

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Nucleic acid-guided genome defense systems from bacteria and archaea

Permalink

<https://escholarship.org/uc/item/56k4j43r>

Author

Strutt, Steven C

Publication Date

2018

Peer reviewed|Thesis/dissertation

Nucleic acid-guided genome defense systems from bacteria and archaea

by

Steven C Strutt

A dissertation submitted in partial satisfaction
of the requirements for the degree of
Doctor of Philosophy
in
Molecular and Cell Biology
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:

Professor Jennifer A. Doudna, Chair

Professor Donald Rio

Professor Lin He

Professor Britt Koskella

Spring 2018

ABSTRACT

Nucleic acid-guided genome defense systems from bacteria and archaea

by

Steven C Strutt

Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

Professor Jennifer A. Doudna, Chair

Bacterial and archaeal genomes are under constant threat by genetic invaders. The need to maintain genomic and cellular integrity has driven the evolution of numerous and diverse genome defense systems. A common theme in prokaryotic defense strategies is interference of foreign DNA and RNA on the sequence level. Clustered regularly interspaced short palindromic repeats (CRISPR)-CRISPR-associated (Cas) systems confer adaptive immunity to previously encountered genetic invaders. Guided by short RNAs, the main effectors of CRISPR-Cas systems are sequence specific nucleases that catalyze degradation of exogenous nucleic acids. At the center of a similar method of genome defense to CRISPR-Cas systems, but operating through non-homologous proteins and pathways, prokaryotic Argonaute proteins (pAgo) have been proposed as sequence specific defense systems. However, our mechanistic knowledge of both CRISPR-Cas and pAgo systems stems from a small subset of the total genetic diversity of these systems. Here, we address this limited understanding through analysis of new CRISPR-Cas and pAgo systems, as well as describe novel activity for previously identified members.

CRISPR-Cas9 has rapidly been adopted as a programmable platform for sequence-specific DNA targeting with endless applications in gene editing, genome-wide screening, and transcriptional control. Current applications draw upon the biochemical activities of a few common Cas9 enzymes. The study of diverse homologs has potential to yield novel Cas9 proteins with desired traits such as increased efficiency or specificity. Surveying a vast metagenomic database, we report the first Cas9 from archaea, expanding the occurrence of CRISPR-Cas9 systems to a new domain of life.

DNA targeting is a hallmark of CRISPR-Cas9 systems. Engineering SpyCas9 to bind and target RNA has been difficult and suffers from reduced efficiency. We sought to identify Cas9 homologs with a natural ability to target RNA molecules. Using *in vitro* purification and biochemical assays, we discovered Cas9 enzymes that efficiently cleave RNA. Furthermore, we show that this activity can be harnessed to reduce phage infection and mediate gene repression *in vivo*, expanding the toolkit of CRISPR-Cas nucleases.

Analogous to CRISPR-Cas systems, Argonaute proteins are well known, RNA-guided nucleases that operate in eukaryotic RNA-interference. Motivated by initial observations that Argonaute homologs in prokaryotes constitute a nucleic acid-guided genome defense system, we studied the physiological role and biochemical activities of a novel clade of pAgos. We offer the first experimental evidence of complex formation between natural, two-piece Argonaute proteins. Preliminary *in vivo* observations implicate this split-pAgo in maintaining motility under conditions induced by introduction of an exogenous plasmid. Together with our studies on CRISPR-Cas systems, our work highlights unexpected functional diversity across divergent nucleic acid-guided genome defense systems.

Acknowledgements

Similar to the proverb “It takes a village to raise a child”, I believe that the same is true for a grad student. I owe a debt of gratitude to all those that helped me along the way, including not only my scientific mentors, collaborators, and labmates, but also my support network outside of science that has kept me balanced.

First of all, I could not be more thankful to Jennifer Doudna for agreeing to mentor a student with essentially zero background in biochemistry. I appreciate the freedom she gave me in exploring scientific questions and allowing me to continue with my pet project, even in the more challenging moments. Her passion for science is infectious and more important than just teaching me about biochemistry, she taught me how to ask good questions and to never lose sight of the big picture. Lastly, I would have been lost without the great team of people that Jennifer assembled – especially Kaihong and Lily/Nan/Jenny who made the Doudna lab an easy place to do science.

As a fresh graduate student with little biochemical experience, I cannot express enough gratitude to Emine Kaya for her guidance and patience during my rotation and first years in the lab. Along with Kevin Doxzen and Audrey Lapinaite, I will miss Team Ago, despite all the struggles. Similarly, a big shout out to Philip Kranzusch, Ross Wilson, and Megan Hochstrasser for holding my hand through crystallography. My RNA brainstorming buddies Mitch O’Connell, Alex Seletsky, Gavin Knott, and Das De Silva, helped me piece together my projects when they were about to fall apart. A special thanks to Alex for being a great friend and serving as a sounding board to talk through my grad school doubts and worries. To Akshay Tambe for answering my data analysis questions and Ben Oakes for providing me with reagents and experimental advice. Many thanks to my bay mates over the years – Emine, Kevin, Sam, Tina, Gogo, and Abhishek – for tolerating my interruptions and unique brand of humor.

Science is rarely done alone and my grad school journey was no exception. One standout collaboration was with David “Dudu” Burstein and Jill Banfield, who were so gracious in sharing their data with me and being so patient when experiments failed. Dudu, I will always remember our think-sessions fondly and I can’t wait to see you light up my PubMed alerts! I am also grateful to Oscar Negrete who served as an excellent mentor during my foray into the RNA-targeting and eukaryotic world. It was great to see how someone in a different field approaches a project and assembles a manuscript. Finally, thanks to Rachel Torrez for being a stellar summer pseudo-undergrad and putting in long hours to help me collect data.

Lastly, grad school would have been rougher if it weren’t for all those people keeping me sane. Many thanks to Eleri Syverson for spending hours on the phone talking about literally nothing and suffering through my rants about failed experiments. To Ricky, I struggle to expression how grateful I am to have you in my life. I’m not sure I would have survived the last year and a half without your support. Thanks for enduring my crazy schedule and keeping my stress in check.

Dedication

To my mother, Nancy Brandt

*None of my success would have
been possible without your support.*

*Your creativity, dedication, and resilience have
and will continue to guide me through life.*

TABLE OF CONTENTS

Abstract.....	1
Acknowledgements.....	i
Dedication.....	ii
Table of Contents.....	iii
List of Figures.....	vi
List of Tables.....	viii

Chapter 1: Introduction to prokaryotic genome defense systems

1.1 Overview of genome defense systems in prokaryotes	2
1.2 CRISPR-Cas systems	5
1.2.1 Spacer acquisition.....	7
1.2.2 CRISPR RNA biogenesis and processing	10
1.2.3 Interference	13
1.2.4 CRISPR-Cas applications	18
1.3 Discovery and overview of RNA interference pathway.....	18
1.3.1 Small RNA processing and loading.....	21
1.3.2 Argonaute proteins: effectors of RNA-interference	22
1.3.3 Prokaryotic Argonautes in genome defense	25

Chapter 2: Discovery of CRISPR-Cas9 systems in archaea

2.1 Chapter Summary	28
2.2 Introduction	28
2.3 Methods	29
2.3.1 Metagenomics and metatranscriptomics.....	29
2.3.2 CRISPR–Cas computation analyses	30
2.3.3 ARMAN-Cas9 protein expression and purification.....	31
2.3.4 RNA <i>in vitro</i> transcription and oligonucleotide purification	32
2.3.5 <i>In vitro</i> cleavage assays.....	33
2.3.6 <i>In vivo E. coli</i> interference assays.....	34
2.4 Results	34
2.4.1 Identification of archaeal Cas9.....	34
2.4.2 Description of archaeal CRISPR-Cas9 system from ARMAN-1	36
2.4.3 ARMAN-1 CRISPR array targets viral and mobile genetic element targets.....	37
2.4.4 CRISPR-Cas9 from ARMAN-4 is a reduced system.....	38
2.4.5 Identification of targeting requirements	38
2.4.6 Biochemical reconstitution of archaeal Cas9	41
2.4.7 Apparent recombination in archaeal Cas9 CRISPR system	42
2.5 Discussion and Conclusion.....	44

Chapter 3: RNA-dependent RNA targeting by CRISPR-Cas9

3.1 Chapter Summary	47
3.2 Introduction	47
3.3 Methods	48
3.3.1 Phylogenetic tree construction and RNA folding	48
3.3.2 Protein purification	48
3.3.3 Oligonucleotide purification and radiolabeling	48
3.3.4 <i>In vitro</i> cleavage assays	51
3.3.5 Filter binding and electrophoretic mobility shift assays	51
3.3.6 MS2 screen and plaque assay	51
3.3.7 MS2 survival and mismatch analysis	52
3.3.8 <i>E. coli in vivo</i> GFP repression	52
3.4 Results	53
3.4.1 Cas9 catalyzes PAM-independent RNA-guided RNA cleavage	53
3.4.2 Cleavage efficiency is impaired by duplex regions in target RNA ...	59
3.4.3 SauCas9 confers <i>in vivo</i> protection against RNA phage	61
3.4.4 SauCas9 represses gene expression in <i>E. coli</i>	68
3.5 Discussion and Conclusion	72

Chapter 4: A natural two-piece Argonaute influences cellular motility

4.1 Chapter Summary	75
4.2 Introduction	75
4.3 Methods	76
4.3.1 Phylogenetic and bioinformatic analyses	76
4.3.2 Protein purification and <i>in vitro</i> reconstitution	77
4.3.3 Strain construction	77
4.3.4 Total nucleic acid extraction and PCR assays	78
4.3.5 Complex immunoprecipitation and nucleic acid analyses	79
4.3.6 Analysis and mapping of S3Piwi RNA fragment	80
4.3.7 Phenotype assays	80
4.4 Results	80
4.4.1 Identification of a short pAgo system in <i>Shewanella sp.</i> ANA-3	80
4.4.2 S3Piwi and S3Apaz form a complex <i>in vitro</i> and <i>in vivo</i>	83
4.4.3 S3Piwi binds heterologous RNAs	85
4.4.4 S3Ago influences cellular motility in the presence of an exogenous plasmid	87
4.5 Discussion and Conclusion	90

Chapter 5: Summary and future directions

5.1 Chapter Summary	93
5.2 Functional and mechanistic comparison of CRISPR-Cas and pAgo systems	93
5.3 Outstanding questions about nucleic acid-guided genome defense systems	94
References.....	96

List of Figures

Figure 1.1 Overview of prokaryotic genome defense systems	2
Figure 1.2 Common themes in toxin-antitoxin systems	4
Figure 1.3 Functional modules of diverse CRISPR-Cas systems.....	6
Figure 1.4 Three stages of CRISPR-Cas interference pathway	7
Figure 1.5 Host factors influence specificity of spacer integration	9
Figure 1.6 Processing of crRNAs.....	11
Figure 1.7 Similarities in Type I and Type III CRISPR interference complex architecture.....	14
Figure 1.8 Mechanism of Cas9 binding and cleavage	16
Figure 1.9 Diverse effectors of Class 2 CRISPR systems	17
Figure 1.10 RNA interference pathway	20
Figure 1.11 Clades of eukaryotic Argonaute proteins	23
Figure 2.1 Multiple sequence alignment of newly described Cas9 proteins	35
Figure 2.2 ARMAN-1 CRISPR array diversity and identification of the ARMAN-1 Cas9 PAM sequence.....	36
Figure 2.3 ARMAN-1 spacers map to genomes of archaeal community members.....	37
Figure 2.4 Archaeal Cas9 from ARMAN-4 with a degenerate CRISPR array is found on numerous contigs	38
Figure 2.5 Predicted structures of guide RNA and purification schema for <i>in vitro</i> biochemistry studies	40
Figure 2.6 Novelty of the reported CRISPR–Cas systems	43
Figure 2.7 Evolutionary tree of Cas9 homologues	44
Figure 3.1 SauCas9 cleaves single-stranded RNA without a PAMmer	54
Figure 3.2 RNA is cleaved by SauCas9 and CjeCas9.....	55
Figure 3.3 ssRNA cleavage is similar to canonical dsDNA cleavage by Cas9 ...	56
Figure 3.4 SauCas9 cleavage of different nucleic acid substrates	57
Figure 3.5 SauCas9 prefers a complementary region of 23nt for binding and cleavage	58
Figure 3.6 <i>In vitro</i> RNA cleavage is impaired by strong secondary structure ...	59
Figure 3.7 RNA cleavage is limited by the RNA target	60
Figure 3.8 SauCas9 confers <i>in vivo</i> protection against an RNA phage	62
Figure 3.9 Enriched guides do not display sequence bias and cluster to regions on the MS2 genome	64
Figure 3.10 Enriched MS2 targeting guides mapped to MS2 genome structure	65
Figure 3.11 Confirmation that enriched guides from the MS2 screen confer protection against MS2 infection	66
Figure 3.12 Effect of single-nucleotide mismatches on ssRNA targeting	68
Figure 3.13 SauCas9 repression of a GFP reporter <i>in vivo</i>	69

Figure 3.14 Repression of GFP mRNA.....	71
Figure 4.1 Phylogenetic tree of pAgos.....	81
Figure 4.2 Organization of S3Ago operon and transcriptional coupling.....	82
Figure 4.3 S3Piwi and S3Apaz form a complex <i>in vitro</i>	83
Figure 4.4 <i>In vivo</i> complex formation and bound nucleic acids.....	84
Figure 4.5 Nucleic acid mapping and stability of interaction with S3Piwi.....	86
Figure 4.6 RNA-binding of S3Piwi MID-domain mutants.....	87
Figure 4.7 Conjugation efficiency into S3 strains.....	88
Figure 4.8 S3Apaz influences motility of S3.....	89

List of Tables

Table 1.1 Experimentally determined Ago guide and target preferences	26
Table 2.1 Archaeal CRISPR-Cas9 homolog summary	31
Table 2.2 DNA and RNA sequences used in this study.....	32
Table 2.3 <i>In vitro</i> cleavage conditions assayed for Cas9 from ARMAN-1 and ARMAN-4.....	41
Table 2.4 <i>In vivo</i> <i>E. coli</i> targeting assays.....	42
Table 3.1 List of sequences used in this study	48
Table 4.1 Strains used in this study	78
Table 4.2 Plasmids and primers used in this study.....	78

Chapter 1

Introduction to prokaryotic genome
defense systems

1.1 Overview of genome defense systems in prokaryotes

Bacteria and archaea must protect themselves from an onslaught of predators. Viruses are by far the most abundant biological entities in the world, outnumbering their microbial counterparts by a factor of 10 to 100 (Wommack and Colwell 2000). Common estimates place viral-mediated turnover of marine bacteria at around 15% (Suttle 1994), which corresponds to the death of $\sim 10^{28}$ bacteria daily (Whitman et al. 1998). To mitigate viral predation, bacteria and archaea have evolved a multitude of genome defense systems to ensure survival (van Houte et al. 2016). However, viruses have evolved a series of counter-measures to overcome host immunity systems, locking the microbe and virus in a veritable arms race for survival (Stern and Sorek 2011). Short replication cycles for viruses and bacteria fuel rapid selection and diversification of prokaryotic genome defense systems, which also display widespread horizontal gene transfer and recombination (Makarova et al. 2011c; Juhas et al. 2009; Jeltsch and Pingoud 1996). The result is a repertoire of systems that vary broadly in terms of composition and mechanism of action (Figure 1.1). Here, we delineate systems by their ability to act directly on foreign nucleic acids or through alternative mechanisms.

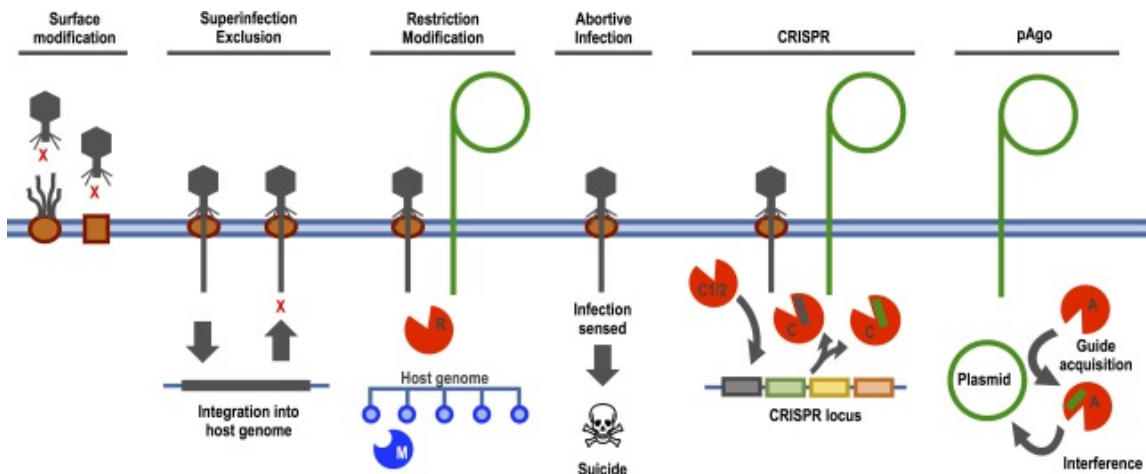


Figure 1.1 Overview of prokaryotic genome defense systems

Genome defense systems can act directly on foreign nucleic acid (CRISPR, pAgo, and restriction modification) or through alternative mechanisms (surface modification, superinfection exclusion, and abortive infection). R, restriction endonuclease; M, methylase; C, Cas nuclease; C1/2, Cas1/Cas2; A, Argonaute. Figure adapted from (van Houte et al. 2016).

Bacteriophage gain entry into cells via adherence onto molecules and receptors, such as lipopolysaccharide (LPS) or pili, displayed on the surface of bacteria (Labrie et al. 2010; Guerrero-Ferreira et al. 2011). One method to prevent or attenuate phage infections is to either reduce the expression of surface receptors or mask them through modification (Seed 2015). After exposure to phage, surviving *Vibrio cholerae* cells acquired mutations in the surface-

expressed, phage attachment protein, OmpU (Seed et al. 2014). This immune strategy might come at a cost however, as many phage receptors are also critical to the host's ability to sense or compete in a resource-limited environment (van Houte et al. 2016; Buckling and Brockhurst 2012; Mattick 2002). Similarly, lysogenic phage residing in the genome of the host sometime encode factors that block DNA injection by other phage, thus excluding superinfection (Labrie et al. 2010). A well-studied example comes from *Streptococcus thermophilus* phage that produce a lipoprotein (Ltp) that is anchored in the cytoplasmic membrane and prevents DNA entry (Ali et al. 2014).

If a virus gains entry into the cell, programmed cell death systems can attenuate viral replication and provide 'altruistic' protection within a bacterial population (Samson et al. 2013). Abortive infection (Abi) systems respond to a wide variety of signatures of phage infection and offer layered protection for their host. Astonishingly, as many as 20 Abis have been detected in a single organism (Chopin et al. 2005). The cellular suicide occurs through diverse mechanisms including depolarization of the membrane, induced lysis, and transcription/translation inhibition (Snyder 1995; Durmaz and Klaenhammer 2007; Dy et al. 2014).

Certain abortive infections are mediated through toxin-antitoxin (TA) systems. In these two component systems, an unstable antitoxin neutralizes a stable toxin. Upon infection, the antitoxin is selectively destabilized, unleashing the toxin and results in death of the cell or growth arrest (Page and Peti 2016). TA systems are incredible widespread among bacteria (Anantharaman and Aravind 2003; Pandey and Gerdes 2005) and fall into six major types depending on the identity of the TA pair (Figure 1.2) (Page and Peti 2016). One well-characterized system in *E. coli* is the mazF ribonuclease that induces cleavage of all cellular RNAs containing the motif 'ACA' upon proteolytic depletion of the mazE antitoxin in response to phage infection (Aizenman et al. 1996). Some systems act much more specifically, like the VapC toxin (VapBC TA module) that cleaves only the initiator tRNA (tRNA^{fMet}) to induce translational arrest (Lopes et al. 2014). Despite a few well-studied examples, there is a general lack of knowledge on the vast majority of putative TA systems.

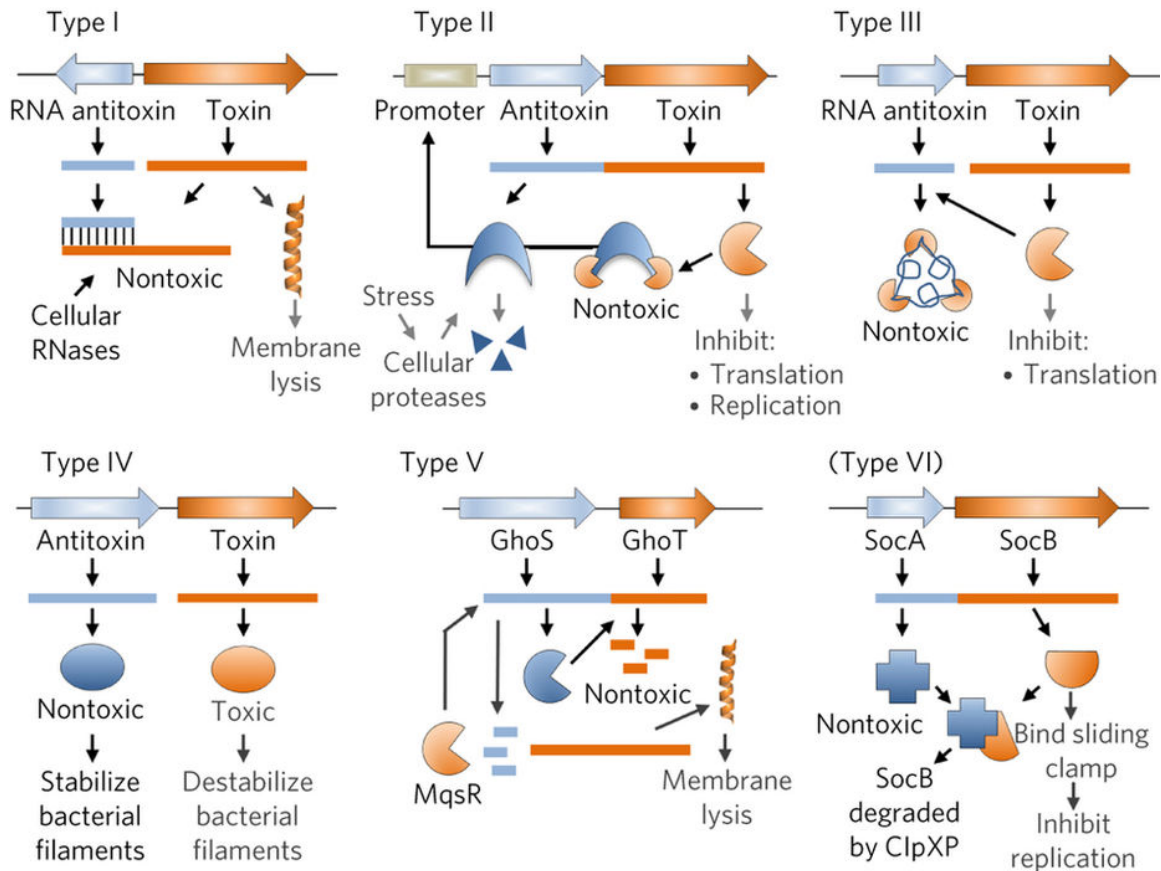


Figure 1.2 Common themes in toxin-antitoxin systems

Systems are delineated into types based on the identity of their toxin and antitoxin components. Antitoxin RNA and protein components (blue) are generally unstable and subject to selective degradation. Toxins (orange) act broadly to inhibit key cellular processes or destabilize the structural integrity of the cell. Figure adapted from (Page and Peti 2016).

Instead of shutting down cellular processes completely, another immune strategy is specific interference of foreign nucleic acids. Restriction-modification (RM) systems were described as early as the 1950's (Luria and Human 1952) and now fall into four major types depending on genetics and biochemistry of their DNA methyltransferase and endonuclease proteins (Loenen et al. 2014; Roberts et al. 2003). Restriction endonucleases (REase) of Type II RM systems consist of a single-protein that often dimerizes to bind and cleave a palindromic recognition site. To prevent double-strand breaks in the host genome, a cognate DNA methyltransferase covalently modifies the same DNA sequence, hereby preventing cleavage of the recognition site by the endonuclease (Murray 2002; Mucke et al. 2003). In this scenario, phage DNA injected into the cell does not contain the protective methylation pattern and is cut by the REase. The ability of Type II REases to cleave defined sequences has led their adoption as laboratory tools for molecular cloning and manipulating DNA (Loenen et al. 2014).

In addition to innate strategies to broadly interfere with invaders, bacteria and archaea also encode sophisticated systems for adaptive immunity. Cluster regularly interspaced short palindromic repeats (CRISPR)-associated (Cas) systems leverage prior exposure to invading nucleic acids to specifically interfere with future infections (Wright et al. 2016; Hille and Charpentier 2016; Rath et al. 2015). CRISPR-Cas systems capture foreign invading DNA and store it in specialized, repetitive loci in their genome. The sequence is transcribed and processed into a short RNA that binds to an effector protein forming an RNP (ribonucleoprotein) (Hochstrasser and Doudna 2015). Upon exposure to the same or similar exogenous DNA, the RNP mediates sequence-specific interference and mitigates detrimental effects of foreign pathogens. A stunning diversity of CRISPR-Cas systems exists, spread across diverse lineages of bacteria and archaea (Koonin et al. 2017a; 2017b; Shmakov et al. 2017). Several effector proteins, namely Cas9, Cas12, and Cas13, have been adopted as programmable platforms for DNA and RNA binding and cleavage, enabling facile applications in genome editing and manipulation (Donohoue et al. 2018; Choi and Lee 2016; Murugan et al. 2017).

Argonaute proteins are central to another proposed system for prokaryotic genome defense (Makarova et al. 2009). DNA-guided DNA interference was demonstrated in several prokaryotic Argonaute (pAgo) systems *in vitro* (Swarts et al. 2014a; 2015a; Zander et al. 2017; Willkomm et al. 2017). *In vivo*, presence of these pAgos was correlated with a reduction in natural transformation efficiency (Swarts et al. 2014a; 2015a; Zander et al. 2017). An additional pAgo system from *R. sphaeroides* (RsAgo) also interfered with exogenous plasmids but the mechanism of DNA degradation is unclear due to absence of an intact nuclease active site in RsAgo (Olovnikov et al. 2013). Recently, one study showed that *T. thermophilus* Ago (TtAgo) generates guides using a non-specific DNase activity and subsequently uses a select set of those guides for specific DNA degradation (Swarts et al. 2017a). In contrast to CRISPR-Cas systems, sequence-specific pAgo interference is not inheritable.

The life and death struggle between bacteria and viruses has driven the evolution of diverse systems for prokaryotic genome defense. With a staggering amount of new genomic and metagenomic data available each year (Narihiro and Kamagata 2017), more novel and divergent systems are sure to be found. In addition to recently identified BREX (Chaudhary 2018; Goldfarb et al. 2015) and DISARM (Ofir et al. 2018) systems, ten new predicted genome defense systems were reported (Doron et al. 2018). Here, we focus on nucleic-acid guided genome defense systems, including both CRISPR-Cas and pAgo systems.

1.2 CRISPR-Cas systems

CRISPR-Cas systems are widespread among bacterial and archaeal phyla and display remarkable genetic diversity (Haft et al. 2005; Makarova et al. 2006; Burstein et al. 2016). As with many mobile genetic elements, CRISPR-Cas systems display patterns of rampant horizontal gene transfer (van Elsas and Bailey 2002). The lack of a universal *cas* gene and recombination between *cas* genes of divergent systems complicate classification by typical methods

(Shmakov et al. 2017). Instead, a combination of comparative genomics and phylogenetics of conserved ‘signature’ genes is used to cluster CRISPR-systems. The current classification system divides *cas* operons into two classes of either multi- (Class 1) or single-component (Class 2) effector complexes (Figure 1.3). These two classes are each further divided into types (and subtypes) based on their sequence similarity and the presence or absence of certain *cas* genes and ancillary proteins. This classification scheme is continually evolving as we gain more insight into the distinct mechanisms of these CRISPR systems and as more new and diverse systems are identified (East-Seletsky et al. 2017; Shmakov et al. 2015; Burstein et al. 2017).

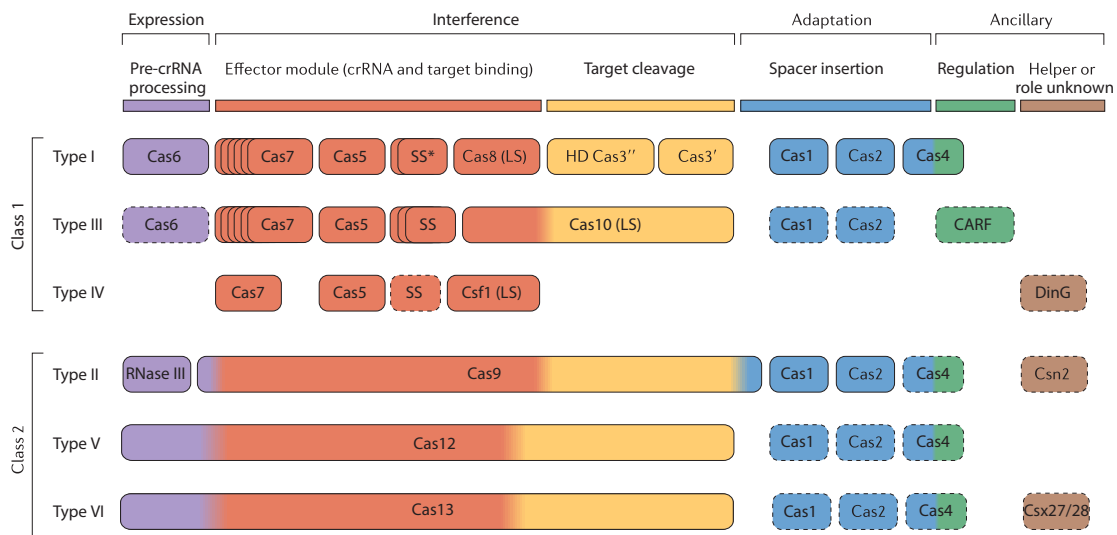


Figure 1.3 Functional modules of diverse CRISPR-Cas systems

Subtypes of CRISPR-Cas systems are grouped in two broader classes based on the presence of a multi-subunit (Class 1) or single (Class 2) interference module. Coloring of proteins reflects their functional role in the CRISPR-Cas pathway as noted above. SS, small-subunit; LS, large-subunit, CARF, CRISPR-associated Rossmann fold. Figure adapted from (Makarova et al. 2015).

Since initial experimental evidence implicating CRISPR-Cas systems in adaptive immunity (Barrangou et al. 2007), mechanistic studies have delineated the CRISPR interference pathway into three major steps. First, an adaptation step occurs wherein a short DNA sequence (termed protospacer) from an invading phage or plasmid is captured and integrated into a CRISPR array located within the genome of the cell. In a second step, the array is transcribed and processed to produce mature CRISPR RNAs (crRNAs). Finally, these crRNAs associate with protein effector complexes and serve as guides for recognition of complementary sequences. These protein-RNA complexes act as sentinels of the cell, and upon re-infection by the same or similar phage, the effectors bind and neutralize the invader. This process depends on target sequence being adjacent to a PAM (a protospacer adjacent motif), which is an

additional element that ensures targeting of foreign elements and preventing lethal cleavage of the host genome (Marraffini and Sontheimer 2010). The overall process is summarized in Figure 1.4.

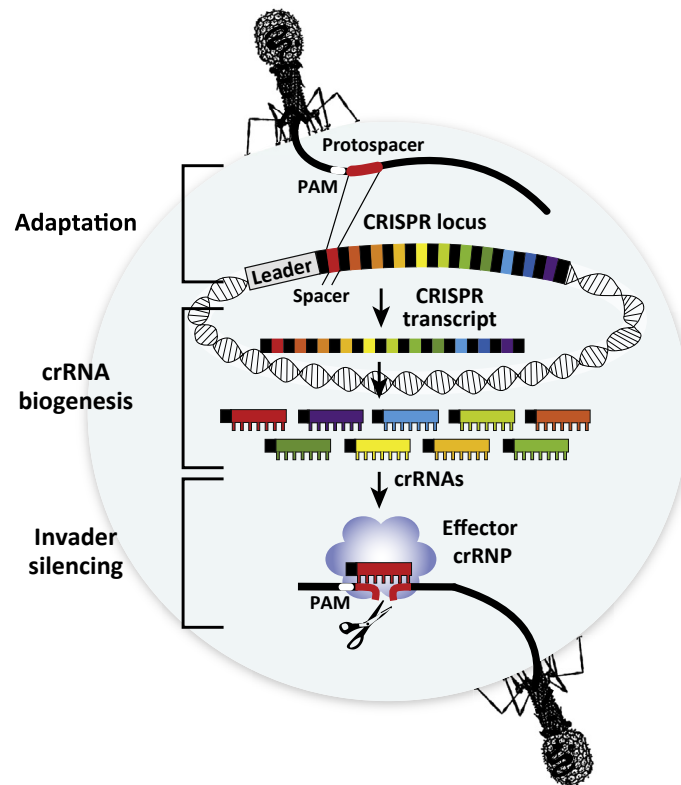


Figure 1.4 Three stages of CRISPR-Cas interference pathway

CRISPR-Cas systems capture a piece of foreign DNA (adaptation), generate a corresponding crRNA (crRNA biogenesis), and associate with Cas proteins to destroy phage and viruses (invader silencing). PAM, protospacer adjacent motif; crRNP, crRNA-ribonucleoprotein. Figure is adapted from (Terns and Terns 2014).

1.2.1 Spacer acquisition

As a system of adaptive immunity to mitigate future phage predation events, CRISPR-Cas systems must recognize and record past infections (Sternberg et al. 2016; Wright et al. 2016; Amitai and Sorek 2016). Early experimental evidence from Barrangou and colleagues (2007) demonstrated that strains of *Streptococcus thermophilus* that were exposed to phage acquired spacers corresponding to the phage genome. These spacers were responsible for conferring protection during subsequent infection and confirmed bioinformatic predictions of the adaptive nature of CRISPR-Cas systems (Makarova et al. 2006; Mojica et al. 2005; Pourcel et al. 2005; Bolotin et al. 2005). Subsequently, acquisition from both plasmids and bacteriophage was reported in a range of bacterial and archaeal systems (Yosef et al. 2012; Díez-Villaseñor et al. 2013;

Datsenko et al. 2012; Wei et al. 2015a; Heler et al. 2015; Wei et al. 2015b; Erdmann et al. 2014; Erdmann and Garrett 2012; Richter et al. 2014).

The molecular details of spacer acquisition in CRISPR systems remained elusive until experimental systems for monitoring adaptation were developed. Using the type I-E CRISPR system from *E. coli* as a model, it was established that Cas1 and Cas2 are necessary for adaptation (Yosef et al. 2012; Díez-Villaseñor et al. 2013). In *E. coli in vivo* screens, new spacers of ~33bp are integrated almost exclusively at the leader proximal end of the array (Datsenko et al. 2012; Yosef et al. 2012; Swarts et al. 2012). Cas1 DNase activity was well-established (Wiedenheft et al. 2009) but initial reports varied in regards to the nature of Cas2's catalytic activity. Initial tests indicated nuclease activity on ssRNA substrates (Beloglazova et al. 2008) but follow-up studies found activity on dsDNA (Ka et al. 2014; Nam et al. 2012b) or no activity at all (Samai et al. 2010). The respective roles of Cas1 and Cas2 in acquisition were resolved after the crystal structure of the complex was obtained (Nuñez et al. 2014). In the apo state, Cas1-Cas2 form a heterohexamer between two Cas1 dimers and one Cas2 dimer 'bridge' (Nuñez et al. 2014). Cas1 active site mutation abolished spacer acquisition *in vivo* while Cas2 mutations had no effect, thus confirming Cas1's catalytic role in acquisition. Nevertheless, Cas2 is necessary for complex formation, which is in turn essential for CRISPR-locus binding and spacer acquisition (Nuñez et al. 2014).

While the apo structure pinpointed the individual roles of each protein, the mechanism of spacer integration into the CRISPR array remained unclear. Through a combination of protospacer-bound Cas1-Cas2 structures and *in vitro* assays reconstituting spacer integration (Nuñez et al. 2015a; Wang et al. 2015; Rollie et al. 2015; Wright and Doudna 2016), it was determined that protospacer integration proceeded by direct nucleophilic attack of the leader-repeat junction by the 3'-OH of one end of a dsDNA protospacer followed by a second nucleophilic attack by the other end to complete the integration (Figure 1.5). These results correlated with integration intermediates previously captured *in vivo* (Arslan et al. 2014). Interestingly, high-throughput profiling of integration products revealed acquisition at all repeat junctions, implicating that additional factors might be involved in driving acquisition at the leader end *in vivo* (Datsenko et al. 2012; Swarts et al. 2012; Yosef et al. 2012; Díez-Villaseñor et al. 2013; Levy et al. 2015). Indeed, integration host factor (IHF) was discovered to play a crucial role in altering DNA topology and facilitating Cas1 interactions with an upstream recognition motif to increase integration fidelity (Figure 1.5) (Wright et al. 2017; Nuñez et al. 2016). More studies in other CRISPR types will determine if acquisition in *E. coli* is exceptional or if co-option and potentially co-evolution with ancillary host factors is a theme.

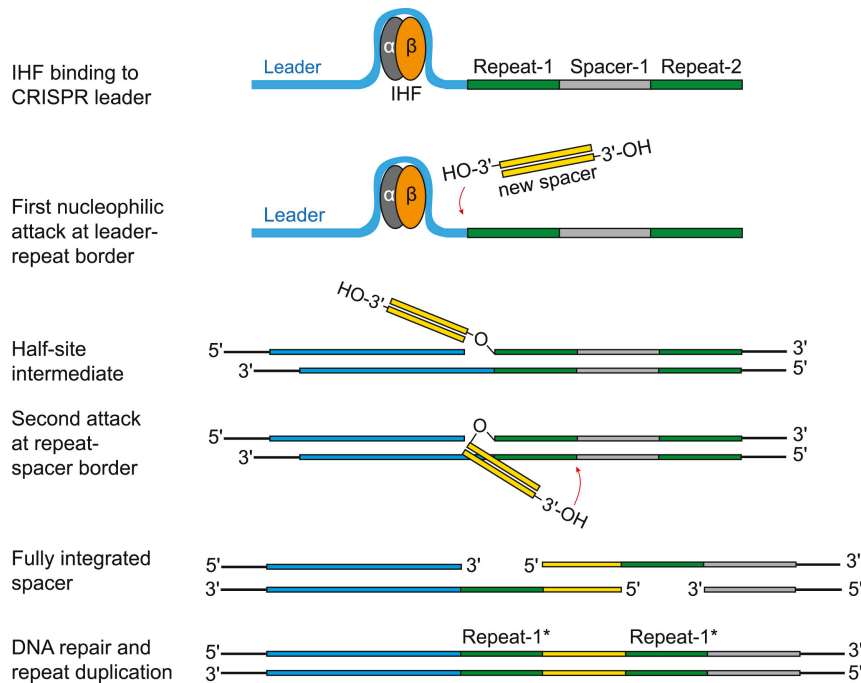


Figure 1.5 Host factors influence specificity of spacer integration

During CRISPR adaptation in *E. coli*, integration host factor (IHF) induces favorable conditions for Cas1-Cas2 mediated nucleophilic attack at the leader proximal repeat sequence (Repeat-1). The free end of the new spacer attacks at the next repeat-spacer junction to complete integration. Gaps in the chromosome are repaired via cellular machinery. Figure adapted from (Nuñez et al. 2016).

During adaptation, it is essential to acquire spacers that are adjacent to the appropriate PAM to ensure functional interference (Marraffini and Sontheimer 2010; Sashital et al. 2011; Semenova et al. 2011; Westra et al. 2013). While the Cas1-Cas2 complex is necessary and sufficient for PAM-adjacent spacer acquisition in type I-E systems (Yosef et al. 2013; 2012), data from type II studies suggest that PAM-recognition by Cas9 influences protospacer selection (Heler et al. 2015). Recently, Cas4 has been implicated in selection, protection, and end-processing of protospacers with the correct PAM sequence (Kieper et al. 2018; Rollie et al. 2018).

Posing a lethal autoimmune threat, the acquisition machinery must discriminate against genome-derived spacers. Naïve adaptation greatly favored spacers of extrachromosomal origin (Yosef et al. 2012). Recent evidence from a type I-F system implicates suicidal selection against these spacers (Staals et al. 2016). This phenomenon may occur across diverse CRISPR systems as abrogation of the Cas9's (type II) catalytic activity results in acquisition of an overrepresentation of self-targeting spacers (Wei et al. 2015b). These data suggest that chromosomal spacers are acquired but drop-out of the population rapidly during interference. Another hypothesis to explain self versus non-self discrimination is the observation that replication stall sites, such as Ter sites,

constitute hotspots of spacer acquisition (Levy et al. 2015). As plasmids have a higher copy number than the chromosome, additional replication and termination events could favor plasmid-derived spacers (Levy et al. 2015; Sternberg et al. 2016).

Besides naïve acquisition, type I CRISPR systems exhibit 'primed' acquisition, wherein the cell preferentially acquires spacers from genetic elements that contain sequences with perfect or partial complementarity to a spacer already present in the array (Sternberg et al. 2016; Wright et al. 2016). Through association with components of the interference complex, Cas1-Cas2 are activated for hyperacquisition of additional spacers (Datsenko et al. 2012; Li et al. 2014; Richter et al. 2014; Savitskaya et al. 2013; Swarts et al. 2012; Fineran et al. 2014). Evidence suggests that Cascade-dependent binding to protospacers with mutated sequences or PAMs causes the Cas3 nuclease/helicase to translocate bi-directionally along the target DNA and potentially activate Cas1-Cas2 for acquisition (Redding et al. 2015; Blosser et al. 2015; Richter et al. 2014). In type I-F where Cas2 and Cas3 are fused, the helicase and nuclease activities of Cas3 are dispensable for acquisition and Cas3-nuclease activity is stimulated by the Csy complex (Fagerlund et al. 2017; Rollins et al. 2017). Interestingly, in *Pectobacterium atrosepticum*, stronger acquisition occurs with targeting spacers than through priming spacers (Staals et al. 2016).

Acquisition by Cas1 and Cas2 is mediated at the DNA level (Nuñez et al. 2014). However, certain classes of Cas1 are fused to a reverse-transcriptase (Silas et al. 2016). This RT-Cas1 catalyzes capture of both DNA and RNA and offers hosts containing these systems a possible mechanism of acquiring resistance to RNA viruses (Silas et al. 2016). RNA-spacer capture seems particularly critical to Cas13 systems, which only seem to target RNA (Abudayyeh et al. 2016; East-Seletsky et al. 2016; Smargon et al. 2017; Yan et al. 2018). Future studies on acquisition in Cas13-containing systems will elucidate if DNA-acquisition is sufficient to protect the cell from parasitic RNA elements or if additional host factors are involved in specific RNA acquisition (Nuñez et al. 2016).

1.2.2 CRISPR RNA biogenesis and processing

CRISPR spacers stored in the genome of the host must be transcribed and processed into mature crRNAs in order to guide interference complexes efficiently (Hochstrasser and Doudna 2015; Charpentier et al. 2015; Li 2015). Transcription generally proceeds from AT-rich leader sequence across the CRISPR array to form a long precursor termed the pre-crRNA before subsequent processing into individual crRNAs (Figure 1.6) (Brouns et al. 2008; Carte et al. 2008). Host factors including H-NS and LeuO, have been shown to exert transcriptional control of the CRISPR array in some organisms (Hommais et al. 2001; Oshima et al. 2006; Pougach et al. 2010; Pul et al. 2010; Westra et al. 2010; Medina-Aparicio et al. 2011).

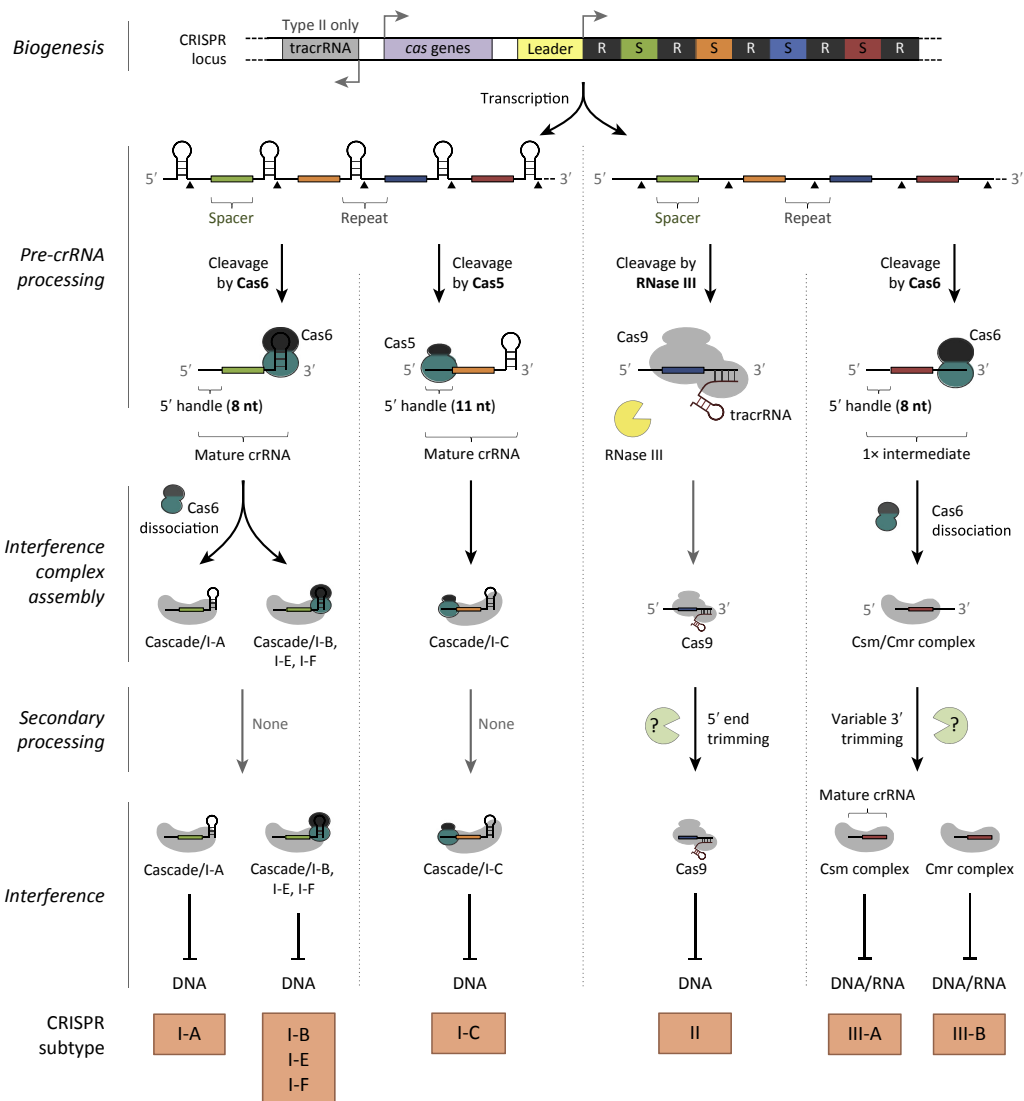


Figure 1.6 Processing of crRNAs

Biogenesis of crRNAs involves transcription of long precursors that are processed by Cas (Cas6, Cas5) or host (RNase III) endoribonucleases. Additional trimming occurs in some systems as a secondary processing step. Figure adapted from (Hochstrasser and Doudna 2015).

