Title
Uncovering the diversity of CRISPR-Cas systems

Permalink
https://escholarship.org/uc/item/7h2869gw

Author
Harrington, Lucas B

Publication Date
2018

Peer reviewed|Thesis/dissertation

Uncovering the diversity of CRISPR-Cas systems

by

Lucas B Harrington

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 Doudna, Chair
Professor Donald Rio
Professor Dirk Hockemeyer
Professor Jillian Banfield

Fall 2018

ABSTRACT

Uncovering the diversity of CRISPR-Cas systems

by

Lucas B Harrington

Doctor of Philosophy in Molecular and Cell Biology
University of California, Berkeley

Professor Jennifer A. Doudna, Chair

CRISPR has revolutionized the speed and efficiency of genome editing. This powerful tool originates from an adaptive immune system found in prokaryotes that protects against viruses and other nefarious nucleic acids. In these systems, genetic memories of prior infections are transcribed into guide RNAs which program an interference complex, such as Cas9, to make a targeted DNA break, halting the infection. These systems are present in nearly half of all prokaryotes and are highly diverse. This dissertation focuses on the diversity of CRISPR-Cas systems, the evolutionary pressures that have led to these elegant immune systems, and applications resulting from this investigation.

Although thousands of Cas9 orthologs have been sequenced, biochemical and genome editing experiments have largely focused on a small subset of representatives. To investigate this diversity further we conducted biochemical interrogation of various Cas9 orthologs and found that all Cas9 proteins tested are robust single stranded DNA (ssDNA) cutters. Moreover, we found that many smaller orthologs had limited ability to interrogate double stranded DNA (dsDNA), explaining their unsuccessful use for genome editing. In this process, we recognized a new Cas9 variant from thermophilic environments, GeoCas9. We developed this ortholog for genome editing in human cells and found that it was more resistant to degradation in human plasma compared to the widely used Cas9 from *S. pyogenes*, expanding CRISPR applications to thermophilic hosts. Our newfound understanding of Cas9 diversity led to curiosity about what factors were driving the diversification of this protein. Anti-CRISPRs (Acrs) are small, viral encoded proteins that have evolved to inactivate CRISPR in this microbial arms race. We studied three different Acrs using biochemical, structural and genome editing experiments. Our results showed that these three Acrs had distinct targeting mechanisms to inactivate Cas9. In addition to allowing precise control of CRISPR-Cas9 in cells, these results provided a window into one of the driving forces steering Cas9 evolution.

We next turned our attention to metagenomic datasets, with the hypothesis that Cas9 was not the only single effector CRISPR system. Searching through terabase scale metagenomic data, we identified two new types of CRISPR systems that we called CasX and CasY (reclassified as Cas12e and Cas12d, respectively). These new systems included the most compact CRISPR systems described to date and provided new, streamlined proteins for genome editing with less restrictive sequence requirements. Continuing our search of metagenomic data, we were surprised to find 8 new varieties of CRISPR proteins that pushed the limit of size even further, at just a third of the size of typical CRISPR effector. We found that these systems, in addition to Cas12a, were targeted ssDNA shredders, a property which we developed into a robust platform for rapid, single nucleotide resolution genotyping. Together the investigation of CRISPR adaptive immunity described here gives insight into the pressures that

have shaped these immune systems and in turn provides new tools that enable us to edit genomes, control CRISPR cutting and diagnose disease.

ACKNOWLEDGEMENTS

The work described in this dissertation would not have been possible without the support of my collaborators, mentors, friends, and family. I would first like to thank Jennifer Doudna for providing me the freedom to pursue my research interests and supporting me during my graduate work and beyond. I am also grateful to my Thesis Committee, Don, Jill, and Dirk, for not kicking me out of graduate school despite my skipping most of the MCB requirements other than filing this dissertation.

Many thanks to all the members of the Doudna Lab and my collaborators for providing a supportive and rigorous environment to learn and do research. Foremost, thanks to the Younglings Bay, primarily Addison and Janice, for making science fun, laughing at our mistakes and pushing each other to improve as researchers. Thanks to Stephen and Philip, your rigor and scientific integrity are unmatched, and I am lucky to have overlapped with you in the lab, both of you are inspirations for what scientists should be. Thanks to Fuguo, Sam, Brett, Gavin, Christof, Ben, Romain and Happy Steve for helpful Cas9 discussions and collaborations. To the D-lab crew of yesteryear, including Megan, Kevin, Romain, April, James, Steven, Alex, Yun, Kyle, Akshay, and Ross thanks for making the lab such a welcoming and fun place to do science. Kaihong, Julie, Molly, and Lisa, the lab wouldn't be able to run without you, thanks for taking care of the behind the scenes work so everyone can focus on their research.

Enbo, beginning in my first year and continuing to this day, you have been an incredible teacher and collaborator. From running gels to dealing with research conflict and career advice, I am fortunate to have been able to learn from you. Josh, your passion for fundamental science, hostility towards applied research, and relentlessly positive attitude is endearing and a welcome addition to the lab. We look forward to seeing your continued accomplishments. James, you are an invaluable mentor and friend and one of the kindest people I know, graduate school would have been a lot more difficult and a lot less fun without you and your unwavering support. Dudu (David), it has been an absolute pleasure to work with you and to have you as a role model. I can't wait to see what discoveries you'll make with your own team in Tel Aviv. Thanks also to Karen Maxwell, Alan Davidson, Erik Sontheimer and Jillian Banfield for being awesome collaborators.

My baymate, mentor, and now co-founder Janice Chen has been the guiding force of my Ph.D. Janice, your candid opinion is always welcome and challenges me to be a better scientist and person. I cannot imagine my graduate work, Sierra, or Mammoth going anywhere without you (and by extension Rust and Riley, not Remy) and I am thrilled to be able to continue doing science with you after Berkeley. Let the record show that Janice's candid feedback on this acknowledgments section is, "not enough snark and sass, sounds like someone else wrote it for you."

Thanks to all of my fishing friends, in particular, Jake and Ryan, for always being ready to drive up into the Sierras when I need a break from the bench and for letting me borrow your more capable vehicles. I would also like to thank everyone that has encouraged and enabled me to pursue my passion for biological sciences in the past including Jeff Mumm, Sandy Burkhard, Chelsea McIntosh, Bob Blankenship, Joe Jez, and Getu Duguma.

Lastly, I cannot thank my parents and sisters enough. To my parents, Ken and Lea, you are always there for your children and have fostered creativity and independence in us since the day we were born. My sisters, Matti, Gabe and Parker, you are all amazing people and I am fortunate to be related to you and to have your support.

To all of you, and those that couldn't fit on this page, thank you for inspiring, challenging and supporting me. This dissertation would not be possible without you.

