

Pommer, A.J., Cal, S., Keeble, A.H., Walker, D., Evans, S.J., Kühlmann, U.C., Cooper, A., Connolly, B.A., Hemmings, A.M., Moore, G.R., et al. (2001). Mechanism and cleavage specificity of the H-N-H endonuclease colicin E9. J Mol Biol *314*, 735–749.

Pougach, K., Semenova, E., Bogdanova, E., Datsenko, K.A., Djordjevic, M., Wanner, B.L., and Severinov, K. (2010). Transcription, processing and function of CRISPR cassettes in Escherichia coli. Mol Microbiol *77*, 1367–1379.

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

Przybilski, R., Richter, C., Gristwood, T., Clulow, J.S., Vercoe, R.B., and Fineran, P.C. (2011).

Csy4 is responsible for CRISPR RNA processing in Pectobacterium atrosepticum. RNA Biol *8*.

Pul, U., Wurm, R., Arslan, Z., Geißen, R., Hofmann, N., and Wagner, R. (2010). Identification and characterization of E. coli CRISPR-cas promoters and their silencing by H-NS. Mol Microbiol *75*, 1495–1512.

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

Qi, Y., Spong, M.C., Nam, K., Banerjee, A., Jiralerspong, S., Karplus, M., and Verdine, G.L. (2009). Encounter and extrusion of an intrahelical lesion by a DNA repair enzyme. Nature *462*, 762–766.

Radermacher, M., Wagenknecht, T., Verschoor, A., and Frank, J. (1987). Three-dimensional reconstruction from a single-exposure, random conical tilt series applied to the 50S ribosomal subunit of Escherichia coli. J Microsc *146*, 113–136.

Raines, R.T. (1998). Ribonuclease A. Chem Rev *98*, 1045–1066.

Ran, F.A., Hsu, P.D., Lin, C.-Y., Gootenberg, J.S., Konermann, S., Trevino, A.E., Scott, D.A., Inoue, A., Matoba, S., Zhang, Y., et al. (2013). Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. Cell *154*, 1380–1389.

Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., Pasić, L., Thingstad, T.F., Rohwer, F., and Mira, A. (2009). Explaining microbial population genomics through phage predation. Nat Rev Microbiol *7*, 828–836.

Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010). Origins of specificity in protein-DNA recognition. Annu Rev Biochem *79*, 233–269.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA shape in protein-DNA recognition. Nature *461*, 1248–1253.

Roseman, A.M. (2004). FindEM--a fast, efficient program for automatic selection of particles from electron micrographs. J. Struct. Biol. *145*, 91–99.

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. Mol Cell *52*, 124–134.

Rousseau, C., Gonnet, M., Le Romancer, M., and Nicolas, J. (2009). CRISPI: a CRISPR interactive database. Bioinformatics *25*, 3317–3318.

Rupert, P.B., and Ferré-D'Amaré, A.R. (2001). Crystal structure of a hairpin ribozyme-inhibitor complex with implications for catalysis. Nature *410*, 780–786.

Rupert, P.B., Massey, A.P., Sigurdsson, S.T., and Ferré-D'Amaré, A.R. (2002). Transition state stabilization by a catalytic RNA. Science *298*, 1421–1424.

Sampson, T.R., and Weiss, D.S. (2014). Exploiting CRISPR/Cas systems for biotechnology. Bioessays *36*, 34–38.

Sampson, T.R., Saroj, S.D., Llewellyn, A.C., Tzeng, Y.-L., and Weiss, D.S. (2013). A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. Nature *497*, 254–257.

Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Res *39*, 9275–9282.

Sashital, D.G., Jinek, M., and Doudna, J.A. (2011). An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. Nat Struct Mol Biol *18*, 680–687.

Sashital, D.G., Wiedenheft, B., and Doudna, J.A. (2012). Mechanism of foreign DNA selection in a bacterial adaptive immune system. Mol Cell *46*, 606–615.