Type I and type III systems encode dedicated nucleases for crRNA processing, Cas6 or Cas5c, the latter of which is only found in subtype I-C (Makarova et al. 2011b; 2011a). Type I systems often contain palindromic repeats that form hairpins that are recognized by Cas6 through extremely tight interactions with a C-terminal RNA-recognition motif (Brouns et al. 2008;

Haurwitz et al. 2010; Gesner et al. 2011; Sashital et al. 2011; Niewoehner et al. 2014; Wang et al. 2011; Shao and Li 2013; Carte et al. 2008; Jackson et al. 2014). Base contacts between Cas6 and the hairpin play a role in specificity and positioning the stem loop structure for cleavage at the base of the stem (Haurwitz et al. 2010). Mutations in the hairpin can weaken or abrogate binding (Wang et al. 2012; Niewoehner et al. 2014). Catalysis proceeds through a general acid-base, metal-independent mechanism and liberates crRNAs with a 5' handle and a 3' stem loop (Haurwitz et al. 2010; 2012; Carte et al. 2010; Sternberg et al. 2012). Cas5c, formerly Cas5d, differs from Cas6 in that it recognizes the base of the stem loop and the 3' overhang (the future 5' handle) and leaves an extended 5' handle (Garside et al. 2012; Nam et al. 2012a; Hochstrasser et al. 2016).

While additional stability between the Cas6 and crRNA is conferred through interactions with repeat upstream of the stem loop (Jackson et al. 2014; Zhao et al. 2014; Mulepati et al. 2014), repeat-protein contacts are especially important for non-structured repeats (Shao and Li 2013; Wang et al. 2011; 2012). Repeats lacking stem loops are sometimes bound to Cas6 dimers (Shao et al. 2016; Reeks et al. 2013) and non-structured repeats of type I-A and type III undergo Cas6 dissociation and 'trimming' prior to interference complex assembly (Lintner et al. 2011; Plagens et al. 2014; Rouillon et al. 2013; Hatoum-Aslan et al. 2013; Hale et al. 2009; Zhang et al. 2012). This is in contrast to the majority of type-I Cascade complexes where Cas6 remains bound to the processed crRNA and nucleates complex assembly (Figure 1.6) (Sashital et al. 2011; Niewoehner et al. 2014; Carte et al. 2010; Wiedenheft et al. 2011b; 2011a).

An outstanding question in the field was the mechanism of crRNA processing in type II systems, which contain genes for acquisition and interference but lack a dedicated *cas* nuclease for guide maturation (Makarova et al. 2011a). An illuminating study in 2011 identified the involvement of RNase III and a *trans*-activating crRNA (tracrRNA) in processing in *S. pyogenes* (Deltcheva et al. 2011) and where subsequently validated in a range of type II systems (Chylinski et al. 2013; 2014). The tracrRNA is encoded in the vicinity of the *cas* operon and hybridizes with the CRISPR repeat region of the pre-crRNA. The resultant duplex RNA is recognized and cleaved by the endoribonuclease RNase III (Figure 1.6) (Deltcheva et al. 2011). Cas9 binds the tracr:crRNA duplex and may stabilize the pair for processing, while the dual-RNA serves to scaffold and activate Cas9 for cleavage (Deltcheva et al. 2011; Jinek et al. 2012). The absence of RNase III in archaea was hypothesized to explain the apparent lack of type II CRISPR systems in these organisms (Makarova et al. 2013; 2015). More recently, internal promoters in the direct repeats of CRISPR arrays were found to drive the transcription of crRNAs in select type II-C systems (Zhang et al. 2013). CrRNAs were still co-processed with a tracrRNA by RNase III but this maturation event was dispensable for RNA-guided interference by Cas9.

The emergence of Class 2 systems that did not encode tracrRNAs similarly raised the question of what proteins were responsible for guide maturation (Zetsche et al. 2015; Abudayyeh et al. 2016; Shmakov et al. 2015). In type V systems, the interference protein Cas12a (formerly Cpf1) harbors intrinsic nuclease activity for cleavage of the pre-crRNA (Fonfara et al. 2016; Swarts et al.

2017b). Similarly, in type VI systems, Cas13a (formerly C2c2) contains a distinct RNase active site responsible for guide maturation (East-Seletsky et al. 2016; Liu et al. 2017b; East-Seletsky et al. 2017; Knott et al. 2017). Subsequent work on Cas13b and Cas13d confirmed that self-contained guide processing was a general feature of Type-VI systems (Smargon et al. 2017; Yan et al. 2018).

1.2.3 Interference

The final step of CRISPR-Cas immunity is targeted destruction of genomic invaders. Interference modules are diverse in their composition and organization (Makarova et al. 2015; Koonin et al. 2017b). Indeed, the structure of interference complex forms the basis for the CRISPR-Cas class division – where Class 1 systems contain multi-subunit complexes while Class 2 systems are composed of a large, single effector protein (Makarova et al. 2013; 2015; Shmakov et al. 2015; 2017). Here, we will focus on the key discoveries and molecular mechanisms of CRISPR interference.

Cascade, from type I systems, forms a multi-subunit interference complex with a crRNA. *E. coli* Cascade is typically composed of five proteins of uneven stoichiometry (Cse1₁Cse2₂Cas7₆Cas5₁Cas6e₁) (Jore et al. 2011); however, longer Cascade complexes can form on extended crRNAs by incorporation of additional backbone units (Cas7) (Luo et al. 2016). DNA surveillance by Cascade is largely random but slows to interrogate PAM sites (Xue et al. 2017; Redding et al. 2015; Jung et al. 2017). Following PAM engagement, formation of an RNA-DNA heteroduplex begins in the seed region of the crRNA (Hayes et al. 2016; Xiao et al. 2017) and propagates directional (towards the 5' end of the target DNA strand) for complete R-loop formation (Semenova et al. 2011; Szczelkun et al. 2014; Rutkauskas et al. 2015). Bona fide targets allow for Cse1-dependent Cas3 recruitment. Possessing nuclease and helicase activity, Cas3 nicks the DNA and translocates in the 5' direction, degrading target DNA through an ATP- and metal-dependent mechanism (Hochstrasser et al. 2014; Huo et al. 2014; Mulepati and Bailey 2013; Redding et al. 2015; Sinkunas et al. 2011; Westra et al. 2012). Incomplete pairing of the crRNA and target DNA may trigger Cas1-Cas2 dependent primed acquisition of novel spacers (Sternberg et al. 2016; Wright et al. 2016).

Type III Csm and Cmr complexes exhibit remarkable structural similarity to Cascade (Figure 1.7) (Zhang et al. 2012; Rouillon et al. 2013; Spilman et al. 2013; Staals et al. 2013; 2014). As with type I systems, the size of the complex is correlated with the length of the crRNA, but it appears that the number of backbone subunits (Csm3/Cmr4) dictates the extent of crRNA trimming by non-cas, cellular nucleases (Staals et al. 2014; Tamulaitis et al. 2014; Walker et al. 2017). Initially, Csm and Cmr were thought to cleave only ssDNA and ssRNA, respectively (Marraffini and Sontheimer 2008; Hale et al. 2009). Discrepancies between *in vitro* and *in vivo* targeting preferences were resolved when it was determined that Type III systems cleave DNA and RNA co-transcriptionally (Samai et al. 2015). Binding of type III complexes to RNAs from active transcription complexes recruits and activates the Cas10 nuclease for non-specific ssDNA cleavage. The Csm3/Cmr4 subunits affect cleavage of RNA

transcripts complementary to the crRNA (Tamulaitis et al. 2017; Samai et al. 2015; Liu et al. 2017c; Elmore et al. 2016; Estrella et al. 2016). The RNA-activation of a non-discriminate DNase activity has been proposed to act in defense by specifically interfering with phages undergoing active transcription (Goldberg et al. 2018).

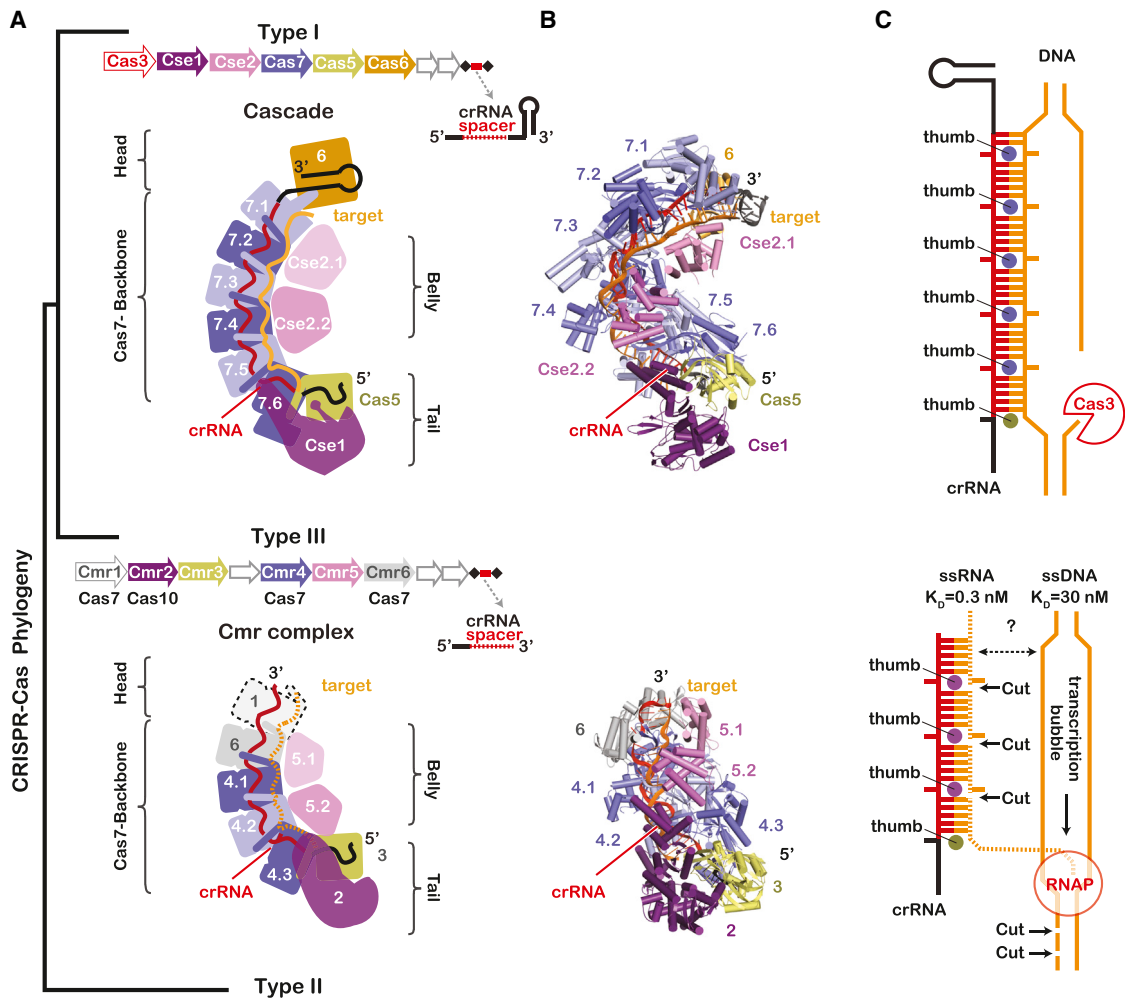


Figure 1.7 Similarities in Type I and Type III CRISPR interference complex architecture

(A) Schematized Cas operons and relative structure of multi-subunit effector complexes composed of common head, backbone, belly and tail features. (B) Crystal structures of Cascade and Cmr complexes. (C) Diagram of nucleic acid targeting in each system shows similarity of protein thumb-mediated base flipping. RNAP, RNA polymerase. Figure adapter from (Jackson and Wiedenheft 2015).

Type III CRISPR systems are exceptional in their lack of PAM recognition during interference (Maniv et al. 2016; Silas et al. 2017; Elmore et al. 2016). Debate still shrouds the mechanism behind self versus non-self discrimination in

these systems but one group suggested that pairing of the 5' handle of the crRNA with the RNA transcript protects the host (Kazlauskienė et al. 2016). However, another group argues for specific recognition of an RNA PAM by the Cmr complex (Elmore et al. 2016). The necessity and role of the auxiliary Cas proteins Csm6 and Csx1 are also in question for Csm and Cmr-containing systems. *In vitro* assays revealed nuclease activity (Niewoehner and Jinek 2016; Sheppard et al. 2016) and are crucial for certain *in vivo* systems (Deng et al. 2012; Jiang et al. 2016b). Interestingly, a feedback mechanism was recently discovered that involved allosteric activation of Csm6 ribonuclease activity by Cas10-dependent cyclic oligonucleotides (Kazlauskienė et al. 2017; Niewoehner et al. 2017). Future studies will hopefully explore the full extent of this pathway and its role in CRISPR immunity.

From the first and best-characterized Class 2 system, Cas9 is a single-effector nuclease and signature gene of type II systems. Cas9 employs two guide RNAs (tracrRNA and crRNA) to generate a double-strand break in target dsDNA containing a PAM sequence (Deltcheva et al. 2011; Jinek et al. 2012; Gasiunas et al. 2012). An engineered single-guide RNA (sgRNA) fuses the tracr:crRNA pair and simplifies reconstitution of active Cas9 molecules (Jinek et al. 2012) and has led to rapid adoption of Cas9 as a tool for fully-programmable genome engineering (Mali et al. 2013; Cong et al. 2013; Jinek et al. 2013; Doudna and Charpentier 2014). Structural and biophysical studies revealed that Cas9 is a bilobed enzyme that scans target molecules for PAM-sites and upon recognition employs a similar binding pathway as Cascade (Figure 1.8) (Jiang and Doudna 2017; Jiang et al. 2015; 2016a; Nishimasu et al. 2014; 2015; Sternberg et al. 2014). crRNA:target seed pairing licenses directional R-loop formation (Sternberg et al. 2014; Singh et al. 2016). The α -helical lobe is responsible for sensing pairing at the 5' end of the guide and triggers activation of Cas9's nuclease lobe for concomitant cleavage of the target and non-target strands by the HNH and RuvC nuclease domains, respectively (Sternberg et al. 2015; Chen et al. 2017; Dagdas et al. 2017).

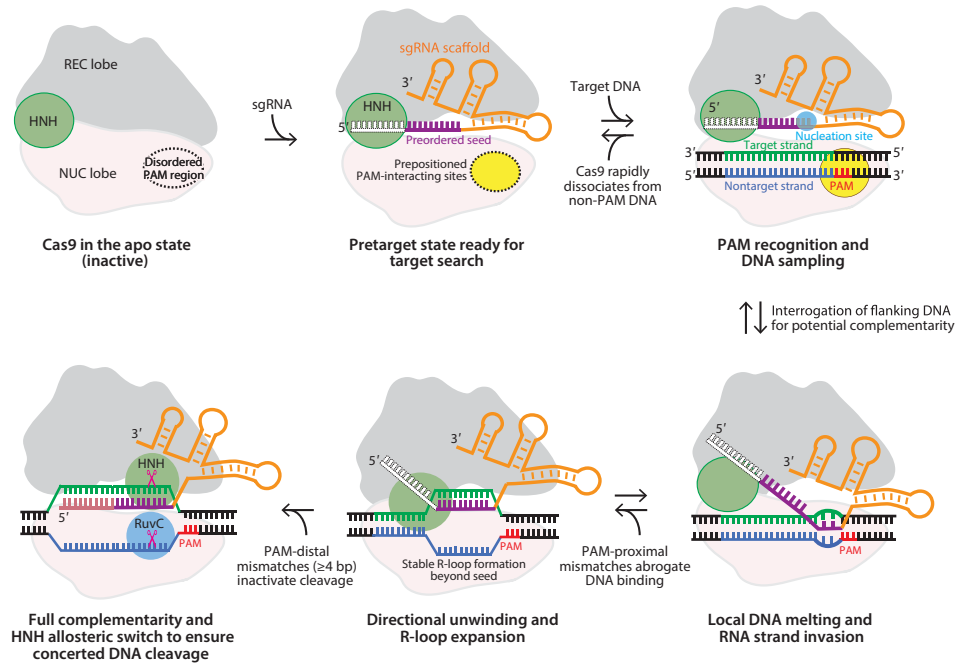


Figure 1.8 Mechanism of Cas9 binding and cleavage

The PAM interacting site of Cas9 aids in scanning DNA for an appropriate target. R-loop formation begins along the preordered seed region of the guide and additional complementarity allows full guide:target pairing and activation of dual nuclease sites. Mismatches in the PAM, seed, and seed-proximal regions disfavor binding; however, PAM-distal mismatches allow binding but trap Cas9 in an catalytically-inactive state. Figure adapted from (Jiang and Doudna 2017).

The utility of Cas9 as a programmable nuclease motivated the mining of additional Class2 CRISPR effectors from the microbial diversity (Figure 1.9). Cas12a (formerly Cpf1) was the first instance of a type V system (Zetsche et al. 2015). Unlike Cas9, Cas12a guides dsDNA cleavage using only a crRNA – dispensing with the need for a tracrRNA, and makes a staggered cut, leaving 5' overhangs. Cas12a still requires a PAM although it is AT-rich (compared to the G-rich PAM of Cas9), but the PAM is near the 5' end of the complementary spacer region. crRNA processing is an activity intrinsic to the protein as Cas12a is necessary and sufficient for processing precrRNAs from an array (Fonfara et al. 2016; Swarts et al. 2017b). With only a single RuvC active site, Cas12a is thought to act through a sequential model to cleave both strands of the DNA although the exact mechanism remains unclear (Yamano et al. 2016; Gao et al. 2016; Dong et al. 2016). Recently, a trans-DNase activity was described for Cas12a wherein binding of Cas12a activates shredding of ssDNA (Li et al. 2018; Chen et al. 2018) and may cleave open transcription bubbles as small as 2 nt (Li et al. 2018).

		Nuclease domains	tracrRNA	PAM	Substrate	Cleavage pattern
Type II Cas9		RuvC and HNH	Yes	3', GC-rich	dsDNA	Blunt ends
Type V-A Cas12a (Cpf1)		RuvC and Nuc	No	5', AT-rich	dsDNA	Staggered ends, 5' overhangs
Type V-B Cas12b (C2c1)		RuvC	Yes	5', AT-rich	dsDNA	Staggered seven-nucleotide cut of target DNA
Type VI-A Cas13a (C2c2)		2 HEPN domains	No	5', non-G PFS	ssRNA	Cleaves ssRNA near uracil and trans activity

Figure 1.9 Diverse effectors of Class 2 CRISPR systems

Current Class 2 CRISPR effectors are diagrammed on the left with general guide requirements, PAM location and cleavage site (red arrows). Molecular details for PAM, substrate preferences, and cleavage pattern are listed if known. Figure adapted from (Shmakov et al. 2017).

Two other Class 2, type V CRISPR systems were described recently – Cas12b and Cas12c (formerly C2c1 and C2c3 respectively). Cas12b uses two guides RNAs like Cas9 but possess a single-active site for dsDNA interference. While atomic structures for Cas12b exist, there is a lack of mechanistic details on how the enzyme cleaves DNA (Yang et al. 2016; Liu et al. 2017a; Wu et al. 2017).

Cas13a (Type VI, formerly C2c2) is different from most CRISPR effectors in that the relevant target is RNA, not DNA (Abudayyeh et al. 2016; East-Seletsky et al. 2016; Smargon et al. 2017; Yan et al. 2018). These enzymes contain two HEPN (higher eukaryote and prokaryote nuclease) domains that are known to cleave RNA (Figure 1.9) (Anantharaman et al. 2013). Similar to Cas12a, Cas13a can process its own guide from a pre-crRNA, although processing is not always necessary for yielding a functional complex (East-Seletsky et al. 2016; 2017). Upon ternary complex formation with a complementary target, Cas13a is activated for robust *trans* RNA cleavage. Two different subclasses of Cas13 (Cas13b and Cas13d) have been described and

are similar to Cas13a in their ability to process their own guide and cleave RNA in *trans* (Smargon et al. 2017; Yan et al. 2018). Interestingly, the RNA targeting ability of Cas13b appears to be modulated by other ancillary factors such as Csx27 and Csx28, which repress and enhance cleavage, respectively (Smargon et al. 2017). The relevance of Cas13's *trans* cleavage activity in the context of genome defense awaits further study in their native hosts.

1.2.4 CRISPR-Cas applications

The minimal architecture and programmability of Class 2 CRISPR-Cas effectors have led to their rapid adoption as platforms for nucleic acid binding and cleavage (Rath et al. 2015; Wright et al. 2016). Catalytically-active Cas9/Cas12a are used to generate double-strand breaks in DNA for genome editing applications in a wide-range of organisms (Doudna and Charpentier 2014; Ma and Liu 2015). One emerging application is the use of catalytically-dead variants of Cas9 (dCas9) fused to base-modification domains that catalyze DNA base transitions and transversions without cutting the DNA (Kim et al. 2017b; Gaudelli et al. 2017; Komor et al. 2016). As an orthogonal system to RNAi, dCas9/dCas12a are employed as powerful transcriptional repressors for specific gene knockdown without modifying the genome (Lo and Qi 2017; Qi et al. 2013; Gilbert et al. 2014; Zhang et al. 2017). In their own right, various Cas13 homologs have been adapted for applications in transcriptome modulation and as sensitive detection systems but the absence of *trans* cleavage in mammalian remains enigmatic (Abudayyeh et al. 2017; Gootenberg et al. 2017; East-Seletsky et al. 2016). The utility of Class 2 CRISPR-Cas systems as programmable tools has motivated our efforts to both discover new effector proteins and also identify novel biochemical activities of known homologs.

1.3 Discovery and overview of RNA interference pathway

In 1990, researchers reported that introduction of a transgene to increase purple color in petunias actually resulted in flower albinism and variegation, a phenomenon they termed cosuppression (Napoli et al. 1990). Related observations in yeast (Romano and Macino 1992) were reported shortly thereafter; however, the connection between these two studies began to emerge after seminal work by Fire and colleagues in *C. elegans* (Fire et al. 1998). They demonstrated that introduction of long double-strand RNA (dsRNA) triggered efficient gene repression and coined the term 'RNA interference', or RNAi (Fire et al. 1998). This catalyzed an explosion of studies focusing on dissecting the biochemical pathways and molecular machines involved in RNA suppression.

The connection between the introduced dsRNA and target RNA suppression was a large outstanding question. Clarity came from observations that plants undergoing RNAi (or post-transcription gene silencing) produced a species of short ~20-25 nt RNAs (Hamilton and Baulcombe 1999) that seemed to originate from the dsRNA trigger molecule (Zamore et al. 2000; Hammond et al. 2000). Indeed, introduction of short 20-21 nt RNA duplexes was sufficient for potent RNAi (Elbashir et al. 2001). Subsequently the RNase III domain-containing proteins Drosha and Dicer were described as the main protein

components responsible for processing the long dsRNA precursor into small function RNA duplexes (Lee et al. 2003; Bernstein et al. 2001). Finally, it was identified that processed short RNAs bind to members of the Argonaute family and guide interference of complementary RNA targets (Song et al. 2004; Liu et al. 2004), via a complex called the RNA-induced silencing complex (RISC). Short-interfering RNAs (siRNAs) and microRNAs (miRNAs) are delineated by full or partial complementarity to the target RNA and dictate cleavage or translational repression of the RNA, respectively (Meister et al. 2004). Analogously, a germline-specific subclade of Argonaute proteins, termed Piwi proteins, associate with Piwi-interacting RNAs (piRNAs) to suppress transposon destabilization of the genome (Siomi et al. 2011). Continued research over nearly 15 years has sharpened our understanding of the molecular and mechanistic details at the heart of the RNAi pathway (Figure 1.10) (Wilson and Doudna 2013).

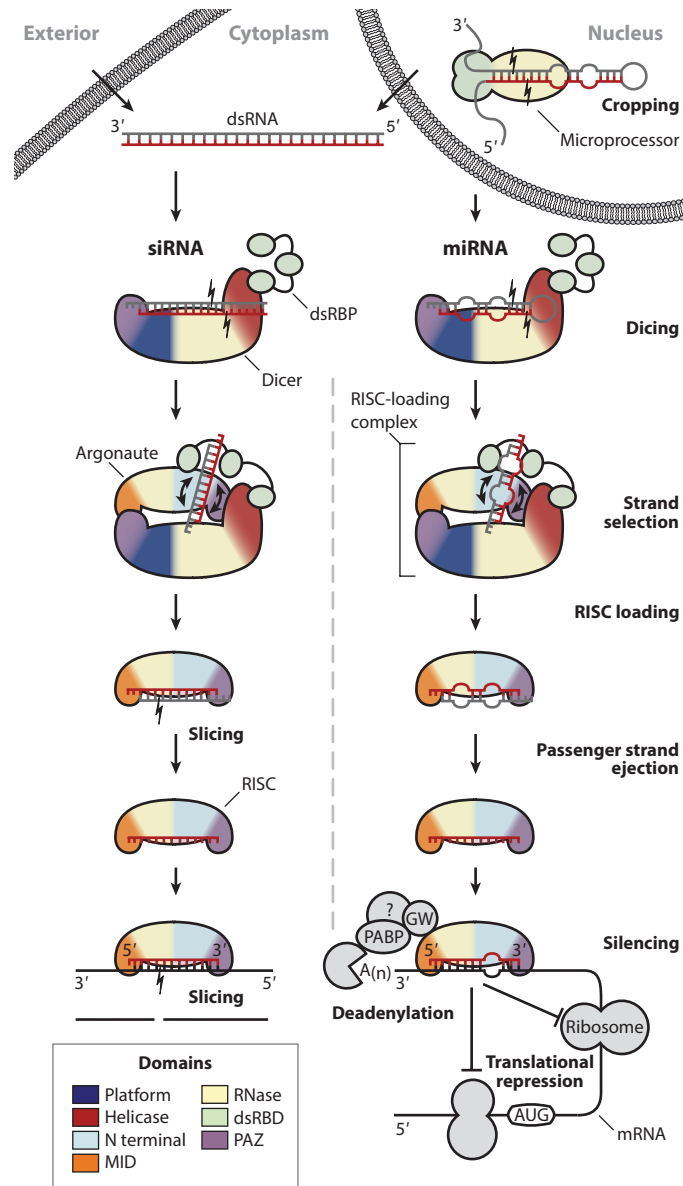


Figure 1.10 RNA interference pathway

Both perfectly complementary (siRNA - left) and partially-duplexed (miRNA - right) small RNAs are processed through similar mechanisms. miRNAs are processed by Microprocessor in the nucleus prior to export and dicing in the cytoplasm. siRNAs are substrates for Dicer directly and often bypass nuclear processing. The RISC-loading complex hands off the duplex to Argonaute, which retains only one guide strand. Silencing is mediated through slicing or a recruitment of cellular factors for general translational repression. Here, an emphasis is placed on the functional domains of the various endonucleases, represented through domain coloring. Adapted from (Wilson and Doudna 2013).

1.3.1 Small RNA processing and loading

RNAs that feed into RNAi pathways can originate from a number of sources (Carthew and Sontheimer 2009). miRNAs are generally transcribed from the genome as long (up to ~1 kbp) primary miRNA (pri-miRNA) transcripts by RNA polymerase II (Saini et al. 2007). However, cytoplasmic introduction of long exogenous dsRNAs can trigger RNAi (Zamore et al. 2000; Hammond et al. 2000) and shortcuts one step of the canonical pathway (Figure 1.10). Bypassing processing completely, short siRNAs, introduced into the cell directly, are potent activators of RNAi activity (Elbashir et al. 2001). Like miRNAs, piRNAs are produced from longer transcripts but maturation involves a complex set of endo- and exonucleases (Siomi et al. 2011; Weick and Miska 2014; Huang et al. 2017). Here, we focus on the canonical miRNA processing pathway and highlight a few interesting exceptions.

Structurally, the pri-miRNA contains a partially double-stranded hairpin that is recognized by the microprocessor complex, a nuclear heterotrimer composed of one Drosha and two DGCR8 (DiGeorge syndrome critical region gene 8) subunits (Han et al. 2004). Microprocessor senses both the upper hairpin and lower ssRNA junction through a combination of a ruler mechanism and specific primary sequence determinants (Zeng et al. 2005; Auyeung et al. 2013; Han et al. 2006; Ma et al. 2013). DGCR8 helps to correctly position the pri-miRNA while Drosha is the nucleases that directs cleavage at the base of hairpin through the action of dual RNase-III domains, liberating a ~70nt precursor miRNA (pre-miRNA) (Han et al. 2006; Nguyen et al. 2015). Recently, a structure of human Drosha bound to C-terminal fragments of DGCR8 revealed similarity to Dicer, suggesting a shared evolutionary history despite unique biochemical activities (Kwon et al. 2016; MacRae et al. 2006).

Pri-miRNAs that are processed by Drosha into pre-miRNAs must be transported out of the nucleus for subsequent maturation events. Exportin-5 was identified as the main transporter for processed pre-miRNAs (Yi et al. 2003; Bohnsack et al. 2004; Lund et al. 2004). Cleavage by Drosha generates pre-miRNAs with a ~2 nt 3' overhang that is specifically recognized by exportin-5 (Lund et al. 2004) and serves as a checkpoint to reduce premature pri-miRNAs from escaping the nucleus (Lund and Dahlberg 2006). In the RNAi pathway, RNA transport out of the nucleus appears to be limiting as overexpression of exportin-5 enhances interference (Yi et al. 2005).

Once in the cytoplasm, pre-miRNAs undergo further processing by another large endoribonuclease appropriately called Dicer (Bernstein et al. 2001). The core functional domains of Dicer consist of a PAZ (Piwi-Argonaute-Zwille) domain that binds the 2nt 3' overhang and 5' phosphate ends of the Drosha-processed hairpin, tandem RNase III domains, and a ruler domain that separates the PAZ and RNase domains (MacRae et al. 2006; Lau et al. 2012; Tian et al. 2014). Additionally, some homologs contain a N-terminal helicase domain that is implicated in regulation of RNA processing and translocation of Dicer along longer dsRNA substrates (Kidwell et al. 2014; Cenik et al. 2011; Welker et al. 2011). To generate miRNAs of defined length, Dicer binds the end of the pre-miRNA in the PAZ domain and orients the RNA along its length with

the dual-RNase III domains cleaving at the beginning of the loop region of the hairpin, liberating a ~22nt mature miRNA duplex (MacRae et al. 2006; Zhang et al. 2002; Hammond 2005). Cleavage is off-set and generates a second 2 nt 3' overhang that mirrors the Drosha-generated end of the pre-miRNA (Ando et al. 2011).

Additional proteins factors can interact with Dicer and influence RNA processing or interactions between the downstream effector of RNAi, Argonaute (Wilson and Doudna 2013; Lee et al. 2013). In humans, two double-stranded RNA binding proteins (dsRBPs) differentially affect processing and selection (Lee et al. 2013). *Trans*-activation response RNA-binding protein (TRBP) influences miRNA length and plays in substrate binding and discrimination (Lee and Doudna 2012; Chakravarthy et al. 2010; Chendrimada et al. 2005; Haase et al. 2005). Protein activator of PKR (PACT) interacts with Dicer and biases selection of miRNA isoforms (isomiRs) (Lee et al. 2013). Both proteins can form a large, RISC-loading complex (RLC) with Dicer and Argonaute and can influence duplex presentation and asymmetric loading into Ago (Figure 1.10), which ultimately decides target-binding specificities (Noland and Doudna 2013; Noland et al. 2011; MacRae et al. 2008; Takahashi et al. 2013).

Argonaute proteins are loaded with duplex-miRNA/siRNA complexes (Nakanishi 2016). Successful assembly of the RISC complex requires the RLC or association with additional proteins like Hsc70/Hsp90 (Iki et al. 2010; Iwasaki et al. 2010; Miyoshi et al. 2010; Liu et al. 2003; Tomari et al. 2004b; 2004a) in some species, while Dicer is unnecessary for RISC loading in mammals (Betancur and Tomari 2012). Thermodynamic asymmetry of the duplex helps dictate correct guide versus passenger strand (Khvorova et al. 2003; Schwarz et al. 2003). The 5' end of the duplex that is less thermodynamically stable is loaded into a binding pocket of Ago, sometimes aided by dsRBPs (Noland and Doudna 2013; Noland et al. 2011; Wang et al. 2009a). Passenger strand ejection is still not well understood but activation of Ago requires removal of the passenger strand and is thought to proceed by cleavage of the non-guide strand by catalytically active Agos or nuclease activity by the endonuclease C3PO (Matranga et al. 2005; Ye et al. 2011).

Some miRNAs destined for Ago shortcut through one or multiple steps of the canonical processing pathway. Short-hairpin RNAs (shRNAs) expressed endogenously or exogenously skip processing by Drosha and are directly exported into the cytoplasm by exportin-5 (Babiarz et al. 2008; Siolas et al. 2005). Interestingly, miR-451 appears to be cleaved by Drosha but is independent of Dicer activity. Catalytically-active human Ago 2 (hAgo2) binds the miR-451 hairpin directly and cleaves the passenger strand to produce a functional miRNA (Cifuentes et al. 2010; Cheloufi et al. 2010).

1.3.2 Argonaute proteins : effectors of RNA-interference

Argonaute proteins are central to RNA-interference pathways in a wide variety of eukaryotic systems. In human, eight proteins belonging to two clades of Argonaute proteins exist: four Argonaute proteins (hAgo1 - hAgo4) and four Piwi clade proteins (Höck and Meister 2008). Of the Argonaute proteins, it is typically

thought that only hAgo2 retains catalytic activity (Liu et al. 2004; Schürmann et al. 2013), although one group has recently reported slicing activity of hAgo3 when loaded with specific guides (Park et al. 2017). Outside of Ago and Piwi homologs, an additional subclade of Ago proteins (WAGO) is limited to worms (Yigit et al. 2006; Hutvagner and Simard 2008) (Figure 1.11)

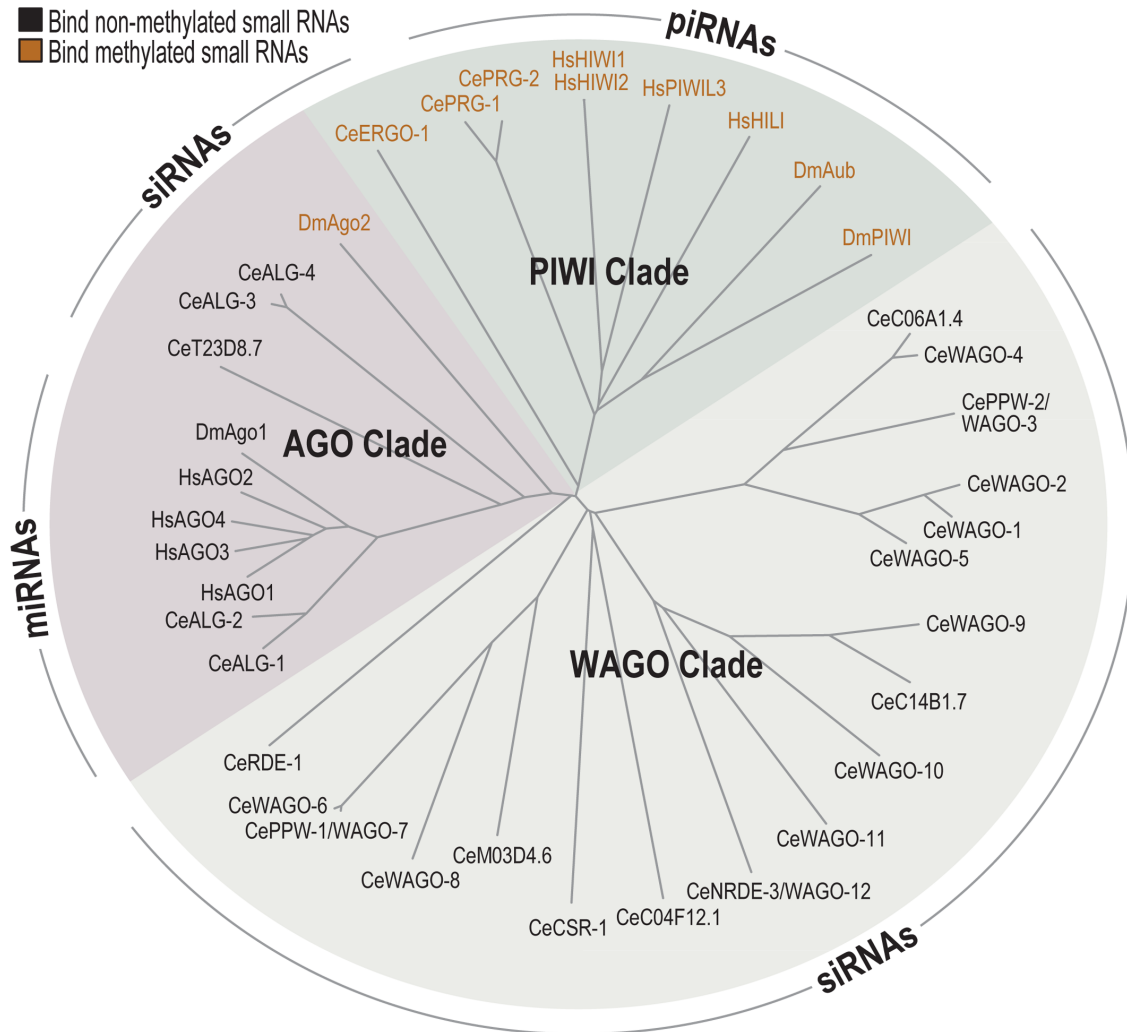


Figure 1.11 Clades of eukaryotic Argonaute proteins

Eukaryotic Argonautes split into three major groups. The Ago clade is characterized by miRNA and siRNA binding towards RNAi, while the Piwi clade binds piRNAs for transposon silencing. Lastly, the WAGO clade is specific to worms and bind siRNAs for RNA silencing. Leaves are labeled with an abbreviated name for the individual protein. Figure adapter from (Montgomery et al. 2012).

Structural and biochemical information for Ago proteins originally came from studies on prokaryotic Ago (pAgo) homologs (Song et al. 2004; Tian et al. 2014; Rivas et al. 2005; Rashid et al. 2007), although these typically bind DNA

instead of RNA guides. Structures of Ago bound to a DNA guide revealed pre-ordering of nucleotides 2-8 in an α -helical arrangement proposed to facilitate target strand annealing (Wang et al. 2008a; 2008b; 2009b; Parker et al. 2004; 2009). Ago from yeast revealed a 'glutamate finger' that plugs into the active site upon guide binding completing the catalytic tetrad (Nakanishi et al. 2012). Structures of hAgo2, the active human slicer, were solved in guide and target bound states (Schirle and MacRae 2012; Schirle et al. 2014; Elkayam et al. 2012). Comparison of prokaryotic, yeast, and human Agos demonstrates the remarkable overall conservation of structure amongst homologs from across the tree of life (Sasaki and Tomari 2012). Differences in structures are mainly restricted to loop regions that serve as platforms for species-specific interactions with partner proteins that aid in translational repression (Takimoto et al. 2009; Till et al. 2007).

Sharing a common architecture, Agos consist of four domains (Song et al. 2004; Schirle and MacRae 2012; Jinek and Doudna 2009) that serve two major functions. Binding to the 5' and 3' end of the guide are the middle (MID) and PAZ domains (Parker et al. 2005; Yuan et al. 2005; Ma et al. 2005), which is similar to the PAZ domain in Dicer (MacRae et al. 2006). Slicing activity is mediated through an RNase H-like fold in the PIWI domain (P element-induced wimpy testis) if it possesses an intact catalytic tetrad – DEDX, where X is either D/H/N (Meister et al. 2004; Nakanishi et al. 2012; Kaya et al. 2016; Sheng et al. 2014). Both guide binding and cleavage activity require divalent metal cations (Sheng et al. 2014; Jinek and Doudna 2009). The fourth domain is a N-terminal domain (N) that can act a wedge to facilitate passenger strand ejection during RISC assembly (Kwak and Tomari 2012). Overall, Ago forms a bilobed architecture with the N and PAZ domains constituting one lobe and the MID and PIWI domains forming the other.

Comparison of guide and target bound Ago structures as well as biochemical and single-molecular FRET studies have elucidated the general targeting mechanism of Ago proteins (Herzog and Ameres 2015; Wee et al. 2012; Salomon et al. 2015; Chandradoss et al. 2015; Jo et al. 2015; Yao et al. 2015). Ago divides its guide into several distinct portions, including the anchor (nt 1), the seed region (nt 2 – 8), the central region (nt 9 – 12), the 3' supplementary region (nt 13 – 16), and the 3' tail (nt 17 – 21) (Wee et al. 2012). A target search model proposes that Ago scans cellular RNA with nt 2 – 5 of the seed, facilitated by one-dimensional diffusion (Salomon et al. 2015; Chandradoss et al. 2015; Jo et al. 2015). Productive binding with this 'sub-seed' sequence induces conformational changes in Ago to allow extended pairing with the target (Schirle and MacRae 2012; Schirle et al. 2014). Ago-target binding with weak seed pairing can be strengthened by interactions with the 3' supplementary region (Salomon et al. 2015; Helwak et al. 2013). Full pairing of the guide and target requires release of the 3' end of the guide from the PAZ domain (Jung et al. 2013; Wang et al. 2009b), and causes protein rearrangements that activate the catalytic site (Schirle et al. 2014; Wang et al. 2009b). Release of the products is unordered but ejection of one half of the target facilitates removal of the second (Yao et al. 2015; Salomon et al. 2015; Jo et al. 2015). Overall, compared to the

kinetics of protein-free RNA binding, Ago improves the speed of target identification by ~250-fold (Salomon et al. 2015; Chandradoss et al. 2015; Jo et al. 2015).

1.3.3 Prokaryotic Argonautes in genome defense

Bacteria and archaea also contain Ago homologs; however, they do not possess other proteins involved in eukaryotic RNA interference pathways (Makarova et al. 2009; Swarts et al. 2014b; Burroughs et al. 2014; 2013). In contrast to our detailed understanding of eukaryotic Agos (eAgos), the physiological function of prokaryotic Agos (pAgos) has remained a mystery. From early structural work, pAgo proteins seem to preferentially associate with guide DNAs, whereas eAgos are exclusively RNA-guided (Wang et al. 2008a; 2008b; Yuan et al. 2005). Present in ~10% of bacteria and archaea, pAgo homologs often occur in genome defense islands and were thus proposed to act as key components of nucleic acid-guided genome defense systems (Makarova et al. 2009; Swarts et al. 2014b).

Initial evidence to support a proposed role in genome defense came from Ago in the bacteria, *Rhodobacter sphaeroides* (RsAgo) (Olovnikov et al. 2013). Analysis of co-purifying nucleic acids revealed that RsAgo binds a broad range of RNAs from the transcriptome and a population of DNAs complementary to the RNAs. A strong U-preference in the 5' end of the RNAs suggested these were the guides (Olovnikov et al. 2013), as Agos typically display a bias at this position due to specific interactions with the MID-domain guide binding pocket (Kawamata and Tomari 2010). DNA molecules were enriched in plasmid-derived sequences and mobile genetic elements, supporting a role in genome defense. Likewise, expression of RsAgo in *E. coli* caused degradation of plasmid DNA (Olovnikov et al. 2013); however, RsAgo does not possess an active catalytic site and how it is mediating plasmid interference is a mystery. Structural work demonstrated RsAgo's preference for a heteroduplex is dictated by recognition of duplex topology in the seed region of the guide (Miyoshi et al. 2016).

Two additional pAgo homologs were demonstrated to have DNA-guided DNA-interference and reduce rates of natural competency in their native hosts (Swarts et al. 2014a; 2015a). Unlike RsAgo, Ago from *Thermus thermophilus* (TtAgo) possesses an intact catalytic site and was shown to nick or linearize plasmids *in vitro* using DNA guides (Swarts et al. 2014a). TtAgo does not seem to act in RNA interference or affect gene regulation (Swarts et al. 2014a; 2015b). Similarly, the same group showed that an archaeal Ago protein interferes with foreign plasmids *in vivo*, but also reports guide-independent DNA cleavage activity *in vitro* (Swarts et al. 2015a).

Prokaryotic Argonaute proteins might interface with other genome defense systems in the cell (Makarova et al. 2013; 2012). Indeed, pAgo from *Marinitoga piezophila* (MpAgo) forms an operon with the adaptation module of a CRISPR-Cas system (Kaya et al. 2016). MpAgo displays a unique binding preference for 5'-hydroxylated guides, opposed to 5'-phosphorylated guides typical to all other Ago proteins. Furthermore, this Ago displays uncommon flexibility in preferences for nucleic acid targeting, using RNA guides for single-stranded DNA or RNA

binding and cleavage (Kaya et al. 2016; Lapinaite et al. 2018; Doxzen and Doudna 2017). Demonstrated guide and target preferences for various Agos are summarized in Table 1.1.

Argonaute	Guide	Target	Reference
Eukaryotic	RNA	RNA	(Meister 2013)
<i>R. sphaeroides</i> (RsAgo)	RNA	DNA	(Olovnikov et al. 2013; Miyoshi et al. 2016)
<i>A. aeolicus</i> (AaAgo)	DNA	RNA	(Yuan et al. 2005)
<i>T. thermophilus</i> (TtAgo)	DNA	DNA/RNA	(Swarts et al. 2014a)
<i>M. jannaschii</i> (MjAgo)	DNA	DNA	(Zander et al. 2017; Willkomm et al. 2017)
<i>P. furiosus</i> (PfAgo)	DNA	DNA	(Swarts et al. 2015a)
<i>M. piezophila</i> (MpAgo)	RNA	DNA/RNA	(Kaya et al. 2016)
<i>T. profunda</i> (TpAgo)	RNA	DNA	(Kaya et al. 2016)

Table 1.1: Experimentally determined Ago guide and target preferences

Outstanding questions in their field were how pAgos generate and select guides as well as how pAgos preferentially target foreign genetic elements. Recently, a dual-DNase activity for TtAgo was reported (Swarts et al. 2017a). In addition to DNA-guided DNA interference activity (Swarts et al. 2014a), TtAgo possesses a general DNase activity that chops foreign DNA, generating a pool of potential guides that are selectively loaded onto TtAgo for guided interference. A similar dual-DNase activity was reported in an archaeal Ago from *Methanocaldococcus jannaschii* (MjAgo) (Zander et al. 2017; Willkomm et al. 2017); however, chromatinization of DNA renders the genome resistant to MjAgo non-guided DNA endonuclease activity (Zander et al. 2017).

Chapter 2

Discovery of CRISPR-Cas9 systems in archaea.

A portion of the content presented in this chapter has been previously published as part of the following research article: Burstein, D., Harrington, L.B., Strutt, S.C., Probst, A.J., Anantharaman, K., Thomas, B.C., Doudna, J.A., and Banfield, J.F. (2017). New CRISPR-Cas systems from uncultivated microbes. *Nature* 542, 237–241.

For published studies: D.B., L.B.H., and S.C.S. were designated co-first authors. D.B., L.B.H., S.C.S., J.A.D. and J.F.B. designed the study and wrote the manuscript. A.J.P., K.A., J.F.B., B.T.C. and D.B. assembled the data and reconstructed the genomes. D.B., L.B.H., S.C.S. and J.F.B. computationally analysed the CRISPR–Cas systems. S.C.S. designed and executed the experimental work with ARMAN Cas9.

2.1 Chapter Summary

CRISPR–Cas systems provide microbes with adaptive immunity by employing short DNA sequences, termed spacers, that guide Cas proteins to cleave foreign DNA (Barrangou et al. 2007; Sorek et al. 2008). Class 2 CRISPR–Cas systems are streamlined versions, in which a single RNA-bound Cas protein recognizes and cleaves target sequences (Makarova et al. 2015; Shmakov et al. 2015). The programmable nature of these minimal systems has enabled researchers to repurpose them into a versatile technology that is broadly revolutionizing biological and clinical research (Barrangou and Doudna 2016). However, current CRISPR–Cas technologies are based solely on systems from isolated bacteria, leaving the vast majority of enzymes from organisms that have not been cultured untapped. Metagenomics, the sequencing of DNA extracted directly from natural microbial communities, provides access to the genetic material of a huge array of uncultivated organisms (Sharon and Banfield 2013; Brown et al. 2015). Here, using genome-resolved metagenomics, we identify the first reported Cas9 in the archaeal domain of life, to our knowledge. This divergent Cas9 protein was found in little-studied nanoarchaea as part of an active CRISPR–Cas system. *In silico* reconstruction and analysis of these archaeal CRISPR-systems revealed highly diverse CRISPR spacer arrays with many spacers matching viruses and mobile genetic elements of other community members. Despite computational identification of necessary DNA and RNA elements for targeting, cleavage activity could not be demonstrated through *in vitro* and *in vivo* experiments. Nevertheless, interrogation of environmental microbial communities allows us to access an unprecedented diversity of genomes, the content of which will expand the repertoire of microbe-based biotechnologies.