TABLE OF CONTENTS

Abstract **1**
Acknowledgements .... **i**
Table of Contents .... **ii**
List of Figures .... **v**
List of Tables .... **vi**

Chapter 1. Introduction to CRISPR–Cas systems and applications **1**
1.1 Introduction 2

Chapter 2. Single-Stranded DNA Cleavage by Divergent CRISPR-Cas9 Enzymes **4**
2.1 Abstract 5
2.2 Introduction 5
2.3 Materials and Methods 6
2.3.1 Cas9 orthologs used in this study 6
2.3.2 Cloning and purification of Cas9 proteins 6
2.3.3 Construction of single guide RNA (sgRNA) 6
2.3.4 *in vitro* DNA cleavage assays 6
2.3.5 Filter binding assays 7
2.3.6 Electrophoretic mobility-shift assay 7
2.3.7 Limited proteolysis of Cas9 complexes 7
2.4 Results **8**
2.4.1 Single-stranded DNA cleavage by highly divergent Cas9 enzymes 8
2.4.2 programmable and PAM-independent Cas9-Catalyzed DNA cleavage 9
2.4.3 DNA unwinding capability differs among Cas9 variants 10
2.4.4 Divergent Cas9s have varying affinities for guide RNAs and substrates 11
2.4.5 Cas9 conformational changes reveal differences in nucleic acid binding modes 12
2.5 Discussion 25
2.6 Acknowledgements 27
2.7 Author contributions 27

Chapter 3. A thermostable Cas9 with increased lifetime in human plasma **28**
3.1 Abstract 29
3.2 Introduction 29
3.3 Materials and Methods 30
3.3.1 Identification of Cas9 orthologs and generation of plasmids 30
3.3.2 Cas9 purification 30

3.3.3 Differential scanning calorimetry 30
3.3.4 Biochemical cleavage assays 31
3.3.5 Small RNA sequencing 31
3.3.6 HEK293T EGFP disruption assays and indel analysis 31
3.4 Results 32
3.5 Discussion 42
3.6 Acknowledgements 43
3.7 Author contributions 43
Chapter 4. Broad spectrum inhibitor of CRISPR-Cas9 44
4.1 Abstract 45
4.2 Introduction 45
4.3 Material and Methods 46
4.3.1 Phylogenetic analyses 46
4.3.2 Protein expression and purification 46
4.3.3 Cleavage assays 46
4.3.4 HEK293T transfection and indel analysis 46
4.3.5 Filter binding 47
4.3.6 Gel shift assays and Fluorescence polarization 47
4.3.7 Isothermal titration calorimetry 48
4.3.7 Size exclusion binding assays 48
4.3.7 In vivo assay of AcrIIC1 activity 48
4.3.7 Generation of AcrIIC1 mutants 48
4.3.7 Crystallography 48
4.3.7 Small-angle X-ray scattering 49
4.3.7 Electron microscopy 49
4.4 Results 50
4.5 Discussion 68
4.6 Acknowledgements 69
4.7 Author contributions 69
Chapter 5. New CRISPR-Cas system from uncultivated microbes 70
5.1 Abstract 71
5.2 Material and Methods 71
5.2.1 Metagenomics and metatranscriptomics 72
5.2.2 CRISPR-Cas computation analysis 72
5.2.3 Generation of heterologous plasmids 74
5.2.4 PAM depletion assay 74
5.2.5 Plasmid Interference 74
5.2.6 Northern Blots 74
5.3 Results and Discussion 75
5.4 Acknowledgements 83
5.5 Author contributions 83
Chapter 6. Programmed DNA destruction by miniature CRISPR-Cas14 enzymes 84
6.1 Abstract 85
6.2 Introduction 85
6.3 Material and Methods 85

6.3.1 Metagenomics and metatranscriptomics 85
6.3.2 CRISPR-Cas computation analysis 86
6.3.3 Generation of expression plasmids, RNA and DNA substrates 86
6.3.4 E. coli RNAseq 86
6.3.5 PAM depletion assays 87
6.3.6 Protein purification 87
6.3.7 *In vitro* cleavage assays 87
6.3.8 Cas14 and Cas12a-DETECTR assays 88
6.3 Results and Discussion 89
6.4 Acknowledgements 107
6.5 Author contributions 107

Bibliography 108

LIST OF FIGURES

Figure 1.1 Overview of CRISPR–Cas adaptive immunity 2

Figure 2.1 Diverse DNA cleavage activity of divergent Cas9 enzymes 13
Figure 2.2 Programmable and PAM-independent Cas9-catalyzed DNA cutting 15
Figure 2.3 Substrate recognition varies among divergent Cas9 enzymes................. 16
Figure 2.4 Binding affinity of Cas9 proteins for sgRNA and DNA. 17
Figure 2.5 Limited proteolysis and structural comparison of Spy and Cdi Cas9 proteins. 18
Figure 2.6 Cas9 proteins used in this work 19
Figure 2.7 Cas9 proteins used in this work and their cleavage activity on various substrates. 20
Figure 2.8 In vitro Cas9 cleavage activity on supercoiled plasmid and ssDNA substrates. 21
Figure 2.9 In vitro cleavage assays on budged substrates. 22
Figure 2.10 Binding to nucleic acids and limited proteolysis of Cas9 enzyme............ 23

Figure 3.1 GeoCas9 is a small, thermostable Cas9 homolog 36
Figure 3.2 PAM identification and engineering of GeoCas9 37
Figure 3.3 Small RNA-seq and sgRNA engineering for GeoCas9 38
Figure 3.4 Genome editing activity of GeoCas9 in mammalian cells 38
Figure 3.5 Thermostability of GeoCas9 and longevity in human plasma 39
Figure 3.6 Alignment of Cselected Cas9 homologs 39
Figure 3.7 Targets used to generate logos in figure 2b and 2c 40
Figure 3.8 Mismatch cleavage by GeoCas9 and degradation of SpyCas9 in mammalian plasma 41
Figure 3.9 Uncropped images of gels for Figure 4 and 5 41

Figure 4.1 AcrIIC1 inhibits diverse Cas9 orthologs while AcrIIC2 and AcrIIC3 are highly specific 55
Figure 4.2 AcrIIC1 traps the DNA-bound Cas9 complex 56
Figure 4.3 AcrIIC1 binds to Cas9 HNH domain 56
Figure 4.4 Structure of AcrIIC1 bound to the NmeCas9 HNH domain 57
Figure 4.5 AcrIIC3 blocks DNA binding and dimerizes Cas9 58
Figure 4.6 Model of AcrIIC1 and AcrIIC3 inhibition of Cas9 59
Figure 4.7 AcrIIC1 inhibits diverse Cas9 orthologs, while AcrIIC2 and AcrIIC3 are highly specific, related to Figure 1 60
Figure 4.8 Impact of AcrIIC1 and AcrIIC3 on Cas9 RNA and DNA binding, related to Figures 2 and 5. 61
Figure 4.9 AcrIIC1 Binds to the Cas9 HNH domain, related to Figure 3. 62

Figure 4.10 Alignment of HNH domains and AcrIIC1 orthologs and structure of AcrIIC1–HNH, related to Figure 4 63
Figure 4.11 Conservation of the Cas9 HNH domain in the context of full length Cas9, related to Figure 4. 65
Figure 4.12 AcrIIC3 is monomeric and dimerizes NmeCas9, related to Figure 5 66

Figure 5.1 Novel identified CRISPR-Cas systems from uncultivated organisms **78**
Figure 5.2 CasX mediates programmable DNA interference in *E. coli*. 79
Figure 5.3 CasX is a dual-guided CRISPR complex 79
Figure 5.4 Expression of a CasY locus in *E. coli* is sufficient for DNA interference**.** 80
Figure 5.5 The newly identified CRISPR-Cas in context of known systems 81
Figure 5.6 Newly identified CRISPR-Cas systems compared to known proteins 81
Figure 5.7 Programmed DNA interference by CasX 82