Sánchez-Rivera, F.J., Papagiannakopoulos, T., Romero, R., Tammela, T., Bauer, M.R., Bhutkar, A., Joshi, N.S., Subbaraj, L., Bronson, R.T., Xue, W., et al. (2014). Rapid modelling of cooperating genetic events in cancer through somatic genome editing. Nature.

Scheres, S.H.W., Núñez-Ramírez, R., Sorzano, C.O.S., Carazo, J.M., and Marabini, R. (2008). Image processing for electron microscopy single-particle analysis using XMIPP. Nat Protoc *3*, 977–990.

Schneider, T.R., and Sheldrick, G.M. (2002). Substructure solution with SHELXD. Acta Crystallogr D Biol Crystallogr *58*, 1772–1779.

Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., van der Oost, J., Brouns, S.J.J., and Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci USA *108*, 10098–10103.

Semenova, E., Nagornykh, M., Pyatnitskiy, M., Artamonova, I.I., and Severinov, K. (2009). Analysis of CRISPR system function in plant pathogen Xanthomonas oryzae. FEMS Microbiol Lett *296*, 110–116.

Shalem, O., Sanjana, N.E., Hartenian, E., Shi, X., Scott, D.A., Mikkelsen, T.S., Heckl, D., Ebert, B.L., Root, D.E., Doench, J.G., et al. (2014). Genome-scale CRISPR-Cas9 knockout screening in human cells. Science *343*, 84–87.

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. Nat Biotechnol *31*, 686–688.

Shen, B.W., Landthaler, M., Shub, D.A., and Stoddard, B.L. (2004). DNA binding and cleavage by the HNH homing endonuclease I-HmuI. J Mol Biol *342*, 43–56.

Simon, M.D., Wang, C.I., Kharchenko, P.V., West, J.A., Chapman, B.A., Alekseyenko, A.A., Borowsky, M.L., Kuroda, M.I., and Kingston, R.E. (2011). The genomic binding sites of a noncoding RNA. Proc Natl Acad Sci USA *108*, 20497–20502.

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. Embo J *30*, 1335–1342.

Sinkunas, T., Gasiunas, G., Waghmare, S.P., Dickman, M.J., Barrangou, R., Horvath, P., and Siksnys, V. (2013). In vitro reconstitution of Cascade-mediated CRISPR immunity in Streptococcus thermophilus. Embo J *32*, 385–394.

Smith, C., Gore, A., Yan, W., Abalde-Atristain, L., Li, Z., He, C., Wang, Y., Brodsky, R.A., Zhang, K., Cheng, L., et al. (2014). Whole-genome sequencing analysis reveals high specificity of CRISPR/Cas9 and TALEN-based genome editing in human iPSCs. Cell Stem Cell *15*, 12–13.

Snoussi, K., and Leroy, J.L. (2001). Imino proton exchange and base-pair kinetics in RNA duplexes. Biochemistry *40*, 8898–8904.

Snyder, J.C., Bateson, M.M., Lavin, M., and Young, M.J. (2010). Use of cellular CRISPR (clusters of regularly interspaced short palindromic repeats) spacer-based microarrays for detection of viruses in environmental samples. Appl Environ Microbiol *76*, 7251–7258.

Song, J., Rechkoblit, O., Bestor, T.H., and Patel, D.J. (2011). Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. Science *331*, 1036–1040.

Sorek, R., Kunin, V., and Hugenholtz, P. (2008). CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat Rev Microbiol *6*, 181–186.

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

Spilman, M., Cocozaki, A., Hale, C., Shao, Y., Ramia, N., Terns, R., Terns, M., Li, H., and Stagg, S. (2013). Structure of an RNA Silencing Complex of the CRISPR-Cas Immune System. Mol Cell *52*, 146–152.

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. Mol Cell *52*, 135–145.