2.2 Introduction

The sequencing of microbial genomes provides access to a large inventory of microbial genes, many of which encode hypothetical proteins or -RNAs of undetermined function. The CRISPR-Cas system, an example of a pathway that was unknown to science prior to the DNA sequencing era, is now understood to confer bacteria and archaea with acquired immunity against phage and viruses (Barrangou et al. 2007; Sorek et al. 2008). CRISPR-Cas systems consist of Cas proteins, which are involved in acquisition, targeting and cleavage of foreign genetic material, and a CRISPR array comprising direct repeats flanking short spacer sequences that guide Cas complexes to their targets. Class 2 CRISPR-Cas systems are streamlined versions in which a single Cas protein bound to RNA is responsible for recognizing and cleaving a targeted sequence (Makarova et al. 2015; Shmakov et al. 2015). The programmable nature of these minimal systems has enabled their use as a versatile technology that is revolutionizing the fields of genetics and molecular biology, animal and plant biology and clinical medicine (Barrangou and Doudna 2016).

Metagenomics, an approach involving the sequencing of DNA extracted from complex natural samples, has brought into focus the enormous variety of proteins for which functional insights are lacking (Brown et al. 2015; Jaroszewski

et al. 2009). Ecosystems that are now being probed by metagenomics include extreme environments and the deeper subsurface, which are populated by many organisms that have never been studied in a laboratory setting. Genomes reconstructed from metagenomic data capture sequences of microbial defense systems and the entities that they target (e.g., phage and plasmids). As such, these methods are ideal to explore both diversity and functionality of CRISPR-Cas systems. Here we analyzed metagenomic samples of various microbial communities from groundwater, sediments, and acid mine drainage seeking Class 2 CRISPR-Cas systems that are not represented among cultured organisms. This led to the identification of the first Cas9 proteins in Archaea, expanding the occurrence of these proteins to a new domain of life. The archaeal Cas9 homologs are encoded exclusively by organisms belonging to lineages that have no known isolated representatives. Despite identification of necessary components for DNA targeting, no activity for the archaeal Cas9 proteins can be demonstrated *in vitro* or *in vivo*.

2.3 Methods

2.3.1 Metagenomics and metatranscriptomics

Metagenomic samples from three different sites were analysed: (1) AMD samples collected between 2006 and 2010 from the Richmond Mine, Iron Mountain, California (Denef and Banfield 2012; Miller et al. 2011). (2) Groundwater and sediment samples collected between 2007 and 2013 from the Rifle Integrated Field Research (IFRC) site, adjacent to the Colorado River near Rifle, Colorado (Brown et al. 2015; Anantharaman et al. 2016). (3) Groundwater collected in 2009 and 2014 from Crystal Geyser, a cold, CO₂-driven geyser on the Colorado Plateau in Utah (Probst et al. 2017).

For the AMD data, DNA extraction methods and short read sequencing were as described (Denef and Banfield 2012; Miller et al. 2011). For the Rifle data, DNA extraction, sequencing, assembly and genome reconstruction were as described (Brown et al. 2015; Anantharaman et al. 2016). For samples from Crystal Geyser, methods were as described (Probst et al. 2017; Emerson et al. 2016). Rifle metatranscriptomic data were used from ref. (Brown et al. 2015).

In brief, DNA was extracted from samples using the PowerSoil DNA Isolation Kit (MoBio Laboratories Inc.). RNA was extracted from 0.2- μ m filters collected from six 2011 Rifle groundwater samples. Following RNA extraction using the Invitrogen TRIzol reagent, DNA was removed with the Qiagen RNase-Free DNase Set and Qiagen Mini RNeasy kits, and cDNA template library was generated using the Applied Biosystems SOLiD Total RNA-Seq kit. DNA was sequenced on Illumina HiSeq2000 platform, and Metatranscriptomic cDNA on 5500XL SOLiD platform after emulsion clonal bead amplification using the SOLiD EZ Bead system (Life Technologies). For the Crystal Geyser data and reanalysis of the AMD data, sequences were assembled using IDBA-UD (Peng et al. 2012). DNA and RNA (cDNA) read-mapping used to determine sequencing coverage and gene expression, respectively, was performed using Bowtie2 (Langmead and Salzberg 2012). Open reading frames were predicted on assembled

scaffolds using Prodigal (Hyatt et al. 2010). Scaffolds from the Crystal Geyser dataset were binned on the basis of differential coverage abundance patterns using a combination of ABAWACA (Brown et al. 2015), ABAWACA2 (<https://github.com/CK7>), Maxbin2 (Wu et al. 2016) and tetranucleotide frequency using Emergent Self-Organizing Maps (ESOM) (Dick et al. 2009). Genomes were manually curated using percent GC content, taxonomic affiliation and genome completeness. Scaffolding errors were corrected using ra2.py (<https://github.com/christophertbrown>).

2.3.2 CRISPR–Cas computation analyses

The assembled contigs from the various samples were scanned for known Cas proteins using hidden Markov model (HMM) profiles, which were built using the HMMer suite (Finn et al. 2011), based on alignments from refs (Makarova et al. 2015; Shmakov et al. 2015). CRISPR arrays were identified using a local version of the CrisprFinder software (Grissa et al. 2007). Loci that contained both Cas1 and a CRISPR array were further analysed if one of the ten open reading frames adjacent to the cas1 gene encoded for an uncharacterized protein larger than 800 amino acids, and no known cas interference genes were identified on the same contig. These large proteins were further analysed as potential class 2 Cas effectors. The potential effectors were clustered to protein families based on sequence similarities using MCL (Enright et al. 2002). These protein families were expanded by building HMMs representing each of these families, and using them to search the metagenomic datasets for similar Cas proteins. To compare the identified protein families to known proteins, homologues were searched using BLAST (Camacho et al. 2009) against the NCBI non-redundant (nr) and metagenomic (env_nr) protein databases, as well as HMM searches against the UniProt KnowledgeBase (UniProt Consortium 2015; Finn et al. 2011). Only proteins with no full-length hits (>25% of the protein's length) were considered novel proteins. Distant homology searches of the putative Cas proteins were performed using HHpred from the HH-suite (Remmert et al. 2011). High scoring HHpred hits were used to infer domain architecture based on comparison to solved crystal structures (Dong et al. 2016; Yamano et al. 2016), and secondary structure that was predicted by JPred4 (Drozdetskiy et al. 2015). Protein modeling was performed using Phyre2 (Kelley et al. 2015).

Spacer sequences were determined from the assembled data using CrisprFinder (Grissa et al. 2007). CRASS (Skenneron et al. 2013) was used to locate additional spacers in short DNA reads of the relevant samples. Spacer targets (protospacers) were then identified by BLAST (Camacho et al. 2009) searches (using 'task blastn-short') against the relevant metagenomic assemblies for hits with ≤ 1 mismatch to spacers. Hits belonging to contigs that contained an associated repeat were filtered out (to avoid identifying CRISPR arrays as protospacers). Identified Cas9 proteins and their CRISPR-array summaries can be found in Table 2.1. Protospacer adjacent motifs (PAMs) were identified by aligning regions flanking the protospacers and visualized using WebLogo (Crooks et al. 2004). In cases where one spacer had multiple putative protospacers with different compositions of flanking nucleotides, each distinct

combination of protospacer and downstream nucleotides was taken into account for the logo calculation. RNA structures were predicted using mFold (Zuker 2003). Average nucleotide identity was computed with the pyani Python module (<https://github.com/widdowquinn/pyani>), using the Mummer (Kurtz et al. 2004) method. CRISPR array diversity was analysed by manually aligning spacers, repeats and flanking sequences from the assembled data. Manual alignments and contig visualizations were performed with Geneious 9.1.

For the phylogenetic analyses of Cas1 and Cas9, we used proteins of the newly identified systems along with the proteins from refs (Makarova et al. 2015; Shmakov et al. 2015). A non-redundant set was compiled by clustering together proteins with $\geq 90\%$ identity using CD-HIT (Fu et al. 2012). Alignments were produced with MAFFT (Katoch and Standley 2013), and maximum-likelihood phylogenies were constructed using RAxML (Stamatakis 2014) with PROTGAMMALG as the substitution model and 100 bootstrap samplings. Cas1 tree were rooted using the branch leading to casposons. Trees were visualized using FigTree 1.4.1 (<http://tree.bio.ed.ac.uk/software/figtree/>) and iTOL v3 (Letunic and Bork 2016).

Protein	NCBI accession	Coordinates	Repeat length	# of spacers	Spacer avg. length
ARMAN-1 Cas9	MOEG00000013	1827..7130	36	271*	34.5
ARMAN-4 Cas9	KY040241	11779..14900	36	1	36

Table 2.1 Archaeal CRISPR-Cas9 homolog summary

* ARMAN-1 spacers were reconstructed from 16 samples

2.3.3 ARMAN-Cas9 protein expression and purification

Expression constructs for Cas9 from ARMAN-1 (AR1) and ARMAN-4 (AR4) were assembled from gBlocks (Integrated DNA Technologies) that were codon-optimized for *E. coli*. The assembled genes were cloned into a pET-based expression vector as an N-terminal His₆-MBP or His₆ fusion protein. Expression vectors were transformed into BL21(DE3) *E. coli* cells and grown in LB broth at 37 °C. For protein expression, cells were induced during mid-log phase with 0.4 mM IPTG (isopropyl β -D-1-thiogalactopyranoside) and incubated overnight at 16 °C. All subsequent steps were conducted at 4 °C. Cell pellets were resuspended in lysis buffer (50 mM Tris-HCl, 500 mM NaCl, 1 mM TCEP, 10 mM Imidazole, 0.5% Triton X-100; pH 8) and supplemented with Complete protease inhibitor mixture (Roche) before lysis by sonication. Lysate was clarified by centrifugation at 15,000g for 40 min and applied to Superflow Ni-NTA agarose (Qiagen) in batch. The resin was washed extensively with wash buffer A (50 mM Tris-HCl, 500 mM NaCl, 1 mM TCEP, 10 mM imidazole; pH 8) followed by 5 column volumes of wash buffer B (50 mM Tris-HCl, 1 M NaCl, 1 mM TCEP, 10 mM imidazole; pH 8). Protein was eluted off of Ni-NTA resin with elution buffer (50 mM Tris-HCl, 500 mM NaCl, 1 mM TCEP, 300 mM Imidazole; pH 8). The His₆-MBP tag was removed by TEV protease during overnight dialysis against

wash buffer A. Cleaved Cas9 was removed from the affinity tag through a second Ni-NTA agarose column. The protein was dialysed into IEX buffer A (50 mM Tris-HCl, 300 mM NaCl, 1 mM TCEP, 5% glycerol; pH 7.5) before application to a 5-ml Heparin HiTrap column (GE Life Sciences). Cas9 was eluted over a linear NaCl (0.3–1.5 M) gradient. Fractions were pooled and concentrated with a 30-kDa spin concentrator (Thermo Fisher). When applicable, Cas9 was further purified via size-exclusion chromatography on a Superdex 200-pg column (GE Life Sciences) and stored in IEX buffer A for subsequent cleavage assays. For yeast expression, AR1-Cas9 was cloned into a Gal1/10 His₆-MBP TEV Ura S. *cerevisiae* expression vector (Addgene plasmid 48305). The vector was transformed into a BY4741 URA3 strain and cultures were grown in synthetic media (5 g L⁻¹ ammonium sulfate, 1.7 g L⁻¹ nitrogen base (Sunrise Science), 0.72 g L⁻¹ complete supplement mixture – ura (Sunrise Science), 20 g L⁻¹ glucose, 1.5% glycerol, 2% lactic acid) at 30 °C. At an OD₆₀₀ of approximately 0.6, protein expression was induced with 2% (w/v) galactose and incubated overnight at 16 °C. Protein purification was performed as above.

2.3.4 RNA *in vitro* transcription and oligonucleotide purification

In vitro transcription reactions were performed as previously described (Sternberg et al. 2012) using synthetic DNA templates containing a T7 promoter sequence. All *in vitro* transcribed putative guide RNA sequences and target RNA or DNA were purified via denaturing PAGE. Double-stranded target RNA and DNA were hybridized in 20 mM Tris HCl (pH 7.5) and 100 mM NaCl by incubation at 95 °C for 1 min, followed by slow-cooling to room temperature. Hybrids were purified by native PAGE. RNA and DNA sequences used in this study are listed in Table 2.2.

Designation	Sequence (5' to 3' unless noted)
Ar1-crRNA DNA targets	GGUAACCGAAGUAAGAUUAACCUUACAAUCGACACUAAAUAUU UGCAUGUGUAAG
Tracr-69 Ar1	GGCAUGGACCAUAUCCAGGUGUUGAUUGUAAACACUUAGCGGGUG AAAUUAUAUAUGUUUGUAAUAUCUUC
Tracr-104 Ar1	GGCAUGGACCAUAUCCAGGUGUUGAUUGUAAACACUUAGCGGGUG AAAUUAUAUAUGUUUGUAAUAUCUUCACUAUCCAAAGUUAUCUCUG GUUUUGGUUUGGUA
Tracr-179 Ar1	GGCAUGGACCAUAUCCAGGUGUUGAUUGUAAACACUUAGCGGGUG AAAUUAUAUAUGUUUGUAAUAUCUUCACUAUCCAAAGUUAUCUCUG GUUUUGGUUUGGUAAGCUUCACUUCACUAUUGUUUUCACUCCCAA UUUGAGUAUACUUGACUUUAACCAUGC UUUCGGGGAGUGCUUUUA
Ar1-crRNA GFP <i>in vivo</i>	CAUCUAAUUCAACAAGAAUUCUUACAAUCGACACUUGCACCGAGUC GGUGCUUUUUUU
sgRNA-69 Ar1	CAUCUAAUUCAACAAGAAUUCUUACAAUCGACACUAAAACAGGUGU UGAUUGUAAAGAAACACUUAGCGGGUGAAAUAUAUAUGUUUGUAA UAUCUUCGCACCGAGUCGUGCUUUUUUU
sgRNA-104 Ar1	CAUCUAAUUCAACAAGAAUUCUUACAAUCGACACUAAAACAGGUGU UGAUUGUAAAGAAACACUUAGCGGGUGAAAUAUAUAUGUUUGUAA UAUCUUCACUAUCCAAAGUUAUCUCUGGUUUUGGUUUGGUAAGCAC CGAGUCGGUGCUUUUUUU

sgRNA-179 Ar1	CAUCUAAUUCACCAAGAAUUCUUUCAAAUAAACAAAUAAAUCUUAGU AAUAUGUAAACGCACCCGAGUCGGUGCUUUUUUUU
Ar4-crRNA GFP <i>in vivo</i>	CAUCUAAUUCACCAAGAAUUCUUUCAAAUAAACAAAUAAAUCUUAGU AAUAUGUAAACGCACCCGAGUCGGUGCUUUUUUUU
sgRNA-75 Ar4	CAUCUAAUUCACCAAGAAUUCUUUCAAAUAAACAAAUAAAAACUUAU UUGUUUAAUUGAAAGAAGCCUAGACGUUAGGCACCCGAGUCGGUGCU UUUUUU
sgRNA-122 Ar4	CAUCUAAUUCACCAAGAAUUCUUUCAAAUAAACAAAUAAAAACUUAU UUGUUUAAUUGAAAGAAGCCUAGACGUUAGGGUUCGCGUGCAUGUU AGGCUCACGACAGGUACCUCCGUUUAACCUAGCACCCGAGUCGGUG CUUUUUUU
dsDNA target	non-target strand 5'-3' AGCAGAAATCTCTGCTGGCCACCTTCGTTTATAACCGAAGTAAGATT ATAATGGAGTACAAACGTCAGCT target-strand 3'-5' TCGTCTTTAGAGACGACCGGGTGAAGCAAATATTGGCTTCATTCTAA TATTACCTCATGTTTGCACTCGA
ssDNA target	TCGTCTTTAGAGACGACCGGGTGAAGCAAATATTGGCTTCATTCTAA TATTACCTCATGTTTGCACTCGA
dsDNA bubble target	non-target strand 5'-3' AGCAGAAATCTCTGCTCCGGGTGAAGCAAATATTGGCTTCATTCTAA TATTTGGAGTACAAACGTCAGCT target-strand 3'-5' TCGTCTTTAGAGACGACCGGGTGAAGCAAATATTGGCTTCATTCTAA TATTACCTCATGTTTGCACTCGA
dsRNA target	non-target strand 5'-3' GGAGCAGAAAUCUCUGCUGGCCACCUUCGUUUUAUACCGAAGUAA GAUUUAUUGGAGUACAAACGUCAGCUCC target-strand 3'-5' CCUCGUCUUUAGAGACGACCGGGUGGAAGCAAUUAUUGGCUUCAU UCUAAUAUJACCUCUUGUUGCAGUCGAGG
ssRNA target	CCUCGUCUUUAGAGACGACCGGGUGGAAGCAAUUAUUGGCUUCAU UCUAAUAUJACCUCUUGUUGCAGUCGAGG

Table 2.2: DNA and RNA sequences used in this study

Sequence in bold indicates region of crRNA and sgRNA that are complementary to the appropriate target.

2.3.5 *In vitro* cleavage assays

Purified DNA and RNA oligonucleotides were radiolabelled using T4 polynucleotide kinase (NEB) and [γ - 32 P] ATP (Perkin-Elmer) in 1× PNK buffer for 30 min at 37 °C. PNK was heat inactivated at 65 °C for 20 min and free ATP was removed from the labelling reactions using illustra Microspin G-25 columns (GE Life Sciences). crRNA and tracrRNA were mixed in equimolar quantities in 1× refolding buffer (50 mM Tris HCl pH 7.5, 300 mM NaCl, 1 mM TCEP, 5% glycerol) and incubated at 70 °C for 5 min and then slow-cooled to room temperature. The reactions were supplemented to 1 mM final metal concentration and subsequently heated at 50 °C for 5 min. After slow-cooling to room temperature, refolded guides were placed on ice. Unless noted for buffer or salt concentration, Cas9 was reconstituted with an equimolar amount of guide in 1× cleavage buffer (50 mM Tris HCl pH 7.5, 300 mM NaCl, 1 mM TCEP, 5% glycerol, 5 mM divalent

metal) at 37 °C for 10 min. Cleavage reactions were conducted in 1× cleavage buffer with a 10× excess of Cas9-guide complex over radiolabelled target at 37 °C or the indicated temperature. Reactions were quenched in an equal volume of gel loading buffer supplemented with 50 mM EDTA. Cleavage products were resolved on 10% denaturing PAGE and visualized by phosphorimaging.

2.3.6 *In vivo E. coli* interference assays

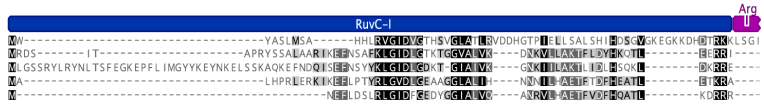
E. coli transformation assays for ARMAN-1 Cas9 and ARMAN-4 Cas9 were conducted as previously published (Oakes et al. 2016). Briefly, *E. coli* transformed with plasmids expressing guide RNA sequences were made electrocompetent. Cells were then transformed with 9 fmol of plasmid encoding wild-type or catalytically inactive Cas9 (dCas9). A dilution series of recovered cells was plated on LB plates with selective antibiotics. Colonies were counted after 16 h at 37 °C.

2.4 Results

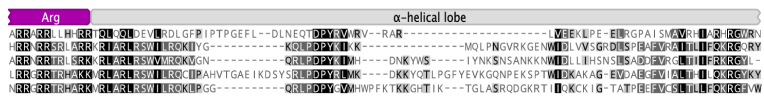
2.4.1 Identification of archaeal Cas9

We sought to identify previously unknown class 2 CRISPR–Cas systems in terabase-scale metagenomic datasets from groundwater, sediment, acid-mine drainage (AMD) biofilms, soil, infant gut, and other microbial communities. Our analyses targeted large uncharacterized genes proximal to a CRISPR array and *cas1*, the universal CRISPR integrase (Levy et al. 2015; Yosef et al. 2012; Nuñez et al. 2015b). Among the 155 million protein-coding genes analysed, we identified the first Cas9 proteins in domain Archaea. One of the hallmarks of CRISPR–Cas9 (type II) systems was their presumed presence only in the bacterial domain (Chylinski et al. 2014; Makarova et al. 2015). We were therefore surprised to discover Cas9 proteins encoded in genomes of the nanoarchaea ‘*Candidatus Micrarchaeum acidiphilum* ARMAN-1’ and ‘*Candidatus Parvarchaeum acidiphilum* ARMAN-4’ (Baker et al. 2010; 2006) in AMD metagenomic datasets (Table 2.1 and Figure 2.1). These findings expand the occurrence of Cas9-containing CRISPR systems to another domain of life.

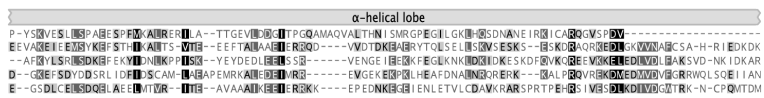
1. cas9|Actinomyces_naeslundii|EJN84392.1
2. novel_Cas9|ARMAN1|IM0E00000013|1827..4676
3. novel_Cas9|ARMAN4|KY040241|11779..14682
4. novel_Cas9|Deltaproteobacteria|OGP65397.1
5. novel_Cas9|Lindowbacteria|OGH57617.1



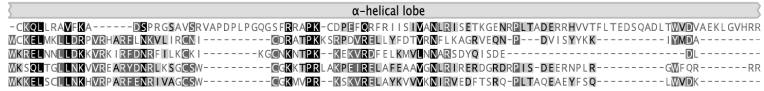
1. cas9|Actinomyces_naeslundii|EJN84392.1
2. novel_Cas9|ARMAN1|IM0E00000013|1827..4676
3. novel_Cas9|ARMAN4|KY040241|11779..14682
4. novel_Cas9|Deltaproteobacteria|OGP65397.1
5. novel_Cas9|Lindowbacteria|OGH57617.1



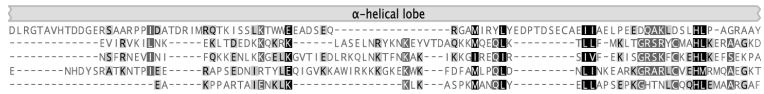
1. cas9|Actinomyces_naeslundii|EJN84392.1
2. novel_Cas9|ARMAN1|IM0E00000013|1827..4676
3. novel_Cas9|ARMAN4|KY040241|11779..14682
4. novel_Cas9|Deltaproteobacteria|OGP65397.1
5. novel_Cas9|Lindowbacteria|OGH57617.1



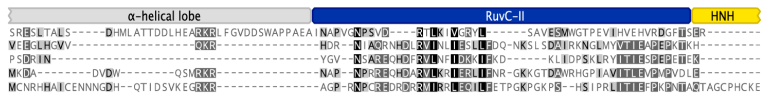
1. cas9|Actinomyces_naeslundii|EJN84392.1
2. novel_Cas9|ARMAN1|IM0E00000013|1827..4676
3. novel_Cas9|ARMAN4|KY040241|11779..14682
4. novel_Cas9|Deltaproteobacteria|OGP65397.1
5. novel_Cas9|Lindowbacteria|OGH57617.1



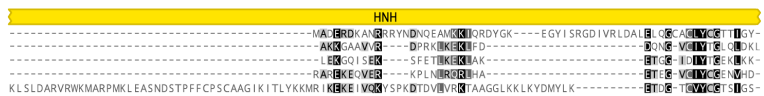
1. cas9|Actinomyces_naeslundii|EJN84392.1
2. novel_Cas9|ARMAN1|IM0E00000013|1827..4676
3. novel_Cas9|ARMAN4|KY040241|11779..14682
4. novel_Cas9|Deltaproteobacteria|OGP65397.1
5. novel_Cas9|Lindowbacteria|OGH57617.1



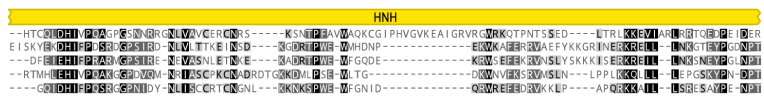
1. cas9|Actinomyces_naeslundii|EJN84392.1
2. novel_Cas9|ARMAN1|IM0E00000013|1827..4676
3. novel_Cas9|ARMAN4|KY040241|11779..14682
4. novel_Cas9|Deltaproteobacteria|OGP65397.1
5. novel_Cas9|Lindowbacteria|OGH57617.1



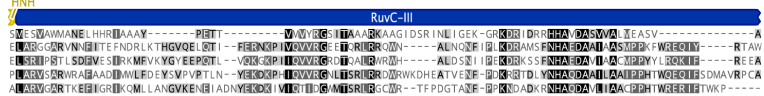
1. cas9|Actinomyces_naeslundii|EJN84392.1
2. novel_Cas9|ARMAN1|IM0E00000013|1827..4676
3. novel_Cas9|ARMAN4|KY040241|11779..14682
4. novel_Cas9|Deltaproteobacteria|OGP65397.1
5. novel_Cas9|Lindowbacteria|OGH57617.1



1. cas9|Actinomyces_naeslundii|EJN84392.1
2. novel_Cas9|ARMAN1|IM0E00000013|1827..4676
3. novel_Cas9|ARMAN4|KY040241|11779..14682
4. novel_Cas9|Deltaproteobacteria|OGP65397.1
5. novel_Cas9|Lindowbacteria|OGH57617.1



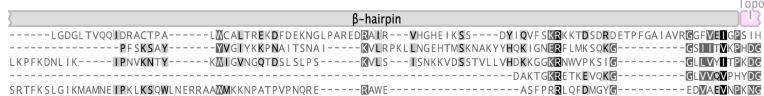
1. cas9|Actinomyces_naeslundii|EJN84392.1
2. novel_Cas9|ARMAN1|IM0E00000013|1827..4676
3. novel_Cas9|ARMAN4|KY040241|11779..14682
4. novel_Cas9|Deltaproteobacteria|OGP65397.1
5. novel_Cas9|Lindowbacteria|OGH57617.1



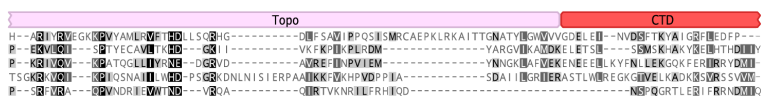
1. cas9|Actinomyces_naeslundii|EJN84392.1
2. novel_Cas9|ARMAN1|IM0E00000013|1827..4676
3. novel_Cas9|ARMAN4|KY040241|11779..14682
4. novel_Cas9|Deltaproteobacteria|OGP65397.1
5. novel_Cas9|Lindowbacteria|OGH57617.1



1. cas9|Actinomyces_naeslundii|EJN84392.1
2. novel_Cas9|ARMAN1|IM0E00000013|1827..4676
3. novel_Cas9|ARMAN4|KY040241|11779..14682
4. novel_Cas9|Deltaproteobacteria|OGP65397.1
5. novel_Cas9|Lindowbacteria|OGH57617.1



1. cas9|Actinomyces_naeslundii|EJN84392.1
2. novel_Cas9|ARMAN1|IM0E00000013|1827..4676
3. novel_Cas9|ARMAN4|KY040241|11779..14682
4. novel_Cas9|Deltaproteobacteria|OGP65397.1
5. novel_Cas9|Lindowbacteria|OGH57617.1



1. cas9|Actinomyces_naeslundii|EJN84392.1
2. novel_Cas9|ARMAN1|IM0E00000013|1827..4676
3. novel_Cas9|ARMAN4|KY040241|11779..14682
4. novel_Cas9|Deltaproteobacteria|OGP65397.1
5. novel_Cas9|Lindowbacteria|OGH57617.1

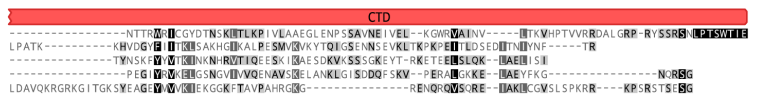


Figure 2.1 Multiple sequence alignment of newly described Cas9 proteins
 Alignment of Cas9 proteins from ARMAN-1 and ARMAN-4, as well as two closely related Cas9 proteins from uncultivated bacteria, to the *Actinomyces naeslundii* Cas9, whose structure has been solved (2014).

2.4.2 Description of archaeal CRISPR-Cas9 system from ARMAN-1

The CRISPR-Cas locus in ARMAN-1 includes large CRISPR arrays adjacent to the *cas1*, *cas2*, *cas4* and *cas9* genes. This system was found on highly similar contigs (an average nucleotide identity of 99.7% outside the CRISPR array) reconstructed independently from 16 different samples. We reconstructed numerous alternative ARMAN-1 CRISPR arrays with a largely conserved end (probably comprised of the oldest spacers) and a variable region into which many distinct spacers have been incorporated (Figure 2.2 A). Given the polarity of the array, we predict that the approximately 200-bp region between the end of the *Cas9* gene and the variable end of the array is likely to contain the leader sequence and transcriptional start site. On the basis of the hypervariability in spacer content, we conclude that the ARMAN-1 CRISPR-Cas9 system is active in the sampled populations.

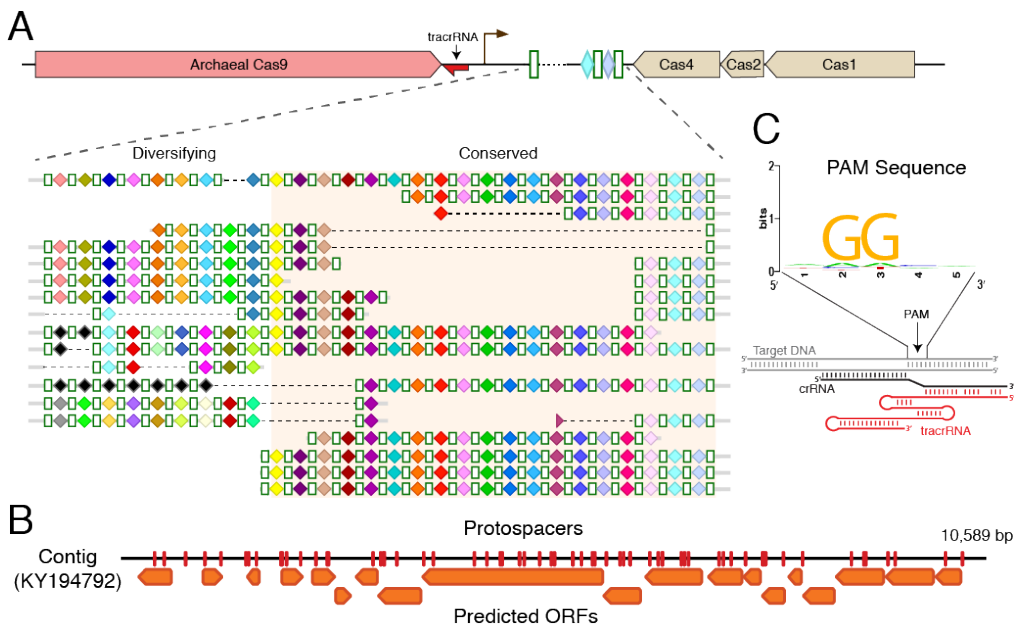


Figure 2.2 ARMAN-1 CRISPR array diversity and identification of the ARMAN-1 Cas9 PAM sequence (A) CRISPR arrays reconstructed from AMD samples. White boxes indicate repeats, colored diamonds indicate spacers (identical spacers are similarly colored; unique spacers are black). The conserved region of the array is highlighted. The diversity of recently acquired spacers (on the left) indicates that the system is active. (B) A single circular, putative viral contig contains 56 protospacers (red vertical bars) from the ARMAN-1 CRISPR arrays. (C) Sequence analysis of 240 protospacers revealed a conserved 'NGG' PAM downstream of the protospacers. ORF, open reading frame.

2.4.3 ARMAN-1 CRISPR array targets viral and mobile genetic element targets

Of the spacers of the ARMAN-1 CRISPR–Cas9 system, 56 target a 10-kbp circular sequence that encodes mostly short hypothetical proteins, and is probably an ARMAN-1 virus (Figure 2.2 B). Indeed, cryo-electron tomographic reconstructions have often identified viral particles attached to ARMAN cells (Baker et al. 2010). ARMAN-1 protospacers also derived from a putative transposon within the genome of ARMAN-2 (another nanoarchaeon (Baker et al. 2006) and a putative mobile element in the genomes of *Thermoplasmales archaea*, including that of I-plasma (Yelton et al. 2013) from the same ecosystem (Figure 2.3). Direct cytoplasmic ‘bridges’ were observed between ARMAN and *Thermoplasmales* cells, implying a close relationship between them (Baker et al. 2010). The ARMAN-1 CRISPR–Cas9 may therefore defend against transposon propagation between these organisms, a role that is reminiscent of piwi-interacting-RNA-mediated defence against transposition in the eukaryotic germ line (Vagin et al. 2006).

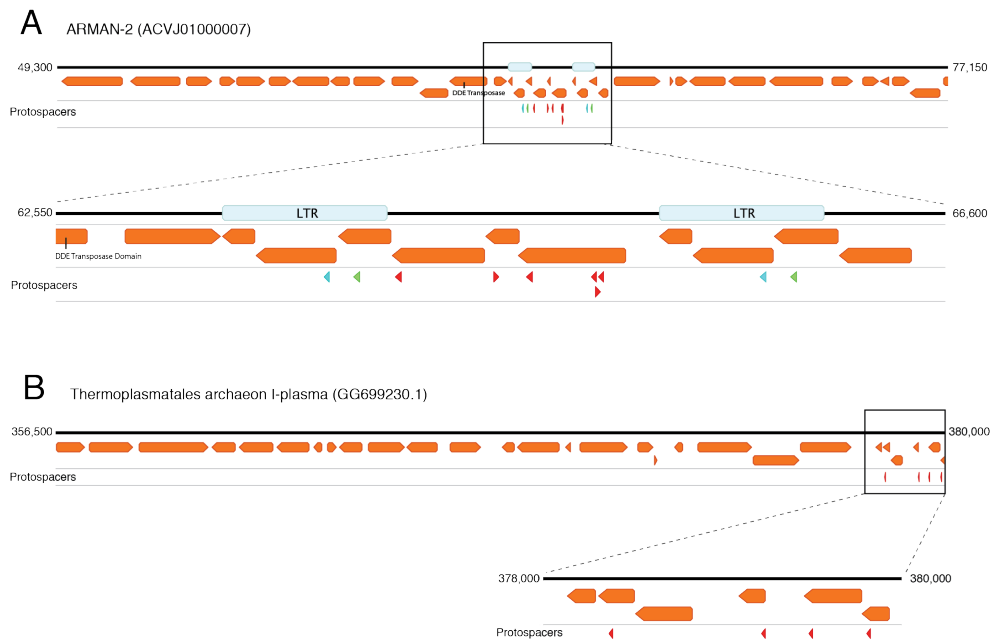


Figure 2.3 ARMAN-1 spacers map to genomes of archaeal community members (A) Protospacers from ARMAN-1 map to the genome of ARMAN-2, a nanoarchaeon from the same environment. Six protospacers (red arrowheads) map uniquely to a portion of the genome flanked by two long-terminal repeats (LTRs), and two additional protospacers match perfectly within the LTRs (blue and green arrowheads). This region is likely to be a transposon, suggesting that the CRISPR–Cas system of ARMAN-1 plays a role in suppressing mobilization of this element. (B) Protospacers also map to a *Thermoplasmales* archaeon (I-plasma), another member of the Richmond Mine ecosystem that is found in the same samples as ARMAN organisms. The protospacers cluster within a region of

the genome encoding short, hypothetical proteins, suggesting this might also represent a mobile element. NCBI accession codes are provided in parentheses.

2.4.4 CRISPR-Cas9 from ARMAN-4 is a reduced system

Unlike the ARMAN-1 CRISPR–Cas system, the ARMAN-4 cas9 gene has only one adjacent CRISPR repeat-spacer unit and no other cas genes in its vicinity (Figure 2.4). The lack of a typical CRISPR array and of cas1 points to a system with no capacity to acquire additional spacers. No target could be identified for the spacer sequence, but given the conservation of the locus in samples collected over several years, we cannot rule out the possibility that it is functional as a ‘single-target’ CRISPR–Cas system. Conservation of a single spacer may indicate that the ARMAN-4 Cas9 exerts an alternative role, such as gene regulation (Stern et al. 2010) or involvement in cell–cell interactions (Zegans et al. 2009).

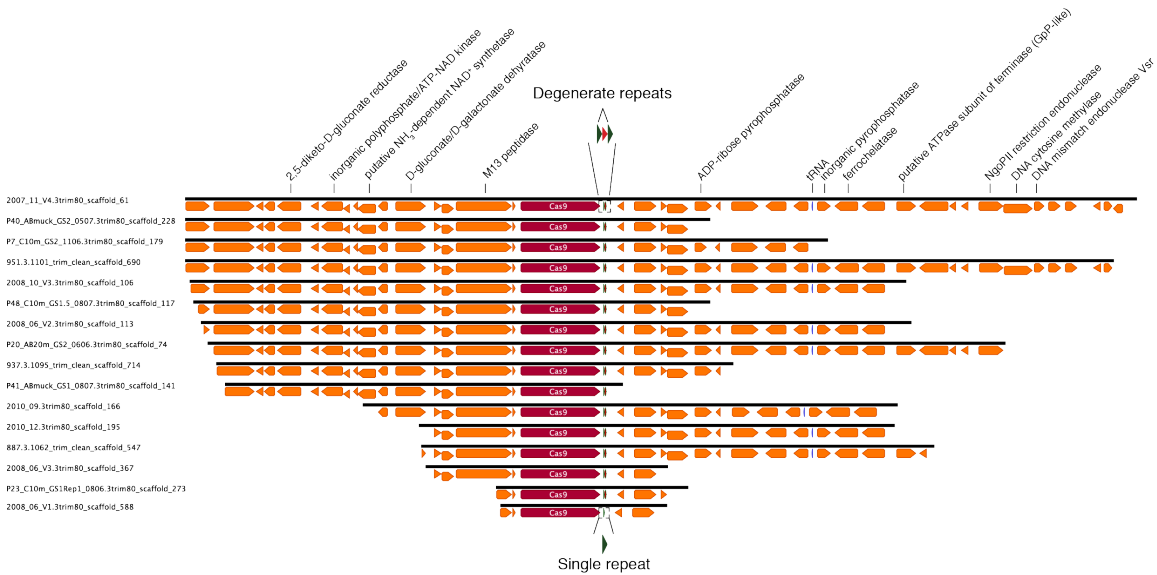


Figure 2.4 Archaeal Cas9 from ARMAN-4 with a degenerate CRISPR array is found on numerous contigs Cas9 from ARMAN-4 is highlighted in dark red on 16 nearly identical contigs from different samples. Proteins with putative domains or functions are labeled, whereas hypothetical proteins are unlabeled. Fifteen of the contigs contain two degenerate direct repeats (36 nucleotides long with one mismatch) and a single conserved spacer of 36 nucleotides. The remaining contig contains only one direct repeat. Unlike ARMAN-1, no additional Cas proteins are found adjacent to Cas9 in ARMAN-4.

2.4.5 Identification of targeting requirements

Active DNA-targeting CRISPR–Cas systems use 2–4-nucleotide protospacer-adjacent motifs (PAMs) located next to target sequences for self-versus-non-self discrimination (Shah et al. 2013; Anders et al. 2014). Examining

sequences adjacent to the genomic target sequences revealed a strong 'NGG' PAM preference in ARMAN-1 (Figure 2.2 C). Cas9 also employs two separate transcripts, CRISPR RNA (crRNA) and trans-activating CRISPR RNA (tracrRNA), for RNA-guided DNA cleavage (Jinek et al. 2012). We identified a putative tracrRNA in the vicinity of both ARMAN-1 and ARMAN-4 CRISPR–Cas9 systems (Figure 2.5 A-D). It has previously been suggested that type II CRISPR systems were absent from archaea owing to a lack of the host factor, RNase III, responsible for crRNA–tracrRNA guide complex maturation (Chylinski et al. 2014; Deltcheva et al. 2011). Notably, no RNase III homologues were identified in the ARMAN-1 genome (estimated to be 95% complete) and no internal promoters have been predicted for the CRISPR array (Zhang et al. 2015), suggesting an as-yet-undetermined mechanism of guide RNA production.

AR1-Cas9 and AR4-Cas9 purifications were separated on a 10% SDS-PAGE gel.

2.4.6 Biochemical reconstitution of archaeal Cas9

Biochemical experiments to test the cleavage activity of ARMAN-1 and ARMAN-4 Cas9 proteins purified from both *E. coli* and yeast (Figure 2.5 E and F) did not reveal any detectable activity, nor did *in vivo E. coli*-targeting assays (Table 2.3 and Table 2.4). Given the unique physiology and ecological niche of these nanoarchaea, the lack of activity may be due to a post-translational modification or a co-factor absent from the experimental expression systems.

Protein Purification	Buffer	Salt (mM)	Metal	Guide	Target	Temp. °C
AR1-Cas9 #1	Tris pH 7.5	300	Mg ²⁺ Mn ²⁺ Zn ²⁺	crRNA cr:69 cr:104 cr:179	dsDNA ssDNA DNA Bubble ssRNA dsDNA	37
AR1-Cas9 #1	Tris pH 7.5	100-500	Mg ²⁺	cr:69 cr:104 cr:179	dsDNA	37
AR1-Cas9 #1	Tris pH 7.5	300	Mg ²⁺ Mn ²⁺ Zn ²⁺	cr:69 cr:104 cr:179	dsDNA	30-48
AR1-Cas9 #1	MOPS: pH 6 pH 6.5 pH 7.0 pH 7.5	300	Mg ²⁺	cr:69 cr:104 cr:179	dsDNA	37
AR1-Cas9 #1	Citrate: pH 5 pH 5.5 pH 6	300	Mg ²⁺	cr:69 cr:104 cr:179	dsDNA	37
AR1-Cas9 #1	Tris pH 7.5	300	Mg ²⁺ Mn ²⁺ Zn ²⁺	cr:69 cr:104 cr:179	plasmid	37-50
AR1-Cas9 #2	Tris pH 7.5	300	Mg ²⁺ Mn ²⁺ Zn ²⁺	cr:69 cr:104 cr:179	dsDNA	37
AR1-Cas9 #3	Tris pH 7.5	300	Mg ²⁺ Mn ²⁺ Zn ²⁺	cr:69 cr:104 cr:179	dsDNA	37
AR1-Cas9 #4	Tris pH 7.5	300	Mg ²⁺ Mn ²⁺ Zn ²⁺	cr:69 cr:104 cr:179	dsDNA	37
AR1-Cas9 #5	Tris pH 7.5	300	Mg ²⁺ Mn ²⁺ Zn ²⁺	cr:69 cr:104 cr:179	dsDNA	37
AR1-Cas9 #6	Tris pH 7.5	300	Mg ²⁺ Mn ²⁺ Zn ²⁺	cr:69 cr:104 cr:179	ssDNA dsDNA	37

AR4-Cas9 #1	Tris ph7.5	300	Mg ²⁺ Mn ²⁺ Zn ²⁺	sgRNA- 122	dsDNA	37
AR4-Cas9 #2	Tris ph7.5	300	Mg ²⁺ Mn ²⁺ Zn ²⁺	sgRNA- 122	dsDNA	37
AR4-Cas9 #3	Tris ph7.5	300	Mg ²⁺ Mn ²⁺ Zn ²⁺	sgRNA- 122	dsDNA	37
AR4-Cas9 #4	Tris ph7.5	300	Mg ²⁺ Mn ²⁺ Zn ²⁺	sgRNA- 122	dsDNA	37

Table 2.3 *In vitro* cleavage conditions assayed for Cas9 from ARMAN-1 and ARMAN-4. Note: grey boxes highlight the differences from standard assays.

Protein	Guide
AR1-Cas9 dAR1-Cas9	crRNA sgRNA-69 sgRNA-104 sgRNA-179
AR4-Cas9 dAR4-Cas9	crRNA sgRNA-75 sgRNA-122

Table 2.4 *In vivo E. coli* targeting assays

2.4.7 Apparent recombination in archaeal Cas9 CRISPR system

Archaeal Cas9 homologs are among the smallest examples (~950aa) of this diverse protein family described yet (Makarova et al. 2015; Chylinski et al. 2014). Phylogenetic analysis of Cas1 collected from all CRISPR systems (See Methods) suggests that this archaeal CRISPR–Cas system does not clearly fall into any existing type II subtype (Figure 2.6 A). The presence of cas4 affiliates it with type II-B systems (Chylinski et al. 2014; Makarova et al. 2015), yet the Cas9 sequence is more similar to type II-C proteins (Figure 2.7). Thus, the archaeal type II system may have arisen as a fusion of type II-C and II-B systems (Figure 2.7 B).

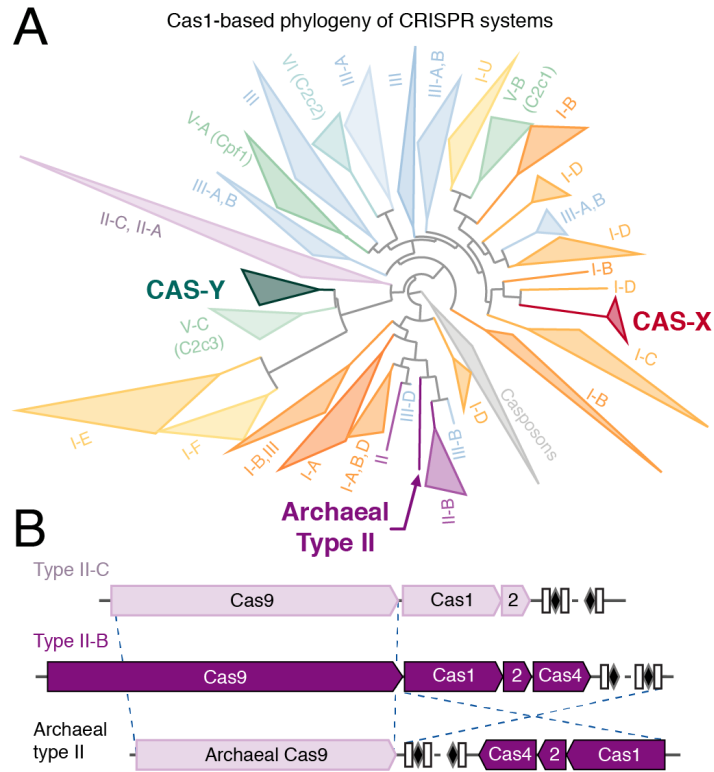


Figure 2.6 Novelty of the reported CRISPR–Cas systems (A) Simplified phylogenetic tree of the universal Cas1 protein. CRISPR types of known systems are noted on the wedges and branches; the newly described archaeal Cas9 systems are indicated. (B) Proposed evolutionary scenario that gave rise to the archaeal type II system as a result of a recombination between type II-B and type II-C loci.