Figure 6.1 Architecture and phylogeny of CRISPR-Cas14 genomic loci. 92
Figure 6.2 CRISPR-Cas14a actively adapts and encodes a tracrRNA 93
Figure 6.3 CRISPR-Cas14a is an RNA-guided DNA-endonuclease 94
Figure 6.4 High fidelity ssDNA SNP detection by CRISPR-Cas14a 95
Figure 6.5 Phylogenetic analysis of Cas14 orthologs 96
Figure 6.6 Maximum likelihood tree for Cas1 from known CRISPR systems 97
Figure 6.7 Acquisition of new spacers by CRISPR-Cas14 systems 98
Figure 6.8 Metatranscriptomics for CRISPR-Cas14 loci. 99
Figure 6.9 RNA processing and heterologous expression by CRISPR-Cas14 100
Figure 6.10 Plasmid depletion by Cas14a1 and SpCas9 101
Figure 6.11 Degradation of ssDNA by Cas14a1 102
Figure 6.12 Kinetics of Cas14a1 cleavage of ssDNA with various guide RNA components 103
Figure 6.13 Optimization of Cas14a1 guide RNA components 104
Figure 6.14 Impact of various activators on Cas14a1 cleavage rate 105
Figure 6.15 Diversity of CRISPR-Cas14 systems 106

LIST OF TABLES

Table 2.1 Cas9 orthologues used in this study. 24
Table 2.2 Sequences of DNA and guide RNA used in this paper. 25

Table 4.1 Data collection and refinement statistics 67

Chapter 1

Introduction to CRISPR–Cas systems

1.1 Introduction

Prokaryotes are in a constant battle with invasive nucleic acids from phages, plasmids and other mobile genetic elements. Adaptive immune systems involving clustered regularly interspaced short palindromic repeats (CRISPRs) have evolved to defend against these invaders. CRISPR systems are found in an estimated 50% of bacteria and nearly 90% of archaea(K S Makarova et al., 2015). Each CRISPR locus encodes CRISPR-associated (Cas) proteins and a CRISPR array composed of short repeating elements intervened by spacers, which serve as genetic memories of previous foreign encounters.

CRISPR immunity is divided into three stages: acquisition, RNA processing and interference (Figure 1.1). In the acquisition stage, foreign DNA is captured and integrated into the CRISPR array by Cas1 and Cas2, the only two proteins conserved across all active CRISPR loci (Nuñez, Harrington, Kranzusch, Engelman, & Doudna, 2015; Nuñez et al., 2014b; Nuñez, Lee, Engelman, & Doudna, 2015; Yosef, Goren, & Qimron, 2012). During RNA processing, the CRISPR locus is transcribed to produce a long precursor CRISPR RNA (pre-crRNA) that is further processed by host factors(Elitza Deltcheva et al., 2011) or dedicated Cas proteins(R. Wang, Preamplume, Terns, Terns, & Li, 2011) to produce multiple crRNA sequences. Each mature crRNA consists of a spacer sequence and a portion of the repeat sequence. In the interference stage, a crRNA is loaded into a Cas protein complex that binds and cleaves foreign

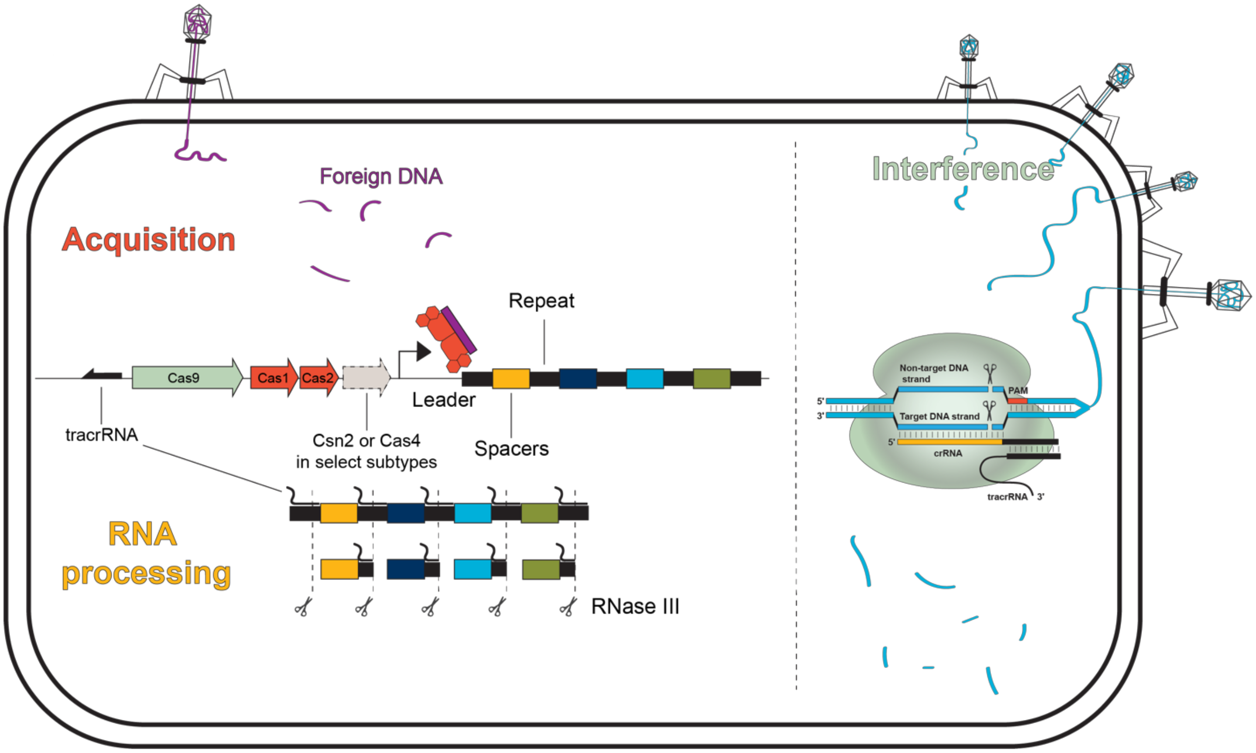


Figure 1.1 | Outline of Type II CRISPR adaptive immunity. CRISPR immunity is divided into three stages, acquisition, RNA processing and interference. In acquisition, foreign DNA is integrated into the CRISPR array by the Cas1-Cas2 complex (red). The array is transcribed and processed to yield multiple mature crRNA sequences. The crRNA is loaded into the Cas9 interference complex (green) which uses it to identify and cleave invading nucleic acids.

nucleic acids based on DNA-RNA(Garneau et al., 2010; M Jinek et al., 2012; Samai et al., 2015a) or RNA-RNA(W. Jiang, Samai, & Marraffini, 2016; Samai et al., 2015a) base pairing with the crRNA.