Stern, A., and Sorek, R. (2011). The phage-host arms race: shaping the evolution of microbes. Bioessays *33*, 43–51.

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.

Suloway, C., Pulokas, J., Fellmann, D., Cheng, A., Guerra, F., Quispe, J., Stagg, S., Potter, C.S., and Carragher, B. (2005). Automated molecular microscopy: the new Leginon system. J. Struct. Biol. *151*, 41–60.

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.

Tanenbaum, M.E., Gilbert, L.A., Qi, L.S., Weissman, J.S., and Vale, R.D. (2014). A Protein-Tagging System for Signal Amplification in Gene Expression and Fluorescence Imaging. Cell *159*, 635–646.

Tang, G., Peng, L., Baldwin, P.R., Mann, D.S., Jiang, W., Rees, I., and Ludtke, S.J. (2007). EMAN2: an extensible image processing suite for electron microscopy. J. Struct. Biol. *157*, 38–46.

Tang, T.-H., Bachellerie, J.-P., Rozhdestvensky, T., Bortolin, M.-L., Huber, H., Drungowski, M., Elge, T., Brosius, J., and Hüttenhofer, A. (2002). Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. Proc Natl Acad Sci USA *99*, 7536–7541.

Tang, T.-H., Polacek, N., Zywicki, M., Huber, H., Brugger, K., Garrett, R., Bachellerie, J.P., and Hüttenhofer, A. (2005). Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus. Mol Microbiol *55*, 469–481.

Tebas, P., Stein, D., Tang, W.W., Frank, I., Wang, S.Q., Lee, G., Spratt, S.K., Surosky, R.T., Giedlin, M.A., Nichol, G., et al. (2014). Gene editing of CCR5 in autologous CD4 T cells of persons infected with HIV. N. Engl. J. Med. *370*, 901–910.

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

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

Terwilliger, T. (2004). SOLVE and RESOLVE: automated structure solution, density modification and model building. J Synchrotron Radiat *11*, 49–52.

Torres, R., Martin, M.C., Garcia, A., Cigudosa, J.C., Ramirez, J.C., and Rodriguez-Perales, S. (2014). Engineering human tumour-associated chromosomal translocations with the RNA-guided CRISPR-Cas9 system. Nat Commun *5*, 3964.

Tsai, S.Q., Wyvekens, N., Khayter, C., Foden, J.A., Thapar, V., Reyon, D., Goodwin, M.J., Aryee, M.J., and Joung, J.K. (2014). Dimeric CRISPR RNA-guided FokI nucleases for highly specific genome editing. Nat Biotechnol *32*, 569–576.

Tyson, G.W., and Banfield, J.F. (2008). Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. Environ Microbiol *10*, 200–207.

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.

van Gelder, C.W., Gunderson, S.I., Jansen, E.J., Boelens, W.C., Polycarpou-Schwarz, M., Mattaj, I.W., and van Venrooij, W.J. (1993). A complex secondary structure in U1A pre-mRNA that binds two molecules of U1A protein is required for regulation of polyadenylation. Embo J *12*, 5191–5200.

van Heel, M., Harauz, G., Orlova, E.V., Schmidt, R., and Schatz, M. (1996). A new generation of the IMAGIC image processing system. J. Struct. Biol. *116*, 17–24.

Veres, A., Gosis, B.S., Ding, Q., Collins, R., Ragavendran, A., Brand, H., Erdin, S., Talkowski, M.E., and Musunuru, K. (2014). Low incidence of off-target mutations in individual CRISPR-Cas9 and TALEN targeted human stem cell clones detected by whole-genome sequencing. Cell Stem Cell *15*, 27–30.

Visnapuu, M.-L., and Greene, E.C. (2009). Single-molecule imaging of DNA curtains reveals intrinsic energy landscapes for nucleosome deposition. Nat Struct Mol Biol *16*, 1056–1062.