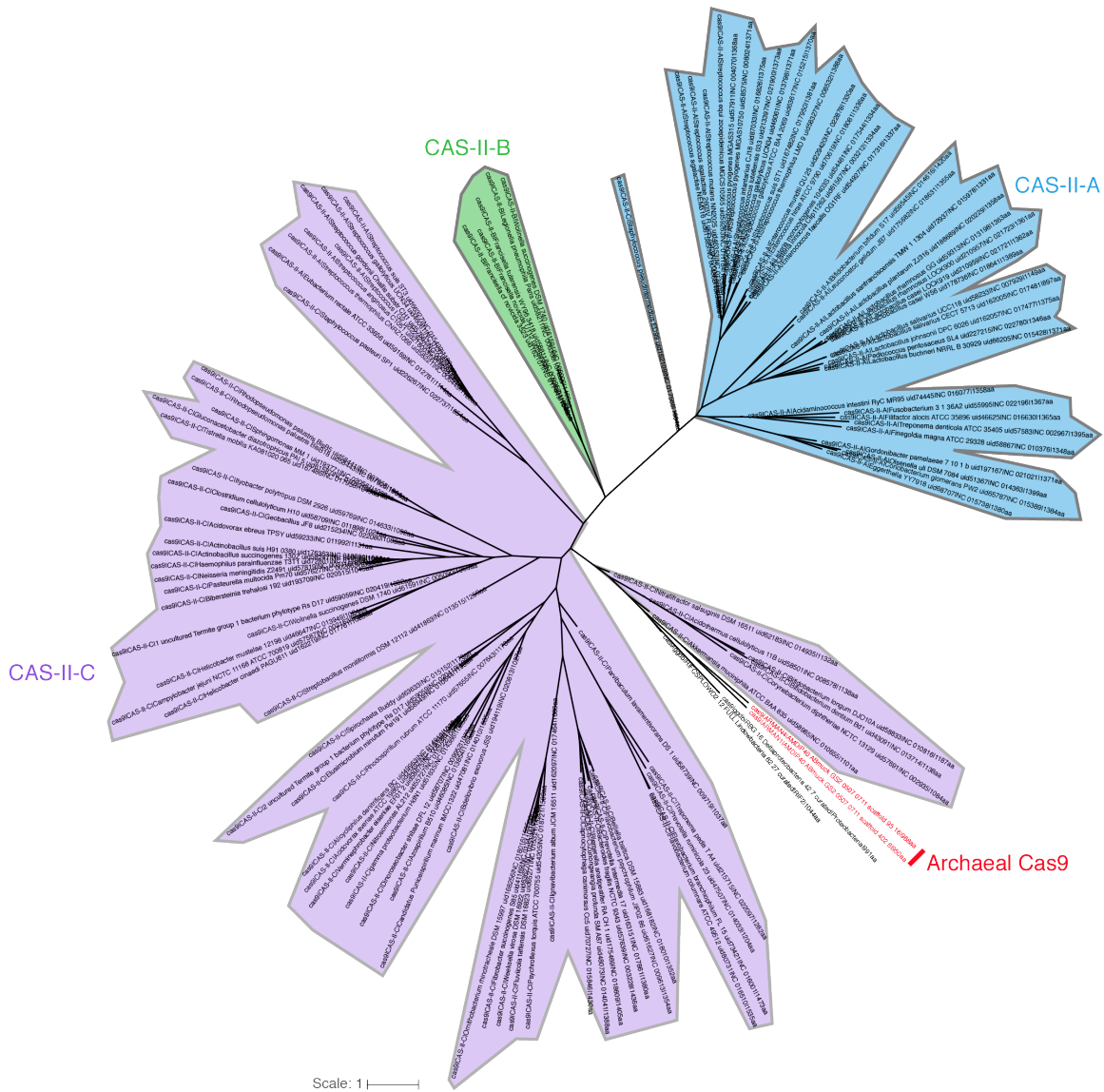


Figure 2.7 Evolutionary tree of Cas9 homologues Maximum-likelihood phylogenetic tree of Cas9 proteins, showing the previously described systems coloured based on their type. II-A, blue; II-B, green; II-C, purple. The archaeal Cas9s (red) cluster with type II-C CRISPR–Cas systems, together with two newly described bacterial Cas9 from uncultivated bacteria.

2.5 Discussion and Conclusion

The new CRISPR-Cas9 systems described here expand our understanding of the diversity and evolution of CRISPR adaptive immunity. We describe an apparently active system that targets viruses and mobile genetic elements that co-occur in the microbial community (Figure 2.2 B and Figure 2.3). Spacer content in the CRISPR system from ARMAN-1 is highly dynamic, showing acquisition of different spacers over time and likely reflecting a high rate

of viral infection in this environment (Figure 2.2 A). This is in contrast to the lone spacer in ARMAN-4 that appears to be static over the same sampling period (Figure 2.4). The Cas1/Cas2 adaptation module is noticeably absent from the latter and explains the apparent lack of new spacer acquisition; however, conservation of the single spacer hints at a potential regulatory role in this organism. No activity could be detected for these systems despite the ability to pinpoint a PAM sequence and predict the tracrRNA (Figure 2.2 B, Figure 2.5, Table 2.3 and Table 2.4). It is tempting to speculate that archaeal Cas9 systems are very rare given the paucity of RNase III homologs in archaea (Makarova et al. 2015; Chylinski et al. 2014) and that some unknown consequence of the unique evolutionary history (Figure 2.6 B) of the AR1-Cas9 CRISPR system has allowed persistence in this archaeon.

The discovery of Cas9 in the archaeal domain of life highlights the power of metagenomics to explore microbial diversity beyond what has been identified using cultured organisms. These Cas9 proteins are the most compact observed to date and the archaea in which they are found have similarly small genomes. Since such organisms often depend on other community members for basic metabolic requirements, they are difficult to study by traditional cultivation-based methods. While no activity could yet be demonstrated, the small size of Cas9 proteins described here make these systems especially valuable for the development of new genome editing tools. In addition to exploring difficult to culture organisms, these techniques enable us to examine new genomes at a rate unmatched by sequence analysis of isolates. Given that virtually all environments where life exists can now be probed by metagenomic methods, this combined computational-experimental approach can greatly expand the diversity of known CRISPR-Cas systems and in turn provide new technologies for biological research and clinical applications.

Chapter 3

RNA-dependent RNA targeting by CRISPR-Cas9

Content presented in this chapter has been previously published as part of the following research article: Strutt, S.C., Torrez, R.M., Kaya, E., Negrete, O.A., Doudna, J.A. (2018). RNA-dependent RNA targeting by CRISPR-Cas9. *Elife* 7: e32724.

For published studies: S.C.S., O.A.N., and J.A.D. designed the study and wrote the manuscript. S.C.S., R.M.T., and E.K. conducted experiments with supervision from O.A.N. and J.A.D.

3.1 Chapter Summary

Double-stranded DNA (dsDNA) binding and cleavage by Cas9 is a hallmark of type II CRISPR-Cas bacterial adaptive immunity. All known Cas9 enzymes are thought to recognize DNA exclusively as a natural substrate, providing protection against DNA phage and plasmids. Here we show that Cas9 enzymes from both subtypes II-A and II-C can recognize and cleave single-stranded RNA (ssRNA) by an RNA-guided mechanism that is independent of a protospacer-adjacent motif (PAM) sequence in the target RNA. RNA-guided RNA cleavage is programmable and site-specific, and we find that this activity can be exploited to reduce infection by single-stranded RNA phage *in vivo*. We also demonstrate that Cas9 can direct PAM-independent repression of gene expression in bacteria. These results indicate that a subset of Cas9 enzymes have the ability to act on both DNA and RNA target sequences, and suggest the potential for use in programmable RNA targeting applications.

3.2 Introduction

Prokaryotic clustered regularly interspaced short palindromic repeat (CRISPR)-CRISPR-associated (Cas) systems provide immunity against plasmids and bacteriophage by using foreign DNA stored as CRISPR spacer sequences together with Cas nucleases to stop infection (Wright et al. 2016; Mohanraju et al. 2016). One such nuclease, Cas9 of the type II systems, employs a CRISPR RNA (crRNA) and a trans-activating crRNA (tracrRNA) to target spacer-complementary regions (protospacers) on the foreign genetic element to guide double-stranded DNA cleavage (Jinek et al. 2012). A protospacer adjacent motif (PAM) must also be present for the Cas9-RNA complex to bind and cleave DNA (Jinek et al. 2012; Gasiunas et al. 2012; Anders et al. 2014; Szczelkun et al. 2014). Combining the crRNA and tracrRNA into a chimeric, single-guide RNA (sgRNA) simplified the system for widespread adoption as a versatile genome editing technology (Jinek et al. 2012).

To date, both genetic and biochemical data support the conclusion that *in vivo*, Cas9 is exclusively a DNA-targeting enzyme. Nonetheless, multiple studies have harnessed Cas9 for RNA targeting under specific circumstances. For example, the *S. pyogenes* Cas9 (SpyCas9) can be supplied with a short DNA oligo containing the PAM sequence (a PAMmer) to induce single-stranded RNA (ssRNA) binding and cutting (O'Connell et al. 2014; Nelles et al. 2016). More recently, it was demonstrated that SpyCas9 could be used to target repetitive RNAs and repress translation in certain mRNAs in the absence of a PAMmer (Liu et al. 2016; Batra et al. 2017). A different Cas9 homolog from *Francisella novicida* (FnoCas9) has been implicated in degradation of a specific mRNA but through a mechanism independent of RNA-based cleavage (Sampson et al. 2013). Together with evidence that some Cas9 homologs can target single-stranded DNA substrates under some conditions (Ma et al. 2015; Zhang et al. 2015), these studies raised the possibility that certain Cas9 enzymes might have intrinsic RNA-guided RNA cleavage activity.

To determine whether evolutionarily divergent Cas9 homologs have a native capacity for programmable RNA targeting, we compared biochemical

behavior of enzymes from the three major Cas9 subtypes. This analysis revealed that certain type II-A and II-C Cas9s can bind and cleave single-stranded RNA sequences with no requirement for a PAM or PAMmer. Furthermore, we found that this activity can inhibit gene expression and confer moderate protection against infection by ssRNA phage through a mechanism reminiscent of RNA-guided DNA targeting. These results establish the utility of Cas9 for facile RNA-guided RNA targeting and suggest that this activity may have biological relevance in bacteria.

3.3 Methods

3.3.1 Phylogenetic tree construction and RNA folding

Cas9 homolog sequences were obtained from Chylinski and colleagues (Chylinski et al. 2014). A structure-guided alignment was produced using PROMALS3D (Pei et al. 2008) and a maximum-likelihood tree was inferred using PHYML3.0 (Guindon et al. 2010). The structure of the pUC ssRNA target was predicted using Mfold (Zuker 2003).

3.3.2 Protein purification

All proteins were expressed as His-MBP fusions (Addgene vector #29706) in *E. coli* strain BL21(DE3). Cells were grown to an OD₆₀₀ of 0.6-0.8, induced with 0.4M IPTG, and then incubated overnight at 16°C with shaking. Proteins were purified using Superflow Ni-NTA affinity resin (Qiagen, Valencia, CA), followed by a HiTrap HP Heparin column (GE Healthcare, Pittsburgh, PA) and gel filtration on a Superdex S200 (GE Healthcare, Pittsburgh, PA), as previously described (Jinek et al. 2012).

3.3.3 Oligonucleotide purification and radiolabeling

DNA oligonucleotides were synthesized by IDT (Coralville, IA). Target RNAs and sgRNAs were transcribed *in vitro* as previously described (Sternberg et al. 2012). DNA targets and *in vitro* transcribed RNAs were gel purified by 7M urea denaturing PAGE. Target RNAs and DNAs were 5' end-labeled with [γ -P32-ATP] by treatment with PNK (NEB, Ipswich, MA). T1 sequencing and hydrolysis ladders were prepared according to manufacturer's directions (Ambion, Grand Island, NY). A list of all sgRNAs and targets can be found in Table 3.1.

Name	Sequence from 5' -> 3'
<i>In vitro targets</i>	
ssRNA	GGGCAGCAUUCAACCAUAUGGCAUCCGCUUACAGCCAAGCUGUGA CC
dsRNA - TS	GGGCAGCAUUCAACCAUAUGGCAUCCGCUUACAGCCAAGCUGUGA CC
dsRNA - NTS	GGUCACAGCUUGGCUGUAAGCGGAUGCCAUAUGGUUGAAUGCUG CCC
ssDNA	GGGCAGCATTCAACCATATGGCATCCGCTTACAGCCAAGCTGTGAC C
dsDNA - TS	GGGCAGCATTCAACCATATGGCATCCGCTTACAGCCAAGCTGTGAC C

dsDNA - NTS	GGTCACAGCTTGGCTGTAAGCGGATGCCATATGGTTGAATGCTGCC C
PAMmer	CCATATGGTTGAATGCTGCC
PAMmer control	GAAGCCCTGAAAGACGCGCAG
NTS ssRNA	GGUCACAGCUUGGCUGUAAGCGGAUGCCAUAUGGUUGAAUGCUG CCC
ssRNA mutant anti-PAM	GGGCAGCCAAGUCCCAUAUGGCAUCCGCUUACAGCCAAGCUGUGA CC
dsRNA mutant anti-PAM – TS	GGGCAGCCAAGUCCCAUAUGGCAUCCGCUUACAGCCAAGCUGUGA CC
dsRNA mutant anti-PAM – NTS	GGUCACAGCUUGGCUGUAAGCGGAUGCCAUAUGGUUGAAUGCUG CCC
2nt 3' end of guide NTS	GGUCACAGCUUGGCUGUAAGCGGAUGCCAUAUAAUUGAAUGCUGC CC
6nt 3' end of guide NTS	GGUCACAGCUUGGCUGUAAGCGGAUGCCUUAUAAAUUGAAUGCUGC CC
12nt 3' end of guide NTS	GGUCACAGCUUGGCUGUAAGCGUACAGUAUAAAUUGAAUGCUGC CC
23nt Bubble/mutant anti- PAM - NTS	GGUCACAGCUUAAUGAAUUCAAUACAGUAUAAAUUGAAUGCUGC CC
2nt 5' end of guide NTS	GGUCACAGCUUAAACUGUAAGCGGAUGCCAUAUGGUUGAAUGCUGC CC
6nt 5' end of guide NTS	GGUCACAGCUUAAUGAAAAGCGGAUGCCAUAUGGUUGAAUGCUGC CC
12nt 5' end of guide NTS	GGUCACAGCUUAAUGAAUUCAAUUGCCAUAUGGUUGAAUGCUGC CC
pUC target RNA	GGGCAGCAUUCAACCAUAUGGCAUCCGCUUACAGCCAAGCUGUGA CC
ON target RNA	GGGAUUCAACCAUAUGGCAUCCGCUUACAGCCAAGCUG
OFF target RNA	GGGUUUUAUUUUGGAUUUGGAAACGAGAGUUUCUGGUCAUGAA
GFP2 dsDNA - TS	GGGCAGCTGTCCCCCATATGGCATCCGCTTACAGCCAAGCTGTGAC C
GFP2 dsDNA - NTS	GGTCACAGCTTGGCTGTAAGCGGATGCCATATGGGGACAGCTGCC C
GFP4 dsDNA - TS	GGGCAGCCTGTCCCCCATATGGCATCCGCTTACAGCCAAGCTGTGAC C
GFP4 dsDNA - NTS	GGTCACAGCTTGGCTGTAAGCGGATGCCATATGGGGACAGGCTGCC C
GFP6 dsDNA - TS	GGGCAGCTATTTGCCATATGGCATCCGCTTACAGCCAAGCTGTGAC C
GFP6 dsDNA - NTS	GGTCACAGCTTGGCTGTAAGCGGATGCCATATGGCAAATAGCTGCC C
<i>In vitro sgRNAs</i>	
Spy	GGCUGUAAGCGGAUGCCAUAAGUUUAGAGCUAGAAAUAGCAAGUU AAAUAAGGCUAGUCCGUUAUCAACUUGAAAAGUGGCACCGAGU CGGUGCUUUUUU
Sau	GGCUGUAAGCGGAUGCCAUAUGGUUUUAGUACUCUGGAAACAGA AUCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUAACUUG UUGGCGAGAUUUUU
Cje	GGCUGUAAGCGGAUGCCAUAUGUUUUAGUCCCUUUUUAAAUUUCU UUAUGGUAAAUAUAAUCUCAUAAGAAUUUAAAAGGGACUAAA AUAAGAGUUUGCGGGACUCUGCGGGUUACAUAUCCCUAAAACC GCUU

Fno	GGCUGUAAGCGGAUGCCAUAUGUUUCAGUUGCGCCGAAAGGCGC UCUGUAAUCAUUUAAAAGUAUUUUGAACGGACCUCUGUUUGACAC GUCUG
GFP repression targets	
Spy PAM adjacent	AATTCTTGTTGAATTAGATGGTGA
None PAM adjacent	CACTACTGGAAACTACCTGTTC
Sau PAM adjacent	ACCTATGGTGTTCATGCTTTTC
Dual PAM adjacent	ATTGGCGATGGCCCTGTCCTTT
GFP tiled sgRNAs	
RBS	CAUGGUACCUUUCUCCUCUUUAAGUUUUAGUACUCUGGAAACAGA AUCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUCAACUUG UUGGCGAGAUUUUUU
GFP1	CCAGUGAAAAGUUCUUCUCCUUUGUUUUAGUACUCUGGAAACAGA AUCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUCAACUUG UUGGCGAGAUUUUUU
GFP2	CACCAUCUAAUUCACAAGAAUUGUUUUAGUACUCUGGAAACAGAA UCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUCAACUUGU UGGCGAGAUUUUUU
GFP3	CAGAAAAUUUGUGCCCAUUAACAGUUUUAGUACUCUGGAAACAGAA UCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUCAACUUGU UGGCGAGAUUUUUU
GFP4	AGCAUCACCUUCACCCUCUCCACGUUUUAGUACUCUGGAAACAGA AUCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUCAACUUG UUGGCGAGAUUUUUU
GFP5	AAUUUAAGGGUGAGUUUUCGGUUGUUUUAGUACUCUGGAAACAGA AUCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUCAACUUG UUGGCGAGAUUUUUU
GFP6	GAACAGGUAGUUUCCAGUAGUGGUUUUAGUACUCUGGAAACAGA AUCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUCAACUUG UUGGCGAGAUUUUUU
GFP7	CGGGCAUGGCACUCUUGAAAAAGGUUUUAGUACUCUGGAAACAGA AUCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUCAACUUG UUGGCGAGAUUUUUU
GFP8	AAAGAUUAGUGCGUCCUGUACGUUUUAGUACUCUGGAAACAGA AUCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUCAACUUG UUGGCGAGAUUUUUU
GFP9	ACGCGUCUUGUAGGUCCGUCUAGUUUUAGUACUCUGGAAACAGA AUCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUCAACUUG UUGGCGAGAUUUUUU
GFP10	UUUAAAAUCAAUACCCUUUAACUGUUUUAGUACUCUGGAAACAGAA UCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUCAACUUGU UGGCGAGAUUUUUU
GFP11	GUUGAACGGAACCAUCUUAACGGUUUUAGUACUCUGGAAACAGA AUCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUCAACUUG UUGGCGAGAUUUUUU
GFP12	AUGUGUAAUCCAGCAGCAGUUAGUUUUAGUACUCUGGAAACAGA AUCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUCAACUUG UUGGCGAGAUUUUUU
GFP2 noncoding	AAUUCUUGUUGAAUUAGAUGGUGGUUUUAGUACUCUGGAAACAGA AUCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUCAACUUG UUGGCGAGAUUUUUU
GFP6 noncoding	CACUACUGGAAACUACCGUUCGUUUUAGUACUCUGGAAACAGA AUCUACUAAAACAAGGCAAAAUGCCGUGUUUAUCUCGUCAACUUG UUGGCGAGAUUUUUU

Table 3.1 List of sequences used in this study

3.3.4 *In vitro* cleavage assays

Cas9 was reconstituted with equimolar sgRNA in 1x cleavage buffer (20 mM Tris-HCl – pH 7.5, 200 mM KCl, 1 mM TCEP, 5% glycerol, 5 mM MgCl₂) for 10 min at 37°C, then immediately placed on ice. Cleavage reactions were conducted with 1 nM target and 10 nM reconstituted Cas9-sgRNA in 1x cleavage buffer unless otherwise noted. Structured RNA substrates were prepared by annealing two separate *in vitro* transcribed RNAs. The target strand was annealed with 10-fold excess of the non-target strand to ensure that all target is complexed prior to the cleavage reaction. Reactions were incubated at 37°C for the indicated time and quenched in Heparin-EDTA buffer (10 µg/ml heparin, 25 mM EDTA) at 25°C for 5 min. Reactions were diluted with 2x Formamide loading buffer and incubated at 95°C for 5 min prior to separation on a 15% denaturing 7M urea PAGE gel. Gels were dried overnight and exposed to a phosphor imaging screen (Amersham/GE Healthcare, Pittsburgh, PA). Results were visualized on a Typhoon (GE Healthcare, Pittsburgh, PA) and quantified in ImageQuantTL (v8.1, GE Healthcare, Pittsburgh, PA). The cleaved fraction of total signal was calculated independently for three separate experiments and were fit with a one-phase exponential decay model in Prism7 (GraphPad Software, La Jolla, CA).

3.3.5 Filter binding and electrophoretic mobility shift assays

Binding reactions consisted of 750 nM catalytically inactive SauCas9 reconstituted with sgRNA to the final concentrations indicated. Radiolabeled target RNA was added to a final concentration of 1nM and the reactions were incubated at 37°C for one hour. Bound probe was separated from unbound using a three-filter system on a vacuum manifold (Rio 2012). Membranes were allowed to dry prior to phosphor imaging and quantification. EMSAs were performed in the presence of 300 nM dSauCas9 and 1 nM radiolabeled target strand DNA pre-annealed in the presence of 10x non-target strand. Complexes were incubated at 37°C for one hour prior to separation on 6% non-denaturing PAGE. Gels were dried prior to phosphor imaging. Three independent experiments were performed and the fraction of bound out of total signal was calculated in ImageQuantTL. Binding isotherms were determined in Prism7 using a one-site binding model.

3.3.6 MS2 screen and plaque assay

All guides of length 20-23 nt antisense to the MS2 bacteriophage genome were synthesized (CustomArray Inc., Bothell, WA) and cloned into a guide expression vector (Oakes et al. 2016) modified with the SauCas9 sgRNA scaffold. XL1-Blue *E. coli* cells with a vector containing a tetracycline-inducible wtSauCas9 construct were made electrocompetent and transformed with the MS2-guide plasmid library in triplicate. Approximately 1 x 10⁶ transformants were grown for 30 min at 37°C with shaking prior to addition of antibiotics and 10nM anhydrotetracycline (aTc) (Sigma, St. Louis, MO) for protein induction. After an

additional 30 min of growth, cultures were split into three equal pools and treated with none, 3.3×10^6 , or 3.3×10^7 MS2 bacteriophage. After three hours of infection, cells were plated on LB-agar supplemented with antibiotics and incubated at 37°C for 16 hours. Plates were scraped with LB and plasmids were isolated using a MidiPrep kit (Qiagen, Valencia, CA), according to the manufacturer's protocol. High-throughput sequencing libraries were prepared by PCR amplification of the variable region of the guide plasmid. Dual unique-molecular identifiers (UMIs), included to separate true single-nucleotide mismatches, as well as duplicates, from PCR artifacts (Kou et al. 2016), were incorporated during a single round of PCR. Excess UMIs were removed by Exol digestion (Thermo Scientific, Waltham, MA) prior to library amplification and barcoding. Individual guides (Supplemental File 1) were cloned using oligonucleotides synthesized by IDT and co-transformed into XL1-Blue *E. coli* cells with the SauCas9 vector. Resistance to MS2 bacteriophage was conducted using a soft-agar overlay method (Abudayyeh et al. 2016) and plaque forming units (PFUs) were calculated. To minimize variability in plaquing efficiency, the same phage dilutions were used for all experiments.

3.3.7 MS2 survival and mismatch analysis

After applying a low-pass filter, reads were trimmed using cutadapt v. 1.14 (Martin 2011) and paired-end overlapping reads were merged using pandaseq (Masella et al. 2012) for error correction. Reads were mapped to the MS2 genome with bowtie2 v2.3.0 (Langmead and Salzberg 2012) using the “very-sensitive” option and de-duplicated based on the dual-UMI (Smith et al. 2017). Feature counts were obtained using HTSeq-count (Anders et al. 2015). Differential expression was calculated using standard pipelines implemented in “edgeR” (Robinson et al. 2010; McCarthy et al. 2012). Significantly enriched guides were defined as those with an FDR-corrected p-value < 0.05. Guides with a positive fold-change compared to the control were mapped to the MS2 genome and visualized using the “Sushi” package (Phanstiel et al. 2014). To examine for nucleotide composition bias, sequences of guides with a significant positive enrichment were aligned at the 3' end (PAM-proximal) and motifs were analyzed using the WebLogo server (Crooks et al. 2004). The distribution of log₂ fold-change values of significantly enriched guides plotted as box and whisker plots in Prism. The secondary structure of the MS2 genome was obtained from (Dai et al. 2017) and reads were mapped and visualized in Forna (Kerpedjiev et al. 2015). Log₂ fold-change values of single-nucleotide mismatch (SNP) guides for each treatment were partitioned by length and averaged at each position. High-throughput sequencing data accompanying this paper are available through the Sequencing Read Archive under the BioProject accession number PRJNA413805.

3.3.8 *E. coli* in vivo GFP repression

Based on the system outlined previously, SauCas9 was cloned into a tetracycline-inducible vector, while individual guides are under control of a constitutive promoter (Oakes et al. 2016). Plasmids were transformed into an *E.*

coli strain with a GFP reporter gene integrated into the chromosome (Qi et al. 2013). Cultures were grown in M9 medium supplemented with 0.4% w/v glucose to mid-log phase and diluted to an OD₆₀₀ of 0.05 prior to transfer to a Tecan Microplate reader (Tecan Systems, San Jose, CA). Protein expression was induced with 10 nM aTc. GFP and OD₆₀₀ were measured every ten minutes for at least 18 hours. Curves of GFP expression over time were fit with a logistic growth model in Prism. At 80% of the maximum value, or at least after 16 hours of growth, the GFP signal was normalized by cell density at OD₆₀₀. To account for effects of guide and protein expression, GFP/OD₆₀₀ was normalized to a null guide or null protein culture, respectively. As expression of different guides change GFP expression levels, the ratio between normalized RNP and guide values was taken to allow comparison of RNP-based repression across different guides. All experiments were conducted in triplicate and all graphing and quantitative analyses were conducted in Prism. Guide and target sequences can be found in Table 3.1.

3.4 Results

3.4.1 Cas9 catalyzes PAM-independent RNA-guided RNA cleavage

To assess whether divergent Cas9 enzymes can catalyze binding to and cleavage of RNA substrates by a mechanism distinct from that of double-stranded DNA cleavage, we tested homologs from the three major subtypes of Cas9 proteins for their ability to cleave single-stranded RNA *in vitro* (Figure 3.1 A, B and Figure 3.2 A-C). When programmed with a cognate sgRNA, *S. aureus* Cas9 (SauCas9) and *C. jejuni* Cas9 (CjeCas9) direct cleavage of RNA in the absence of a PAMmer (Figure 3.1 and Figure 3.2). No RNA cleavage was detected using SpyCas9, which requires a PAMmer for efficient RNA cleavage *in vitro* (O'Connell et al. 2014), or using *F. novicida* Cas9 (FnoCas9). While the cleavage efficiencies for both SauCas9 and CjeCas9 are indistinguishable (Figure 3.2 D), we focused on the activity of SauCas9 due to the abundance of mechanistic and structural data for this enzyme (Nishimasu et al. 2015; Ran et al. 2015; Kleinstiver et al. 2015).

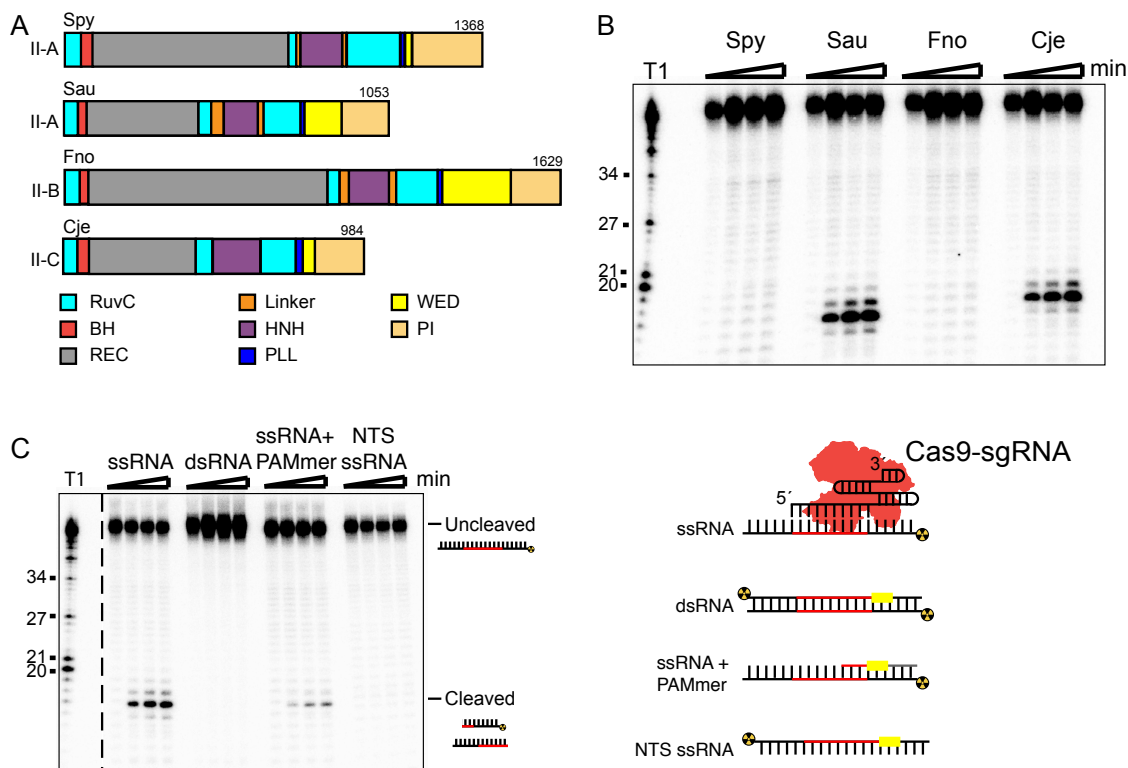


Figure 3.1 SauCas9 cleaves single-stranded RNA without a PAMmer
 (A) Schematic of Cas9 proteins tested for sgRNA mediated RNA cleavage. RuvC, RuvC nuclease domain; BH, bridge-helix; REC, recognition domain; HNH, HNH nuclease domain; PLL, phosphate-lock loop; WED, wedge domain; PI, PAM-interacting domain. Adapted from (Nishimasu et al. 2014; 2015; Hirano et al. 2016; Yamada et al. 2017). (B) Representative *in vitro* cleavage of ssRNA by Cas9-sgRNA RNP complexes of homologs in (A). Radiolabeled pUC target RNA was incubated with Cas9 RNP at 37°C and time points were taken at 0, 10, 30, and 60 min. Full time course is presented in Figure 3.2 B. T1 indicates size markers generated by RNase T1 digestion of ssRNA target. Size in nucleotides is indicated on the left. (C) (Left) *In vitro* cleavage assay of various RNA substrates (Right). Full time course is presented in Figure 3.4.

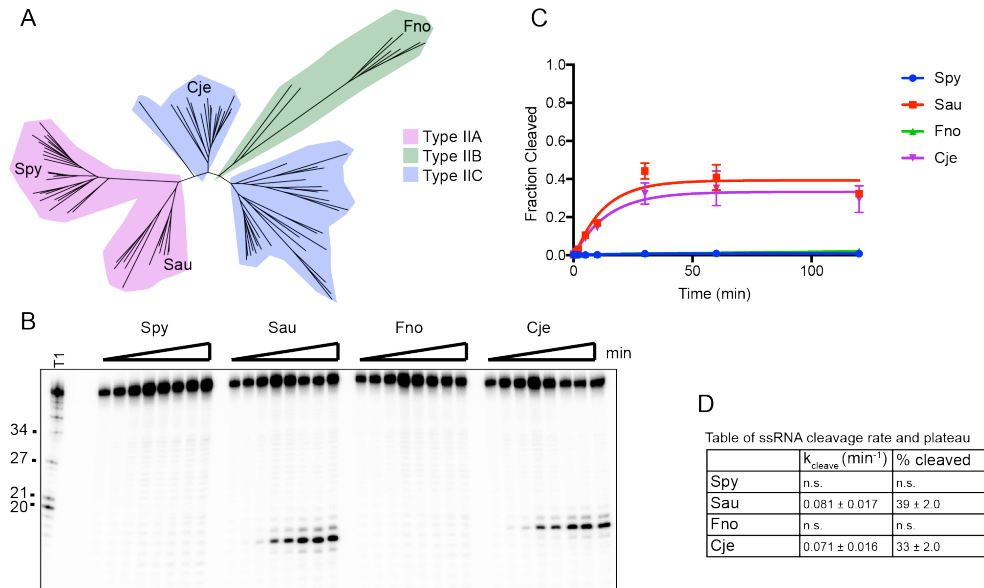


Figure 3.2 RNA is cleaved by SauCas9 and CjeCas9

(A) Phylogenetic tree of Cas9 homologs assayed for ssRNA cleavage activity. Tree was generated using homologs gathered from (Chylinski et al. 2014). Only homologs tested for activity are highlighted as leaves on the tree. Clades are colored by Cas9 sub-type. (B) Representative *in vitro* cleavage gel for ssRNA targeting by various Cas9 homologs in (A). Target used for cleavage was the pUC ssRNA. Time points are 0, 1, 2, 5, 10, 30, 60, and 120 mins. T1 RNase digest size fragments are given on the left. (C) Quantification of fraction cleaved in (B). Fit was determined in Prism using a single-exponential decay model. Error bars represent the mean \pm S.D. ($n=3$). (D) Apparent pseudo-first order fit parameters of the data in (C) where ‘% cleaved’ indicates the fraction of substrate cleaved when the reaction plateaus (mean \pm S.D.).

RNA cleavage activity and products were similar to those of canonical Cas9-mediated DNA cleavage activity *in vitro*. RNA targeting by SauCas9 requires the presence of a guide RNA and a catalytically-active protein, as both apo protein lacking the guide RNA and a catalytically inactive mutant (D10A and N580A) do not cleave RNA (Figure 3.3 A). Furthermore, addition of EDTA to chelate divalent metal ions abolished RNA cleavage, verifying that divalent metal ions are necessary for catalysis. As with DNA substrates (Sternberg et al. 2014), incubation of SauCas9 with an excess of RNA target demonstrated that cleavage is single-turnover (Figure 3.3 B and C). Hydrolysis mapping of the cleavage product revealed that the predominant RNA cleavage site is shifted by one nucleotide compared to the site of DNA cleavage (Garneau et al. 2010; Jinek et al. 2012; Gasiunas et al. 2012) (Figure 3.3 D and E). The shift is consistent with that observed for PAM-dependent SpyCas9 RNA-cleavage (O’Connell et al. 2014) and is likely due to the more compact geometry of an RNA-RNA helix relative to an RNA-DNA hybrid helix (Wang et al. 1982).

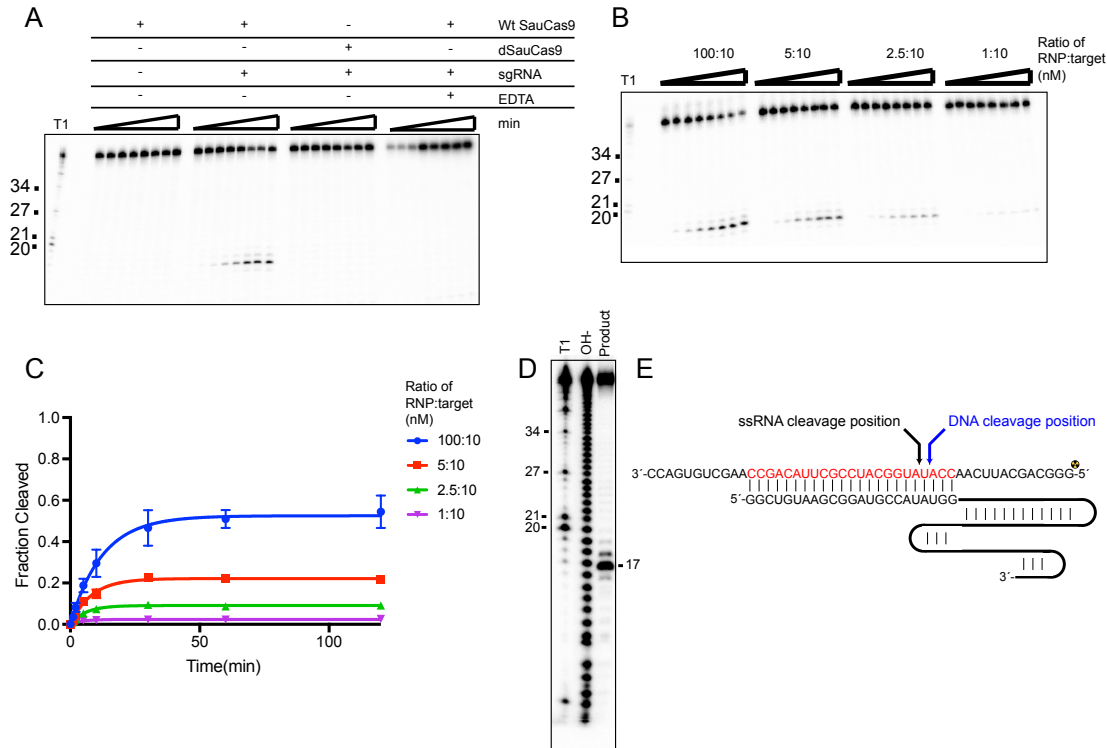


Figure 3.3 ssRNA cleavage is similar to canonical dsDNA cleavage by Cas9
 (A) *In vitro* SauCas9 cleavage assay of ssRNA. Reactions were incubated with wild-type (Wt SauCas9) or catalytically-inactive dSauCas9 (D10A and N580A) in the presence or absence of sgRNA as indicated above the reactions. EDTA was included at 25mM where applicable. (B) SauCas9 ssRNA cleavage is single-turnover. SauCas9 RNP was incubated with the RNA target in the various ratios indicated. (A and B) Time points are 0, 1, 2, 5, 10, 30, 60, and 120 mins. T1 RNase digest size fragments are given on the left. Target used for cleavage was the pUC ssRNA. (C) Graphical representation of ssRNA fraction cleaved of reactions in (B). Fit was determined in Prism using a single-exponential decay model. Error bars represent the mean \pm S.D. (n=3). (D) Mapping of SauCas9 ssRNA cleavage site. Reaction products from a 2-hr incubation of SauCas9 RNP with the pUC ssRNA target were separated on a 15% denaturing PAGE gel with a hydrolysis and T1 digest ladder to determine exact site of the major cleavage product. (E) Diagram of canonical DNA cleavage position and ssRNA cleavage position as determined in (D).

SauCas9 targets ssRNA in the absence of a PAMmer, a contrast to SpyCas9 targeting of ssRNA (O'Connell et al. 2014). Testing SauCas9 *in vitro* ssRNA cleavage in the presence of a PAMmer (30x molar excess over ssRNA target) revealed that turn-over was two-fold slower than the reaction with only target ssRNA (Figure 3.1 C and Figure 3.4 C). SauCas9 ssRNA cleavage conducted in the presence of a non-complementary, control DNA oligo did not

yield a similar reduction in cleavage rate (Figure 3.4 C), indicating that the complementary PAMmer impairs RNA cleavage activity. Consistent with cleavage being guide-dependent, single-stranded RNA that is not complementary to the sgRNA is not cleaved (Figure 3.1 and Figure 3.4). Double-stranded RNA (dsRNA) is also not a substrate for SauCas9.

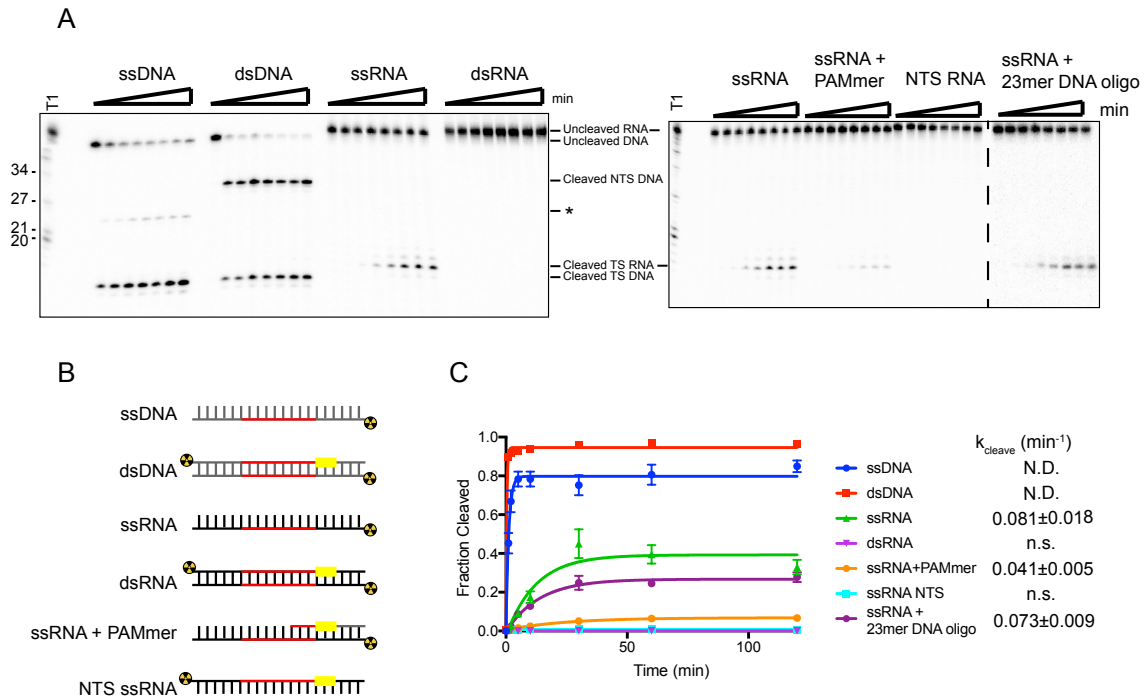


Figure 3.4 SauCas9 cleavage of different nucleic acid substrates.

(A) Representative cleavage assay of nucleic acid substrates diagramed in (B) by SauCas9. Asterisk denotes an off-target cleavage site. Time points are 0, 1, 2, 5, 10, 30, 60, and 120 mins. T1 RNase digest size fragments are given on the left. (C) Quantification of results in (A). Fit was determined in Prism using a single-exponential decay model. Error bars represent the mean \pm S.D. ($n=3$). Apparent pseudo-first order rate constant ($k_{\text{cleave}} \pm$ S.D.) is given to the right of the substrate legend. N.D. indicates that an accurate rate cannot be determined due to the reaction reaching completion before the second time point. N.s., not significant.

Given that Cas9 proteins are active with different length guide RNA segments (~20 - 24 nt) (Chylinski et al. 2013; Ran et al. 2015; Kim et al. 2017a), we tested whether longer guide segments might enhance ssRNA targeting activity. Increasing the length of the targeting region of the guide up to 23 nt results in tighter binding and more efficient cleavage (Figure 3.5), mirroring the

preference for longer guides for DNA cleavage (Ran et al. 2015). Extending the guide strand complementarity to the target beyond 23 nt did not increase RNA target binding or cleavage efficiency, indicating that 23 nt is the optimal length for *in vitro* binding and targeting applications. The apparent dissociation constant ($K_{d,app}$) of the *Sau*Cas9-sgRNA complex (23 nt targeting region) for the ssRNA target is 1.8 ± 0.09 nM (Figure 3.5 D), which is $\sim 5x$ weaker than the 0.34 ± 0.03 nM binding affinity measured for a dsDNA substrate of the same sequence.

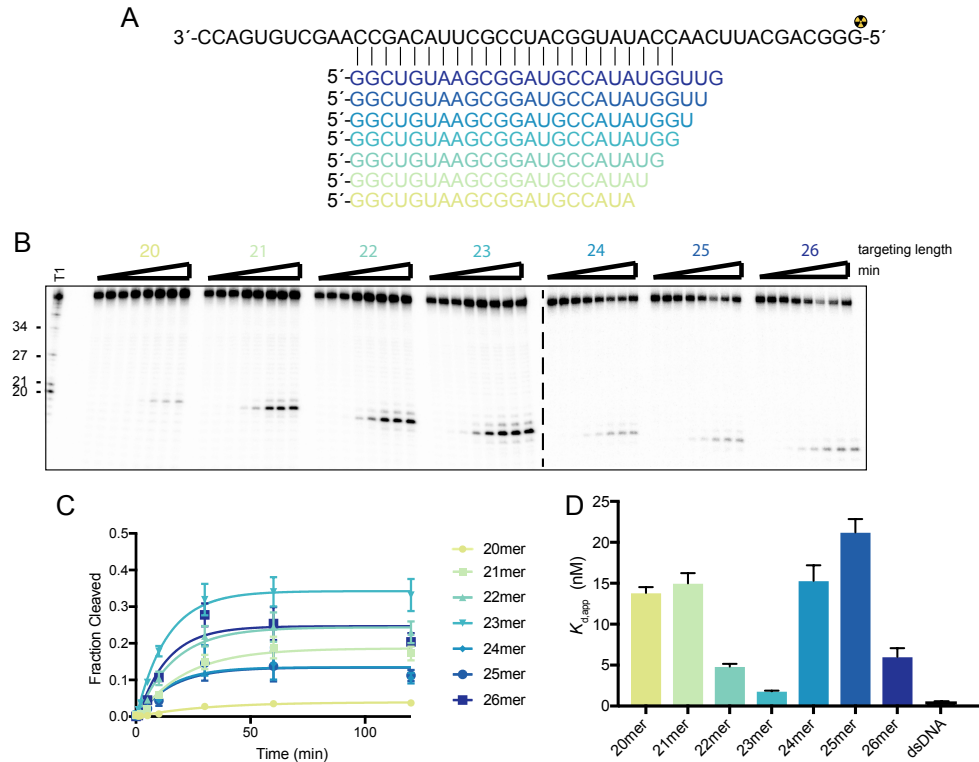


Figure 3.5 *Sau*Cas9 prefers a complementary region of 23nt for binding and cleavage

(A) Diagram of pUC ssRNA target and regions of complementary for the different length sgRNAs. (B) Representative *in vitro* cleavage assays using sgRNAs with a complementary region to the target of the indicated lengths. Time points are 0, 1, 2, 5, 10, 30, 60, and 120 mins. T1 RNase digest size fragments are given on the left. (C) Quantification of cleavage products from reactions in (B). Length of targeting region of the sgRNA given as *n*-mer. Fit was determined in Prism using a single-exponential decay model. Error bars represent the mean \pm S.D. ($n=3$). (D) Filter binding data for *dSau*Cas9 and the structured RNA substrates were fit in Prism using a one-site binding model and the apparent dissociation constant ($K_{d,app}$) was determined. Bars represent the mean \pm S.D. ($n=3$).

3.4.2 Cleavage efficiency is impaired by duplex regions in target RNA

We noted that SauCas9-catalyzed ssRNA cleavage is limited to ~30% fraction cleaved (Figure 3.4), compared to >80% fraction cleaved for ssDNA and dsDNA targets. Greater thermodynamic stability of RNA secondary structures, relative to those in ssDNA (Bercy and Bockelmann 2015), might occlude SauCas9-sgRNA binding to an ssRNA target sequence, a possibility that we tested using a panel of partially duplexed RNA substrates (Figure 3.6). Previously, introduction of a short segment of mismatched base pairs to mimic partially unwound dsDNA substrates was shown to enhance the ability of type II-C Cas9s (including CjeCas9) to unwind and cleave dsDNA (Ma et al. 2015). Here, we found that duplex-RNA substrates containing a 2- or 6-base pair mismatched segment located near the 5' or 3' end of the 23 nt guide RNA region of the sgRNA could not be cleaved (Figure 3.6 A-C, substrates 5, 6, 10, and 11). However, when the unpaired region was increased to 12-base pairs, SauCas9 was able to cleave the target strand. There was a slight cleavage preference for RNA substrates in which the 12-base pair mismatched segment is located near the 5' end of the guide sequence of the sgRNA (Figure 3.6 A-C, substrates 7 and 12).