CRISPR loci are classified according to the Cas proteins flanking the CRISPR array(K S Makarova et al., 2015). The majority of this dissertation will focus on Type II and Type V CRISPR systems, which include a single interference proteins Cas9 and Cas12 respectively. In addition to the crRNA, Type II systems also require a trans-activating crRNA (tracrRNA)(Chylinski, Le Rhun, & Charpentier, 2013; M Jinek et al., 2012). The tracrRNA base pairs with the repeat portion of the crRNA and the tracrRNA:crRNA duplex is processed by host RNase III(Elitza Deltcheva et al., 2011). Once loaded with this dimeric RNA, Cas9 binds sequences containing a 2-5 bp protospacer adjacent motif (PAM) flanking the targeted region in the DNA(Cencic et al., 2014; Deveau et al., 2008; Mojica, Diez-Villasenor, et al., 2009), at which point the crRNA invades the DNA duplex to form an R-loop(Fuguo Jiang et al., 2016; Szczelkun et al., 2014). Upon binding to the correct target, the HNH and RuvC nucleases are activated to cleave the target and non-target strands, respectively(M Jinek et al., 2012).

Type II CRISPR systems are further divided into three subtypes (A, B and C) based on other Cas proteins found in the CRISPR locus. It is believed that the ancestral Type II locus had only three proteins (Cas9, Cas1 and Cas2). This simple CRISPR locus architecture still exists in Type II–C systems – the most abundant subtype(Chylinski, Makarova, Charpentier, & Koonin, 2014a). Type II–B diverged from Type II–C systems by acquiring a new Cas1 and Cas2 and the additional protein Cas4. Type II–B subtypes are rare and the Cas9 proteins found in these loci are typically larger(Chylinski et al., 2014a). The model Type II–B system is from *Francisella novicida* and has been proposed to have diversified functions(Hirano et al., 2016; Sampson, Saroj, Llewellyn, Tzeng, & Weiss, 2013). After Type II–B divergence, the Type II–C lineage further diversified by addition of Csn2 to form the Type II–A subtype(Chylinski et al., 2014a; K S Makarova et al., 2015). This subtype includes the system from *Streptococcus pyogenes*, the first Type II system to be functionally reconstituted.

Type V CRISPR systems were more recently discovered and are distinguished by the presence of a single nuclease domain called the RuvC(Shmakov et al., 2015; Zetsche et al., 2015). Despite only having a single nuclease, many Type V systems have been shown to cleave double stranded DNA (dsDNA). Prior to the work described here, three distinct subtypes of Type V systems have been described (Type V-a, b and c). Type V systems are diverse and likely originated from the transposase associated protein TnpB (Shmakov et al., 2017a). There is no known function for this ancestral protein and it is not essential for transposition (Bao & Jurka, 2013; He et al., 2015).

For my graduate work, I aimed to understand the diversity of CRISPR-Cas systems and the evolutionary pressures that have led to these elegant immune systems. In the process of studying this microbial warfare, we have developed new tools that enable us to edit genomes, control CRISPR cutting and diagnose disease.

Chapter 2

Single-stranded DNA cleavage by divergent CRISPR-Cas9 enzymes

Part of the work presented in this chapter has previously been published in the following manuscript: Ma, E., Harrington, L. B.*, O'Connell, M. R., Zhou, K., & Doudna, J. A. (2015). Single-Stranded DNA Cleavage by Divergent CRISPR-Cas9 Enzymes. *Molecular Cell*, *60*(3), 398–407.

2.1 Abstract

Double-stranded DNA (dsDNA) cleavage by Cas9 is a hallmark of type II CRISPR-Cas immune systems. Cas9–guide RNA complexes recognize 20-base-pair sequences in DNA and generate a site-specific double-strand break, a robust activity harnessed for genome editing. DNA recognition by all studied Cas9 enzymes requires a protospacer adjacent motif (PAM) next to the target site. We show that Cas9 enzymes from evolutionarily divergent bacteria can recognize and cleave single-stranded DNA (ssDNA) by an RNA-guided, PAM-independent recognition mechanism. Comparative analysis shows that in contrast to the type II-A S. pyogenes Cas9 that is widely used for genome engineering, the smaller type II-C Cas9 proteins have limited dsDNA binding and unwinding activity and promiscuous guide-RNA specificity. These results indicate that inefficiency of type II-C Cas9 enzymes for genome editing results from limited ability to cleave dsDNA, and suggest that ssDNA cleavage was an ancestral function of the Cas9 enzyme family.

2.2 Introduction

Bacteria and archaea use CRISPR (clustered regularly interspaced short palindromic repeats) systems composed of Cas (CRISPR-associated) proteins and short RNA guides to provide adaptive immunity against invasive nucleic acids (Doudna & Charpentier, 2014; W. Jiang & Marraffini, 2015; van der Oost, Westra, Jackson, & Wiedenheft, 2014). Cas9, a protein component of Type II CRISPR-Cas systems, is a programmable, RNA-guided DNA endonuclease whose specificity is determined by RNA-DNA base pairing (Gasiunas, Barrangou, Horvath, & Siksnys, 2012; Martin Jinek et al., 2012a). The *Streptococcus pyogenes* Cas9 (Spy Cas9) has been employed widely for genome engineering based on its ability to generate site-specific double-stranded DNA breaks at sequences abutting an NGG PAM sequence (Doudna & Charpentier, 2014; Sontheimer & Barrangou, 2015; Terns & Terns, 2015). In nature, a large and diverse collection of *cas9* genes have been identified, and a few different Cas9 enzymes have been characterized (Jinek et al., 2012; Chylinski et al., 2014; Esvelt et al., 2013; Fonfara et al., 2014; Ran et al., 2015). In all cases studied so far, PAM binding plays a key role in double-stranded DNA (dsDNA) target sequence recognition, although PAM sequences can differ between Cas9 variants. The nucleic-acid cleavage activities of more evolutionarily divergent Cas9 proteins, and whether all Cas9 proteins employ the same DNA recognition and cleavage mechanism have not been determined.

To investigate the function of highly divergent Cas9 proteins, we studied the DNA binding and cleavage activities of seven different Cas9 proteins belonging to the type II-A and type II-C subclasses of the Cas9 superfamily. Surprisingly, all of these enzymes were found to possess PAM-independent single-stranded DNA (ssDNA) cleavage activity, and the type II-C Cas9 proteins could utilize a range of guide RNA constructs for productive ssDNA binding and cleavage. We show that the type II-C Cas9s possess little dsDNA unwinding capability and consequently do not have robust dsDNA cleavage activity. These results suggest that ssDNA cleavage is an intrinsic activity of the Cas9 enzyme family that involves a distinct mode of substrate binding and catalytic domain organization. This activity may have evolved to target particular kinds of bacterial pathogens, and may be useful for applications that involve ssDNA binding or cleavage due to the lack of target sequence constraints.

2.3 Materials and Methods

2.3.1 Cloning and purification of Cas9 proteins

In this study, seven Cas9 proteins have been used. The sources and molecular weights of these Cas9s are summarized in Table S1. To produce the Cas9 proteins for *in vitro* studies, we amplified each ORF from genomic DNA via PCR with primers tagged with a common sequence for ligase-independent cloning (LIC), we then cloned the PCR products into the 2CT expression vector (His$_6$-MBP-N$_{10}$-tev-ORF, from UC Berkeley MacroLab). All Cas9 proteins were expressed in *Rosetta* cells (Novagen), purified with Ni^{2+} resin, cleaved with TEV protease to remove the MBP tag, and further purified using Superdex 200 size-exclusion column (Figure 2.6B). The Cas9 proteins were stored at −80 °C in Cas9 storage buffer (20 mM HEPES (pH 7.5), 0.2 M KCl and 5% Glycerol).