Vonrhein, C., Blanc, E., Roversi, P., and Bricogne, G. (2007). Automated structure solution with autoSHARP. Methods Mol Biol *364*, 215–230.

Voss, N.R., Yoshioka, C.K., Radermacher, M., Potter, C.S., and Carragher, B. (2009). DoG Picker and TiltPicker: software tools to facilitate particle selection in single particle electron microscopy. J. Struct. Biol. *166*, 205–213.

Voss, N.R., Lyumkis, D., Cheng, A., Lau, P.-W., Mulder, A., Lander, G.C., Brignole, E.J., Fellmann, D., Irving, C., Jacovetty, E.L., et al. (2010). A toolbox for ab initio 3-D reconstructions in single-particle electron microscopy. J. Struct. Biol. *169*, 389–398.

Voytas, D.F., and Gao, C. (2014). Precision genome engineering and agriculture: opportunities and regulatory challenges. PLoS Biol. *12*, e1001877.

Wang, F., Redding, S., Finkelstein, I.J., Gorman, J., Reichman, D.R., and Greene, E.C. (2013a). The promoter-search mechanism of Escherichia coli RNA polymerase is dominated by three-dimensional diffusion. Nat Struct Mol Biol *20*, 174–181.

Wang, H., Yang, H., Shivalila, C.S., Dawlaty, M.M., Cheng, A.W., Zhang, F., and Jaenisch, R. (2013b). One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell *153*, 910–918.

Wang, R., Preamplume, G., Terns, M.P., Terns, R.M., and Li, H. (2011). Interaction of the Cas6

Riboendonuclease with CRISPR RNAs: Recognition and Cleavage. Structure *19*, 257–264.

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014a). Genetic screens in human cells using the CRISPR-Cas9 system. Science *343*, 80–84.

Wang, Y., Wang, Z., and Tanaka Hall, T.M. (2013c). Engineered proteins with Pumilio/fem-3 mRNA binding factor scaffold to manipulate RNA metabolism. Febs J *280*, 3755–3767.

Wang, Y., Cheng, X., Shan, Q., Zhang, Y., Liu, J., Gao, C., and Qiu, J.-L. (2014b). Simultaneous editing of three homoeoalleles in hexaploid bread wheat confers heritable resistance to powdery mildew. Nat Biotechnol *32*, 947–951.

Weeks, K.M., and Crothers, D.M. (1993). Major groove accessibility of RNA. Science *261*, 1574–1577.

Weinbauer, M.G. (2004). Ecology of prokaryotic viruses. FEMS Microbiol. Rev. *28*, 127–181.

Westra, E.R., Pul, U., Heidrich, N., Jore, M.M., Lundgren, M., Stratmann, T., Wurm, R., Raine, A., Mescher, M., Van Heereveld, L., et al. (2010). H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. Mol Microbiol *77*, 1380–1393.

Wiedenheft, B., Lander, G.C., Zhou, K., Jore, M.M., Brouns, S.J.J., van der Oost, J., Doudna, J.A., and Nogales, E. (2011a). Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature *477*, 486–489.

Wiedenheft, B., Sternberg, S.H., and Doudna, J.A. (2012). RNA-guided genetic silencing systems in bacteria and archaea. Nature *482*, 331–338.

Wiedenheft, B., van Duijn, E., Bultema, J.B., Bultema, J., Waghmare, S.P., Waghmare, S., Zhou, K., Barendregt, A., Westphal, W., Heck, A.J.R., et al. (2011b). RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proc Natl Acad Sci USA *108*, 10092–10097.

Wiedenheft, B., Zhou, K., Jinek, M., Coyle, S.M., Ma, W., and Doudna, J.A. (2009). Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. Structure *17*, 904–912.

Wu, H., Lima, W.F., and Crooke, S.T. (1999). Properties of cloned and expressed human RNase H1. J Biol Chem *274*, 28270–28278.

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. Nat Biotechnol *32*, 670–676.