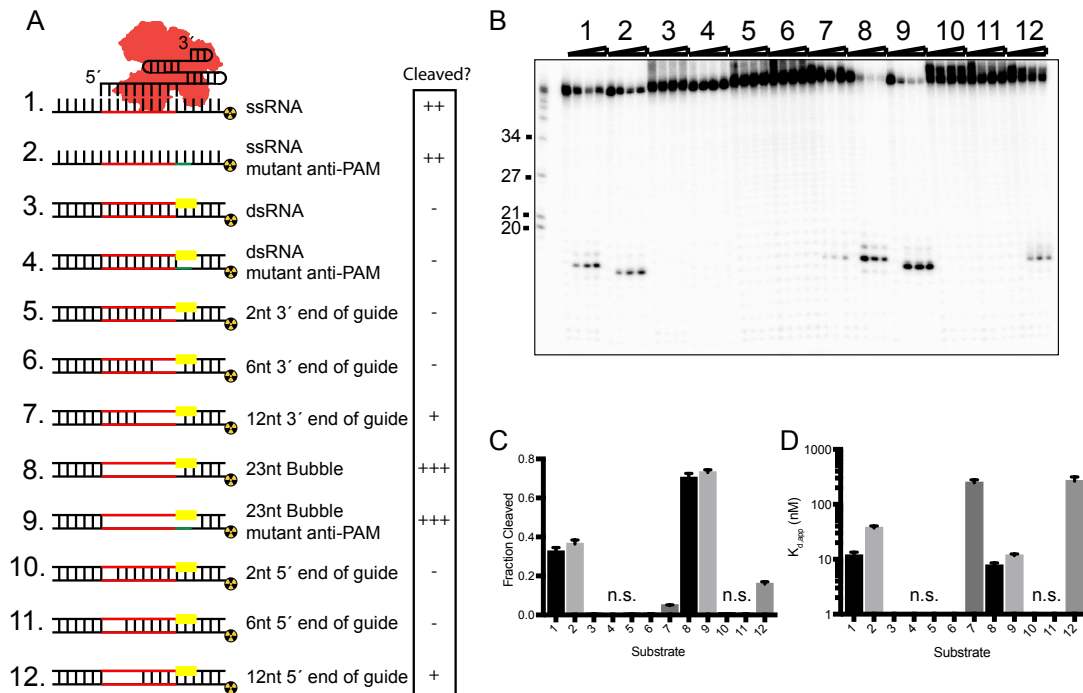


Figure 3.6 *In vitro* RNA cleavage is impaired by strong secondary structure (A) Schematic representation of structured RNA targets for *in vitro* cleavage assays. Symbols on right indicate relative level of cleavage activity for each substrate: “-”, no cleavage; “+”, low cleavage; “++”, medium cleavage; “+++” high cleavage. (B) Representative cleavage assay of partially-duplexed RNA targets

diagrammed in (A). T1 indicates size markers generated by RNase T1 digestion of ssRNA target. Size in nucleotides is indicated on the left. (C and D) Fraction of target cleaved (C) and $K_{d,app}$ (D) for substrates diagrammed in (A). Fits were determined in Prism using a single-exponential decay and a one-site binding model, respectively. Bars represent mean \pm S.D. (n=3). N.s. denotes no significant cleavage or binding.

Interestingly, the 23-base pair mismatched segment RNA substrates ('Bubble' substrates 8 and 9) are targeted more efficiently than their ssRNA counterparts (substrates 1 and 2) (Figure 3.6 C). We measured the binding affinity of all substrates and found that both the 23-base pair mismatched segment RNA and ssRNA substrates are bound with similar affinity (Figure 3.6 D). Furthermore, the apparent difference in cleavage efficiency was not due to the presence of a double-stranded PAM sequence, as mutating the PAM region does not impair cleavage (Figure 3.6 C, compare substrates 8 and 9). We hypothesize that RNA containing a mismatched segment presents a more accessible substrate to the Cas9-sgRNA complex due to stable annealing between the ends of the non-target and target strands, whereas the ssRNA substrate alone has ends that are predicted to stabilize a conformation that is partially structured and therefore inaccessible (Figure 3.7 A).

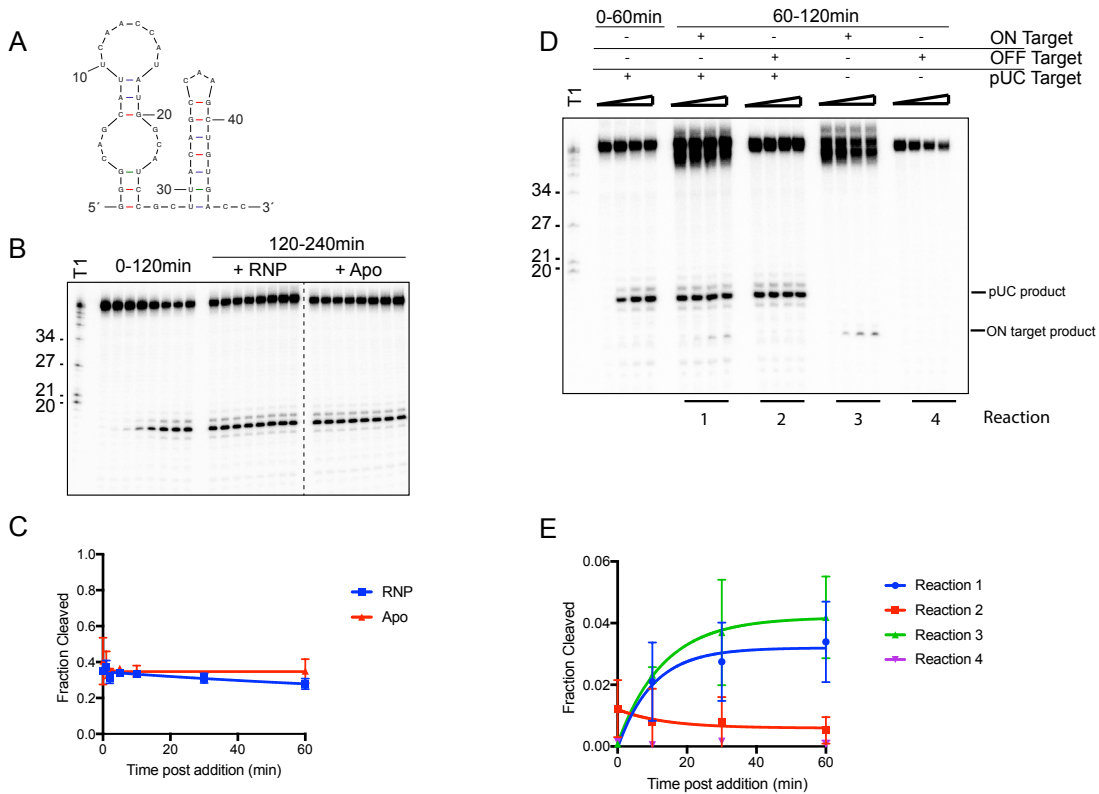


Figure 3.7 RNA cleavage is limited by the RNA target

(A) Predicted secondary structure of target RNA used in this study. (B) *In vitro* cleavage assay of ssRNA with SauCas9 was conducted for 2 hr (time points: 0, 1, 2, 5, 10, 30, 60, 120 mins). The reaction was split and SauCas9-sgRNA RNP or apo SauCas9 were added. The reaction was further incubated at 37°C and additional time points were taken to check for additional cleavage of the target. Time points were taken at 0, 1, 2, 5, 10, 30, 60, and 120 mins post-RNP/apo SauCas9 addition. (C) Fit for data in (B) was determined in Prism using a single-exponential decay model. Error bars represent the mean \pm S.D. (n=3). (D) *In vitro* cleavage assay of two ssRNA targets added sequentially. After 60 min incubation of SauCas9 with the pUC target, another target containing either the same recognition sequence (ON target – reaction 1) or an unrelated sequence (OFF target – reaction 2) were added to the reaction. Cleavage was assayed for an additional 60 mins (time points: 0, 10, 30, 60 min). Reactions containing only the second target (Reactions 3 and 4) were conducted with SauCas9 RNP that was incubated for 60min at 37°C prior to addition to the cleavage reaction. (E) Quantification of cleavage of second target in (D) for time points after addition. Fit was determined in Prism using a single-exponential decay model. Error bars represent the mean \pm S.D. (n=3).

An alternative hypothesis to explain the limited cleavage of ssRNA substrates is that SauCas9 enzyme inactivation occurs over the course of the reaction, even with SauCas9 protein-sgRNA (ribonucleoprotein, RNP) present in 10-fold excess relative to the ssRNA substrate. To test this, we spiked reactions with fresh SauCas9 protein alone or SauCas9 RNP after reactions reached equilibrium; however, we did not observe an increase in the amount of ssRNA cleavage (Figure 3.7 B and C). We also tested whether the SauCas9 RNP was able to cleave a second ssRNA substrate that was added to the reaction after it reached completion (Figure 3.7 D and E). After 1hr of incubation, the addition of a second target ssRNA complementary to the guide RNA resulted in a burst of cleavage activity, whereas a non-complementary ssRNA substrate did not stimulate cleavage. The second target ssRNA is cleaved to a comparable extent to that observed when this second target was the only substrate in the reaction (Figure 3.7 D and E, compare reactions 1 and 3). These observations suggest that SauCas9 RNP is still competent and available for cleavage at the end of the reaction and that a property intrinsic to the ssRNA substrate is the limiting factor. We propose that the observed difference in cleavage extents for various RNA substrates reflects the fraction of molecules that are structurally accessible for cleavage by the SauCas9 RNP.

3.4.3 SauCas9 confers *in vivo* protection against RNA phage

Based on the biochemical ability of SauCas9 RNP to bind and cleave ssRNA substrates, we wondered whether this activity might provide protection against RNA phage infection in bacteria. To test this, we generated a plasmid library encoding sgRNAs containing guide sequences complementary to the genome of MS2, a single-stranded RNA phage that can infect *E. coli*. A subset of these sgRNAs contained scrambled guide sequences that should not target

MS2, providing negative controls. Another sgRNA subset included single-nucleotide mismatches introduced at each position of a target sequence to test for mismatch sensitivity in ssRNA recognition. This plasmid library, comprising 18,114 sgRNAs, was co-transformed into *E. coli* along with a vector encoding a catalytically active version of SauCas9 and the population of transformants was subjected to infection by bacteriophage MS2 (Figure 3.8 A). The experiment was performed in biological triplicate and included an untreated control population and two experimental conditions (multiplicities of infection (MOIs) of 10 and 100). After selection, plasmids were recovered from surviving colonies and sequenced (Figure 3.8 A).

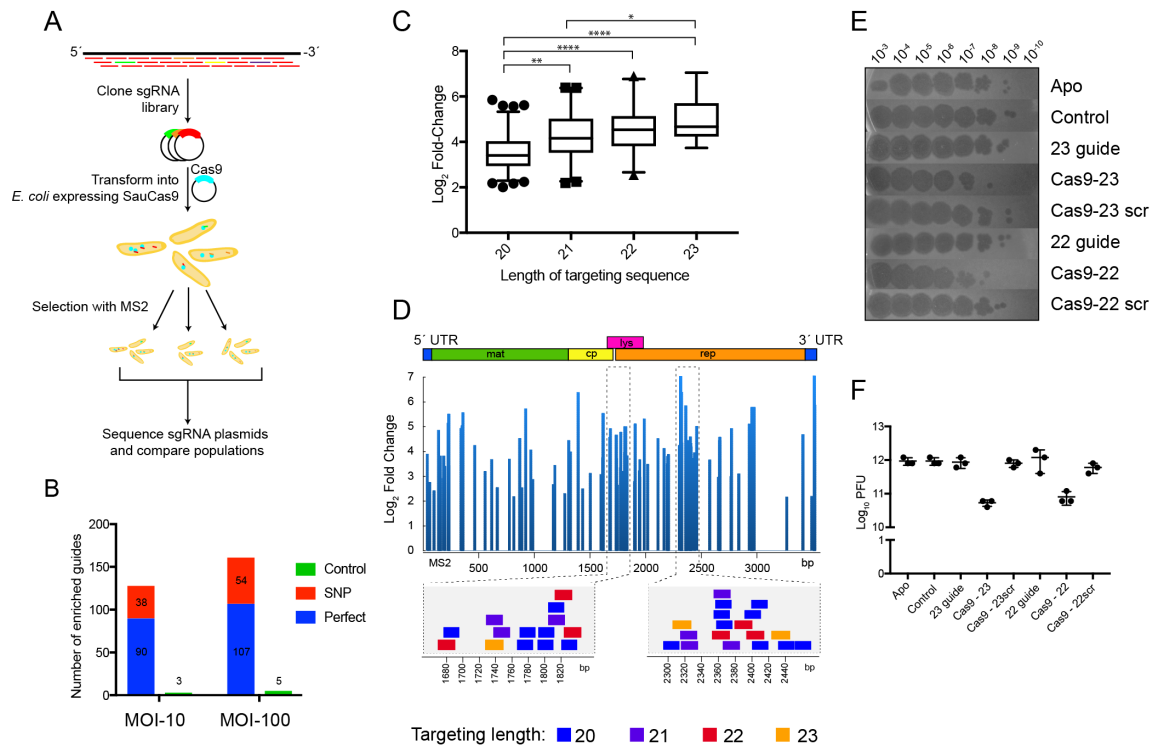


Figure 3.8 SauCas9 confers *in vivo* protection against an RNA phage
 (A) Overview of MS2 targeting screen. Guides tiled against the library were cloned into sgRNA expression plasmids and co-transformed into *E. coli* with a plasmid containing wild-type SauCas9 under inducible control. Plasmids from surviving colonies after MS2 selection were recovered and sequenced. For more detail, see Methods. (B) Number of guides with significant positive enrichment from three biological experiments. SNP, guides with single-nucleotide mismatch. (C) Box and whiskers plot of average \log_2 fold-change of perfect MS2 guides by length. Whiskers represent 5% and 95% values with outliers graphed as points. * $p < 0.05$, ** $p < 0.01$, **** $p < 0.0001$, by one-way ANOVA. (D) (Upper) \log_2 fold-change of guides with an FDR-corrected p -value < 0.05 mapped to the MS2 genome for MOI-100 treatment. Schematic of MS2 genome is provided above. (Lower) Individual guides mapped to highlighted regions of MS2 genome. Other

graphs for MOI-10 and -100 treatments are presented in Figure 3.9. (E) Representative plaque assay of SauCas9 *in vivo* protection. *E. coli* containing constructs on the right are spotted with various phage dilutions as indicated. Scr signifies that the targeting portion of the guide has been scrambled to serve as a non-targeting control. (F) Relative plaque forming units (PFU) (mean \pm S.D., n=3) from results in (E). More guides and controls are presented in Figure 3.11.

We identified between 131 and 166 sgRNAs that were significantly enriched (false discovery rate (FDR)-adjusted p-value < 0.05) in the two different MS2 infection conditions (Figure 3.8 B). The majority of these sgRNAs were perfectly complementary to the MS2 genome, and only three and five control sgRNAs (out of 708 total control sgRNAs) for the MOI-10 and -100 conditions, respectively, were enriched (Figure 3.9 B). The lengths of enriched guide sequences were skewed towards shorter targeting lengths (Figure 3.9 A, left); however, this likely reflects bias in the cloned input library since the ratio between the enriched guide sequences and those of the library without phage selection are similar (Figure 3.9 A, right). When comparing the degree of enrichment between the different guide lengths, the 23-nt guide segment sgRNAs were preferentially enriched over those of shorter length (Figure 3.8 C), consistent with the *in vitro* observation that longer guides are more efficient for directing ssRNA cleavage (Figure 3.5 C). To assess whether there was any sequence bias within the enriched guides, we aligned guide sequences of all lengths at their 3' end. These alignments showed no specific sequence bias in the enriched guides relative to those in the unselected library (Figure 3.9 B). This is consistent with the crystal structure of an SauCas9-sgRNA-DNA bound complex which revealed the absence of base-specific contacts of Cas9 to the target strand (Nishimasu et al. 2015).

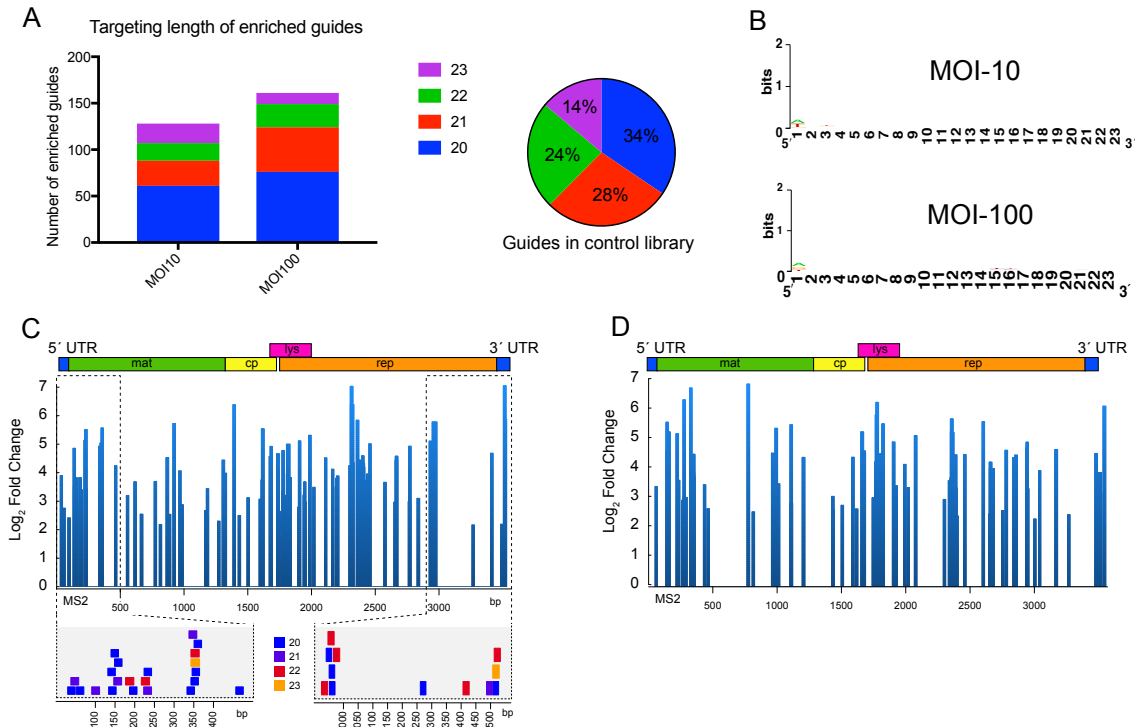


Figure 3.9 Enriched guides do not display sequence bias and cluster to regions on the MS2 genome

(A) (Left) Stacked bar graph of positively enriched guides (FDR-adjusted p-value < 0.05) for perfectly complementary and single-nucleotide mismatch (SNP) guides for a multiplicities of infection (MOI) of 10 and 100. (Right) Percentages of perfect and SNP guides by length averaged across the control libraries (n=3). (B) WebLogo (Crooks et al. 2004) representation of positively enriched guides (perfect complementarity, FDR-adjusted p-value < 0.05) for MOI-10 (n=84) and MOI-100 (n=107). Different length guides were aligned at their 3' end, which contains the pre-ordered 'seed' region (Jiang et al. 2015). (C) (Upper) Log₂ fold-change of positively enriched guides (FDR-adjusted p-value < 0.05) mapped to the MS2 genome for MOI-100 treatment. Schematic of MS2 genome is provided above. (Lower) Individual guides mapped to highlighted regions of MS2 genome. (D) Log₂ fold-change of guides with an FDR-adjusted p-value < 0.05 mapped to the MS2 genome for MOI-10 treatment.

Strikingly, mapping enriched guide sequences onto the MS2 genome showed that enriched sgRNAs were clustered at specific regions, which were consistent across both experimental conditions (Figure 3.8 D, Figure 3.9 C and D). Together with our biochemical data suggesting that SauCas9 cannot bind or cleave structured RNAs (Figure 3.8), we interpret these targeting "hotspots" to be regions of low structural complexity. It is important to note that sgRNAs containing different guide segment lengths overlap at these regions, possibly

indicating that increases in targeting efficiency due to guide length are secondary to target accessibility to the Cas9 RNP. We mapped the enriched guide sequences onto the published secondary structure of the MS2 genome determined through cryoelectron microscopy (Dai et al. 2017) (Figure 3.10). Guides targeted not only single-stranded, accessible regions but also those that form apparently stable secondary structures. The structure of the MS2 genome was determined on the intact phage particle, however, and may not represent the RNA structure(s) relevant to the infection stage during which SauCas9-mediated protection is crucial.

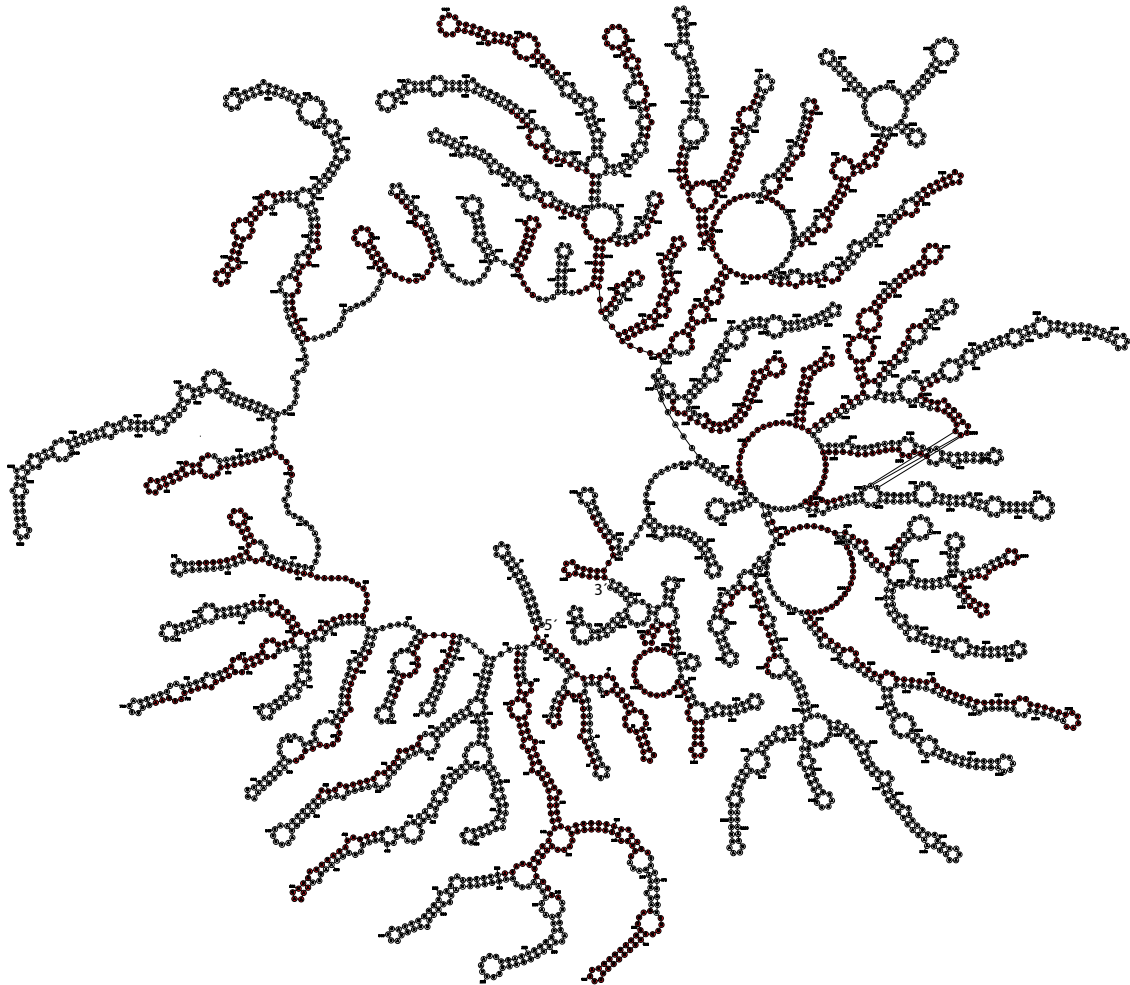


Figure 3.10 Enriched MS2 targeting guides mapped to MS2 genome structure

Structure of the MS2 genome inside the viral particle was obtained from a recently published EM structure (Dai et al. 2017) and guides (red) significantly enriched in the MOI-100 treatment (FDR-adjusted p -value < 0.05) were mapped to the MS2 genome and subsequently visualized in Forna (Kerpedjiev et al. 2015).

Highly enriched sgRNAs from the screen were confirmed for their ability to confer protection against MS2 phage infection through a soft-agar plaque assay. Reconstitution of SauCas9 with a targeting guide confers approximately a ten-fold protection against the RNA phage (Figure 3.8 E and F). No protection was observed in the absence of an sgRNA or SauCas9 protein. Scrambling the sequence of the guide also abrogates protection, confirming that sequence complementary is necessary for phage elimination. Guide segments of all lengths tested (20-23 nts) conferred protection to a similar level (Figure 3.11 A and B), consistent with the result from the MS2 screen that guide segments of all lengths were enriched in 'hotspot' regions (Figure 3.8 D and Figure 3.9 C). Two 'control' guides were enriched in both the MOI-10 and -100 treatments. Interestingly, both guides conferred protection but their scrambled counterparts did not (Figure 3.11 C and D). Whereas a possible off-target binding site was found for one guide (#14238) within the MS2 genome (Figure 3.11 E), it remains unclear how guide #14210 confers protection. Possibly this sgRNA acts by targeting an *E. coli* host factor that is necessary for infection.

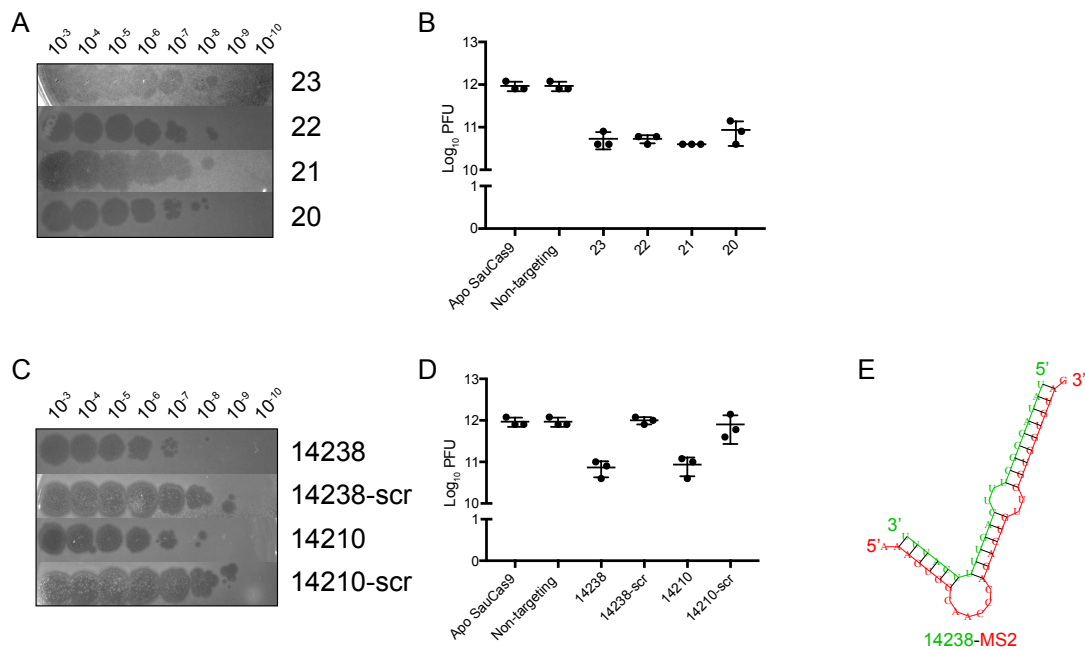


Figure 3.11 Confirmation that enriched guides from the MS2 screen confer protection against MS2 infection

(A) Representative plaque assay for lawns of *E. coli* expressing wtSauCas9 and sgRNA of different length spotted with phage dilutions indicated. Here, the sgRNA with the highest fold-change in both MOI-10 and -100 samples was chosen for each length. The 23-mer sgRNA produces hazy plaques for an unknown reason. All other guides tested, including a different 23-mer sgRNA,

produced clear plaques. (B) Quantification of relative plaque forming units (PFU, mean \pm S.D., n=3) from data in (A). (C) Same as in (A). Guides were two 'control' guides that were significantly enriched in both MOI-10 and -100 treatments during phage selection. Scrambled (scr) indicates random shuffling of the target sequence to serve as a non-targeting control. Scrambled sequences were verified against the MS2 genome and its reverse-complement to ensure no partial matches. (D) Quantification of relative PFU as in (C). Guides 14238 and 14210 confer \sim 10-fold protection over their scrambled counterparts. The level of protection is similar to perfectly complementary guides (B, and see Figure 3.8 E and F). (E) Predicted binding of guide 14238 (green) to a fragment of the MS2 genome (red, nts: 1533-1563) using RNAhybrid (Rehmsmeier et al. 2004).

Screening against the MS2 genome was also used to test the effect of single-nucleotide mismatches on SauCas9's targeting ability. We computed an average fold change (between phage treated and untreated samples) for all sgRNAs that contained a mismatch at the same position, and obtained average values for mismatches at each position across the guide. We observed a pronounced gradient of increasing guide stringency with length. On average, short guides were less sensitive to mismatches, while mismatches in longer sgRNAs led to decreased recovery compared to control samples (Figure 3.12 A, B). Previous work and models suggest that shorter guide segments should be more sensitive to mismatches and lead to higher fidelity Cas9 targeting (Fu et al. 2014; Bisaria et al. 2017). Further study is needed to thoroughly examine this unexpected pattern of RNA-targeting stringency, as one shortcoming of this experiment is that mismatched guides were not designed, *a priori*, to recognize accessible parts of the MS2 genome. Nevertheless, despite potential noise introduced in this analysis due to guide segments that target inaccessible MS2 regions, we observe an interesting correlation between mismatches in the MS2 screen and *in vitro* biochemical cleavage assays for the sgRNA with a 23 nt guide segment sequence (Figure 3.12 C and D). The first few nucleotides in the 'seed' region (guide 3' end proximal) are sensitive to mismatches, while a central region of sensitivity is also observed, similar to previously demonstrated regions of sensitivity for SpyCas9 DNA cleavage (Cong et al. 2013; Jiang et al. 2013; Fu et al. 2016; Gorski et al. 2017).

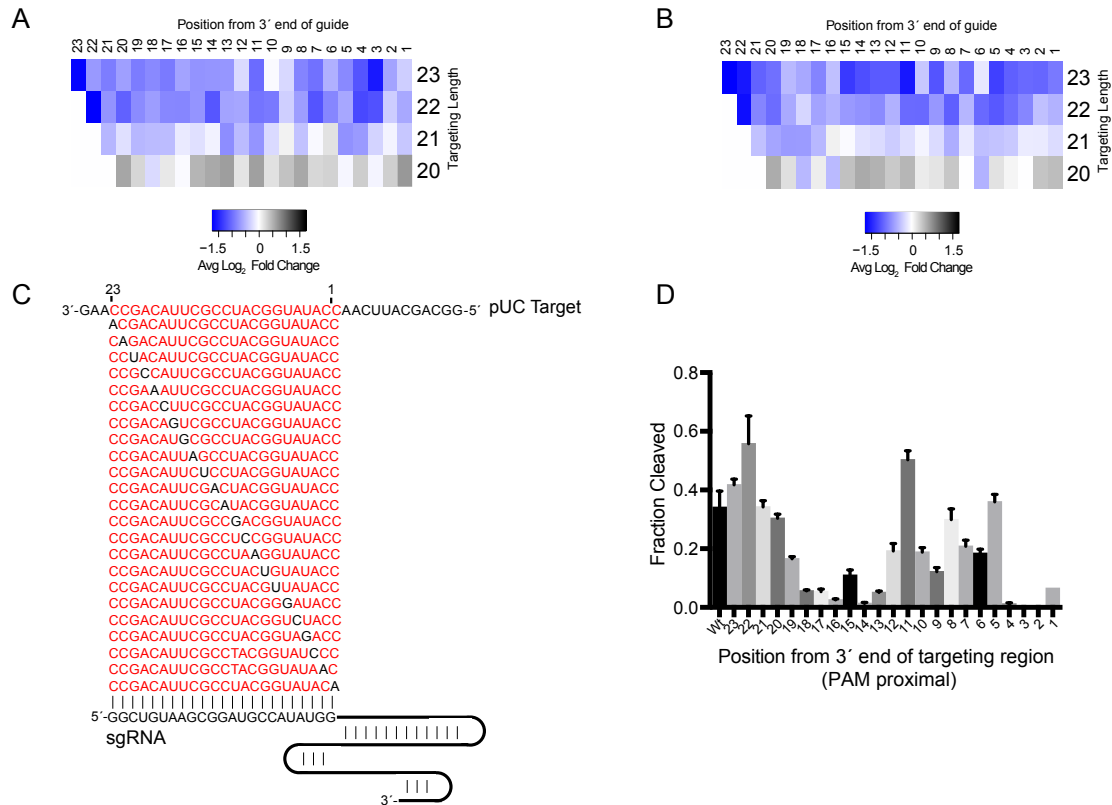


Figure 3.12 Effect of single-nucleotide mismatches on ssRNA targeting

(A-B) Heatmap of average log₂ fold-change for all single-nucleotide mismatch (SNP) guides in MOI-10 (A) and MOI-100 (B) treatment. Deeper blue represents greater negative selection of guides indicating greater sensitivity to mismatches at that position. While deeper black represents greater positive selection indicating that mismatches at that position are more tolerated. Positions are given as distance from 3' end of the targeting region of the sgRNA. (C) Diagram of target ssRNAs with SNPs for *in vitro* cleavage assays. Red highlights the region complementary to the guide while black nucleotides indicate the mismatched base in the targeting region. Numbering of nucleotides is labeled from 1 to 23 to reflect positions in (A) and (B). (D) Quantification of *in vitro* cleavage assays with mismatched targets in (C). Bars represent the mean ± S.D. (n=3). 'Wt' indicates 23nt of perfect complementarity between the sgRNA and the target.

3.4.4 SauCas9 represses gene expression in *E. coli*

An efficient RNA-targeting Cas9 could serve as an important tool in regulating gene expression *in vivo*. To test the ability of SauCas9 to mediate repression of host gene expression, we targeted dSauCas9 and dSpyCas9 RNPs to a GFP reporter sequence encoded in the *E. coli* chromosome (Qi et al. 2013). Catalytically inactive versions of Cas9 were used to prevent cleavage of the bacterial chromosome when targeting a site adjacent to a PAM. As

expression of Cas9 and sgRNA exerts metabolic stress on *E. coli*, GFP fluorescence values were normalized by the OD₆₀₀ value to account for differences in cell growth between cultures (Oakes et al. 2016). When using sgRNAs designed to recognize a sequence in the GFP gene adjacent to the appropriate PAM for SauCas9 (NNGRRT) or SpyCas9 (NGG), GFP expression is significantly reduced (Figure 3.13 A) consistent with CRISPR-interference (CRISPRi) (Qi et al. 2013; Gilbert et al. 2014). When sgRNAs were designed to recognize GFP sequences not flanked by a PAM, dSauCas9 but not dSpyCas9 was able to repress GFP expression. The SauCas9-mediated GFP repression was dependent on sgRNAs that target the coding strand; sgRNAs that recognize the non-coding strand did not result in reduced GFP expression (Figure 3.14 A). The length of the targeting sequence *in vivo* corroborates *in vitro* data, with longer guides working more efficiently (Figure 3.13 B).

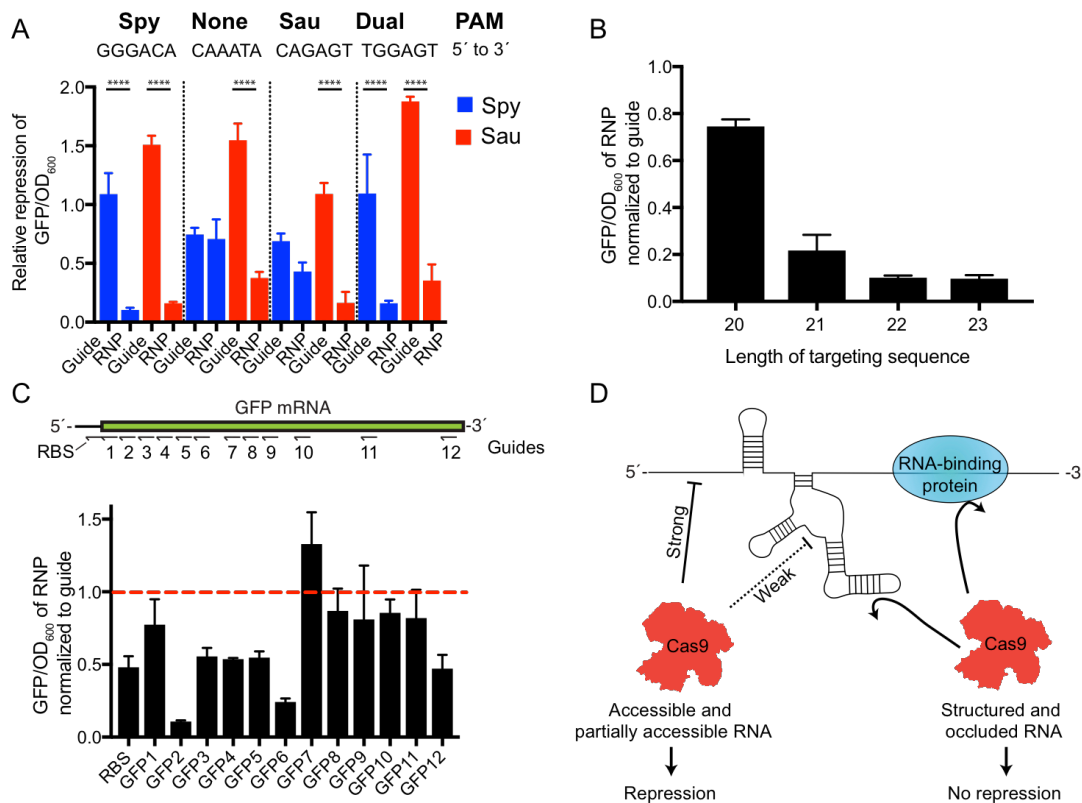


Figure 3.13 SauCas9 repression of a GFP reporter *in vivo*

(A) Comparison of dSpy and dSauCas9 to repress GFP expression on the DNA and RNA level. GFP signal is normalized to OD₆₀₀ to control for difference in cell density between samples. GFP/OD₆₀₀ ratios for guide alone and RNP are normalized to values for a non-targeting guide vector and an Apo protein control, respectively. Target sites were chosen to be adjacent to PAM sites for Spy, Sau, both, or neither as indicated. Note: the slight GFP repression observed with dSpyCas9 using the target sequence adjacent to the Sau PAM (CAGAGT) likely results from the ability of SpyCas9 to use an NAG PAM, albeit with reduced efficiency (Hsu et al. 2013). *****p* < 0.0001 by one-way ANOVA. (B) Relative

expression of GFP using guides with different length targeting sequences. Target site here is the GFP2 sequence chosen for its robust targeting activity. (C) (Upper) Diagram of targeting sequences across the GFP mRNA and ribosome binding site (RBS). (Lower) Relative expression of GFP of SauCas9 RNP normalized to sgRNA alone for targeting sequences across the GFP reporter. Dashed red line indicates that the sgRNA alone is as efficient as the RNP for GFP repression. (A-C) Bars represent mean \pm S.D. (n=3). (D) Model for observed SauCas9 ssRNA targeting activity. We propose that accessible RNA is cleaved or repressed efficiently while structured and protein-bound RNA is not targeted by SauCas9.

Different guide sequences display variable efficiencies of targeting. We tiled sgRNAs across the GFP mRNA sequence to test the robustness of dSauCas9 to repress GFP expression (Figure 3.13 C). As no sites are adjacent to PAM sequences, all repression presumably occurs on the mRNA level. The efficiency of dSauCas9-mediated GFP repression varied according to the target sequence, with some dSauCas9 RNPs reducing GFP signal to 15-30% of that observed in the presence of the sgRNA alone (Figure 3.13 C, GFP2 and 6) and others showing no ability to repress GFP expression (GFP7 and 9). Electrophoretic mobility shift assays support the conclusion that repression is not occurring at the dsDNA level by promiscuous PAM binding (Figure 3.14 B). Repression is largely equivalent between catalytically active and inactive forms of SauCas9 (Figure 3.14 C), suggesting that binding of the Cas9-sgRNA complex to the mRNA is sufficient for repression and consistent with *in vitro* data showing that the enzyme does not catalyze multiple-turnover RNA cleavage. While we speculate that the Cas9-RNP blocks the ribosome directly (either at initiation or during elongation), our data do not rule out the possibility that Cas9 is otherwise destabilizing the mRNA transcript through an unknown mechanism.

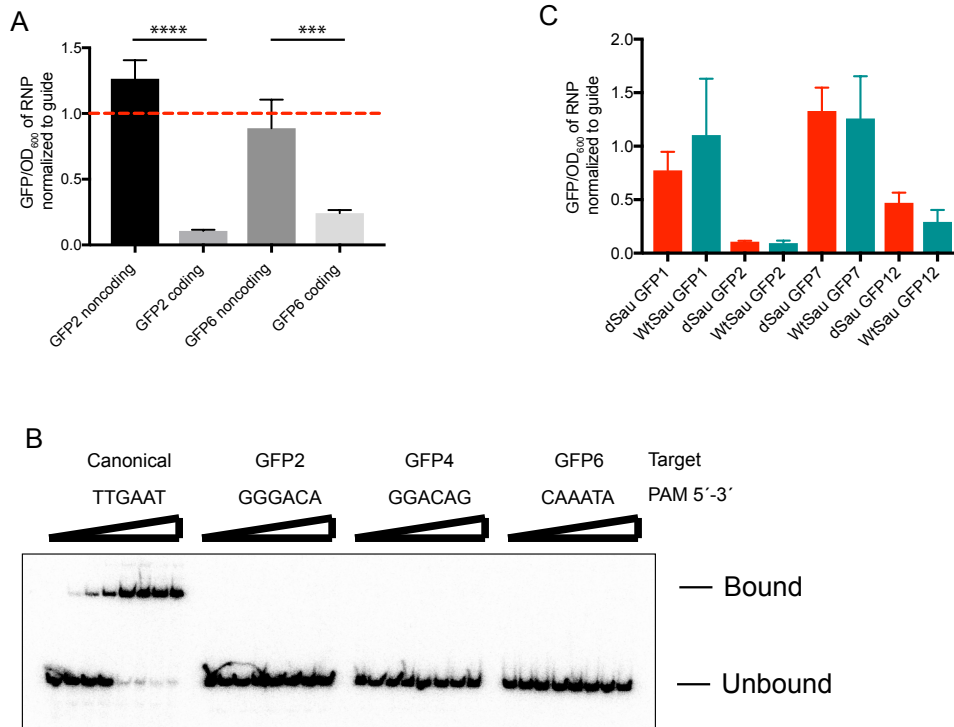


Figure 3.14 Repression of GFP mRNA

(A) dSauCas9-sgRNA directed against a GFP mRNA (coding) or antisense (noncoding) sequence. GFP2 and GFP6 refer to guides diagrammed in Figure 4C. Dashed red line indicates that the sgRNA alone is as efficient as the SauCas9 RNP for GFP repression. Bars represent mean \pm S.D. ($n=3$). $***p < 0.001$, $****p < 0.0001$, by one-way ANOVA. (B) Electrophoretic mobility shift assay (EMSA) confirming that dSauCas9 does not bind dsDNA adjacent to non-canonical PAMs. Targeting sequence is identical for all substrates but with varied PAM sequences as indicated for the guides in Figure 4C. Final concentrations of dSauCas9 from left to right: 0, 0.1, 0.5, 1, 5, 10, 50, 100nM. (C) Comparison of ability of dSau and wtSauCas9 to repress GFP expression *in vivo*. Bars represent mean \pm S.D. ($n=3$).

Together our biochemical and *in vivo* data support a model in which SauCas9 can readily bind and cleave bacteriophage RNA and mRNA sequences that are exposed and unstructured (Figure 3.13 D). Regions that form strong structures are inaccessible to SauCas9 RNP binding, thereby preventing cleavage or repression activity. As Cas9 cleavage activity is limited by target accessibility, we expect that RNA occluded by RNA-binding proteins would also be recalcitrant to cleavage.

3.5 Discussion and Conclusion

Investigation of CRISPR-Cas9 has focused on its function as a double-stranded DNA endonuclease, while the ability of diverse homologs to cleave natural RNA substrates has remained unexplored. Here, we present evidence that type II-A and type II-C Cas9 enzymes can catalyze programmable and PAM-independent single-stranded RNA cleavage. Focusing on SauCas9, we show that this enzyme can be employed both biochemically and in cells to cleave RNA and regulate genes on both the transcriptional and translational level in parallel by accounting for target site PAM proximity. Importantly, SauCas9 ssRNA scission requires only an sgRNA and does not need a PAMmer, thereby simplifying applications (Nelles et al. 2015) and facilitating delivery to cells as a pre-assembled RNP (Zuris et al. 2015; Mout et al. 2017)

The RNA-targeting capability of SauCas9 and related Cas9 enzymes offers the advantage of repressing viruses whose lifecycles do not involve a DNA genome or intermediate, thereby rendering them inaccessible to Cas9-mediated DNA cleavage. We demonstrated that SauCas9 could be programmed to confer protection to *E. coli* against MS2, an RNA bacteriophage with no DNA intermediate. Whether RNA-based viral repression by Cas9 occurs in natural systems is not known, but seems possible based on our results. DNA cleavage by SauCas9 remains more rapid than RNA cleavage, indicating that DNA-targeting is probably the biologically preferred method for phage and plasmid interference. However, Cas9 activity on RNA is PAM-independent and may mitigate the effects of PAM-escape mutants that would evade DNA-level interference (Deveau et al. 2008), thus acting as an additional line of defense.

Intriguingly, ‘hotspots’ of preferential targeting emerged when tiling guides across the genome, but these sites were devoid of sequence bias. In conjunction with *in vitro* cleavage data of partially structured RNAs, we suggest that SauCas9 cleavage efficiency is inversely related to structural complexity of the RNA target. As an alternative to the current approach of screening multiple sgRNAs for activity, experimental knowledge about RNA structure, such as SHAPE-seq data (Loughrey et al. 2014), would simplify target identification for viral targeting and repression experiments. Nevertheless, future work will concentrate on understanding the structural constraints on RNA targeting and methods to improve Cas9 access to duplex RNA regions.

SauCas9 holds promise for a range of RNA targeting applications. We showed that SauCas9 could repress gene expression in *E. coli*. Repression of the reporter occurs in the absence of the PAM and is specific for targeting of the coding strand. Recently, the Type VI CRISPR-Cas system effector, Cas13, has been proposed and demonstrated to target RNA (Shmakov et al. 2015; Abudayyeh et al. 2016; East-Seletsky et al. 2016). ‘Activated’ Cas13 exhibits robust *trans* cleavage of RNAs (Abudayyeh et al. 2016; East-Seletsky et al. 2016; Smargon et al. 2017). While RNA-cleavage by SauCas9 is single-turnover and kinetically less robust than that of Cas13, Cas9 does not cleave RNAs indiscriminately and lends itself to targeting of specific transcripts. A programmable Cas9 capable of repressing genes on the RNA level has potential advantages over CRISPRi DNA-based techniques (Qi et al. 2013; Gilbert et al.

2014). For example, isoform-specific targeting of different transcripts originating from the same transcription start site or resulting from alternative splicing events might be possible. More broadly, due to its intrinsic ssRNA-binding activity, SauCas9 may have utility as a platform for directing other effector proteins to specific RNA molecules, such as proteins or domains that up-regulate translation or RNA base-modifying enzymes for site-specific epigenetic modification of RNAs.