2.3.2 Construction of single guide RNA (sgRNA)

Guide RNAs were designed according to the single-guide RNA construct structure established previously (Jinek et al 2012) in which crRNA and tracrRNA sequences were linked by an RNA tetraloop. We used predictions based on type II CRISPR system and Cas9 protein phylogenies (Chylinski et al., 2014; Fonfara et al., 2014) to generate crRNA and tracrRNA sequences; the Cdi guide RNA was based on an experimentally validated construct (Ran et al., 2015). sgRNA-encoding DNAs were amplified by PCR to add a T7 promoter fragment (TAATACGACTCACTATAGG). The T7 fragment-tagged DNAs were then used as templates for the *in vitro* transcription of sgRNAs as described   The transcribed sgRNAs were purified by denaturing PAGE, ethanol precipitated and folded according to (Lin et al., 2014). All other RNA oligos (and all DNA oligos) were synthesized by Integrated DNA Technologies, Inc. (Coralville, Iowa). DNA and RNA Oligo sequences as well as sgRNA sequences used in this study are listed in Table S2.

2.3.3 In vitro DNA cleavage assays

In vitro cleavage reactions were carried out in a total volume of 10 µl of Cas9 cleavage buffer (150 mM KCl, 20 mM HEPES (pH7.5), 1% glycerol, 5 mM MgCl$_2$ and 1 mM DTT) in the presence of 30 nM Cas9, 30 nM sgRNA and ~2 nM 5'-end ^{32}P-labeled target DNA. Reactions were incubated at 37°C for one hour (unless otherwise stated). The reactions were stopped by the addition of 12 ul of 2X loading buffer (2X loading buffer: 0.2 mg/ml of bromophenol blue, 0.2 mg/ml of xylene cyanol, 40 mM EDTA and 95% of formamide) and incubated at 70°C for 10 min. Cleavage products were analyzed by 10% denaturing PAGE, visualized by phosphorimaging and quantified with ImageQuant TL (GE Healthcare). The fraction of RNA cleaved at each time point was plotted as a function of time, and these data were fit with a single exponential decay curve using Prism 6 (GraphPad Software, Inc., La Jolla, USA), according to the equation: Fraction cleaved =A X (1 – exp(-*k* x t)), where *A* is the amplitude of the curve, *k* is the first-order rate constant, and t is time. All experiments were carried out at least in triplicate, with representative replicates shown in the figure panels. Pseudo first-order rate constants (k_{cleave}) and associated errors are reported in the figure legends.

2.3.4 Filter binding assays

Filter binding was carried out in RNA Binding Buffer (20 mM Tris, pH 7.5, 150 mM KCl, 5 mM MgCl2, 1 mM DTT, 5% glycerol, 0.01% Igepal CA-630, 10µg/ml yeast tRNA and 10 µg/ml BSA). Cas9 was incubated with radiolabeled RNA (<0.02nM) for 1hr at room temperature. Tufryn, Protran and Hybond-N+ were assembled onto a dot-blot apparatus in the order listed above. The membranes were washed with 50µL Equilibration Buffer (20 mM Tris, pH7.5, 150 mM KCl, 5 mM MgCl2, 1 mM DTT, 5% glycerol) twice before the sample was applied to the membranes. Membranes were again washed twice with 50 µL Equilibration Buffer, dried and visualized by phosphorimaging. Data were quantified with ImageQuant TL Software (GE Healthcare) and fit to a binding isotherm using Prism (GraphPad Software). All experiments were carried out at least in triplicate, with representative replicates shown in the figure panels Dissociation constants (K_D) and associated errors are reported in the figure legends.

2.3.5 Electrophoretic mobility-shift assay

Assays were done in Binding Buffer (20 mM Tris, pH 7.5, 150 mM KCl, 5 mM EDTA, 1 mM DTT, 5% (v/v) glycerol, 50 µg/mL heparin, 0.01% Tween 20, and 100 µg/mL BSA). Cas9-sgRNA complexes were preformed by incubating 1µM of Cas9 and sgRNA for 30min at room temperature. Cas9-sgRNA complexes were diluted and DNA (<0.05nM) was added and allowed to incubate for another 30min at room temperature. Samples were then analyzed by 8% polyacrylamide gel containing 0.5x TBE. Gels were imaged by phosphorimaging, quantified using ImageQuant TL Software (GE Healthcare) and fit to a binding isotherm using Prism (GraphPad Software). All experiments were carried out at least in triplicate, with representative replicates shown in the figure panels. Dissociation constants (K_D) and associated errors are reported in the figure legends.

2.3.6 Limited proteolysis of Cas9 complexes

Limited proteolysis was conducted in a 10µL reaction using 5 µM Cas9 with or without guide (1:1.5 molar ratio) and/or DNA (1:3 molar ratio). The Cas9 and guide mixture was pre-incubated for 30min at room temperature in 1x Trypsin Buffer (100mM Tris, pH 7.5, 150 mM KCl, 2% glycerol, 1 mM EDTA and 1 mM DTT). DNA was added and the reaction was incubated for 5 to 30 min before the addition of 0.2µg trypsin per 10µL of Cas9 complex. The reaction was quenched by addition of 5× loading dye (5% β-Mercaptoethanol, 0.02% bromophenol blue, 30% glycerol, 10% sodium dodecyl sulfate, 250 mM Tris-Cl, pH 6.8) to a final concentration of 1x and the products were analyzed on 10% PAGE gels. All experiments were carried out at least in triplicate, with representative replicates shown in the figure panels.

2.4 Results

2.4.1 Single-stranded DNA cleavage by highly divergent Cas9 enzymes

Phylogenetic analysis of the three established type II CRISPR-Cas9 subtypes, type II-A, -B and -C, suggests that type II-C is ancestral to the others (Chylinski et al., 2014b; Fonfara et al., 2014). Of interest, Type II-C Cas9 proteins may harbor unique biochemical properties as they are diverse at the level of primary sequence and are generally smaller than type II-A and II-B Cas9 proteins (Chylinski et al., 2014b; Fonfara et al., 2014). To determine whether biochemical activities of the smaller and ostensibly more ancient type II-C Cas9 proteins diverge from that of Spy Cas9, we selected six type II-C enzymes for analysis (Figure 2.1A). The type II-C Cas9 enzyme from *N. meningitidis*, as well as the type II-A enzyme from *S. pyogenes*, have been shown to function in eukaryotic cells for genome engineering applications, although Spy Cas9 is by far the most widely used to date (reviewed in Doudna and Charpentier, 2014).

All of the selected Cas9 enzymes, found within evolutionarily divergent bacterial strains, share a common chassis consisting of the RuvC and HNH catalytic domains, the bridge helix, the alpha-helical domain (also known as the REC domain; Nishimasu et al., 2014), and the PAM-interaction domain (Figure 2.1B). However, the size of the alpha-helical domain and the primary sequence outside of the catalytic domains vary widely among the different Cas9 proteins tested (Figure 2.6*A*). After expressing and purifying these enzymes (Figure 2.7B), we tested their activities in DNA cleavage assays with various types of substrates. We focus here on the relative activities of the *C. diphtheriae* (Cdi) Cas9 enzyme, compared to the Spy Cas9 as a control. The Cdi Cas9 shares PAM recognition specificity with Spy Cas9, simplifying DNA substrate design, and its guide RNA has been validated (Ran et al., 2015a). The Ana Cas9 bears sequence similarities to Cdi Cas9 and its crystal structure is known (Martin Jinek et al., 2014). In the absence of validated tracrRNA and PAM sequences for Ana Cas9, this enzyme provides a useful benchmark for PAM-independent DNA binding and cleavage behavior. Relevant biochemical data for the other four Cas9 enzymes are shown in supplemental figure panels.