Xia, T., SantaLucia, J., Burkard, M.E., Kierzek, R., Schroeder, S.J., Jiao, X., Cox, C., and Turner, D.H. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry *37*, 14719–14735.

Xiao, A., Wang, Z., Hu, Y., Wu, Y., Luo, Z., Yang, Z., Zu, Y., Li, W., Huang, P., Tong, X., et al. (2013). Chromosomal deletions and inversions mediated by TALENs and CRISPR/Cas in zebrafish. Nucleic Acids Res *41*, e141.

Xie, K., and Yang, Y. (2013). RNA-guided Genome Editing in Plants Using A CRISPR-Cas System. Mol Plant.

Yagi, Y., Nakamura, T., and Small, I. (2014). The potential for manipulating RNA with pentatricopeptide repeat proteins. Plant J. *78*, 772–782.

Yang, C.-G., Yi, C., Duguid, E.M., Sullivan, C.T., Jian, X., Rice, P.A., and He, C. (2008). Crystal structures of DNA/RNA repair enzymes AlkB and ABH2 bound to dsDNA. Nature *452*, 961–965.

Yang, W. (2008). An equivalent metal ion in one- and two-metal-ion catalysis. Nat Struct Mol Biol *15*, 1228–1231.

Yang, W. (2011). Nucleases: diversity of structure, function and mechanism. Q Rev Biophys *44*, 1–93.

Yang, W., Lee, J.Y., and Nowotny, M. (2006). Making and breaking nucleic acids: two-Mg2+-ion catalysis and substrate specificity. Mol Cell *22*, 5–13.

Yin, H., Xue, W., Chen, S., Bogorad, R.L., Benedetti, E., Grompe, M., Koteliansky, V., Sharp, P.A., Jacks, T., and Anderson, D.G. (2014). Genome editing with Cas9 in adult mice corrects a disease mutation and phenotype. Nat Biotechnol.

Yin, P., Li, Q., Yan, C., Liu, Y., Liu, J., Yu, F., Wang, Z., Long, J., He, J., Wang, H.-W., et al. (2013). Structural basis for the modular recognition of single-stranded RNA by PPR proteins. Nature *504*, 168–171.

Zamel, R., Poon, A., Jaikaran, D., Andersen, A., Olive, J., De Abreu, D., and Collins, R.A. (2004). Exceptionally fast self-cleavage by a Neurospora Varkud satellite ribozyme. Proc Natl Acad Sci USA *101*, 1467–1472.

Zegans, M.E., Wagner, J.C., Cady, K.C., Murphy, D.M., Hammond, J.H., and O'Toole, G.A. (2009). Interaction between bacteriophage DMS3 and host CRISPR region inhibits group behaviors of Pseudomonas aeruginosa. J Bacteriol *191*, 210–219.

Zhang, Y., Heidrich, N., Ampattu, B.J., Gunderson, C.W., Seifert, H.S., Schoen, C., Vogel, J., and Sontheimer, E.J. (2013). Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. Mol Cell *50*, 488–503.

Zhou, Y., Zhu, S., Cai, C., Yuan, P., Li, C., Huang, Y., and Wei, W. (2014). High-throughput screening of a CRISPR/Cas9 library for functional genomics in human cells. Nature *509*, 487–491.

Zuris, J.A., Thompson, D.B., Shu, Y., Guilinger, J.P., Bessen, J.L., Hu, J.H., Maeder, M.L., Joung, J.K., Chen, Z.-Y., and Liu, D.R. (2014). Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. Nat Biotechnol.

Zwart, P.H., Afonine, P.V., Grosse-Kunstleve, R.W., Hung, L.-W., Ioerger, T.R., McCoy, A.J., McKee, E., Moriarty, N.W., Read, R.J., Sacchettini, J.C., et al. (2008). Automated structure solution with the PHENIX suite. Methods Mol Biol *426*, 419–435.