Chapter 4

A natural two-piece Argonaute influences cellular motility

4.1 Chapter Summary

The role of Argonaute (Ago) proteins in RNA-interference and post-transcriptional gene silencing in eukaryotes is well established. Ago homologs exist in bacteria and archaea but details of their physiological function and molecular mechanisms remain enigmatic. Experimental inquiry into prokaryotic Ago (pAgo) function has been limited to a small fraction of the total genetic diversity of systems. Here, we investigate one member of a yet-uncharacterized family of pAgos that consists of seemingly two-piece enzymes. We demonstrate that the split-Ago system from *Shewanella* sp. ANA-3 is transcriptionally coupled and the proteins physically associate to form a heterodimer *in vitro* and *in vivo*. Mutational analyses showed that this novel pAgo binds RNA in a similar mechanism to canonical Ago proteins. Initial data supports a role of this pAgo system in maintaining motility of host cells upon addition of an exogenous plasmid although the molecular mechanism of this phenotype remains unknown. Taken together, our results highlight the diversity of protein architectures and potential functions of pAgo systems.

4.2 Introduction

Argonaute proteins span the tree of life. In eukaryotes, Argonaute (Ago) plays a central role in RNA-dependent post-transcriptional gene silencing, termed RNA-interference. Longer precursor RNAs are processed by Drosha and/or Dicer into small interfering RNAs (siRNAs) or microRNAs (miRNAs) ~20-30nt in length that guide Ago to bind complementary sequences in target mRNAs (Carthew and Sontheimer 2009; Hutvagner and Simard 2008; Jinek and Doudna 2009). Complete or incomplete pairing of the guide and target directs cleavage or translational repression of the mRNA, respectively (Meister et al. 2004). The Piwi-subfamily of Ago proteins associate with Piwi-interacting RNAs (piRNAs) and maintain genome stability by suppressing activation of mobile elements in the germline (Klattenhoff and Theurkauf 2008). Highlighting further functional diversity, in plants and *Drosophila*, Agos are also implicated in antiviral immunity (Hamilton and Baulcombe 1999; van Rij et al. 2006).

Structurally, Ago proteins consist of four domains that adopt a bilobed architecture (Schirle and MacRae 2012; Rashid et al. 2007; Song et al. 2004). In the Piwi lobe, the MID (middle) domain is responsible for coordinating the 5' end of the guide while the PIWI (P element-induced wimpy testis) domain catalyzes cleavage of fully complementary target sequences in variants that possess the necessary catalytic tetrad – DED[D/H/N] (Parker et al. 2005; Yuan et al. 2005; Ma et al. 2005; Meister et al. 2004; Nakanishi et al. 2012; Kaya et al. 2016). The Paz lobe consists of an N-terminal (N) domain and PAZ (PIWI-Argonaute-Zwille) domain responsible for duplex unwinding and guide 3' end binding, respectively (Lingel et al. 2004; Yan et al. 2003; Ma et al. 2004; Faehnle et al. 2013; Kwak and Tomari 2012). Agos from both eukaryotes and prokaryotes display remarkable structural homology despite low primary sequence conservation (Swarts et al. 2014b; Elkayam et al. 2012).

Prokaryotic Argonautes (pAgo) are proposed to act in genome defense (Makarova et al. 2009). Initial evidence suggested that pAgos used short, 5'-phosphorylated DNA guides to cleave exogenous DNA (Swarts et al. 2014a), but several pAgo homologs have been described with varying guide and target preferences (Kaya et al. 2016; Swarts et al. 2015a; Olovnikov et al. 2013). For instance, pAgo from *Marinitoga piezophila* (MpAgo) binds 5'-hydroxylated guides for both single-stranded DNA and RNA cleavage (Kaya et al. 2016; Lapinaite et al. 2018). Some systems were also characterized by a non-specific nuclease activity on invading DNA molecules (Swarts et al. 2015a; 2017a). Recently, guide-independent DNA cleavage was reported in *Thermus thermophilus* and was implicated in guide generation from invading plasmids (Swarts et al. 2017a). If this activity is common to all homologs, it complicates proposed use of pAgos systems as effectors for genome engineering (Hegge et al. 2018). Importantly, all functional data for pAgos derive from studies of full-length proteins (Swarts et al. 2014a; 2015a; Olovnikov et al. 2013; Kaya et al. 2016; Swarts et al. 2017a), which constitute only a small fraction of total pAgo diversity (Makarova et al. 2009; Swarts et al. 2014b; Burroughs et al. 2013).

In contrast to well-studied defense mechanisms, such as restriction-modification and CRISPR-Cas systems, major and diverse pAgo-containing systems remain unexplored. Here, we present data on S3Ago a natural two-piece, pAgo protein that forms a high-affinity complex *in vitro* and *in vivo*. One half of this complex, S3Piwi, associates with a tRNA-half when expressed in *E. coli*. We also show that S3Ago influences motility in the presence of exogenous DNA, although S3Piwi seems to be dispensable for maintaining motility. While S3Piwi's partner, S3Apaz, positively influences plasmid content in the native organism, the full molecular details of S3Ago on the motility pathway remain elusive.

4.3 Methods

4.3.1 Phylogenetic and bioinformatic analyses

Using the MID-PIWI domains (aa 388-685) from *T. thermophilus* Ago as a query, we collected all PIWI-containing proteins in the NCBI RefSeq database (as of August 2015) by running iterative PSI-BLAST until convergence (Altschul and Koonin 1998). Homologs were aligned using a structure-based alignment approach in PROMALS3D (Pei et al. 2008) with input structures of TtAgo (PDB: 4n41), MpAgo (PDB: 5I4A) and PfAgo (PDB: 1U04). Alignments were trimmed to only consider the MID-PIWI module using the TtAgo structure as a template. A maximum likelihood-tree was inferred using PhyML3.0 (Guindon et al. 2010). Leaves were collapsed to reflect similar genomic neighborhoods identified by manual inspection. Domains in partner proteins were predicted via HHPred (Söding et al. 2005), BLAST (Altschul et al. 1990), and homology modeling with Phyre2.0 (Kelley et al. 2015). Figures were prepared using the iTOL webserver (Letunic and Bork 2016) and Jalview 2 (Waterhouse et al. 2009).

4.3.2 Protein purification and *in vitro* reconstitution

Genomic DNA was isolated from *Shewanella* sp. ANA-3 (a kind gift from C. Saltikov, UCSC). Genes encoding S3Piwi and S3Apaz were amplified and cloned as His-MBP fusions (Addgene vector #29706) and expressed in *E. coli* strain BL21(DE3). Cells were grown to an OD₆₀₀ of 0.6-0.8, induced with 0.4M IPTG, and then incubated overnight at 16°C with shaking. Cells were resuspended in lysis buffer (50mM Tris-HCl pH 8, 500mM NaCl, 1mM TCEP, 10mM imidazole, 0.01% Triton X-100 and supplemented with protease inhibitor tablets. Following sonication and centrifugation of insoluble components, clarified lysates were applied to Superflow Ni-NTA affinity resin (Qiagen, Valencia, CA) that was equilibrated in was buffer (50mM Tris-HCl pH 8, 500mM NaCl, 1mM TCEP, 10mM imidazole). After extensive washing (>50 column volumes), protein was eluted in 5 volumes of elution buffer (50mM Tris-HCl pH 8, 500mM NaCl, 1mM TCEP, 300mM imidazole). After tag removal by TEV digestion, the protein was dialyzed into ion exchange buffer A (50mM Tris-HCl pH 8, 300mM NaCl, 1mM TCEP, 5% glycerol) at 4°C for only two hours to prevent precipitation. The protein was immediately applied to a HiTrap HP Heparin column (GE Healthcare, Pittsburgh, PA) to remove bound nucleic acids eluted with ion exchange buffer B (50mM Tris-HCl pH 8, 1.5M NaCl, 1mM TCEP, 5% glycerol). Fractions containing the protein were pooled and separated on a Superdex S75 (GE Healthcare, Pittsburgh, PA) in gel filtration buffer (50mM Tris-HCl pH 8, 500mM NaCl, 1mM TCEP, 5% glycerol). Protein was purified to >95% homogeneity as judged by SDS-PAGE analysis and is free of nucleic acids. Protein was concentrated in gel filtration buffer and stored at -80°C.

To assay complex formation, purified S3Apaz was incubated with a slight molar excess of S3Piwi in gel filtration buffer (above) for one hour at 4°C. The binding reaction was separated on a Superdex S200 column and fractions from both peaks analyzed by SDS-PAGE and Coomassie Blue staining. Proteins for isothermal titration calorimetry (ITC) were dialyzed overnight against gel filtration buffer. Samples were diluted to 100uM (S3Apaz) and 10uM (S3Piwi) prior to titration on the MicroCal Auto-ITC200 at 30°C (Malvern Instruments, Worcestershire, UK).

4.3.3 Strain construction

Shewanella chromosomal insertions and deletions were obtained through homologous recombination and selection with a suicide vector as previously described (Murat et al. 2010). Briefly, a desired mutation was flanked by roughly 800bp of homology in the suicide vector backbone (pSR47S) and introduced into *E. coli* strain WM3064, a DAP-auxotroph that contains the RP4 conjugation machinery. One OD₆₀₀ unit of both the plasmid-containing donor and recipient strain were mixed in 50ul of LB and spotted onto a plate containing 3mM DAP. After 6-8 hr of incubation, cells were recovered from the plate and selected on Kan^R plates. Colony formation typically occurs in 24-48 hrs and individual colonies were confirmed for resistance, followed by growth in unselective liquid media (LB with no Kan) overnight. Cultures were serially passaged on LB plates containing 5% sucrose for counter-selection against SacB and colonies were

screened by PCR and sequencing to identify successful conjugants. All strains used in this study are listed in Table 4.1.

Strain	Notes	Source/Citation
Top10	<i>E. coli</i> cloning strain	Thermo Fisher
BL21(DE3)	<i>E. coli</i> protein expression strain	Thermo Fisher
WM3064	<i>E. coli</i> strain with RP4 conjugation machinery, DAP-auxotroph	(Murat et al. 2010), originally from William Metcalf
S3/Wt	<i>Shewanella</i> sp. ANA-3, wildtype	Kind gift from Chad Saltikov, UCSC
DKO	S3 double knockout: Δ S3Apaz Δ S3Piwi,	This study
OE	S3 overexpression: pRPSL-S3Ago	This study
FLAG	S3 FLAG-S3Apaz	This study
OE-FLAG	S3 pRPSL-FLAG-S3Apaz	This study

Table 4.1: Strains used in this study

4.3.4 Total nucleic acid extraction and PCR assays

Total RNA was extracted from *Shewanella* strains at an OD₆₀₀ of 0.6-0.8 using Trizol according to manufacturer's instructions (Thermo Scientific, Waltham, MA) and treated with RQ1 DNase (Promega, Madison, WI). cDNA was reverse-transcribed with SSIII (Thermo Scientific, Waltham, MA) using random hexamers and served as a template for PCR with primers specific to S3Apaz and S3Piwi. Total DNA was extracted using a direct phenol method (Cheng and Jiang 2006). Quantitative-PCR (qPCR) on the Kan^R gene of pBBR1 was performed on 0.1ng input total DNA using the Brilliant II SYBR Green qPCR Master Mix (Agilent, Santa Clara, CA). Quantitative reverse transcription PCR (qRT-PCR) with S3Apaz specific primers was conducted on 1ng of total RNA with the Brilliant III SYBR Green qRT-PCR Master Mix (Agilent, Santa Clara, CA). GyraseB (*gyrB*) was used as an internal control. Data was plotted in Prizm (GraphPad Software, La Jolla, CA). All DNA primers were synthesized by IDT (Coralville, IA). All plasmids and primers used in this study are listed in Table 4.2.

Strain	Notes	Source/Citation
His-MBP-ORF	Expression and purification, Amp ^R	Addgene #29656
His-ORF	Expression and immunoprecipitation, Kan ^R	Addgene #29653
FLAG-ORF	Expression and immunoprecipitation, Kan ^R	Addgene #29662
pSR47S	Suicide vector for	(Monroe et al.

	homologous recombination in S3. Kan ^R	2009)
pBBR1-null	Stable plasmid in S3, ΔmamK-GFP, Kan ^R	(Abreu et al. 2014)
pBBR1-S3Apaz	pBBR1-null with S3Apaz under pTac	This study
P1 - TACTTCCAATCCAATgcaAAGCAGATT GATATTGAAAACC	Primer for Apaz RT-PCR	This study
P2 - TTATCCACTTCCAATgttattaATTTGCA CTGACATCCTC	Primer for Apaz RT-PCR	This study
P3 - TACTTCCAATCCAATgcaCAAATTAAG ATCCTTGAAGAGC	Primer for Piwi RT-PCR	This study
P4 - TTATCCACTTCCAATgttattaCATATAA AAGCTATAGCTGACCT	Primer for Piwi RT-PCR	This study
KanF1 - GCTATGACTGGGCACAACAG KanR1 - CCTCGTCCTGCAGTTCATTC	Primer pair for plasmid qPCR	This study
Apaz F1 - AACAAAGCAGGGAAGGCTTG Apaz R1 - ACAAGATTCCGGGCACTGGTA	Primer pair for Apaz qRT-PCR	This study
gyrB F1 - GACGCGTTACCACAACATCA gyrB R1 - GAATAAAGGCGGCTGAGCAA	Primer pair for gyrase B qPCR/qRT-PCR	This study

Table 4.2: Plasmids and primers used in this study

4.3.5 Complex immunoprecipitation and nucleic acid analyses

N-terminally tagged His-MBP or FLAG constructs of S3Piwi and S3Apaz were co-expressed in *E. coli* BL21(DE3). Cultures were grown to an OD₆₀₀ of ~0.6 and induced with 4mM IPTG for 3 hr at 37°C. Following cell disruption by sonication and lysate clarification, lysates were applied to FLAG M2 affinity resin (Sigma-Aldrich, St. Louis, MO) in TBS (50mM Tris pH 7.5, 300mM NaCl, and 1mM TCEP) and allowed to bind for 1 hr at 4°C. The resin was washed with TBS and protein was eluted by incubation with 3X FLAG peptide for 30min at 4°C. Presence of both proteins was verified by SDS-PAGE and nucleic acids were isolated by phenol-chloroform extraction prior to 15% denaturing 7M urea PAGE analysis and SYBR Gold staining. FLAG-tagged S3Piwi MID-domain mutants were also expressed, isolated, and analyzed as above. For studies on stability of bound RNA fragments, His-S3Piwi was expressed instead with our standard purification wash buffer (50mM Tris, 300mM NaCl, 1mM TCEP, 10mM imidazole) unless otherwise indicated.

Native complexes were isolated by growing the appropriate strain of S3 at 30°C to an OD₆₀₀ of ~0.6. Cells were harvested by centrifugation and protein complexes were co-immunoprecipitated as above. Eluents were split for SDS-

PAGE analysis and western blotting with an anti-FLAG antibody (Sigma-Aldrich, St. Louis, MO). Nucleic acids were purified and analyzed as above.

4.3.6 Analysis and mapping of S3Piwi RNA fragment

After expression and FLAG immunoprecipitation from *E. coli*, the specific band of RNA bound to S3Piwi was gel purified by 15% denaturing 7M urea PAGE. The target RNA was CIP treated (NEB, Ipswich, MA) and was 5' end-labeled with [γ -P32-ATP] by treatment with PNK (NEB, Ipswich, MA). T1 sequencing and hydrolysis ladders were prepared according to manufacturer's directions (Ambion, Grand Island, NY).

For cloning and sequencing, purified RNA was treated by subsequent incubation with PNK and CIP to remove 3' end modifications. *E. coli* poly(A) polymerase (E-PAP, Ambion, Grand Island, NY) was used to tail RNA for 10 min at 37°C. First strand synthesis by reverse transcription using SSIII and an anchored dT primer was performed followed by RNase I digestion (Ambion, Grand Island, NY). An adenylated adapter was ligated onto the cDNA by the Mth Ligase (NEB, Ipswich, MA) for 1hr at 65°C in the presence of manganese. The ligated product was selectively amplified by PCR, cloned into a TA-cloning vector (Thermo Scientific, Waltham, MA) and sent for sequencing.

4.3.7 Phenotype assays

Conjugation efficiency assays were conducted by conjugation of one OD₆₀₀ unit of WM3064 containing pBBR1-null with an equal amount of recipient cells. Colonies were enumerated after 24 hours on selective plates and confirmed for Kan^R. Cultures for motility assays were started from colonies and grown to an OD₆₀₀ of 0.6-0.8. One μ L of culture normalized to OD ~0.6 was spotted on a 0.25% LB-Agar plate and incubated at 30°C. Diameter of swimming motility was quantified after 24 hr. Data was graphed in Prizm (GraphPad Software, La Jolla, CA). All plasmids used in this study are listed in Table 4.2.

4.4 Results

4.4.1 Identification of a short pAgo system in *Shewanella sp. ANA-3*

As all previous experimental work has focused on full-length pAgos (Swarts et al. 2014a; 2015a; Olovnikov et al. 2013; Kaya et al. 2016), the biochemical mechanisms and physiological functions of two-piece pAgo systems have remained uninvestigated. Based on sequence conservation of the MID and PIWI domains for all distant homologs present in NCBI's Refseq database (as of August 2015), phylogenetic analysis showed distinct clustering of PIWI-domain containing proteins into predicted active and inactive variants (Figure 4.1, leaves colored green and grey, respectively), consistent with previous analyses (Makarova et al. 2009; Swarts et al. 2014b; Burroughs et al. 2014). The majority of active variants are full-length pAgo proteins, which are sometimes associated with a nuclease of the Cas4 family. Certain catalytically inactive variants are full-length but roughly 60% contain only the PIWI-MID module. These 'short' pAgo-

proteins appear almost exclusively in operons with partner proteins predicted to contain nucleases domains (Makarova et al. 2009; Swarts et al. 2014b; Burroughs et al. 2014).

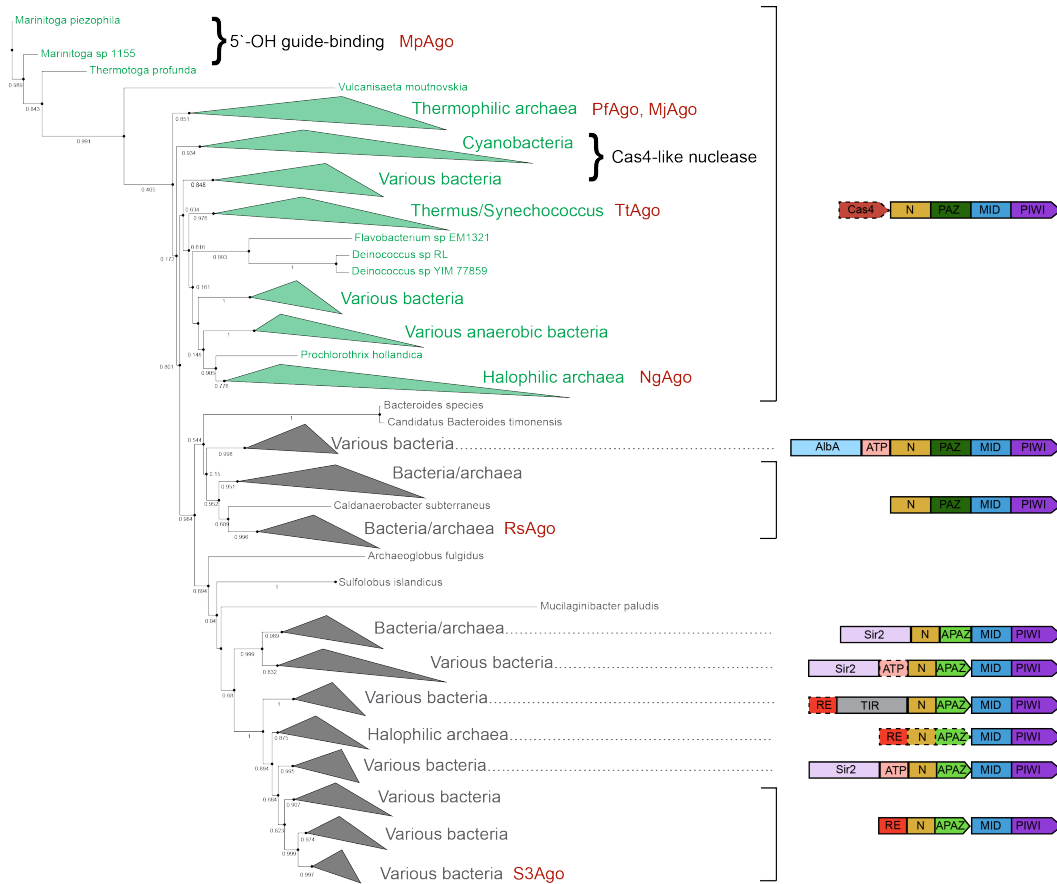


Figure 4.1 Phylogenetic tree of pAgos

PIWI-containing proteins in bacterial and archaeal genomes were aligned by their MID-PIWI domains. Green and grey leaves correspond to predicted active and inactive homologs, respectively. Leaves were collapsed if the pAgo shared a common genetic neighborhood or protein architecture and the organism(s) that contain them are indicated next to the leaves. Homologs that have been structurally or biochemically investigated are highlighted in red and include the subject of this study (S3Ago). pAgo partner protein domain organization for the various clades are diagrammed on the right. Dashed lines indicate common but absolute occurrence of that domain. AlbA, AlbA-like domain; APAZ, Analogous to PAZ; ATP, Schlafen-like ATPase; Cas4, Cas4-like nuclease; MID, middle; N, N-terminal; PAZ, Piwi-Argonaute-Zwille; PIWI, P element-induced wimpy testis; RE, restriction endonuclease; Sir2, Sirtuin 2-like domain; TIR, toll-interleukin receptor.

We sought to identify a short pAgo system amenable for both biochemical and genetic investigation. One candidate system was identified in *Shewanella sp.*

ANA-3, hereafter referred to as S3, that consists of a short pAgo protein (S3Piwi, here differentiated from the PIWI domain) and its partner protein (S3Apaz) (Figure 4.2 A). Generally, *Shewanella* species are easily cultivated under standard laboratory conditions and numerous protocols for genetic manipulation exist (Saltikov et al. 2005; Myers and Myers 1997). S3Piwi contains the MID-PIWI domains with high-confidence, predicted structural homology to other pAgos, including RsAgo and TtAgo (Figure 4.2 A). The central region of S3Apaz shares homology with the N domain of eukaryotic Ago proteins. Domains flanking this region have low predicted similarity to annotated proteins but it is suggested that the “APAZ” domain is analogous to the PAZ domain in Agos, while the “RE” domain shares a fold with restriction-endonucleases (Makarova et al. 2009; Swarts et al. 2014b). S3Apaz has weak predicted structural homology to the C-terminal fragment of a CdiA toxin (Morse et al. 2015), while other partner proteins in short pAgo systems contain predicted N-terminal nucleases of various families (Figure 4.1) (Makarova et al. 2009; Swarts et al. 2014b). As suggested previously (Makarova et al. 2009; Swarts et al. 2014b; Burroughs et al. 2014), the shared homology between the two individual proteins and canonical Ago proteins fuels the hypothesis that these proteins associate to form a complex similar to their full-length counterparts (S3Ago).

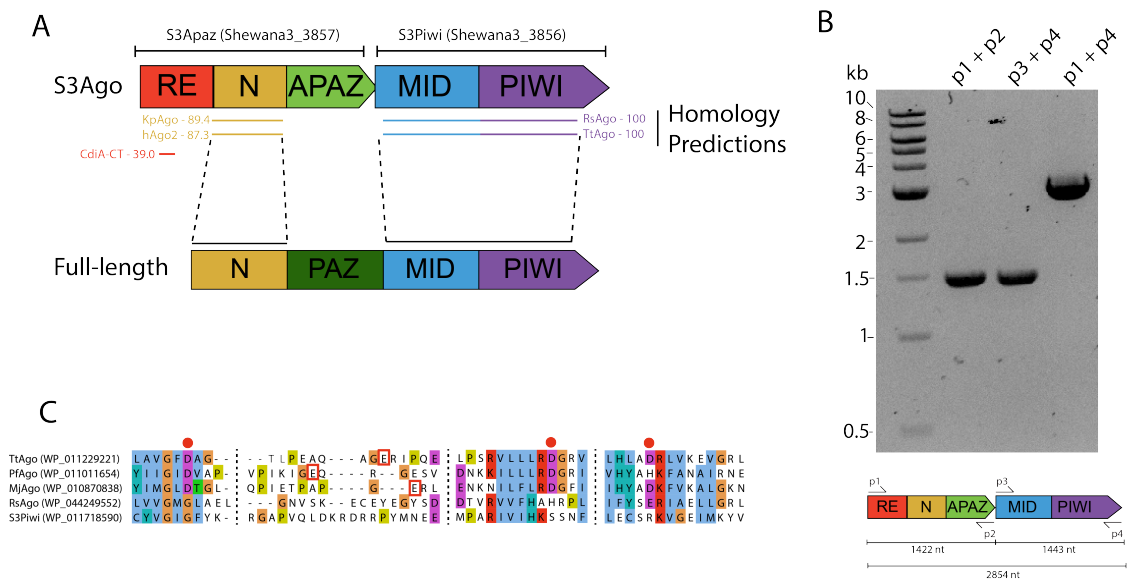


Figure 4.2 Organization of S3Ago operon and transcriptional coupling
S3Ago operon is diagrammed in (A). S3Apaz and S3Piwi are indicated along with their NCBI gene identifiers. Phyre2 homology to known proteins is colored by domain along with their confidence estimate. Individual domain homologies are shown in reference to full-length Ago architecture. (B) RT-PCR results for S3Ago operon using specific primers diagrammed below. (C) Conservation of catalytic residues segments of PIWI domain for select homologs. Catalytic tetrad indicated by red dot or boxed red amino acid for those residue on loops and exhibit poor alignment. The alignment figure was prepared in Jalview and residues are colored according the ClustalX scheme.

Organizationally, the open-reading frames (ORFs) for these two proteins overlap by 11bp, likely indicating expression as a single-transcriptional unit (Figure 4.2 B). Indeed, RT-PCR on total RNA isolated from S3 in mid-log phase revealed a product corresponding to a single-transcript for both proteins (Figure 4.2 B). Consistent with its placement in phylogenetic analyses (Figure 4.1), S3Piwi is not predicted to contain the necessary catalytic residues to effect cleavage (Figure 4.2 C). These results demonstrate that the S3Ago operon is transcriptionally active under standard laboratory conditions and underline the questions of complex formation in short pAgo systems and what role inactive pAgo variants play in their native hosts.

4.4.2 S3Piwi and S3Apaz form a complex *in vitro* and *in vivo*

To answer the question about physical association between short pAgos and their partner proteins, we reconstituted the complex *in vitro* using purified proteins. Individually, S3Piwi and S3Apaz were heterologously expressed in *E. coli* and purified to homogeneity and subsequently mixed together *in vitro*. Size-exclusion chromatography revealed that both proteins elute together at an earlier volume than the S3Piwi protein alone (Figure 4.3 A). SDS-PAGE analysis confirmed the presence of bands corresponding to both S3Piwi and S3Apaz in the faster eluting complex and indicated a 1:1 stoichiometry for each subunit. Isothermal titration calorimetry (ITC) using purified protein revealed an equilibrium binding constant (K_d) of ~ 19 nM, suggesting heterodimer formation is highly-stable (Figure 4.3 B). Supporting a previous hypothesis (Makarova et al. 2009; Swarts et al. 2014b; Burroughs et al. 2014), this is the first experimental evidence that short pAgo systems can form a larger complex with partner proteins and potentially constitutes a native two-piece Argonaute.

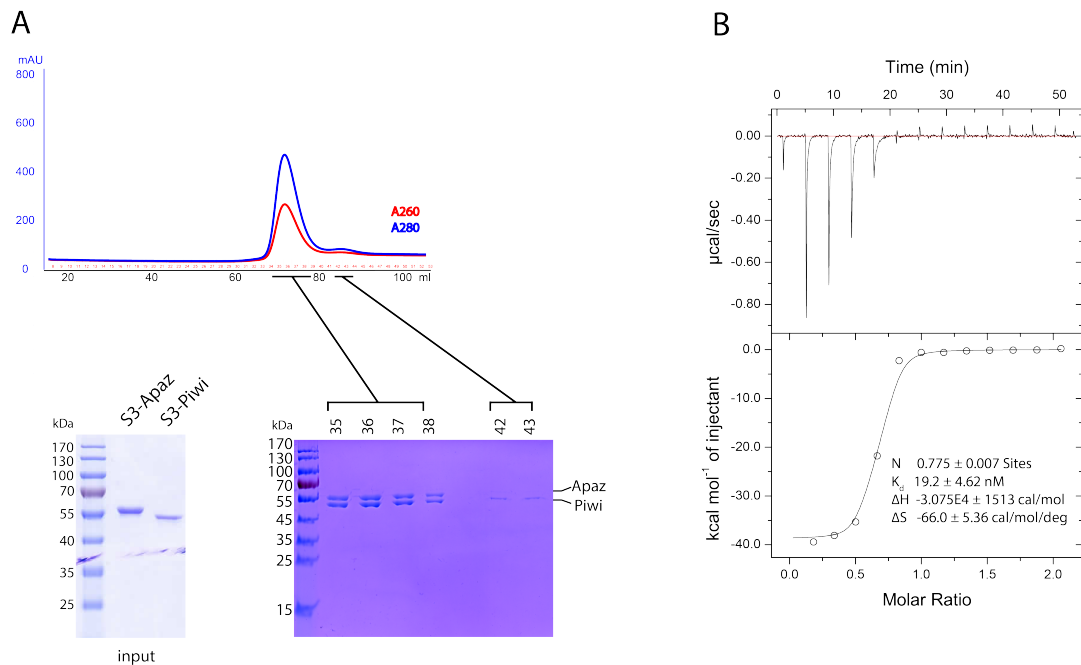


Figure 4.3 S3Piwi and S3Apaz form a complex *in vitro*

(A) Size-exclusion chromatogram of S3Ago reconstitution (top) with SDS-PAGE analysis of fractions (bottom). Size standards of S3Apaz and S3Piwi are shown on the left. (B) Experimental isothermal titration calorimetry (ITC) curve for injection of S3Apaz into S3Piwi at 30°C. Values for stoichiometry, K_d , ΔH , and ΔS are given as mean \pm SEM for three independent titrations.

Leveraging the ability to cultivate *Shewanella* in the lab, we sought to confirm complex association in the native organism. The N-terminus of S3Apaz was FLAG-tagged through homologous recombination and counterselection of a suicide vector introduced by conjugation in the wild-type parental strain (Wt) (Figure 4.4 A). Despite confirmation that the locus is transcribed (Figure 4.2 B) initial results suggested that protein expression was below the limit of detection (Figure 4.4 B). To increase expression, a constitutive promoter from the large ribosomal subunit gene (*rpsL*) of S3 was introduced upstream of the S3Ago operon (Figure 4.4 A). Isolation of FLAG-S3Apaz revealed co-immunoprecipitation (IP) with S3Piwi (Figure 4.4 B) confirming that S3Piwi and S3Apaz also associate *in vivo*.

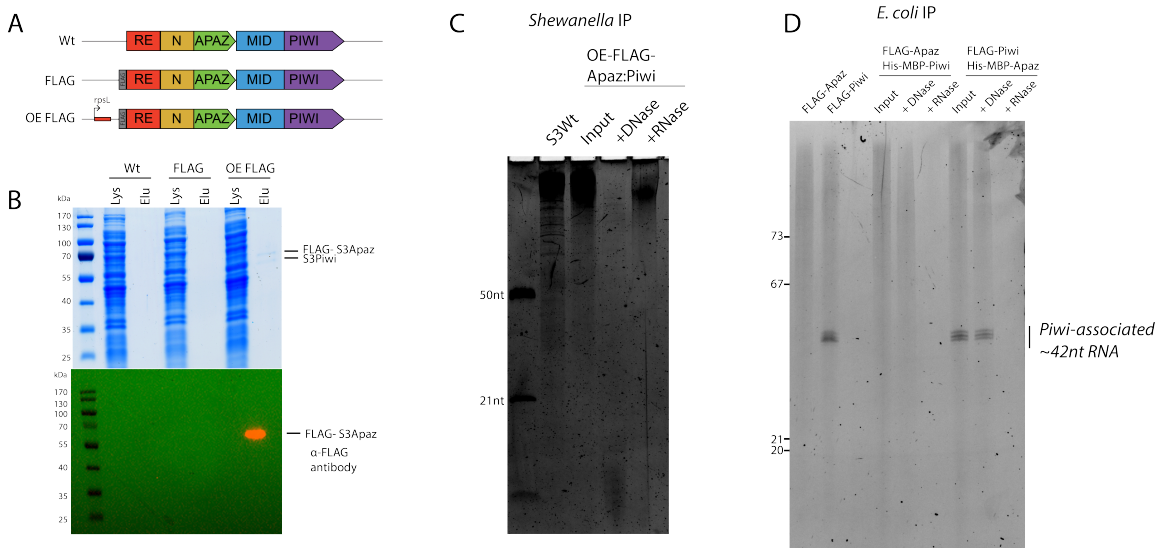


Figure 4.4 *In vivo* complex formation and bound nucleic acids

(A) Diagram of strains for isolation of S3Ago complexes *in vivo*. *rpsL*, large ribosomal subunit promoter. (B) SDS-PAGE analysis of lysate and elution from FLAG immunoprecipitations (top) and confirmation of FLAG-S3Apaz by western blot (bottom). (C,D) Denaturing PAGE-SYBR Gold gel of nucleic acids bound to FLAG-S3Apaz:Piwi from S3 (C) or the indicated proteins in *E. coli* (D). Nucleic acids were treated with the indicated nuclease or nothing (Input). RNA size markers are indicated on the left.

4.4.3 S3Piwi binds heterologous RNAs

One hallmark of Argonaute proteins is their association with short RNA or DNA molecules that serve as guides for target binding and cleavage (Burroughs et al. 2014; Kobayashi and Tomari 2016). Using our overexpression S3 strain, we isolated FLAG-S3Apaz:S3Piwi complexes and extracted co-purifying nucleic acids. No discrete bands were evident in the affinity purified sample and only high-molecular weight DNAs were present but could not be differentiated from an untagged control IP (Figure 4.4 C). The absence of putative nucleic acid guides from the native complex was surprising given an observation that IP of S3Piwi heterologously expressed in *E. coli* demonstrated strong association with short ~42 nt RNAs (Figure 4.4 D). S3Apaz does not appear to bind a specific nucleic acid species but also does not prevent S3Piwi-RNA binding when both proteins are present (Figure 4.4 D). Intriguingly, the length of S3Piwi-bound RNAs (~42 nt) is roughly double that of a canonical Ago guide molecule (~20-22 nt) (Schirle and MacRae 2012), leading to speculation that the RNA may form a duplex to mimic dsRNA substrates bound by eukaryotic Agos (Liu et al. 2003; Lee et al. 2004; Pham et al. 2004; Okamura et al. 2004; Tomari et al. 2004b).

The absence of nucleic acids bound to S3Ago *in vivo* led us to investigate the identity of S3Piwi-bound RNAs from *E. coli* (Figure 4.4 D and Figure 4.5). The distinct banding pattern of isolated S3Piwi RNAs is atypical of pAgo proteins, which usually bind a broader size distribution of RNAs/DNAs (Olovnikov et al. 2013; Swarts et al. 2014a). RNase T1-digest under denaturing and native conditions showed a limited number of cleavage sites (Figure 4.5 A), suggesting a low-complexity population of RNA molecules. However, attempts at adapter ligation for small-RNA sequencing preparation were not successful. Instead, we developed an approach using poly-A tailing, anchored reverse transcription, and DNA adapter ligation (see Methods) to clone these small RNAs. Sanger sequencing revealed that all clones contained inserts that mapped back to the 3' half of the arginine tRNA from *E. coli* (Figure 4.5 B, shown in red). The ~42 nt length roughly places the 5' of the RNA in the anticodon of the tRNA. To account for possible biases in cloning or sequencing, we are preparing high-throughput libraries to more confidently identify the RNA species present in the sample.

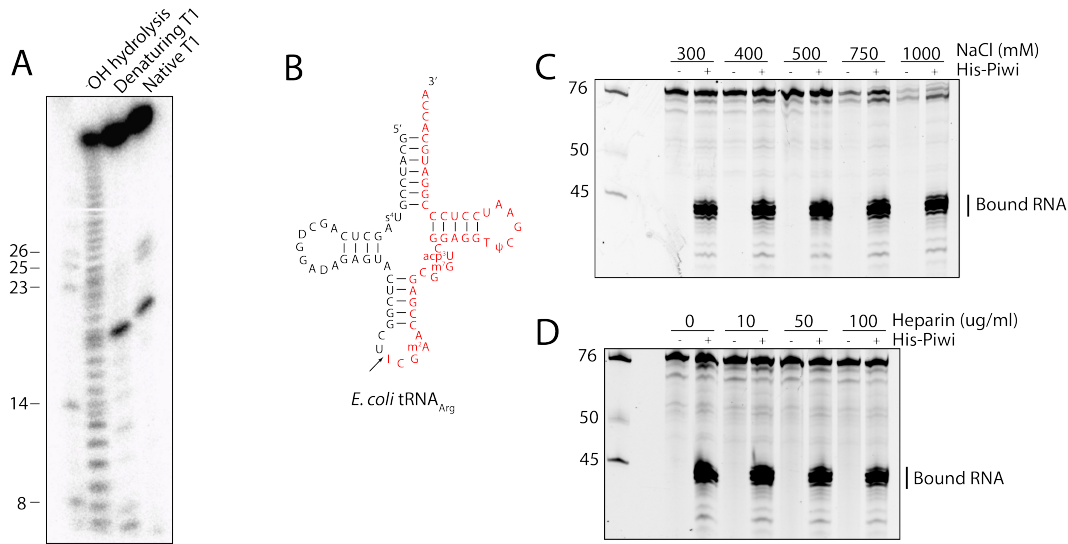


Figure 4.5 Nucleic acid mapping and stability of interaction with S3Piwi

(A). T1 digest and size mapping of radiolabeled nucleic acids isolated from IP of FLAG-S3Piwi. Size markers (nt) are indicated to the left. (B) Map of *E. coli* arginine tRNA was sequenced fragment shown in red and cleavage indicated by an arrow. (C,D) Denaturing PAGE-SYBR Gold gel of nucleic acids bound to S3Piwi under increasingly stringent washing conditions of salt (C) or heparin (D).

Given the abundance of tRNAs in cells, we wanted to confirm that the ~42 nt tRNA half was specifically associated with S3Piwi and not an artifact of insufficiently stringent washing. Increasing the salt concentration from 300mM to 1M did not disrupt binding (Figure 4.5 C). Likewise, the fragment is still recovered after washing with buffer including 100ug/ml heparin (Figure 4.5 D). These results indicate a strong interaction between S3Piwi and its RNA from *E. coli* although the absence of associated nucleic acids from S3 remains enigmatic.

The MID and PAZ domains of Ago are responsible for binding the 5' and 3' end of the guide RNA, respectively (Parker et al. 2004; Rashid et al. 2007; Yuan et al. 2005; Ma et al. 2005). S3Apaz is predicted to have a PAZ-like domain (Figure 4.2 A) but this protein is indispensable for stable RNA binding by S3Piwi (Figure 4.4 D), which contains only a MID-domain. Furthermore, the RNA here is double the size of typical Ago guides (~42 nt vs ~21 nt). We questioned if RNA binding is mediated through canonical MID-domain interactions. Point mutants against conserved residues (motif [R/Y]-K-Q-K) predicted to play a role in metal ion-dependent guide coordination were introduced into the MID-domain of S3Piwi (Figure 4.6 A). All mutants still produce soluble full-length protein with comparable stability to wild-type S3Piwi (Figure 4.6 B). However, analyzing bound nucleic acids to mutant proteins demonstrated that mutants R188A, Q203A, and K236A completely abrogate binding (Figure 4.6 C). Mutant K192A did not disrupt binding and may indicate that this residue less critical for stabilizing divalent metals for guide coordination at the 5' end. These results

conclude that RNA-binding by S3Piwi is largely dependent on canonical MID-domain contacts.

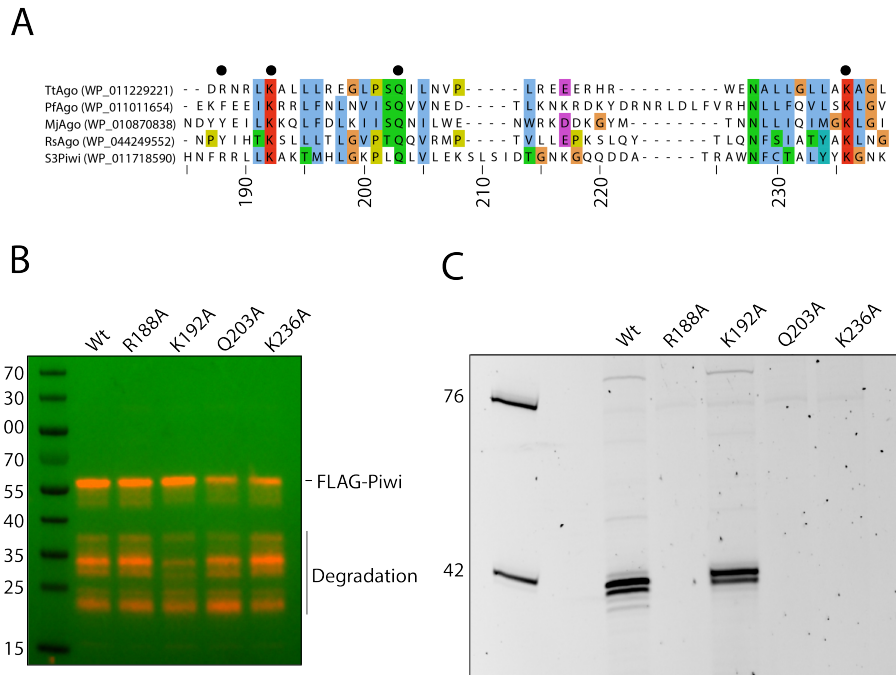


Figure 4.6 RNA-binding of S3Piwi MID-domain mutants

(A) Alignment of guide binding region of MID-domain for select homologs and S3Piwi. Conserved residues for 5' end guide coordination are indicated with a black dot and numbering is based on the S3Piwi protein. The alignment figure was prepared in Jalview 2 and residues are colored according the ClustalX scheme. (B) Western blot analysis of FLAG-tagged mutants isolated from *E. coli*. Full length and degradation products are indicated on the right. (C) Denaturing PAGE-SYBR Gold gel of extracted nucleic acids from FLAG-IP of mutants. Size markers (nt) given on left. Specific amino acid mutations are indicated at the top of (B) and (C).

4.4.4 S3Ago influences cellular motility in the presence of an exogenous plasmid

Initial evidence from catalytically-active pAgos suggested activity in guided interference of plasmid DNA and supported their hypothesized role as a genome defense system (Swarts et al. 2014a; 2015a; Olovnikov et al. 2013). Most of these studies leveraged organisms that are naturally competent to establish a role in defending against invasive genetic elements. Unfortunately, we could not generate competent S3 cells through standard methods or isolate a bacteriophage capable of infecting S3. The only successful method to introduce DNA into S3 is through conjugation from an *E. coli* donor strain. Using this approach, we tested conjugation efficiency of a plasmid into three different S3 strains: wild-type (Wt), double knockout (DKO), and S3Ago overexpression (OE)

(Figure 4.7 A). Overall conjugation efficiencies are low but no significant difference was observed between the strains.

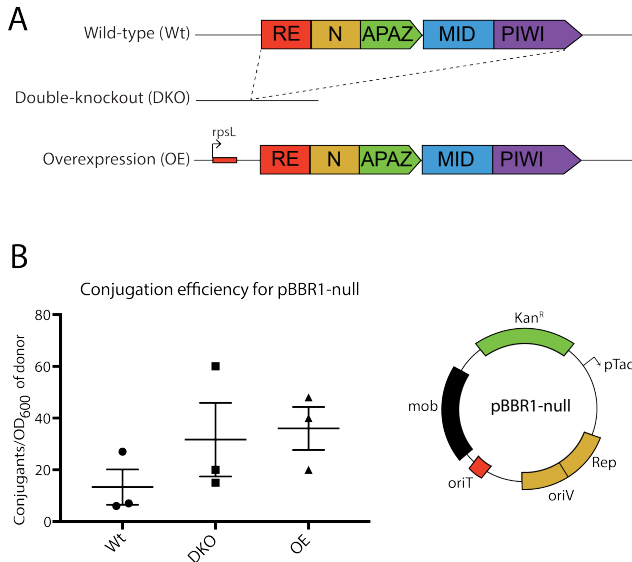


Figure 4.7 Conjugation efficiency into S3 strains

(A) Diagrams of strains test for conjugation efficiency. *rpsL*, large ribosomal subunit promoter. (B) Quantification of conjugants per OD₆₀₀ unit of donor containing pBBR1-null diagrammed on right. Conjugations were carried out in biological triplicate and mean ± SEM is reported. Kan^R, kanamycin-resistance cassette; *mob*, mobility protein; *oriT*, origin of transfer; *oriV*, origin of replication; Rep, replication protein.

Previous bioinformatic analyses postulated that pAgos might associate with toxins and constitute toxin-antitoxin (TA) systems (Makarova et al. 2012). Interestingly, we find that the RE domain of S3Apaz displays distant (weak) structural homology with a CdiA toxin (Figure 4.2 A), involved in contact-dependent inhibition, and was previously reported to be similar to the Mrr family of restriction endonucleases (Makarova et al. 2009; Swarts et al. 2014b). Although we do not observe noticeable toxicity upon individual expression in *E. coli*, we have been unable to clone the entire operon together under the control of either Tac or T7 promoters. However, we are able to obtain sequence-confirmed clones by removal of the promoter. Further work is ongoing to determine the nature of apparent toxicity of these proteins, as the established test of TA systems is to rescue toxin-induced cell death by conditional expression of the antitoxin.

While testing for common phenotypes manifested by TA systems, we noticed an effect on swimming motility. Upon introduction of the pBBR1-null plasmid, the motility of the DKO strain, but not Wt, is drastically reduced as determined by diameter of the swimming area in 0.25% agarose plates (Figure 4.8 A and 4.8 B). The S3Agos genes are not crucial to flagella formation as there

is no difference in the motility of Wt and DKO strains in the absence of the plasmid (Figure 4.8 A and 4.8 B). Plasmid-dependent arrest of motility in the DKO strain appears independent of effects on cell growth as liquid cultures show little difference in growth compared to Wt (Figure 4.8 C).