To explore the substrate selectivity of these Cas9 proteins, we examined cleavage of double-stranded and single-stranded DNA substrates, as well as a double-stranded DNA substrate containing a 20-base-pair mismatched segment along the length of the guide RNA recognition site (Figure 2.1C). Guide RNAs used in these experiments bear the single-guide RNA architecture (Martin Jinek et al., 2012a) consisting of CRISPR RNA (crRNA) and tracrRNA components derived from each bacterial host genome encoding the Cas9 in question (see Methods). Surprisingly, we found that in contrast to Spy Cas9, which has robust dsDNA cleavage activity but slow ssDNA cleavage activity, Cdi Cas9 demonstrated a preference for ssDNA cleavage (Figure 2.1D). Unsurprisingly, Ana Cas9 was unable to cleave dsDNA, due to a non-cognate guide RNA and PAM sequence in the substrate (Figure 2.7B). However, Ana Cas9 was active for ssDNA cleavage. Notably, the cleavage site in the ssDNA substrate was the same as that observed when the dsDNA substrate was used, implying that the mechanism of cleavage site selection does not require the PAM nucleotides, which are present in the non-target strand of the DNA (see Figure 2.1C). Although Spy Cas9 has been shown previously to cleave target ssRNA only in the presence of a PAM-containing DNA oligonucleotide (PAMmer) (O'Connell et al., 2014), we did not observe ssRNA cleavage (in the absence of a PAMmer) by Spy Cas9, Cdi Cas9 or Ana Cas9 (Figure 2.7D).

We next tested the ability of the RuvC and HNH domains to cleave DNA individually by conducting cleavage assays with the mismatched DNA substrate and only one strand of the substrate radiolabeled. Both the Spy and Cdi enzymes contain two active domains, but we did not detect catalytic activity of the RuvC domain in the Ana Cas9 (Figure 2.1D, Figure 2.7B). We also noted that in contrast to the staggered RuvC-domain cleavage characteristic of the Spy Cas9 enzyme, the RuvC domain of Cdi Cas9 produced a single specific DNA cleavage product. Similar results were obtained for the other type II-C Cas9 enzymes tested in this study (Figure 2.7B).

Efficient dsDNA cleavage requires Cas9-mediated local DNA unwinding to promote the formation of an R-loop (Sternberg, Redding, Jinek, Greene, & Doudna, 2014; Szczelkun et al., 2014). This DNA unwinding ability can be detected indirectly by assessing the dsDNA versus ssDNA cleavage activities of each Cas9 enzyme. Substrate cleavage kinetics, compared under single-turnover reaction conditions, quantify the marked difference in ssDNA versus dsDNA cleavage capabilities of the Spy versus Cdi and Ana Cas9 enzymes (Figure 2.1E, Figure 2.7E). Whereas the Spy Cas9 rapidly cleaves a dsDNA substrate, only slow cleavage of this substrate is observed for Cdi Cas9, and as anticipated, no cleavage occurs with the Ana Cas9. In contrast, when the same target DNA strand is presented as a ssDNA substrate, similar rates of cleavage are observed for all three Cas9 enzymes. When the target DNA strand is presented in the context of a "bulged" dsDNA in which the 20-nucleotide target site is mismatched to the complementary DNA strand, the three Cas9 enzymes also show similar rates of cleavage. Analogous results were obtained using a supercoiled plasmid DNA substrate (Figure 2.8A, 2.8B), and control experiments using a guide RNA with a scrambled target-recognition sequence did not support DNA cleavage (Figure 2.7C). These observations imply that although Cas9 enzymes share a fundamental RNA-targeted DNA cleavage activity, they have markedly different abilities to unwind a dsDNA substrate.

2.4.2 Programmable and PAM-independent Cas9-catalyzed DNA cleavage

We next compared the guide RNA specificities of the Spy, Cdi and Ana Cas9 enzymes to support dsDNA versus ssDNA substrate cleavage (see Figure 2.2A for guide RNA schematic). As observed previously, we found that the Spy Cas9 enzyme is highly selective for its cognate guide RNA, and only supports dsDNA or ssDNA cleavage when its own guide RNA is used in the reaction ((Briner et al., 2014; Martin Jinek et al., 2012a); Figure 2B). In contrast, both the Cdi and Ana proteins can use a variety of guide RNAs corresponding to guide RNA sequences corresponding to five different Cas9 enzymes (Figures 2.2B and 2.8D). The Cdi enzyme catalyzed dsDNA cleavage in the presence of either its own guide RNA or the Spy guide RNA, but only ssDNA cleavage in the presence of the other three guide RNAs tested. As expected, Ana Cas9 could only catalyze ssDNA cleavage but could do so using any of the five guide RNAs in these reactions.

We also observed some differences in DNA cleavage site specificity as a function of the guide RNA construct used in the reaction. With some guide RNAs, both the Cdi and Ana enzymes generated two cleavage products corresponding to a site three base pairs from the end of the target sequence (the cognate cleavage site) and a site ~2 nucleotides downstream of this position (Figure 2.2B and 2.8D). For both Cdi and Ana Cas9 enzymes, the two cleavage products were generated only in reactions containing the Spy, Nme or Cje guide RNAs. These results suggest that guide RNA positioning and/or HNH domain docking may differ depending on the length and/or structure of the guide RNA used by these proteins.

We next tested whether ssDNA cleavage is programmable by generating a set of Spy guide RNAs with a common architecture but different 20-nt. target recognition sequences designed to bind successive DNA target sites shifted by two nucleotides in each case (Figure 2.2C). Each of the three Cas9 enzymes tested in this experiment showed programmable DNA cleavage such that the primary product generated with each guide RNA corresponded to cleavage at the phosphodiester bond positioned three nucleotides downstream of the 3′ end of the guide RNA (Figure 2.2D, 2.8C). In addition, we found that the Cdi and Ana enzymes were somewhat less precise in the cleavage reaction, generating two or three cleavage products in each reaction. These results show that like Cas9-catalyzed dsDNA cleavage, ssDNA cleavage is site-specific and programmable. In contrast to Cas9-catalyzed dsDNA cleavage, however, ssDNA cleavage is PAM-independent and the cleavage site appears to be determined largely by measuring from the end of the RNA-DNA hybrid. There may also be sequence-context effects that influence the precision of cleavage, perhaps reflecting differences in docking of the HNH domain that is catalyzing the cleavage reaction in each case.