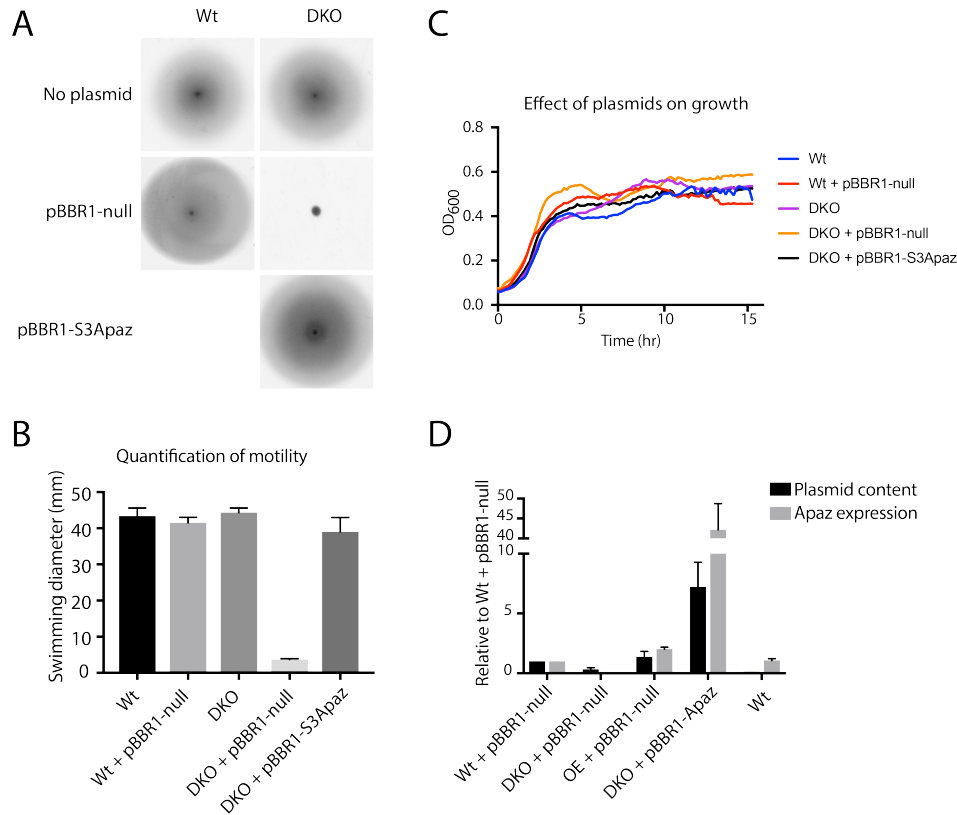


Figure 4.8 S3Apaz influences motility of S3

(A) Swimming diameter of strains indicated (top) with and without exogenous plasmids (left). (B) Quantification of data in (A), reported as mean \pm SEM for biological triplicates. (C) Average growth curves of strains ($n=3$), error bars are omitted for clarity. (D) Relative plasmid content of total DNA determined by qPCR against the Kan^R gene and Apaz expression was determined by qRT-PCR on total RNA and normalized relative to Wt + pBBR1-null condition. Reported values are mean \pm SEM for three independent cultures.

We wanted to determine if either protein alone or the full S3Ago complex could rescue the motility defect in plasmid-containing DKO strains. To test this, we attempted to introduce the pBBR1 vector containing the S3Apaz or S3Piwi genes into the DKO background. No successful conjugants were only obtained for the pBBR1-Piwi construct and we did not have success in cloning the pBBR1-S3Ago construct, as mentioned above. However, pBBR1-S3Apaz was able to rescue motility in the DKO strain (Figure 4.8 A and 4.8 B). One hypothesis to explain this phenomenon is that S3Apaz interferes with the plasmid and reduction of plasmid levels permits motility; however, qPCR on total DNA extracted suggests the opposite effect. Plasmid content in DKO cells is reduced

compared to Wt and there is a large relative increase in plasmid content in the cells overexpressing S3Apaz from the pBBR1 vector (Figure 4.8 D). Taken together, these results suggest a role of S3Apaz in maintaining motility under conditions induced by an exogenous plasmid but not through direct plasmid interference. RNA sequencing of these strains is currently underway to identify potential target pathways affected by introduction of a plasmid and how S3Apaz mitigates those effects.

4.5 Discussion and Conclusion

Studies of prokaryotic Argonaute proteins have been restricted to full-length homologs that all share a canonical four-domain architecture. Here, we investigated the biochemical properties and physiological role of a natural, two-piece Argonaute protein. We show that the short pAgo (S3Piwi) from *Shewanella* sp. ANA-3 is transcribed as a single operon with its partner protein (S3Apaz) and highlight that each protein shares homology with a different lobe of archetypal Ago proteins (Figure 4.2). Importantly, we offer the first experimental evidence that both proteins physically associate to form a heterodimer in the native host (Figure 4.4 A) and reconstitute complex formation *in vitro*, independent of additional factors (Figure 4.3). This ability to form a high-affinity complex and shared homology between two-piece and canonical Agos supports the hypothesis that full-length Agos emerged from two-piece ancestors through fusion of the two ORFs (Swarts et al. 2014b; Burroughs et al. 2014).

S3Piwi associates with RNAs when expressed in *E. coli* but no RNA is detectable when the complex is isolated from the native organism (Figure 4.4B and C). Initial sequencing determined that the fragment is the 3'-half of the arginine tRNA (Figure 4.5 B). The structured nature of tRNA molecules might explain why direct adapter ligation failed and required an alternative cloning approach (Pang et al. 2014; Zheng et al. 2015; Shigematsu et al. 2017). We show that RNA binding is highly stable to S3Piwi and is mediated through established MID-domain interactions (Figure 4.5 C and D, Figure 4.6). Interestingly, previous groups report isolation of RsAgo and various eukaryotic Ago homologs bound to tRNA fragments (Olovnikov et al. 2013; Couvillion et al. 2010; Lee et al. 2009; Cole et al. 2009; Keam and Hutvagner 2015). In the case of Twi12, a Piwi protein from *Tetrahymena thermophila*, the bound tRNA fragment did not act as a guide for RNA interference but instead was critical for nuclear import and activation of RNA processing pathways (Couvillion et al. 2010; 2012). Despite the uncertainty of why S3Piwi is binding a tRNA-half from *E. coli*, it is clear that this protein is competent for RNA-binding and the absence of such activity in *Shewanella* is puzzling, but may indicate that binding or substrate generation is regulated.

Prokaryotic Argonautes have been proposed to play a role in genome defense; however, our inability to transform the S3 strain and the lack of an isolated bacteriophage complicates testing this hypothesis. Conjugation can be used to successfully introduce exogenous DNA into S3, but there was no significant difference in conjugation efficiency in the presence or absence of S3Ago (Figure 4.7). Intriguingly, in a knockout strain, introduction of an

exogenous plasmid induces a motility defect that is not observed in the presence of both proteins or the S3Apaz protein alone (Figure 4.8). Unfortunately, we are unable to obtain conjugants of either S3Piwi or S3Ago to test their motility. Bacterial flagella are exquisitely complex structures that display tight regulation for both expression and assembly (Osterman et al. 2015). As S3Piwi and S3Apaz are not essential for formation of the flagellum (Figure 4.8), we propose they directly or indirectly influence the regulation of flagellar pathways.

Previous work hypothesized that partner proteins to short pAgos serve the role of DNA interference since the PIWI-domain in these systems lacks the necessary residues to catalyze cleavage (Makarova et al. 2009; Swarts et al. 2014b; Olovnikov et al. 2013; Burroughs et al. 2014; 2013). The expectation was that introduction of the S3Apaz partner protein into *Shewanella* would reduce plasmid content, but instead we observe an increase in relative plasmid levels in strains with more S3Apaz expression (Figure 4.8 D). Instead of operating directly on the plasmid, here we suggest that the S3Ago system is playing an indirect role in mitigating stress or toxic effects of plasmid introduction, which also manifests as a change in motility and maintenance of the plasmid at higher levels. Further work is needed to unite all of our observations under one cohesive model.

In eukaryotes, Argonaute proteins seem to act exclusively as RNA-guided effectors for RNA-interference. The functions of pAgos are less well defined, however, and these prokaryotic homologs display greater diversity in guide and target preferences than their eukaryotic partners. Our functional model and understanding of pAgos as a family derive from a limited subset of the total pAgo diversity. Further investigation of genetically divergent systems will determine if pAgos exert functions that extend beyond a generalized role in genome defense as well as possibly lead to the identification of suitable homologs for application as nucleic acid-guided targeting platforms, analogous to CRISPR-Cas9.

Chapter 5

Summary and future directions

5.1 Chapter Summary

A matter of life and death, prokaryotes have evolved an arsenal of defense systems to maintain genomic stability in the face predatory genetic elements. As an alternative to methods that offer non-specific protection to the host, bacteria and archaea possess nucleic acid-guided genome defense systems that confer targeted interference of genomic threats. CRISPR-Cas systems employ small RNAs to direct nucleases for degradation of parasitic DNA and RNA molecules (Wright et al. 2016; Rath et al. 2015). Similarly, pAgos have been proposed as an orthogonal system for small RNA or small DNA-directed interference of invading plasmids (Makarova et al. 2009; Swarts et al. 2014a; Burroughs et al. 2014). Here, we review the similarities and differences between these two nucleic acid-mediated genome defense systems, as well as highlight outstanding questions about both systems.

5.2 Functional and mechanistic comparison of CRISPR-Cas and pAgo systems

Although they operate by different underlying mechanisms and do not share homologous proteins, CRISPR-Cas and pAgo systems both share overall functional similarity in their targeting pathway. Generally, a non-specific nuclease activity precedes a specific targeting step. Adaptation in CRISPR-Cas systems requires generation of short protospacers prior to integration into the CRISPR array, mostly through non-specific Cas or cellular nuclease activity (Levy et al. 2015; Semenova et al. 2016). The Cas interference step uses the information contained in those protospacers to guide specific cleavage of invading DNA. Similarly, pAgos have been implicated in generating their own guides from invading DNA through a non-specific nuclease activity. Ago then binds these guides to mediate specific cleavage (Swarts et al. 2017a), paralleling the multistep process of CRISPR interference.

CRISPR-Cas systems differ from pAgo systems in their source of nucleic acid guides. CRISPR-Cas systems employ crRNAs that are transcribed from previously acquired spacers and require processing by Cas nucleases or other host factors (Hochstrasser and Doudna 2015; Deltcheva et al. 2011; Wright et al. 2016). While detailed mechanisms of guide generation are lacking, some pAgos exhibit autonomous acquisition of guides directly from the invader (Swarts et al. 2017a; Olovnikov et al. 2013). Alternative methods of guide generation also imply a temporal difference in response time to pathogens with CRISPR-Cas systems exhibiting a greater lag time between initial exposure and generation of guided, catalytically-competent Cas nucleases.

In addition to source of their guides, CRISPR-Cas and pAgo systems also differ in identity and composition of their nucleic acid guide molecules. pAgos can associate with small DNA (Swarts et al. 2015a; 2014a; Zander et al. 2017) or small RNA (Olovnikov et al. 2013) guides for *in vivo* plasmid interference. Cas nucleases are guided invariably by RNAs as no CRISPR-Cas system have been identified that are naturally guided by DNA, although a recent study reports efficient CRISPR-Cas9 gene editing in human cells using chimeric RNA:DNA guides (Yin et al. 2018). By virtue of being stored and transcribed from genomic

loci, mature crRNAs contain host-derived nucleotides; however, pAgo guides are apparently only composed of exogenous sequences.

Cas9 and pAgo proteins both have a role in scaffolding and functional division of their guide to facilitate the target search process. Early structural and biochemical work on pAgo homologs revealed that the guide contains a contiguous, preordered seed sequence (Kunne et al. 2014). Remarkably, Cas9 also preorders the seed sequence of its guide RNA (Jiang et al. 2015), highlighting an underlying mechanistic similarity in targeting. Functionally, the seed in both Cas9 and Ago is the nucleation site for target annealing and serves as a discriminator between ON and OFF-target sites to accelerate correct target identification (Sternberg et al. 2014; Ma et al. 2016; Herzog and Ameres 2015; Kunne et al. 2014). Although specific subtypes and homologs display differential preferences, both pAgos and CRISPR-Cas systems can target both RNA and DNA.

5.3 Outstanding questions about nucleic acid-guided genome defense systems

Despite extensive research into CRISPR-Cas systems many outstanding questions still exist. For example, not all prokaryotes contain CRISPR-Cas systems and the lack of these systems in some microbial lineages remains a mystery (Burstein et al. 2016). Furthermore, we have shown that CRISPR-Cas9 systems expand beyond the bacterial domain with the identification of the first archaeal Cas9 homologs. No activity could be demonstrated for these systems and further work is necessary to determine if they operate in a similar manner as their bacterial counterparts.

Interest in CRISPR nucleases as tools for applications in biotechnology has motivated efforts to discover novel CRISPR-Cas systems (Shmakov et al. 2015; Zetsche et al. 2015; Burstein et al. 2017). Exploration of ever-expanding genomic and metagenomic databases is likely to add to the CRISPR-Cas repertoire but what is the limit of Cas gene diversity in nature? Additionally, there is a lack of *in vivo* functional information about CRISPR-Cas system in the context of their native organisms, especially from type IV, type V, and type VI systems.

Besides discovery efforts, investigation into previously identified CRISPR-Cas nucleases is revealing novel biochemical properties. Here, we showed that certain Cas9 homologs display a natural ability to bind and cleave RNA in addition to canonical DNA cleavage. Two other reports also described RNA-dependent RNA targeting in various Cas9 enzymes (Rousseau et al. 2018; Dugar et al. 2018). The extent and role of RNA targeting *in vivo* is not well understood but may protect against parasitic RNA elements or modulate gene expression. Nevertheless, as a platform for RNA binding and targeting, these Cas9 homologs can potentially be repurposed for a range of therapeutic and research applications, including interference of RNA-only viruses or differential targeting of mRNA splice variants.

Few studies have examined the physiological role of pAgos in their native organisms. Initial evidence supports a role in genome defense through direct

cleavage of invading DNA; however, given that the majority of pAgo homologs are inactive nucleases, the generalizability of this role and mechanism is questionable. Our data on a novel, two-piece pAgo system suggests that it functions in an indirect manner to mitigate effects induced by an exogenous plasmid, not through direct cleavage. This system represents only a small minority of total diversity of two-piece pAgo systems, all of which could display different functions based off their divergent partner proteins (Makarova et al. 2009; Swarts et al. 2014b). Early work suggested that pAgos could be used as an alternative for Cas9 in gene editing (Swarts et al. 2014a) but such an application awaits further experimental proof. Given the genetic diversity of pAgos, this family constitutes a wealth of potential nucleic acid-guided nucleases for applications in genome editing and beyond.

References

- Abreu N, Mannoubi S, Ozyamak E, Pignol D, Ginet N, Komeili A. 2014. Interplay between two bacterial actin homologs, MamK and MamK-Like, is required for the alignment of magnetosome organelles in *Magnetospirillum magneticum* AMB-1. *Journal of Bacteriology* **196**: 3111–3121.
- Abudayyeh OO, Gootenberg JS, Essletzbichler P, Han S, Joung J, Belanto JJ, Verdine V, Cox DBT, Kellner MJ, Regev A, et al. 2017. RNA targeting with CRISPR-Cas13. *Nature* **550**: 280–284.
- Abudayyeh OO, Gootenberg JS, Konermann S, Joung J, Slaymaker IM, Cox DBT, Shmakov S, Makarova KS, Semenova E, Minakhin L, et al. 2016. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**: aaf5573.
- Aizenman E, Engelberg-Kulka H, Glaser G. 1996. An *Escherichia coli* chromosomal "addiction module" regulated by guanosine [corrected] 3',5'-bispyrophosphate: a model for programmed bacterial cell death. *Proc Natl Acad Sci USA* **93**: 6059–6063.
- Ali Y, Koberg S, Heßner S, Sun X, Rabe B, Back A, Neve H, Heller KJ. 2014. Temperate *Streptococcus thermophilus* phages expressing superinfection exclusion proteins of the Ltp type. *Front Microbiol* **5**: 98.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410.
- Altschul SF, Koonin EV. 1998. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. *Trends Biochem Sci* **23**: 444–447.
- Amitai G, Sorek R. 2016. CRISPR-Cas adaptation: insights into the mechanism of action. *Nat Rev Microbiol* **14**: 67–76.
- Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, et al. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun* **7**: 13219.
- Anantharaman V, Aravind L. 2003. New connections in the prokaryotic toxin-antitoxin network: relationship with the eukaryotic nonsense-mediated RNA decay system. *Genome Biol* **4**: R81.
- Anantharaman V, Makarova KS, Burroughs AM, Koonin EV, Aravind L. 2013. Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. *Biol Direct* **8**: 15.
- Anders C, Niewoehner O, Duerst A, Jinek M. 2014. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* **513**: 569–573.
- Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166–169.
- Ando Y, Maida Y, Morinaga A, Burroughs AM, Kimura R, Chiba J, Suzuki H, Masutomi K, Hayashizaki Y. 2011. Two-step cleavage of hairpin RNA with 5'

- overhangs by human DICER. *BMC Mol Biol* **12**: 6.
- Arslan Z, Hermanns V, Wurm R, Wagner R, Pul Ü. 2014. Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. *Nucleic Acids Research* **42**: 7884–7893.
- Auyeung VC, Ulitsky I, McGeary SE, Bartel DP. 2013. Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell* **152**: 844–858.
- Babiarz JE, Ruby JG, Wang Y, Bartel DP, Blelloch R. 2008. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes Dev* **22**: 2773–2785.
- Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD, Land ML, Verberkmoes NC, Hettich RL, Banfield JF. 2010. Enigmatic, ultrasmall, uncultivated Archaea. *Proc Natl Acad Sci USA* **107**: 8806–8811.
- Baker BJ, Tyson GW, Webb RI, Flanagan J, Hugenholtz P, Allen EE, Banfield JF. 2006. Lineages of acidophilic archaea revealed by community genomic analysis. *Science* **314**: 1933–1935.
- Barrangou R, Doudna JA. 2016. Applications of CRISPR technologies in research and beyond. *Nat Biotechnol* **34**: 933–941.
- Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**: 1709–1712.
- Batra R, Nelles DA, Pirie E, Blue SM, Marina RJ, Wang H, Chaim IA, Thomas JD, Zhang N, Nguyen V, et al. 2017. Elimination of Toxic Microsatellite Repeat Expansion RNA by RNA-Targeting Cas9. *Cell* **170**: 899–912.e10.
- Beloglazova N, Brown G, Zimmerman MD, Proudfoot M, Makarova KS, Kudritska M, Kochinyan S, Wang S, Chruszcz M, Minor W, et al. 2008. A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *J Biol Chem* **283**: 20361–20371.
- Bercy M, Bockelmann U. 2015. Hairpins under tension: RNA versus DNA. *Nucleic Acids Research* **43**: 9928–9936.
- Bernstein E, Caudy AA, Hammond SM, Hannon GJ. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**: 363–366.
- Betancur JG, Tomari Y. 2012. Dicer is dispensable for asymmetric RISC loading in mammals. *RNA* **18**: 24–30.
- Bisaria N, Jarmoskaite I, Herschlag D. 2017. Lessons from Enzyme Kinetics Reveal Specificity Principles for RNA-Guided Nucleases in RNA Interference and CRISPR-Based Genome Editing. *Cell Syst* **4**: 21–29.
- Blosser TR, Loeff L, Westra ER, Vlot M, Künne T, Sobota M, Dekker C, Brouns SJJ, Joo C. 2015. Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. *Molecular Cell* **58**: 60–70.
- Bohnsack MT, Czapinski K, Gorlich D. 2004. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA* **10**: 185–191.
- Bolotin A, Quinquis B, Sorokin A, Dusko Ehrlich S. 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of

- extrachromosomal origin. *Microbiology (Reading, Engl)* **151**: 2551–2561.
- Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Snijders APL, Dickman MJ, Makarova KS, Koonin EV, van der Oost J. 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**: 960–964.
- Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF. 2015. Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**: 208–211.
- Buckling A, Brockhurst M. 2012. Bacteria-virus coevolution. *Adv Exp Med Biol* **751**: 347–370.
- Burroughs AM, Ando Y, Aravind L. 2014. New perspectives on the diversification of the RNA interference system: insights from comparative genomics and small RNA sequencing. *Wiley Interdiscip Rev RNA* **5**: 141–181.
- Burroughs AM, Iyer LM, Aravind L. 2013. Two novel PIWI families: roles in inter-genomic conflicts in bacteria and Mediator-dependent modulation of transcription in eukaryotes. *Biol Direct* **8**: 13.
- Burstein D, Harrington LB, Strutt SC, Probst AJ, Anantharaman K, Thomas BC, Doudna JA, Banfield JF. 2017. New CRISPR-Cas systems from uncultivated microbes. *Nature* **542**: 237–241.
- Burstein D, Sun CL, Brown CT, Sharon I, Anantharaman K, Probst AJ, Thomas BC, Banfield JF. 2016. Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nat Commun* **7**: 10613.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* **10**: 421.
- Carte J, Pfister NT, Compton MM, Terns RM, Terns MP. 2010. Binding and cleavage of CRISPR RNA by Cas6. *RNA* **16**: 2181–2188.
- Carte J, Wang R, Li H, Terns RM, Terns MP. 2008. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev* **22**: 3489–3496.
- Carthew RW, Sontheimer EJ. 2009. Origins and Mechanisms of miRNAs and siRNAs. *Cell* **136**: 642–655.
- Cenik ES, Fukunaga R, Lu G, Dutcher R, Wang Y, Tanaka Hall TM, Zamore PD. 2011. Phosphate and R2D2 restrict the substrate specificity of Dicer-2, an ATP-driven ribonuclease. *Molecular Cell* **42**: 172–184.
- Chakravarthy S, Sternberg SH, Kellenberger CA, Doudna JA. 2010. Substrate-specific kinetics of Dicer-catalyzed RNA processing. *J Mol Biol* **404**: 392–402.
- Chandradoss SD, Schirle NT, Szczepaniak M, MacRae IJ, Joo C. 2015. A Dynamic Search Process Underlies MicroRNA Targeting. *Cell* **162**: 96–107.
- Charpentier E, Richter H, van der Oost J, White MF. 2015. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS Microbiol Rev* **39**: 428–441.
- Chaudhary K. 2018. Bacteriophage EXclusion (BREX): A novel anti-phage mechanism in the arsenal of bacterial defense system. *J Cell Physiol* **233**: 771–773.
- Cheloufi S, Santos Dos CO, Chong MMW, Hannon GJ. 2010. A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature*

- 465**: 584–589.
- Chen JS, Dagdas YS, Kleinstiver BP, Welch MM, Sousa AA, Harrington LB, Sternberg SH, Joung JK, Yildiz A, Doudna JA. 2017. Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* **550**: 407–410.
- Chen JS, Ma E, Harrington LB, Da Costa M, Tian X, Palefsky JM, Doudna JA. 2018. CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* **546**: eaar6245.
- Chendrimada TP, Gregory RI, Kumaraswamy E, Norman J, Cooch N, Nishikura K, Shiekhattar R. 2005. TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature* **436**: 740–744.
- Cheng H-R, Jiang N. 2006. Extremely rapid extraction of DNA from bacteria and yeasts. *Biotechnol Lett* **28**: 55–59.
- Choi KR, Lee SY. 2016. CRISPR technologies for bacterial systems: Current achievements and future directions. *Biotechnol Adv* **34**: 1180–1209.
- Chopin M-C, Chopin A, Bidnenko E. 2005. Phage abortive infection in lactococci: variations on a theme. *Curr Opin Microbiol* **8**: 473–479.
- Chylinski K, Le Rhun A, Charpentier E. 2013. The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA Biol* **10**: 726–737.
- Chylinski K, Makarova KS, Charpentier E, Koonin EV. 2014. Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Research* **42**: 6091–6105.
- Cifuentes D, Xue H, Taylor DW, Patnode H, Mishima Y, Cheloufi S, Ma E, Mane S, Hannon GJ, Lawson ND, et al. 2010. A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. *Science* **328**: 1694–1698.
- Cole C, Sobala A, Lu C, Thatcher SR, Bowman A, Brown JWS, Green PJ, Barton GJ, Hutvagner G. 2009. Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA* **15**: 2147–2160.
- Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, et al. 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**: 819–823.
- Couvillion MT, Bounova G, Purdom E, Speed TP, Collins K. 2012. A Tetrahymena Piwi bound to mature tRNA 3' fragments activates the exonuclease Xrn2 for RNA processing in the nucleus. *Molecular Cell* **48**: 509–520.
- Couvillion MT, Sachidanandam R, Collins K. 2010. A growth-essential Tetrahymena Piwi protein carries tRNA fragment cargo. *Genes Dev* **24**: 2742–2747.
- Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* **14**: 1188–1190.
- Dagdas YS, Chen JS, Sternberg SH, Doudna JA, Yildiz A. 2017. A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9. *Sci Adv* **3**: eaao0027.
- Dai X, Li Z, Lai M, Shu S, Du Y, Zhou ZH, Sun R. 2017. In situ structures of the genome and genome-delivery apparatus in a single-stranded RNA virus.

- Nature* **541**: 112–116.
- Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E. 2012. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat Commun* **3**: 945.
- Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E. 2011. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**: 602–607.
- Denef VJ, Banfield JF. 2012. In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. *Science* **336**: 462–466.
- Deng L, Kenchappa CS, Peng X, She Q, Garrett RA. 2012. Modulation of CRISPR locus transcription by the repeat-binding protein Cbp1 in *Sulfolobus*. *Nucleic Acids Research* **40**: 2470–2480.
- Deveau H, Barrangou R, Garneau JE, Labonté J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S. 2008. Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*. *Journal of Bacteriology* **190**: 1390–1400.
- Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP, Banfield JF. 2009. Community-wide analysis of microbial genome sequence signatures. *Genome Biol* **10**: R85.
- Díez-Villaseñor C, Guzmán NM, Almendros C, García-Martínez J, Mojica FJM. 2013. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. *RNA Biol* **10**: 792–802.
- Dong D, Ren K, Qiu X, Zheng J, Guo M, Guan X, Liu H, Li N, Zhang B, Yang D, et al. 2016. The crystal structure of Cpf1 in complex with CRISPR RNA. *Nature* **532**: 522–526.
- Donohoue PD, Barrangou R, May AP. 2018. Advances in Industrial Biotechnology Using CRISPR-Cas Systems. *Trends Biotechnol* **36**: 134–146.
- Doron S, Melamed S, Ofir G, Leavitt A, Lopatina A, Keren M, Amitai G, Sorek R. 2018. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**: eaar4120.
- Doudna JA, Charpentier E. 2014. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**: 1258096–1258096.
- Doxzen KW, Doudna JA. 2017. DNA recognition by an RNA-guided bacterial Argonaute. ed. E. Paci. *PLoS ONE* **12**: e0177097.
- Drozdetskiy A, Cole C, Procter J, Barton GJ. 2015. JPred4: A protein secondary structure prediction server. *Nucleic Acids Research* **43**: W389–W394.
- Dugar G, Leenay RT, Eisenbart SK, Bischler T, Aul BU, Beisel CL, Sharma CM. 2018. CRISPR RNA-Dependent Binding and Cleavage of Endogenous RNAs by the *Campylobacter jejuni* Cas9. *Molecular Cell* **69**: 893–905.e7.
- Durmaz E, Klaenhammer TR. 2007. Abortive phage resistance mechanism *AbiZ* speeds the lysis clock to cause premature lysis of phage-infected *Lactococcus lactis*. *Journal of Bacteriology* **189**: 1417–1425.
- Dy RL, Przybilski R, Semeijn K, Salmond GPC, Fineran PC. 2014. A widespread bacteriophage abortive infection system functions through a Type IV toxin-

- antitoxin mechanism. *Nucleic Acids Research* **42**: 4590–4605.
- East-Seletsky A, O'Connell MR, Burstein D, Knott GJ, Doudna JA. 2017. RNA Targeting by Functionally Orthogonal Type VI-A CRISPR-Cas Enzymes. *Molecular Cell* **66**: 373–383.e3.
- East-Seletsky A, O'Connell MR, Knight SC, Burstein D, Cate JHD, Tjian R, Doudna JA. 2016. Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature* **538**: 270–273.
- Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, Tuschl T. 2001. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**: 494–498.
- Elkayam E, Kuhn C-D, Tocilj A, Haase AD, Greene EM, Hannon GJ, Joshua-Tor L. 2012. The structure of human argonaute-2 in complex with miR-20a. *Cell* **150**: 100–110.
- Elmore JR, Sheppard NF, Ramia N, Deighan T, Li H, Terns RM, Terns MP. 2016. Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system. *Genes Dev* **30**: 447–459.
- Emerson JB, Thomas BC, Alvarez W, Banfield JF. 2016. Metagenomic analysis of a high carbon dioxide subsurface microbial community populated by chemolithoautotrophs and bacteria and archaea from candidate phyla. *Environ Microbiol* **18**: 1686–1703.
- Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* **30**: 1575–1584.
- Erdmann S, Garrett RA. 2012. Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. *Mol Microbiol* **85**: 1044–1056.
- Erdmann S, Le Moine Bauer S, Garrett RA. 2014. Inter-viral conflicts that exploit host CRISPR immune systems of *Sulfolobus*. *Mol Microbiol* **91**: 900–917.
- Estrella MA, Kuo F-T, Bailey S. 2016. RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex. *Genes Dev* **30**: 460–470.
- Faehnle CR, Elkayam E, Haase AD, Hannon GJ, Joshua-Tor L. 2013. The making of a slicer: activation of human Argonaute-1. *Cell Rep* **3**: 1901–1909.
- Fagerlund RD, Wilkinson ME, Klykov O, Barendregt A, Pearce FG, Kieper SN, Maxwell HWR, Capolupo A, Heck AJR, Krause KL, et al. 2017. Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. *Proc Natl Acad Sci USA* **114**: E5122–E5128.
- Fineran PC, Gerritzen MJH, Suárez-Diez M, Künne T, Boekhorst J, van Hijum SAFT, Staals RHJ, Brouns SJJ. 2014. Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc Natl Acad Sci USA* **111**: E1629–38.
- Finn RD, Clements J, Eddy SR. 2011. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research* **39**: W29–37.
- Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**: 806–811.
- Fonfara I, Richter H, Bratovič M, Le Rhun A, Charpentier E. 2016. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **532**: 517–521.

- Fu BXH, St Onge RP, Fire AZ, Smith JD. 2016. Distinct patterns of Cas9 mismatch tolerance in vitro and in vivo. *Nucleic Acids Research* **44**: 5365–5377.
- Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**: 3150–3152.
- Fu Y, Sander JD, Reyon D, Cascio VM, Joung JK. 2014. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. *Nat Biotechnol* **32**: 279–284.
- Gao P, Yang H, Rajashankar KR, Huang Z, Patel DJ. 2016. Type V CRISPR-Cas Cpf1 endonuclease employs a unique mechanism for crRNA-mediated target DNA recognition. *Cell Res* **26**: 901–913.
- Garneau JE, Dupuis M-È, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadán AH, Moineau S. 2010. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**: 67–71.
- Garside EL, Schellenberg MJ, Gesner EM, Bonanno JB, Sauder JM, Burley SK, Almo SC, Mehta G, Macmillan AM. 2012. Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. *RNA* **18**: 2020–2028.
- Gasiunas G, Barrangou R, Horvath P, Siksnys V. 2012. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc Natl Acad Sci USA* **109**: E2579–86.
- Gaudelli NM, Komor AC, Rees HA, Packer MS, Badran AH, Bryson DI, Liu DR. 2017. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**: 464–471.
- Gesner EM, Schellenberg MJ, Garside EL, George MM, Macmillan AM. 2011. Recognition and maturation of effector RNAs in a CRISPR interference pathway. *Nat Struct Mol Biol* **18**: 688–692.
- Gilbert LA, Horlbeck MA, Adamson B, Villalta JE, Chen Y, Whitehead EH, Guimaraes C, Panning B, Ploegh HL, Bassik MC, et al. 2014. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**: 647–661.
- Goldberg GW, McMillan EA, Varble A, Modell JW, Samai P, Jiang W, Marraffini LA. 2018. Incomplete prophage tolerance by type III-A CRISPR-Cas systems reduces the fitness of lysogenic hosts. *Nat Commun* **9**: 61.
- Goldfarb T, Sberro H, Weinstock E, Cohen O, Doron S, Charpak-Amikam Y, Afik S, Ofir G, Sorek R. 2015. BREX is a novel phage resistance system widespread in microbial genomes. *Embo J* **34**: 169–183.
- Gootenberg JS, Abudayyeh OO, Lee JW, Essletzbichler P, Dy AJ, Joung J, Verdine V, Donghia N, Daringer NM, Freije CA, et al. 2017. Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* **356**: 438–442.
- Gorski SA, Vogel J, Doudna JA. 2017. RNA-based recognition and targeting: sowing the seeds of specificity. *Nat Rev Mol Cell Biol* **18**: 215–228.
- Grissa I, Vergnaud G, Pourcel C. 2007. CRISPRFinder: A web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research* **35**: W52–7.

- Guerrero-Ferreira RC, Viollier PH, Ely B, Poindexter JS, Georgieva M, Jensen GJ, Wright ER. 2011. Alternative mechanism for bacteriophage adsorption to the motile bacterium *Caulobacter crescentus*. *Proc Natl Acad Sci USA* **108**: 9963–9968.
- Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307–321.
- Haase AD, Jaskiewicz L, Zhang H, Lainé S, Sack R, Gatignol A, Filipowicz W. 2005. TRBP, a regulator of cellular PKR and HIV-1 virus expression, interacts with Dicer and functions in RNA silencing. *Embo Rep* **6**: 961–967.
- Haft DH, Selengut J, Mongodin EF, Nelson KE. 2005. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/cas subtypes exist in prokaryotic genomes. *PLoS Comput Biol* **1**: 0474–0483.
- Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP. 2009. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell* **139**: 945–956.
- Hamilton AJ, Baulcombe DC. 1999. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* **286**: 950–952.
- Hammond SM. 2005. Dicing and slicing: the core machinery of the RNA interference pathway. *Febs Lett* **579**: 5822–5829.
- Hammond SM, Bernstein E, Beach D, Hannon GJ. 2000. An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature* **404**: 293–296.
- Han J, Lee Y, Yeom K-H, Kim Y-K, Jin H, Kim VN. 2004. The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev* **18**: 3016–3027.
- Han J, Lee Y, Yeom K-H, Nam J-W, Heo I, Rhee J-K, Sohn SY, Cho Y, Zhang B-T, Kim VN. 2006. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* **125**: 887–901.
- Hatoum-Aslan A, Samai P, Maniv I, Jiang W, Marraffini LA. 2013. A ruler protein in a complex for antiviral defense determines the length of small interfering CRISPR RNAs. *J Biol Chem* **288**: 27888–27897.
- Haurwitz RE, Jinek M, Wiedenheft B, Zhou K, Doudna JA. 2010. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* **329**: 1355–1358.
- Haurwitz RE, Sternberg SH, Doudna JA. 2012. Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA. *Embo J* **31**: 2824–2832.
- Hayes RP, Xiao Y, Ding F, van Erp PBG, Rajashankar K, Bailey S, Wiedenheft B, Ke A. 2016. Structural basis for promiscuous PAM recognition in type I-E Cascade from *E. coli*. *Nature* **530**: 499–503.
- Hegge JW, Swarts DC, van der Oost J. 2018. Prokaryotic Argonaute proteins: novel genome-editing tools? *Nat Rev Microbiol* **16**: 5–11.
- Heler R, Samai P, Modell JW, Weiner C, Goldberg GW, Bikard D, Marraffini LA. 2015. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. *Nature* **519**: 199–202.
- Helwak A, Kudla G, Dudnakova T, Tollervey D. 2013. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*

153: 654–665.

- Herzog VA, Ameres SL. 2015. Approaching the Golden Fleece a Molecule at a Time: Biophysical Insights into Argonaute-Instructed Nucleic Acid Interactions. *Molecular Cell* **59**: 4–7.
- Hille F, Charpentier E. 2016. CRISPR-Cas: biology, mechanisms and relevance. *Philos Trans R Soc Lond, B, Biol Sci* **371**: 20150496.
- Hirano H, Gootenberg JS, Horii T, Abudayyeh OO, Kimura M, Hsu PD, Nakane T, Ishitani R, Hatada I, Zhang F, et al. 2016. Structure and Engineering of Francisella novicida Cas9. - PubMed - NCBI. *Cell* **164**: 950–961. <https://www.ncbi.nlm.nih.gov/pubmed/26875867>.
- Hochstrasser ML, Doudna JA. 2015. Cutting it close: CRISPR-associated endoribonuclease structure and function. *Trends Biochem Sci* **40**: 58–66.
- Hochstrasser ML, Taylor DW, Bhat P, Guegler CK, Sternberg SH, Nogales E, Doudna JA. 2014. CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proc Natl Acad Sci USA* **111**: 6618–6623.
- Hochstrasser ML, Taylor DW, Kornfeld JE, Nogales E, Doudna JA. 2016. DNA Targeting by a Minimal CRISPR RNA-Guided Cascade. *Molecular Cell* **63**: 840–851.
- Hommais F, Krin E, Laurent-Winter C, Soutourina O, Malpertuy A, Le Caer JP, Danchin A, Bertin P. 2001. Large-scale monitoring of pleiotropic regulation of gene expression by the prokaryotic nucleoid-associated protein, H-NS. *Mol Microbiol* **40**: 20–36.
- Höck J, Meister G. 2008. The Argonaute protein family. *Genome Biol* **9**: 210.
- Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, et al. 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31**: 827–832.
- Huang X, Fejes Tóth K, Aravin AA. 2017. piRNA Biogenesis in Drosophila melanogaster. *Trends Genet* **33**: 882–894.
- Huo Y, Nam KH, Ding F, Lee H, Wu L, Xiao Y, Farchione MD, Zhou S, Rajashankar K, Kurinov I, et al. 2014. Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. *Nat Struct Mol Biol* **21**: 771–777.
- Hutvagner G, Simard MJ. 2008. Argonaute proteins: Key players in RNA silencing. *Nat Rev Mol Cell Biol* **9**: 22–32.
- Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**: 119.
- Iki T, Yoshikawa M, Nishikiori M, Jaudal MC, Matsumoto-Yokoyama E, Mitsuhashi I, Meshi T, Ishikawa M. 2010. In vitro assembly of plant RNA-induced silencing complexes facilitated by molecular chaperone HSP90. *Molecular Cell* **39**: 282–291.
- Iwasaki S, Kobayashi M, Yoda M, Sakaguchi Y, Katsuma S, Suzuki T, Tomari Y. 2010. Hsc70/Hsp90 chaperone machinery mediates ATP-dependent RISC loading of small RNA duplexes. *Molecular Cell* **39**: 292–299.
- Jackson RN, Golden SM, van Erp PBG, Carter J, Westra ER, Brouns SJJ, van der Oost J, Terwilliger TC, Read RJ, Wiedenheft B. 2014. Structural biology.

- Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. *Science* **345**: 1473–1479.
- Jackson RN, Wiedenheft B. 2015. A Conserved Structural Chassis for Mounting Versatile CRISPR RNA-Guided Immune Responses. *Molecular Cell* **58**: 722–728.
- Jaroszewski L, Li Z, Krishna SS, Bakolitsa C, Wooley J, Deacon AM, Wilson IA, Godzik A. 2009. Exploration of uncharted regions of the protein universe. ed. C. Chothia. *PLoS Biol* **7**: e1000205.
- Jeltsch A, Pingoud A. 1996. Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. *J Mol Evol* **42**: 91–96.
- Jiang F, Doudna JA. 2017. *CRISPR-Cas9 Structures and Mechanisms*.
- Jiang F, Taylor DW, Chen JS, Kornfeld JE, Zhou K, Thompson AJ, Nogales E, Doudna JA. 2016a. Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* **351**: 867–871.
- Jiang F, Zhou K, Ma L, Gressel S, Doudna JA. 2015. STRUCTURAL BIOLOGY. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science* **348**: 1477–1481.
- Jiang W, Bikard D, Cox D, Zhang F, Marraffini LA. 2013. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat Biotechnol* **31**: 233–239.
- Jiang W, Samai P, Marraffini LA. 2016b. Degradation of Phage Transcripts by CRISPR-Associated RNases Enables Type III CRISPR-Cas Immunity. *Cell* **164**: 710–721.
- Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**: 816–821.
- Jinek M, Doudna JA. 2009. A three-dimensional view of the molecular machinery of RNA interference. *Nature* **457**: 405–412.
- Jinek M, East A, Cheng A, Lin S, Ma E, Doudna J. 2013. RNA-programmed genome editing in human cells. *Elife* **2**: e00471.
- Jo MH, Shin S, Jung S-R, Kim E, Song J-J, Hohng S. 2015. Human Argonaute 2 Has Diverse Reaction Pathways on Target RNAs. *Molecular Cell* **59**: 117–124.
- Jore MM, Lundgren M, van Duijn E, Bultema JB, Westra ER, Waghmare SP, Wiedenheft B, Pul Ü, Wurm R, Wagner R, et al. 2011. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat Struct Mol Biol* **18**: 529–536.
- Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol Rev* **33**: 376–393.
- Jung C, Hawkins JA, Jones SK, Xiao Y, Rybarski JR, Dillard KE, Hussmann J, Saifuddin FA, Savran CA, Ellington AD, et al. 2017. Massively Parallel Biophysical Analysis of CRISPR-Cas Complexes on Next Generation Sequencing Chips. *Cell* **170**: 35–47.e13.
- Jung S-R, Kim E, Hwang W, Shin S, Song J-J, Hohng S. 2013. Dynamic anchoring of the 3'-end of the guide strand controls the target dissociation of

- Argonaute-guide complex. *J Am Chem Soc* **135**: 16865–16871.
- Ka D, Kim D, Baek G, Bae E. 2014. Structural and functional characterization of *Streptococcus pyogenes* Cas2 protein under different pH conditions. *Biochem Biophys Res Commun* **451**: 152–157.
- Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.
- Kawamata T, Tomari Y. 2010. Making RISC. *Trends Biochem Sci* **35**: 368–376.
- Kaya E, Doxzen KW, Knoll KR, Wilson RC, Strutt SC, Kranzusch PJ, Doudna JA. 2016. A bacterial Argonaute with noncanonical guide RNA specificity. *Proc Natl Acad Sci USA* **113**: 4057–4062.
- Kazlauskienė M, Kostiuk G, Venclovas Č, Tamulaitis G, Siksnys V. 2017. A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. *Science* **357**: 605–609.
- Kazlauskienė M, Tamulaitis G, Kostiuk G, Venclovas Č, Siksnys V. 2016. Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition. *Molecular Cell* **62**: 295–306.
- Keam SP, Hutvagner G. 2015. tRNA-Derived Fragments (tRFs): Emerging New Roles for an Ancient RNA in the Regulation of Gene Expression. *Life (Basel)* **5**: 1638–1651.
- Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* **10**: 845–858.
- Kerpedjiev P, Hammer S, Hofacker IL. 2015. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics* **31**: 3377–3379.
- Khvorova A, Reynolds A, Jayasena SD. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115**: 209–216.
- Kidwell MA, Chan JM, Doudna JA. 2014. Evolutionarily conserved roles of the dicer helicase domain in regulating RNA interference processing. *J Biol Chem* **289**: 28352–28362.
- Kieper SN, Almendros C, Behler J, McKenzie RE, Nobrega FL, Haagsma AC, Vink JNA, Hess WR, Brouns SJJ. 2018. Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. *Cell Rep* **22**: 3377–3384.
- Kim E, Koo T, Park SW, Kim D, Kim K, Cho H-Y, Song DW, Lee KJ, Jung MH, Kim S, et al. 2017a. In vivo genome editing with a small Cas9 orthologue derived from *Campylobacter jejuni*. *Nat Commun* **8**: 14500.
- Kim YB, Komor AC, Levy JM, Packer MS, Zhao KT, Liu DR. 2017b. Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nat Biotechnol* **35**: 371–376.
- Klattenhoff C, Theurkauf W. 2008. Biogenesis and germline functions of piRNAs. *Development* **135**: 3–9.
- Kleinstiver BP, Prew MS, Tsai SQ, Nguyen NT, Topkar VV, Zheng Z, Joung JK. 2015. Broadening the targeting range of *Staphylococcus aureus* CRISPR-Cas9 by modifying PAM recognition. *Nat Biotechnol* **33**: 1293–1298.
- Knott GJ, East-Seletsky A, Cofsky JC, Holton JM, Charles E, O'Connell MR,

- Doudna JA. 2017. Guide-bound structures of an RNA-targeting A-cleaving CRISPR-Cas13a enzyme. *Nat Struct Mol Biol* **24**: 825–833.
- Kobayashi H, Tomari Y. 2016. RISC assembly: Coordination between small RNAs and Argonaute proteins. *Biochim Biophys Acta* **1859**: 71–81.
- Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR. 2016. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**: 420–424.
- Koonin EV, Makarova KS, Wolf YI. 2017a. Evolutionary Genomics of Defense Systems in Archaea and Bacteria. *Annu Rev Microbiol* **71**: 233–261.
- Koonin EV, Makarova KS, Zhang F. 2017b. Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol* **37**: 67–78.
- Kou R, Lam H, Duan H, Ye L, Jongkam N, Chen W, Zhang S, Li S. 2016. Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations. ed. J. Wang. *PLoS ONE* **11**: e0146638.
- Kunne T, Swarts DC, Brouns SJ. 2014. Planting the seed: target recognition of short guide RNAs. *Trends Microbiol* **22**: 74–83 PB – M3 – N1 – UR –.
- Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5**: R12.
- Kwak PB, Tomari Y. 2012. The N domain of Argonaute drives duplex unwinding during RISC assembly. *Nat Struct Mol Biol* **19**: 145–151.
- Kwon SC, Nguyen TA, Choi Y-G, Jo MH, Hohng S, Kim VN, Woo J-S. 2016. Structure of Human DROSHA. *Cell* **164**: 81–90.
- Labrie SJ, Samson JE, Moineau S. 2010. Bacteriophage resistance mechanisms. *Nat Rev Microbiol* **8**: 317–327.
- Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.
- Lapinaite A, Doudna JA, Cate JHD. 2018. Programmable RNA recognition using a CRISPR-associated Argonaute. *Proc Natl Acad Sci USA* **115**: 3368–3373.
- Lau P-W, Guiley KZ, De N, Potter CS, Carragher B, MacRae IJ. 2012. The molecular architecture of human Dicer. *Nat Struct Mol Biol* **19**: 436–440.
- Lee HY, Doudna JA. 2012. TRBP alters human precursor microRNA processing in vitro. *RNA* **18**: 2012–2019.
- Lee HY, Zhou K, Smith AM, Noland CL, Doudna JA. 2013. Differential roles of human Dicer-binding proteins TRBP and PACT in small RNA processing. *Nucleic Acids Research* **41**: 6568–6576.
- Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Rådmark O, Kim S, et al. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**: 415–419.
- Lee YS, Nakahara K, Pham JW, Kim K, He Z, Sontheimer EJ, Carthew RW. 2004. Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell* **117**: 69–81.
- Lee YS, Shibata Y, Malhotra A, Dutta A. 2009. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev* **23**: 2639–2649.
- Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the

- display and annotation of phylogenetic and other trees. *Nucleic Acids Research* **44**: W242–5.
- Levy A, Goren MG, Yosef I, Auster O, Manor M, Amitai G, Edgar R, Qimron U, Sorek R. 2015. CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**: 505–510.
- Li H. 2015. Structural Principles of CRISPR RNA Processing. *Structure* **23**: 13–20.
- Li M, Wang R, Xiang H. 2014. Haloarcula hispanica CRISPR authenticates PAM of a target sequence to prime discriminative adaptation. *Nucleic Acids Research* **42**: 7226–7235.
- Li S-Y, Cheng Q-X, Liu J-K, Nie X-Q, Zhao G-P, Wang J. 2018. CRISPR-Cas12a has both cis- and trans-cleavage activities on single-stranded DNA. *Cell Res* **163**: 759.
- Lingel A, Simon B, Izaurralde E, Sattler M. 2004. Nucleic acid 3'-end recognition by the Argonaute2 PAZ domain. *Nat Struct Mol Biol* **11**: 576–577.
- Lintner NG, Kerou M, Brumfield SK, Graham S, Liu H, Naismith JH, 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 J, Carmell MA, Rivas FV, Marsden CG, Thomson JM, Song J-J, Hammond SM, Joshua-Tor L, Hannon GJ. 2004. Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305**: 1437–1441.
- Liu L, Chen P, Wang M, Li X, Wang J, Yin M, Wang Y. 2017a. C2c1-sgRNA Complex Structure Reveals RNA-Guided DNA Cleavage Mechanism. *Molecular Cell* **65**: 310–322.
- Liu L, Li X, Wang J, Wang M, Chen P, Yin M, Li J, Sheng G, Wang Y. 2017b. Two Distant Catalytic Sites Are Responsible for C2c2 RNase Activities. *Cell* **168**: 121–134.e12.
- Liu Q, Rand TA, Kalidas S, Du F, Kim H-E, Smith DP, Wang X. 2003. R2D2, a bridge between the initiation and effector steps of the Drosophila RNAi pathway. *Science* **301**: 1921–1925.
- Liu TY, Iavarone AT, Doudna JA. 2017c. RNA and DNA Targeting by a Reconstituted Thermus thermophilus Type III-A CRISPR-Cas System. ed. S. Korolev. *PLoS ONE* **12**: e0170552.
- Liu Y, Chen Z, He A, Zhan Y, Li J, Liu L, Wu H, Zhuang C, Lin J, Zhang Q, et al. 2016. Targeting cellular mRNAs translation by CRISPR-Cas9. *Sci Rep* **6**: srep29652. <https://www.nature.com/articles/srep29652>.
- Lo A, Qi L. 2017. Genetic and epigenetic control of gene expression by CRISPR-Cas systems. *F1000Res* **6**: 747.
- Loenen WAM, Dryden DTF, Raleigh EA, Wilson GG, Murray NE. 2014. Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Research* **42**: 3–19.
- Lopes APY, Lopes LM, Fraga TR, Chura-Chambi RM, Sanson AL, Cheng E, Nakajima E, Morganti L, Martins EAL. 2014. VapC from the leptospiral VapBC toxin-antitoxin module displays ribonuclease activity on the initiator

- tRNA. ed. P. Hoskisson. *PLoS ONE* **9**: e101678.
- Loughrey D, Watters KE, Settle AH, Lucks JB. 2014. SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Research* **42**: e165–e165.
- Lund E, Dahlberg JE. 2006. Substrate selectivity of exportin 5 and Dicer in the biogenesis of microRNAs. *Cold Spring Harb Symp Quant Biol* **71**: 59–66.
- Lund E, Güttinger S, Calado A, Dahlberg JE, Kutay U. 2004. Nuclear export of microRNA precursors. *Science* **303**: 95–98.
- Luo ML, Jackson RN, Denny SR, Tokmina-Lukaszewska M, Maksimchuk KR, Lin W, Bothner B, Wiedenheft B, Beisel CL. 2016. The CRISPR RNA-guided surveillance complex in *Escherichia coli* accommodates extended RNA spacers. *Nucleic Acids Research* **44**: 7385–7394.
- LURIA SE, HUMAN ML. 1952. A nonhereditary, host-induced variation of bacterial viruses. *Journal of Bacteriology* **64**: 557–569.
- Ma D, Liu F. 2015. Genome Editing and Its Applications in Model Organisms. *Genomics Proteomics Bioinformatics* **13**: 336–344.
- Ma E, Harrington LB, O'Connell MR, Zhou K, Doudna JA. 2015. Single-Stranded DNA Cleavage by Divergent CRISPR-Cas9 Enzymes. *Molecular Cell* **60**: 398–407.
- Ma H, Tu L-C, Naseri A, Huisman M, Zhang S, Grunwald D, Pederson T. 2016. CRISPR-Cas9 nuclear dynamics and target recognition in living cells. *J Cell Biol* **214**: 529–537.
- Ma H, Wu Y, Choi J-G, Wu H. 2013. Lower and upper stem-single-stranded RNA junctions together determine the Drosha cleavage site. *Proc Natl Acad Sci USA* **110**: 20687–20692.
- Ma J-B, Ye K, Patel DJ. 2004. Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature* **429**: 318–322.
- Ma J-B, Yuan Y-R, Meister G, Pei Y, Tuschl T, Patel DJ. 2005. Structural basis for 5'-end-specific recognition of guide RNA by the *A. fulgidus* Piwi protein. *Nature* **434**: 666–670.
- MacRae IJ, Ma E, Zhou M, Robinson CV, Doudna JA. 2008. In vitro reconstitution of the human RISC-loading complex. *Proc Natl Acad Sci USA* **105**: 512–517.
- MacRae IJ, Zhou K, Li F, Repic A, Brooks AN, Cande WZ, Adams PD, Doudna JA. 2006. Structural basis for double-stranded RNA processing by Dicer. *Science* **311**: 195–198.
- Makarova KS, Anantharaman V, Aravind L, Koonin EV. 2012. Live virus-free or die: coupling of antiviral immunity and programmed suicide or dormancy in prokaryotes. *Biol Direct* **7**: 40.
- Makarova KS, Aravind L, Wolf YI, Koonin EV. 2011a. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol Direct* **6**: 38.
- Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. 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. <http://biologydirect.biomedcentral.com/articles/10.1186/1745-6150-1-7>.
- Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, Moineau S, Mojica FJM, Wolf YI, Yakunin AF, et al. 2011b. Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**: 467–477. <http://www.nature.com/articles/nrmicro2577>.
- Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJJ, Charpentier E, Haft DH, et al. 2015. An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* **13**: 722–736.
- Makarova KS, Wolf YI, Koonin EV. 2013. Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Research* **41**: 4360–4377.
- Makarova KS, Wolf YI, Snir S, Koonin EV. 2011c. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *Journal of Bacteriology* **193**: 6039–6056.
- Makarova KS, Wolf YI, van der Oost J, Koonin EV. 2009. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biol Direct* **4**: 29.
- Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM. 2013. RNA-guided human genome engineering via Cas9. *Science* **339**: 823–826.
- Maniv I, Jiang W, Bikard D, Marraffini LA. 2016. Impact of Different Target Sequences on Type III CRISPR-Cas Immunity. ed. G.A. O'Toole. *Journal of Bacteriology* **198**: 941–950.
- Marraffini LA, Sontheimer EJ. 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**: 1843–1845.
- Marraffini LA, Sontheimer EJ. 2010. Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* **463**: 568–571.
- Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* **17**: 10.
- Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. 2012. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* **13**: 31.
- Matranga C, Tomari Y, Shin C, Bartel DP, Zamore PD. 2005. Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes. *Cell* **123**: 607–620.
- Mattick JS. 2002. Type IV pili and twitching motility. *Annu Rev Microbiol* **56**: 289–314.
- McCarthy DJ, Chen Y, Smyth GK. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40**: 4288–4297.
- Medina-Aparicio L, Rebollar-Flores JE, Gallego-Hernández AL, Vázquez A, Olvera L, Gutiérrez-Ríos RM, Calva E, Hernández-Lucas I. 2011. The CRISPR/Cas immune system is an operon regulated by LeuO, H-NS, and leucine-responsive regulatory protein in *Salmonella enterica* serovar Typhi. *Journal of Bacteriology* **193**: 2396–2407.