2.4.3 DNA unwinding capability differs among Cas9 variants

The different relative activities of Spy, Cdi and Ana Cas9 enzymes for cleavage of dsDNA versus ssDNA substrates implied a difference in DNA unwinding capabilities of these different proteins. To examine this, we tested the ability of each enzyme to cleave dsDNA substrates containing either a completely base-paired target site or varying numbers of mismatched base pairs (2-16) extending from the PAM-proximal end of the target site (Figure 2.3A). Importantly, in these experiments, the radiolabeled target strand of the DNA does not change, only the unlabeled strand to which it is annealed to generate the dsDNA substrate. Remarkably, a two base mismatch in the dsDNA target is required to greatly enhance dsDNA cleavage by Cdi Cas9, unlike Spy Cas9, which cleaves all of the tested substrates robustly (Figures 2.3B and 2.3C). The same behavior is observed for other type II-C enzymes tested with their cognate guide RNAs (Figure 2.9A). For Ana Cas9, at least six bases of the target dsDNA sequence must be mismatched to the DNA complementary strand before cleavage occurs, and there is a notable increase in cleavage product generation when the size of the mismatched DNA region is 12 base pairs or larger. Similar behavior is observed for other type II-C enzymes tested with non-cognate guide RNAs (Figure 2.9A). These results show that the Cdi and other type II-C Cas9 enzymes have limited ability to cleave dsDNA unless the nucleotides located in the target sequence adjacent to the PAM are not base paired. Kinetic assays showed that Cdi Cas9 cleaves dsDNA at least 50-fold slower than Spy Cas9 unless the DNA substrate contains the two-base-pair bulge (Figures 2.3B, 2.3C and 2.9B). To ensure this observation wasn't an artifact caused by the use of an artificial Cdi sgRNA, we repeated ssDNA vs. dsDNA cleavage experiments using the native Cdi dual crRNA:tracrRNA guide and saw a similar effect (Figures 2.9C, 2.9D). Thus, the type II-C proteins lack the robust DNA unwinding capacity observed for the type II-A Spy Cas9 enzyme.

These findings along with the weak selectivity for cognate guide RNA molecules suggested that extremely minimal guide RNA constructs might be sufficient for ssDNA target recognition. We compared ssDNA cleavage by Spy, Cdi and Ana Cas9 using the Spy single-guide RNA, the cognate crRNA or a 20 nucleotide RNA complementary to the ssDNA target sequence (Figure 2.3D). Spy Cas9 can use all three of these RNAs for sequence-specific ssDNA cleavage, although at similar concentrations the sgRNA supports faster cleavage kinetics (Figure 2.3D). In contrast, Cdi Cas9 only worked well with the full-length sgRNA; with the 20-nt guide RNA, a slightly longer cleavage product was slowly generated, and with the crRNA only trace reaction products appeared. Similar results were observed for the Ana Cas9, except that the crRNA construct was not functional in these reactions(Figure 2.9E). We also noted that the ssDNA cleavage site for Cdi and Ana enzymes changed to a 2-nt longer product when the shorter guide RNAs were used. These results suggest that an RNA-DNA hybrid provides some of the DNA cleavage specificity in these reactions, independent of PAM binding or the presence of the full-length guide RNA.

2.4.4 Divergent Cas9 proteins have varying affinities for guide RNAs and substrates

It is possible that the different cleavage activities and guide RNA specificities of the Spy, Cdi and Ana Cas9 enzymes reflect differences in binding affinities of these complexes. To test this, we first used filter-binding assays to test Cas9 protein affinity for cognate versus non-cognate guide RNA constructs. We found that Spy Cas9 binds to its cognate guide RNA with an apparent K_d of ~30 pM, whereas binding affinities for non-cognate guide RNAs were ~10,000-fold weaker (Figure 2.4A). The Cdi Cas9 has a binding affinity of ~1 nM for its cognate guide RNA, whereas binding to a non-cognate guide RNA is >100-fold weaker. The binding affinities of Ana Cas9 are similar to Cdi Cas9 at ~1nM for Cdi sgRNA and ~100nM for Spy sgRNA (Figure 2.10A). Based on these results, binding affinities alone do not account for the differences observed for DNA cleavage activity among the Spy, Cdi and Ana Cas9 enzymes, since cleavage assays were conducted under saturating or near-saturating conditions for the different guide RNAs with the Spy and Cdi enzymes. Instead, we suspect that the large Cas9 protein architecture can accommodate RNAs in distinct binding sites that may differ in their capacity to induce formation of a functional protein-RNA complex. Evidence for alternate guide RNA binding sites within Spy Cas9 was obtained previously (Martin Jinek et al., 2012a).

We next used gel mobility shift assays to test the affinity of pre-formed Cas9-guide RNA complexes for binding to a dsDNA substrate and to each ssDNA strand of this substrate individually (target and non-target) (Figure 2.4B, 2.10B and 2.10C). All three Cas9-guide RNA complexes had a binding affinity of ~10-60 nM for the target-strand ssDNA substrate. However, they had markedly different affinities for the dsDNA and non-target ssDNA molecules. Whereas the Spy Cas9-guide RNA complex bound to dsDNA with a similar ~10 nM affinity, binding to the non-target ssDNA was >1000 nM under these conditions. By contrast, the Cdi Cas9-guide RNA complex bound to dsDNA with a K_d of >1000 nM, and binding to non-target ssDNA was not detectable (Figure 2.4B). The Ana Cas9-guide RNA complex did not bind detectably to either dsDNA or non-target ssDNA (Figure 2.10B). These data show that only Spy Cas9 possesses a robust dsDNA binding capability. Consistent with DNA cleavage behavior for the Spy, Cdi and Ana proteins, this may reflect the superior ability of Spy Cas9 to unwind a target dsDNA to enable guide RNA strand annealing.

2.4.5 Cas9 conformational changes reveal differences in nucleic acid binding modes

Previous studies of Spy Cas9 revealed that the protein undergoes a substantial conformational rearrangement upon binding to guide-RNA and subsequently target DNA that can be detected by susceptibility to proteolytic cleavage (F. Jiang, Zhou, Ma, Gressel, & Doudna, 2015; Martin Jinek et al., 2014; Sternberg et al., 2014). We used trypsin digestion to assess structural changes that might accompany Cdi or Ana Cas9 assembly with guide RNA and DNA substrates. In control experiments we observed that Spy Cas9 alone is rapidly cleaved by trypsin, but in the presence of guide RNA the protein is largely protected (Figure 2.10A; Jiang et al., 2015; Jinek et al., 2014). Similarly, the Cdi Cas9 enzyme is cleaved rapidly by trypsin in the absence of guide RNA but becomes protected upon binding to guide RNA. We also noted that although the addition of non-target-strand ssDNA to Cdi Cas9: guide-RNA results in only small changes in protease sensitivity, Cdi Cas9 becomes more susceptible to proteolysis in the presence of target strand ssDNA (Figure 2.5A). Therefore, there is likely an additional conformational change occurring upon guide RNA-DNA hybridization, perhaps undocking of the mobile HNH domain and subsequent interaction with the substrate. Furthermore, the distinct cleavage patterns of Cdi Cas9 in the presence of target or non-target ssDNA indicate that in contrast to Spy Cas9, Cdi Cas9 maintains its ssDNA sequence specificity even at these elevated concentrations (15μM DNA). By contrast, the Ana Cas9 enzyme did not exhibit much difference in proteolytic digestion patterns in the absence or presence of (non-cognate) guide RNA and DNA (Figure 2.9B). Together, these results suggest that like Spy Cas9, Cdi Cas9 undergoes structural changes as it assembles into a functional RNA-guided DNA-bound complex. The lack of conformational change detected for Ana Cas9, even under conditions that favor guide RNA-DNA hybrid binding, may reflect an alternate mode of nucleic acid recognition in the absence of a cognate guide RNA.