- Meister G. 2013. Argonaute proteins: functional insights and emerging roles. *Nat Rev Genet* **14**: 447–459.
- Meister G, Landthaler M, Patkaniowska A, Dorsett Y, Teng G, Tuschl T. 2004. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Molecular Cell* **15**: 185–197.
- Miller CS, Baker BJ, Thomas BC, Singer SW, Banfield JF. 2011. EMIRGE: reconstruction of full-length ribosomal genes from microbial community short read sequencing data. *Genome Biol* **12**: R44.
- Miyoshi T, Ito K, Murakami R, Uchiumi T. 2016. Structural basis for the recognition of guide RNA and target DNA heteroduplex by Argonaute. *Nat Commun* **7**: 11846.
- Miyoshi T, Takeuchi A, Siomi H, Siomi MC. 2010. A direct role for Hsp90 in pre-RISC formation in *Drosophila*. *Nat Struct Mol Biol* **17**: 1024–1026.
- Mohanraju P, Makarova KS, Zetsche B, Zhang F, Koonin EV, van der Oost J. 2016. Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. *Science* **353**: aad5147.
- Mojica FJM, Díez-Villaseñor C, García-Martínez J, Soria E. 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**: 174–182.
- Monroe KM, McWhirter SM, Vance RE. 2009. Identification of host cytosolic sensors and bacterial factors regulating the type I interferon response to *Legionella pneumophila*. ed. R.R. Isberg. *PLoS Pathog* **5**: e1000665.
- Montgomery TA, Rim Y-S, Zhang C, Downen RH, Phillips CM, Fischer SEJ, Ruvkun G. 2012. PIWI associated siRNAs and piRNAs specifically require the *Caenorhabditis elegans* HEN1 ortholog henn-1. ed. S.K. Kim. *PLoS Genet* **8**: e1002616.
- Morse RP, Willett JLE, Johnson PM, Zheng J, Credali A, Iniguez A, Nowick JS, Hayes CS, Goulding CW. 2015. Diversification of β -Augmentation Interactions between CDI Toxin/Immunity Proteins. *J Mol Biol* **427**: 3766–3784.
- Mout R, Ray M, Yesilbag Tonga G, Lee Y-W, Tay T, Sasaki K, Rotello VM. 2017. Direct Cytosolic Delivery of CRISPR/Cas9-Ribonucleoprotein for Efficient Gene Editing. *ACS Nano* **11**: 2452–2458.
- Mucke M, Krüger DH, Reuter M. 2003. Diversity of type II restriction endonucleases that require two DNA recognition sites. *Nucleic Acids Research* **31**: 6079–6084.
- Mulepati S, Bailey S. 2013. In vitro reconstitution of an *Escherichia coli* RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA target. *J Biol Chem* **288**: 22184–22192.
- Mulepati S, Héroux A, Bailey S. 2014. Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science* **345**: 1479–1484.
- Murat D, Quinlan A, Vali H, Komeili A. 2010. Comprehensive genetic dissection of the magnetosome gene island reveals the step-wise assembly of a prokaryotic organelle. *Proc Natl Acad Sci USA* **107**: 5593–5598.
- Murray NE. 2002. 2001 Fred Griffith review lecture. Immigration control of DNA in

- bacteria: self versus non-self. *Microbiology (Reading, Engl)* **148**: 3–20.
- Murugan K, Babu K, Sundaresan R, Rajan R, Sashital DG. 2017. The Revolution Continues: Newly Discovered Systems Expand the CRISPR-Cas Toolkit. *Molecular Cell* **68**: 15–25.
- Myers CR, Myers JM. 1997. Replication of plasmids with the p15A origin in *Shewanella putrefaciens* MR-1. *Lett Appl Microbiol* **24**: 221–225.
- Nakanishi K. 2016. Anatomy of RISC: how do small RNAs and chaperones activate Argonaute proteins? *Wiley Interdiscip Rev RNA* **7**: 637–660.
- Nakanishi K, Weinberg DE, Bartel DP, Patel DJ. 2012. Structure of yeast Argonaute with guide RNA. *Nature* **486**: 368–374.
- Nam KH, Haitjema C, Liu X, Ding F, Wang H, DeLisa MP, Ke A. 2012a. Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. *Structure* **20**: 1574–1584.
- Nam KH, Huang Q, Ke A. 2012b. Nucleic acid binding surface and dimer interface revealed by CRISPR-associated CasB protein structures. *Febs Lett* **586**: 3956–3961.
- Napoli C, Lemieux C, Jorgensen R. 1990. Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. *Plant Cell* **2**: 279–289.
- Narihiro T, Kamagata Y. 2017. Genomics and Metagenomics in Microbial Ecology: Recent Advances and Challenges. *Microbes Environ* **32**: 1–4.
- Nelles DA, Fang MY, Aigner S, Yeo GW. 2015. Applications of Cas9 as an RNA-programmed RNA-binding protein. *Bioessays* **37**: 732–739.
- Nelles DA, Fang MY, O'Connell MR, Xu JL, Markmiller SJ, Doudna JA, Yeo GW. 2016. Programmable RNA Tracking in Live Cells with CRISPR/Cas9. *Cell* **165**: 488–496. [http://www.cell.com/cell/abstract/S0092-8674\(16\)30204-5](http://www.cell.com/cell/abstract/S0092-8674(16)30204-5).
- Nguyen TA, Jo MH, Choi Y-G, Park J, Kwon SC, Hohng S, Kim VN, Woo J-S. 2015. Functional Anatomy of the Human Microprocessor. *Cell* **161**: 1374–1387.
- Niewoehner O, Garcia-Doval C, Rostøl JT, Berk C, Schwede F, Bigler L, Hall J, Marraffini LA, Jinek M. 2017. Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. *Nature* **548**: 543–548.
- Niewoehner O, Jinek M. 2016. Structural basis for the endoribonuclease activity of the type III-A CRISPR-associated protein Csm6. *RNA* **22**: 318–329.
- Niewoehner O, Jinek M, Doudna JA. 2014. Evolution of CRISPR RNA recognition and processing by Cas6 endonucleases. *Nucleic Acids Research* **42**: 1341–1353.
- Nishimasu H, Cong L, Yan WX, Ran FA, Zetsche B, Li Y, Kurabayashi A, Ishitani R, Zhang F, Nureki O. 2015. Crystal Structure of Staphylococcus aureus Cas9. *Cell* **162**: 1113–1126.
- Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, Ishitani R, Zhang F, Nureki O. 2014. Crystal structure of Cas9 in complex with guide RNA and target DNA. - PubMed - NCBI. *Cell* **156**: 935–949. <https://www.ncbi.nlm.nih.gov/pubmed/24529477?dopt=Abstract>.
- Noland CL, Doudna JA. 2013. Multiple sensors ensure guide strand selection in

- human RNAi pathways. *RNA* **19**: 639–648.
- Noland CL, Ma E, Doudna JA. 2011. siRNA repositioning for guide strand selection by human Dicer complexes. *Molecular Cell* **43**: 110–121.
- Nuñez JK, Bai L, Harrington LB, Hinder TL, Doudna JA. 2016. CRISPR Immunological Memory Requires a Host Factor for Specificity. *Molecular Cell* **62**: 824–833.
- Nuñez JK, Harrington LB, Kranzusch PJ, Engelman AN, Doudna JA. 2015a. Foreign DNA capture during CRISPR-Cas adaptive immunity. *Nature* **527**: 535–538.
- Nuñez JK, Kranzusch PJ, Noeske J, Wright AV, Davies CW, Doudna JA. 2014. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat Struct Mol Biol* **21**: 528–534.
- Nuñez JK, Lee ASY, Engelman A, Doudna JA. 2015b. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* **519**: 193–198.
- Oakes BL, Nadler DC, Flamholz A, Fellmann C, Staahl BT, Doudna JA, Savage DF. 2016. Profiling of engineering hotspots identifies an allosteric CRISPR-Cas9 switch. *Nat Biotechnol* **34**: 646–651.
- Ofir G, Melamed S, Sberro H, Mukamel Z, Silverman S, Yaakov G, Doron S, Sorek R. 2018. DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat Microbiol* **3**: 90–98.
- Okamura K, Ishizuka A, Siomi H, Siomi MC. 2004. Distinct roles for Argonaute proteins in small RNA-directed RNA cleavage pathways. *Genes Dev* **18**: 1655–1666.
- Olovnikov I, Chan K, Sachidanandam R, Newman DK, Aravin AA. 2013. Bacterial argonaute samples the transcriptome to identify foreign DNA. *Molecular Cell* **51**: 594–605.
- Oshima T, Ishikawa S, Kurokawa K, Aiba H, Ogasawara N. 2006. Escherichia coli histone-like protein H-NS preferentially binds to horizontally acquired DNA in association with RNA polymerase. *DNA Res* **13**: 141–153.
- Osterman IA, Dikhtyar YY, Bogdanov AA, Dontsova OA, Sergiev PV. 2015. Regulation of Flagellar Gene Expression in Bacteria. *Biochemistry Mosc* **80**: 1447–1456.
- O’Connell MR, Oakes BL, Sternberg SH, East-Seletsky A, Kaplan M, Doudna JA. 2014. Programmable RNA recognition and cleavage by CRISPR/Cas9. *Nature* **516**: 263–266.
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4268322/>.
- Page R, Peti W. 2016. Toxin-antitoxin systems in bacterial growth arrest and persistence. *Nat Chem Biol* **12**: 208–214.
- Pandey DP, Gerdes K. 2005. Toxin-antitoxin loci are highly abundant in free-living but lost from host-associated prokaryotes. *Nucleic Acids Research* **33**: 966–976.
- Pang YLJ, Abo R, Levine SS, Dedon PC. 2014. Diverse cell stresses induce unique patterns of tRNA up- and down-regulation: tRNA-seq for quantifying changes in tRNA copy number. *Nucleic Acids Research* **42**: e170–e170.
- Park MS, Phan H-D, Busch F, Hinckley SH, Brackbill JA, Wysocki VH, Nakanishi

- K. 2017. Human Argonaute3 has slicer activity. *Nucleic Acids Research* **45**: 11867–11877.
- Parker JS, Parizotto EA, Wang M, Roe SM, Barford D. 2009. Enhancement of the seed-target recognition step in RNA silencing by a PIWI/MID domain protein. *Molecular Cell* **33**: 204–214.
- Parker JS, Roe SM, Barford D. 2004. Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity. *Embo J* **23**: 4727–4737.
- Parker JS, Roe SM, Barford D. 2005. Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. *Nature* **434**: 663–666.
- Pei J, Kim B-H, Grishin NV. 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Research* **36**: 2295–2300.
- Peng Y, Leung HCM, Yiu SM, Chin FYL. 2012. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420–1428.
- Pham JW, Pellino JL, Lee YS, Carthew RW, Sontheimer EJ. 2004. A Dicer-2-dependent 80s complex cleaves targeted mRNAs during RNAi in *Drosophila*. *Cell* **117**: 83–94.
- Phanstiel DH, Boyle AP, Araya CL, Snyder MP. 2014. Sushi.R: flexible, quantitative and integrative genomic visualizations for publication-quality multi-panel figures. *Bioinformatics* **30**: 2808–2810.
- Plagens A, Tripp V, Daume M, Sharma K, Klingl A, Hrle A, Conti E, Urlaub H, Randau L. 2014. In vitro assembly and activity of an archaeal CRISPR-Cas type I-A Cascade interference complex. *Nucleic Acids Research* **42**: 5125–5138.
- Pougach K, Semenova E, Bogdanova E, Datsenko KA, Djordjevic M, Wanner BL, Severinov K. 2010. Transcription, processing and function of CRISPR cassettes in *Escherichia coli*. *Mol Microbiol* **77**: 1367–1379.
- Pourcel C, Salvignol G, 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.
- Probst AJ, Castelle CJ, Singh A, Brown CT, Anantharaman K, Sharon I, Hug LA, Burstein D, Emerson JB, Thomas BC, et al. 2017. Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO₂ concentrations. *Environ Microbiol* **19**: 459–474.
- Pul Ü, Wurm R, Arslan Z, Geissen R, Hofmann N, Wagner R. 2010. Identification and characterization of *E. coli* CRISPR-cas promoters and their silencing by H-NS. *Mol Microbiol* **75**: 1495–1512.
- Qi LS, Larson MH, Gilbert LA, Doudna JA, Weissman JS, Arkin AP, Lim WA. 2013. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**: 1173–1183.
- Ran FA, Cong L, Yan WX, Scott DA, Gootenberg JS, Kriz AJ, Zetsche B, Shalem O, Wu X, Makarova KS, et al. 2015. In vivo genome editing using *Staphylococcus aureus* Cas9. - PubMed - NCBI. *Nature* **520**: 186–191.

- <https://www.ncbi.nlm.nih.gov/pubmed/25830891>.
- Rashid UJ, Paterok D, Koglin A, Gohlke H, Piehler J, Chen JC-H. 2007. Structure of Aquifex aeolicus argonaute highlights conformational flexibility of the PAZ domain as a potential regulator of RNA-induced silencing complex function. *J Biol Chem* **282**: 13824–13832.
- Rath D, Amlinger L, Rath A, Lundgren M. 2015. The CRISPR-Cas immune system: biology, mechanisms and applications. *Biochimie* **117**: 119–128.
- Redding S, Sternberg SH, Marshall M, Gibb B, Bhat P, Guegler CK, Wiedenheft B, Doudna JA, Greene EC. 2015. Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. *Cell* **163**: 854–865.
- Reeks J, Sokolowski RD, Graham S, Liu H, Naismith JH, White MF. 2013. Structure of a dimeric crenarchaeal Cas6 enzyme with an atypical active site for CRISPR RNA processing. *Biochem J* **452**: 223–230.
- Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. 2004. Fast and effective prediction of microRNA/target duplexes. *RNA* **10**: 1507–1517.
- Richter C, Dy RL, McKenzie RE, Watson BNJ, Taylor C, Chang JT, McNeil MB, Staals RHJ, Fineran PC. 2014. Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic Acids Research* **42**: 8516–8526.
- Rio DC. 2012. Filter-binding assay for analysis of RNA-protein interactions. *Cold Spring Harb Protoc* **2012**: 1078–1081.
- Rivas FV, Tolia NH, Song J-J, Aragon JP, Liu J, Hannon GJ, Joshua-Tor L. 2005. Purified Argonaute2 and an siRNA form recombinant human RISC. *Nat Struct Mol Biol* **12**: 340–349.
- Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, Blumenthal RM, Degtyarev SK, Dryden DTF, Dybvig K, et al. 2003. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Research* **31**: 1805–1812.
- Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**: 139–140.
- Rollie C, Graham S, Rouillon C, White MF. 2018. Prespacer processing and specific integration in a Type I-A CRISPR system. *Nucleic Acids Research* **46**: 1007–1020.
- Rollie C, Schneider S, Brinkmann AS, Bolt EL, White MF. 2015. Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. *Elife* **4**: 7884.
- Rollins MF, Chowdhury S, Carter J, Golden SM, Wilkinson RA, Bondy-Denomy J, Lander GC, Wiedenheft B. 2017. Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. *Proc Natl Acad Sci USA* **114**: E5113–E5121.
- Romano N, Macino G. 1992. Quelling: transient inactivation of gene expression in Neurospora crassa by transformation with homologous sequences. *Mol Microbiol* **6**: 3343–3353.
- Rouillon C, Zhou M, Zhang J, Politis A, Beilsten-Edmands V, Cannone G, Graham S, Robinson CV, Spagnolo L, White MF. 2013. Structure of the

- CRISPR interference complex CSM reveals key similarities with cascade. *Molecular Cell* **52**: 124–134.
- Rousseau BA, Hou Z, Gramelspacher MJ, Zhang Y. 2018. Programmable RNA Cleavage and Recognition by a Natural CRISPR-Cas9 System from *Neisseria meningitidis*. *Molecular Cell* **69**: 906–914.e4.
- Rutkauskas M, Sinkunas T, Songailiene I, Tikhomirova MS, Siksnyis V, Seidel R. 2015. Directional R-loop formation by the CRISPR-cas surveillance complex cascade provides efficient off-target site rejection. *Cell Rep* **10**: 1534–1543.
- Saini HK, Griffiths-Jones S, Enright AJ. 2007. Genomic analysis of human microRNA transcripts. *Proc Natl Acad Sci USA* **104**: 17719–17724.
- Salomon WE, Jolly SM, Moore MJ, Zamore PD, Serebrov V. 2015. Single-Molecule Imaging Reveals that Argonaute Reshapes the Binding Properties of Its Nucleic Acid Guides. *Cell* **162**: 84–95.
- Saltikov CW, Wildman RA, Newman DK. 2005. Expression dynamics of arsenic respiration and detoxification in *Shewanella* sp. strain ANA-3. *Journal of Bacteriology* **187**: 7390–7396.
- Samai P, Pyenson N, Jiang W, Goldberg GW, Hatoum-Aslan A, Marraffini LA. 2015. Co-transcriptional DNA and RNA cleavage during type III CRISPR-cas immunity. *Cell* **161**: 1164–1174.
- Samai P, Smith P, Shuman S. 2010. Structure of a CRISPR-associated protein Cas2 from *Desulfovibrio vulgaris*. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **66**: 1552–1556.
- Sampson TR, Saroj SD, Llewellyn AC, Tzeng Y-L, Weiss DS. 2013. A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. *Nature* **497**: 254–257.
- Samson JE, Bélanger M, Moineau S. 2013. Effect of the abortive infection mechanism and type III toxin/antitoxin system AbiQ on the lytic cycle of *Lactococcus lactis* phages. *Journal of Bacteriology* **195**: 3947–3956.
- Sasaki HM, Tomari Y. 2012. The true core of RNA silencing revealed. July 5.
- Sashital DG, Jinek M, Doudna JA. 2011. An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. *Nat Struct Mol Biol* **18**: 680–687.
- Savitskaya E, Semenova E, Dedkov V, Metlitskaya A, Severinov K. 2013. High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in *E. coli*. *RNA Biol* **10**: 716–725.
- Schirle NT, MacRae IJ. 2012. The crystal structure of human Argonaute2. *Science* **336**: 1037–1040.
- Schirle NT, Sheu-Gruttadauria J, MacRae IJ. 2014. Structural basis for microRNA targeting. *Science* **346**: 608–613.
- Schürmann N, Trabuco LG, Bender C, Russell RB, Grimm D. 2013. Molecular dissection of human Argonaute proteins by DNA shuffling. *Nat Struct Mol Biol* **20**: 818–826.
- Schwarz DS, Hutvagner G, Du T, Xu Z, Aronin N, Zamore PD. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115**: 199–208.
- Seed KD. 2015. Battling Phages: How Bacteria Defend against Viral Attack. ed. V.L. Miller. *PLoS Pathog* **11**: e1004847.

- Seed KD, Yen M, Shapiro BJ, Hilaire IJ, Charles RC, Teng JE, Ivers LC, Boncy J, Harris JB, Camilli A. 2014. Evolutionary consequences of intra-patient phage predation on microbial populations. *Elife* **3**: e03497.
- Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, Wanner B, van der Oost J, Brouns SJJ, 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, Savitskaya E, Musharova O, Strotskaya A, Vorontsova D, Datsenko KA, Logacheva MD, Severinov K. 2016. Highly efficient primed spacer acquisition from targets destroyed by the Escherichia coli type I-E CRISPR-Cas interfering complex. *Proc Natl Acad Sci USA* **113**: 7626–7631.
- Shah SA, Erdmann S, Mojica FJM, Garrett RA. 2013. Protospacer recognition motifs: Mixed identities and functional diversity. *RNA Biol* **10**: 891–899.
- Shao Y, Li H. 2013. Recognition and cleavage of a nonstructured CRISPR RNA by its processing endoribonuclease Cas6. *Structure* **21**: 385–393.
- Shao Y, Richter H, Sun S, Sharma K, Urlaub H, Randau L, Li H. 2016. A Non-Stem-Loop CRISPR RNA Is Processed by Dual Binding Cas6. *Structure* **24**: 547–554.
- Sharon I, Banfield JF. 2013. Microbiology. Genomes from metagenomics. *Science* **342**: 1057–1058.
- Sheng G, Zhao H, Wang J, Rao Y, Tian W, Swarts DC, van der Oost J, Patel DJ, Wang Y. 2014. Structure-based cleavage mechanism of Thermus thermophilus Argonaute DNA guide strand-mediated DNA target cleavage. *Proc Natl Acad Sci USA* **111**: 652–657.
- Sheppard NF, Glover CVC, Terns RM, Terns MP. 2016. The CRISPR-associated Csx1 protein of Pyrococcus furiosus is an adenosine-specific endoribonuclease. *RNA* **22**: 216–224.
- Shigematsu M, Honda S, Loher P, Telonis AG, Rigoutsos I, Kirino Y. 2017. YAMAT-seq: an efficient method for high-throughput sequencing of mature transfer RNAs. *Nucleic Acids Research* **45**: e70.
- Shmakov S, Abudayyeh OO, Makarova KS, Wolf YI, Gootenberg JS, Semenova E, Minakhin L, Joung J, Konermann S, Severinov K, et al. 2015. Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. *Molecular Cell* **60**: 385–397.
- Shmakov S, Smargon A, Scott D, Cox D, Pyzocha N, Yan W, Abudayyeh OO, Gootenberg JS, Makarova KS, Wolf YI, et al. 2017. Diversity and evolution of class 2 CRISPR-Cas systems. *Nat Rev Microbiol* **15**: 169–182.
- Silas S, Lucas-Elio P, Jackson SA, Aroca-Crevillén A, Hansen LL, Fineran PC, Fire AZ, Sanchez-Amat A. 2017. Type III CRISPR-Cas systems can provide redundancy to counteract viral escape from type I systems. *Elife* **6**: aaf5573.
- Silas S, Mohr G, Sidote DJ, Markham LM, Sanchez-Amat A, Bhaya D, Lambowitz AM, Fire AZ. 2016. Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science* **351**: aad4234–aad4234.
- Singh D, Sternberg SH, Fei J, Doudna JA, Ha T. 2016. Real-time observation of DNA recognition and rejection by the RNA-guided endonuclease Cas9. *Nat*

- Commun* **7**: 12778.
- Sinkunas T, Gasiunas G, Fremaux C, Barrangou R, Horvath P, 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.
- Siolas D, Lerner C, Burchard J, Ge W, Linsley PS, Paddison PJ, Hannon GJ, Cleary MA. 2005. Synthetic shRNAs as potent RNAi triggers. *Nat Biotechnol* **23**: 227–231.
- Siomi MC, Sato K, Pezic D, Aravin AA. 2011. PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* **12**: 246–258.
- Skenneron CT, Imelfort M, Tyson GW. 2013. Crass: Identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Research* **41**: e105–e105.
- Smargon AA, Cox DBT, Pyzocha NK, Zheng K, Slaymaker IM, Gootenberg JS, Abudayyeh OA, Essletzbichler P, Shmakov S, Makarova KS, et al. 2017. Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. *Molecular Cell* **65**: 618–630.e7.
- Smith T, Heger A, Sudbery I. 2017. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**: 491–499.
- Snyder L. 1995. Phage-exclusion enzymes: a bonanza of biochemical and cell biology reagents? *Mol Microbiol* **15**: 415–420.
- Song J-J, Smith SK, Hannon GJ, Joshua-Tor L. 2004. Crystal structure of argonaute and its implications for RISC slicer activity. *Science* **305**: 1434–1437.
- Sorek R, Kunin V, Hugenholtz P. 2008. CRISPR - A widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol* **6**: 181–186.
- Söding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research* **33**: W244–8.
- Spilman M, Cocozaki A, Hale C, Shao Y, Ramia N, Terns R, Terns M, Li H, Stagg S. 2013. Structure of an RNA silencing complex of the CRISPR-Cas immune system. *Molecular Cell* **52**: 146–152.
- Staals RHJ, Agari Y, Maki-Yonekura S, Zhu Y, Taylor DW, van Duijn E, Barendregt A, Vlot M, Koehorst JJ, Sakamoto K, et al. 2013. Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of *Thermus thermophilus*. *Molecular Cell* **52**: 135–145.
- Staals RHJ, Jackson SA, Biswas A, Brouns SJJ, Brown CM, Fineran PC. 2016. Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. *Nat Commun* **7**: 12853.
- Staals RHJ, Zhu Y, Taylor DW, Kornfeld JE, Sharma K, Barendregt A, Koehorst JJ, Vlot M, Neupane N, Varossieau K, et al. 2014. RNA targeting by the type III-A CRISPR-Cas Csm complex of *Thermus thermophilus*. *Molecular Cell* **56**: 518–530.
- Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-

- analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.
- Stern A, Keren L, Wurtzel O, Amitai G, Sorek R. 2010. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet* **26**: 335–340.
- Stern A, Sorek R. 2011. The phage-host arms race: shaping the evolution of microbes. *Bioessays* **33**: 43–51.
- Sternberg SH, Haurwitz RE, Doudna JA. 2012. Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *RNA* **18**: 661–672.
- Sternberg SH, LaFrance B, Kaplan M, Doudna JA. 2015. Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature* **527**: 110–113.
- Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA. 2014. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**: 62–67.
- Sternberg SH, Richter H, Charpentier E, Qimron U. 2016. Adaptation in CRISPR-Cas Systems. *Molecular Cell* **61**: 797–808.
- Suttle CA. 1994. The significance of viruses to mortality in aquatic microbial communities. *Microb Ecol* **28**: 237–243.
- Swarts DC, Hegge JW, Hinojo I, Shiimori M, Ellis MA, Dumrongkulraksa J, Terns RM, Terns MP, van der Oost J. 2015a. Argonaute of the archaeon *Pyrococcus furiosus* is a DNA-guided nuclease that targets cognate DNA. *Nucleic Acids Research* **43**: 5120–5129.
- Swarts DC, Jore MM, Westra ER, Zhu Y, Janssen JH, Snijders AP, Wang Y, Patel DJ, Berenguer J, Brouns SJJ, et al. 2014a. DNA-guided DNA interference by a prokaryotic Argonaute. *Nature* **507**: 258–261.
- Swarts DC, Koehorst JJ, Westra ER, Schaap PJ, van der Oost J. 2015b. Effects of Argonaute on Gene Expression in *Thermus thermophilus*. ed. L. Randau. *PLoS ONE* **10**: e0124880.
- Swarts DC, Makarova K, Wang Y, Nakanishi K, Ketting RF, Koonin EV, Patel DJ, van der Oost J. 2014b. The evolutionary journey of Argonaute proteins. *Nat Struct Mol Biol* **21**: 743–753.
- Swarts DC, Mosterd C, van Passel MWJ, Brouns SJJ. 2012. CRISPR interference directs strand specific spacer acquisition. ed. I. Mokrousov. *PLoS ONE* **7**: e35888.
- Swarts DC, Szczepaniak M, Sheng G, Chandradoss SD, Zhu Y, Timmers EM, Zhang Y, Zhao H, Lou J, Wang Y, et al. 2017a. Autonomous Generation and Loading of DNA Guides by Bacterial Argonaute. *Molecular Cell* **65**: 985–998.e6.
- Swarts DC, van der Oost J, Jinek M. 2017b. Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. *Molecular Cell* **66**: 221–233.e4.
- Szczelkun MD, Tikhomirova MS, Sinkunas T, Gasiunas G, Karvelis T, Pschera P, Siksnys V, Seidel R. 2014. Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. **111**: 9798–9803. <http://www.pnas.org/lookup/doi/10.1073/pnas.1402597111>.
- Takahashi T, Miyakawa T, Zenno S, Nishi K, Tanokura M, Ui-Tei K. 2013. Distinguishable in vitro binding mode of monomeric TRBP and dimeric PACT with siRNA. ed. A. Hofmann. *PLoS ONE* **8**: e63434.

- Takimoto K, Wakiyama M, Yokoyama S. 2009. Mammalian GW182 contains multiple Argonaute-binding sites and functions in microRNA-mediated translational repression. *RNA* **15**: 1078–1089.
- Tamulaitis G, Kazlauskienė M, Manakova E, Venclovas Č, Nwokeoji AO, Dickman MJ, Horvath P, Siksnys V. 2014. Programmable RNA shredding by the type III-A CRISPR-Cas system of *Streptococcus thermophilus*. *Molecular Cell* **56**: 506–517.
- Tamulaitis G, Venclovas Č, Siksnys V. 2017. Type III CRISPR-Cas Immunity: Major Differences Brushed Aside. *Trends Microbiol* **25**: 49–61.
- Terns RM, Terns MP. 2014. CRISPR-based technologies: prokaryotic defense weapons repurposed. *Trends Genet* **30**: 111–118.
- Tian Y, Simanshu DK, Ma J-B, Park J-E, Heo I, Kim VN, Patel DJ. 2014. A phosphate-binding pocket within the platform-PAZ-connector helix cassette of human Dicer. *Molecular Cell* **53**: 606–616.
- Till S, Lejeune E, Thermann R, Bortfeld M, Hothorn M, Enderle D, Heinrich C, Hentze MW, Ladurner AG. 2007. A conserved motif in Argonaute-interacting proteins mediates functional interactions through the Argonaute PIWI domain. *Nat Struct Mol Biol* **14**: 897–903.
- Tomari Y, Du T, Haley B, Schwarz DS, Bennett R, Cook HA, Koppetsch BS, Theurkauf WE, Zamore PD. 2004a. RISC assembly defects in the *Drosophila* RNAi mutant armitage. *Cell* **116**: 831–841.
- Tomari Y, Matranga C, Haley B, Martinez N, Zamore PD. 2004b. A protein sensor for siRNA asymmetry. *Science* **306**: 1377–1380.
- UniProt Consortium. 2015. UniProt: a hub for protein information. *Nucleic Acids Research* **43**: D204–12.
- Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD. 2006. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313**: 320–324.
- van Elsas JD, Bailey MJ. 2002. The ecology of transfer of mobile genetic elements. *FEMS Microbiol Ecol* **42**: 187–197.
- van Houte S, Buckling A, Westra ER. 2016. Evolutionary Ecology of Prokaryotic Immune Mechanisms. *Microbiol Mol Biol Rev* **80**: 745–763.
- van Rij RP, Saleh M-C, Berry B, Foo C, Houk A, Antoniewski C, Andino R. 2006. The RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity in *Drosophila melanogaster*. *Genes Dev* **20**: 2985–2995.
- Walker FC, Chou-Zheng L, Dunkle JA, Hatoum-Aslan A. 2017. Molecular determinants for CRISPR RNA maturation in the Cas10-Csm complex and roles for non-Cas nucleases. *Nucleic Acids Research* **45**: 2112–2123.
- Wang AH, Fujii S, van Boom JH, van der Marel GA, van Boeckel SA, Rich A. 1982. Molecular structure of r(GCG)_n(TATACGC): a DNA–RNA hybrid helix joined to double helical DNA. *Nature* **299**: 601–604.
- Wang H-W, Noland C, Siridechadilok B, Taylor DW, Ma E, Felderer K, Doudna JA, Nogales E. 2009a. Structural insights into RNA processing by the human RISC-loading complex. *Nat Struct Mol Biol* **16**: 1148–1153.
- Wang J, Li J, Zhao H, Sheng G, Wang M, Yin M, Wang Y. 2015. Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas

- Systems. *Cell* **163**: 840–853.
- Wang R, Preamplume G, Terns MP, Terns RM, Li H. 2011. Interaction of the Cas6 Riboendonuclease with CRISPR RNAs: Recognition and Cleavage. *Structure* **19**: 257–264.
- Wang R, Zheng H, Preamplume G, Shao Y, Li H. 2012. The impact of CRISPR repeat sequence on structures of a Cas6 protein-RNA complex. *Protein Sci* **21**: 405–417.
- Wang Y, Juranek S, Li H, Sheng G, Tuschl T, Patel DJ. 2008a. Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature* **456**: 921–926.
- Wang Y, Juranek S, Li H, Sheng G, Wardle GS, Tuschl T, Patel DJ. 2009b. Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. *Nature* **461**: 754–761.
- Wang Y, Sheng G, Juranek S, Tuschl T, Patel DJ. 2008b. Structure of the guide-strand-containing argonaute silencing complex. *Nature* **456**: 209–213.
- Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**: 1189–1191.
- Wee LM, Flores-Jasso CF, Salomon WE, Zamore PD. 2012. Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. *Cell* **151**: 1055–1067.
- Wei Y, Chesne MT, Terns RM, Terns MP. 2015a. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in *Streptococcus thermophilus*. *Nucleic Acids Research* **43**: 1749–1758.
- Wei Y, Terns RM, Terns MP. 2015b. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev* **29**: 356–361.
- Weick E-M, Miska EA. 2014. piRNAs: from biogenesis to function. *Development* **141**: 3458–3471.
- Welker NC, Maity TS, Ye X, Aruscavage PJ, Krauchuk AA, Liu Q, Bass BL. 2011. Dicer's helicase domain discriminates dsRNA termini to promote an altered reaction mode. *Molecular Cell* **41**: 589–599.
- Westra ER, Pul Ü, Heidrich N, Jore MM, 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.
- Westra ER, Semenova E, Datsenko KA, Jackson RN, Wiedenheft B, Severinov K, Brouns SJJ. 2013. Type I-E CRISPR-cas systems discriminate target from non-target DNA through base pairing-independent PAM recognition. ed. P.H. Viollier. *PLoS Genet* **9**: e1003742.
- Westra ER, van Erp PBG, Künne T, Wong SP, Staals RHJ, Seegers CLC, Bollen S, Jore MM, Semenova E, Severinov K, et al. 2012. CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Molecular Cell* **46**: 595–605.
- Whitman WB, Coleman DC, Wiebe WJ. 1998. Prokaryotes: the unseen majority. *Proc Natl Acad Sci USA* **95**: 6578–6583.
- Wiedenheft B, Lander GC, Zhou K, Jore MM, Brouns SJJ, van der Oost J,

- Doudna JA, Nogales E. 2011a. Structures of the RNA-guided surveillance complex from a bacterial immune system. *Nature* **477**: 486–489.
- Wiedenheft B, van Duijn E, Bultema JB, Bultema J, Waghmare SP, Waghmare S, Zhou K, Barendregt A, Westphal W, Heck AJR, 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 SM, Ma W, Doudna JA. 2009. Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* **17**: 904–912.
- Willkomm S, Oellig CA, Zander A, Restle T, Keegan R, Grohmann D, Schneider S. 2017. Structural and mechanistic insights into an archaeal DNA-guided Argonaute protein. *Nat Microbiol* **2**: 17035.
- Wilson RC, Doudna JA. 2013. Molecular mechanisms of RNA interference. *Annu Rev Biophys* **42**: 217–239.
- Wommack KE, Colwell RR. 2000. Virioplankton: viruses in aquatic ecosystems. *Microbiol Mol Biol Rev* **64**: 69–114.
- Wright AV, Doudna JA. 2016. Protecting genome integrity during CRISPR immune adaptation. *Nat Struct Mol Biol* **23**: 876–883.
- Wright AV, Liu J-J, Knott GJ, Doxzen KW, Nogales E, Doudna JA. 2017. Structures of the CRISPR genome integration complex. *Science* **357**: 1113–1118.
- Wright AV, Nuñez JK, Doudna JA. 2016. Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. *Cell* **164**: 29–44.
- Wu D, Guan X, Zhu Y, Ren K, Huang Z. 2017. Structural basis of stringent PAM recognition by CRISPR-C2c1 in complex with sgRNA. *Cell Res* **27**: 705–708.
- Xiao Y, Luo M, Hayes RP, Kim J, Ng S, Ding F, Liao M, Ke A. 2017. Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. *Cell* **170**: 48–60.e11.
- Xue C, Zhu Y, Zhang X, Shin Y-K, Sashital DG. 2017. Real-Time Observation of Target Search by the CRISPR Surveillance Complex Cascade. *Cell Rep* **21**: 3717–3727.
- Yamada M, Watanabe Y, Gootenberg JS, Hirano H, Ran FA, Nakane T, Ishitani R, Zhang F, Nishimasu H, Nureki O. 2017. Crystal Structure of the Minimal Cas9 from *Campylobacter jejuni* Reveals the Molecular Diversity in the CRISPR-Cas9 Systems. *Molecular Cell* **65**: 1109–1121.e3.
[http://www.cell.com/molecular-cell/pdf/S1097-2765\(15\)00821-7.pdf](http://www.cell.com/molecular-cell/pdf/S1097-2765(15)00821-7.pdf).
- Yamano T, Nishimasu H, Zetsche B, Hirano H, Slaymaker IM, Li Y, Fedorova I, Nakane T, Makarova KS, Koonin EV, et al. 2016. Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. *Cell* **165**: 949–962.
- Yan KS, Yan S, Farooq A, Han A, Zeng L, Zhou M-M. 2003. Structure and conserved RNA binding of the PAZ domain. *Nature* **426**: 468–474.
- Yan WX, Chong S, Zhang H, Makarova KS, Koonin EV, Cheng DR, Scott DA. 2018. Cas13d Is a Compact RNA-Targeting Type VI CRISPR Effector Positively Modulated by a WYL-Domain-Containing Accessory Protein.

Molecular Cell.

- Yang H, Gao P, Rajashankar KR, Patel DJ. 2016. PAM-Dependent Target DNA Recognition and Cleavage by C2c1 CRISPR-Cas Endonuclease. *Cell* **167**: 1814–1828.e12.
- Yao C, Sasaki HM, Ueda T, Tomari Y, Tadokuma H. 2015. Single-Molecule Analysis of the Target Cleavage Reaction by the *Drosophila* RNAi Enzyme Complex. *Molecular Cell* **59**: 125–132.
- Ye X, Huang N, Liu Y, Paroo Z, Huerta C, Li P, Chen S, Liu Q, Zhang H. 2011. Structure of C3PO and mechanism of human RISC activation. *Nat Struct Mol Biol* **18**: 650–657.
- Yi R, Doehle BP, Qin Y, Macara IG, Cullen BR. 2005. Overexpression of exportin 5 enhances RNA interference mediated by short hairpin RNAs and microRNAs. *RNA* **11**: 220–226.
- Yi R, Qin Y, Macara IG, Cullen BR. 2003. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev* **17**: 3011–3016.
- Yigit E, Batista PJ, Bei Y, Pang KM, Chen C-CG, Tolia NH, Joshua-Tor L, Mitani S, Simard MJ, Mello CC. 2006. Analysis of the *C. elegans* Argonaute family reveals that distinct Argonautes act sequentially during RNAi. *Cell* **127**: 747–757.
- Yin H, Song C-Q, Suresh S, Kwan S-Y, Wu Q, Walsh S, Ding J, Bogorad RL, Zhu LJ, Wolfe SA, et al. 2018. Partial DNA-guided Cas9 enables genome editing with reduced off-target activity. *Nat Chem Biol* **14**: 311–316.
- Yosef I, Goren MG, Qimron U. 2012. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Research* **40**: 5569–5576.
- Yosef I, Shitrit D, Goren MG, Burstein D, Pupko T, Qimron U. 2013. DNA motifs determining the efficiency of adaptation into the *Escherichia coli* CRISPR array. *Proc Natl Acad Sci USA* **110**: 14396–14401.
- Yuan Y-R, Pei Y, Ma J-B, Kuryavyi V, Zhadina M, Meister G, Chen H-Y, Dauter Z, Tuschl T, Patel DJ. 2005. Crystal Structure of *A. aeolicus* Argonaute, a Site-Specific DNA-Guided Endoribonuclease, Provides Insights into RISC-Mediated mRNA Cleavage. *Molecular Cell* **19**: 405–419.
- Zamore PD, Tuschl T, Sharp PA, Bartel DP. 2000. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**: 25–33.
- Zander A, Willkomm S, Ofer S, van Wolferen M, Egert L, Buchmeier S, Stöckl S, Tinnefeld P, Schneider S, Klingl A, et al. 2017. Guide-independent DNA cleavage by archaeal Argonaute from *Methanocaldococcus jannaschii*. *Nat Microbiol* **2**: 17034.
- Zegans ME, Wagner JC, Cady KC, Murphy DM, Hammond JH, O'Toole GA. 2009. Interaction between bacteriophage DMS3 and host CRISPR region inhibits group behaviors of *Pseudomonas aeruginosa*. *Journal of Bacteriology* **91**: 210–219.
- Zeng Y, Yi R, Cullen BR. 2005. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *Embo J* **24**: 138–148.
- Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS,

- Essletzbichler P, Volz SE, Joung J, van der Oost J, Regev A, et al. 2015. Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **163**: 759–771.
- Zhang H, Kolb FA, Brondani V, Billy E, Filipowicz W. 2002. Human Dicer preferentially cleaves dsRNAs at their termini without a requirement for ATP. *Embo J* **21**: 5875–5885.
- Zhang J, Rouillon C, Kerou M, Reeks J, Brugger K, Graham S, Reimann J, Cannone G, Liu H, Albers S-V, et al. 2012. Structure and mechanism of the CMR complex for CRISPR-mediated antiviral immunity. *Molecular Cell* **45**: 303–313.
- Zhang X, Wang J, Cheng Q, Zheng X, Zhao G, Wang J. 2017. Multiplex gene regulation by CRISPR-ddCpf1. *Cell Discov* **3**: 17018.
- Zhang Y, Heidrich N, Ampattu BJ, Gunderson CW, Seifert HS, Schoen C, Vogel J, Sontheimer EJ. 2013. Processing-independent CRISPR RNAs limit natural transformation in *Neisseria meningitidis*. *Molecular Cell* **50**: 488–503.
- Zhang Y, Rajan R, Seifert HS, Mondragón A, Sontheimer EJ. 2015. DNase H Activity of *Neisseria meningitidis* Cas9. *Molecular Cell* **60**: 242–255.
- Zhao H, Sheng G, Wang J, Wang M, Bunkoczi G, Gong W, Wei Z, Wang Y. 2014. Crystal structure of the RNA-guided immune surveillance Cascade complex in *Escherichia coli*. *Nature* **515**: 147–150.
- Zheng G, Qin Y, Clark WC, Dai Q, Yi C, He C, Lambowitz AM, Pan T. 2015. Efficient and quantitative high-throughput tRNA sequencing. *Nat Methods* **12**: 835–837.
- Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research* **31**: 3406–3415.
- Zuris JA, Thompson DB, Shu Y, Guilinger JP, Bessen JL, Hu JH, Maeder ML, Joung JK, Chen Z-Y, Liu DR. 2015. Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. *Nat Biotechnol* **33**: 73–80.