Similarities between the Cdi and Ana Cas9 protein sizes (1085 and 1098 amino acids, respectively) and sequences (45.4% sequence identity) enabled creation of a homology model of Cdi Cas9 based on the crystal structure (Jinek et al., 2014) of Ana Cas9 (Figure 2.5B). Comparison to the Spy Cas9 protein structure in the presence of guide RNA (PDB ID: 4ZT9), suggest that the alpha-helical (REC) lobe contributes both to guide RNA recognition specificity (Briner et al., 2014) and possibly to DNA unwinding activity (Figure 2.5B). This analysis suggests that ancestral Cas9 proteins like Cdi or Ana may have lacked robust activity against dsDNA due to a more compact alpha-helical lobe structure that doesn't appear to include additional regions of the alpha-helical lobe found in Spy Cas9 that bind to the stem-loop bulge and nexus regions of the sgRNA, and contribute to guide-RNA specificity (Figure 2.5B) and DNA cleavage efficiency (Briner et al., 2014). Evolutionary pressure may have led to enhanced dsDNA binding and cleavage capability due to acquisition of new protein architectures and consequent guide-RNA specificity.

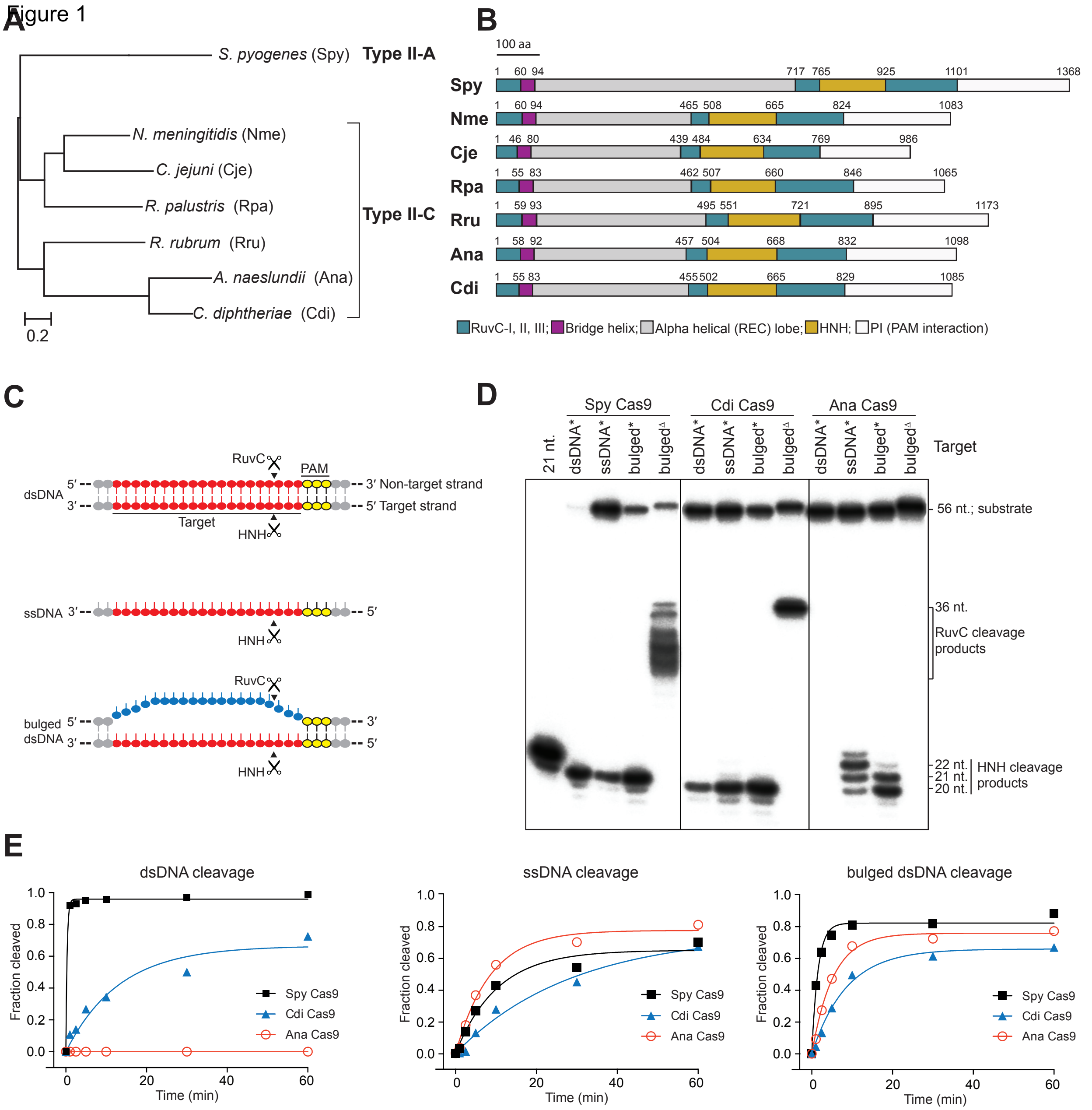


Figure 2.1. Diverse DNA cleavage activity of divergent Cas9 enzymes. (A). Phylogenetic tree of the seven Cas9 proteins used in this study, generated using MAFFT (Katoh, 2002). (B). Secondary structures and sizes of divergent Cas9 enzymes in this study. Domains are colored and drawn proportionally to the full length of its protein sequence. (C). Schematic presentation of three DNA substrates used for the figures of D and E. Target sequence is presented in red and underlined. PAM (AGG) is shown in yellow. (D). In vitro cleavage of DNA substrates. Cleavages of single-stranded DNA (ssDNA), double-strand DNA (dsDNA) and bulged DNA (bulged) mediated by Spy, Cdi and Ana Cas9 proteins are shown. * denotes that

the target strand is labeled; Δ indicates the non-target strand is labeled. Substrate and product sizes are labeled on the right. Vertical lines indicate the border between two separate gels (E). Quantification of DNA cleavage activity on dsDNA, ssDNA and bulged substrates with the 5′- end of the target strand radiolabeled. Cleavage assays were conducted in at least triplicate, as described in the Methods, and the quantified data were fitted with single-exponential decays to obtain pseudo first-order rate constants (kcleave¬) for each reaction. kcleave ± SD values for dsDNA cleavage by Spy, Cdi and Ana Cas9 are 3.5 ± 0.9 min-1, 0.041 ± 0.003 min-1 and 0.020 ± 0.001 min-1, respectively. kcleave ± SD values for ssDNA cleavage by Spy, Cdi and Ana Cas9 are 0.27 ± 0.09 min-1, 0.16 ± 0.1 min-1 and 0.11 ± 0.001 min-1, respectively. kcleave ± SD values for bulged dsDNA cleavage by Spy, Cdi and Ana Cas9 are 1.0 ± 0.5 min-1, 0.59 ± 0.4 min-1 and 0.73 ± 0.5 min-1, respectively.

Figure 2

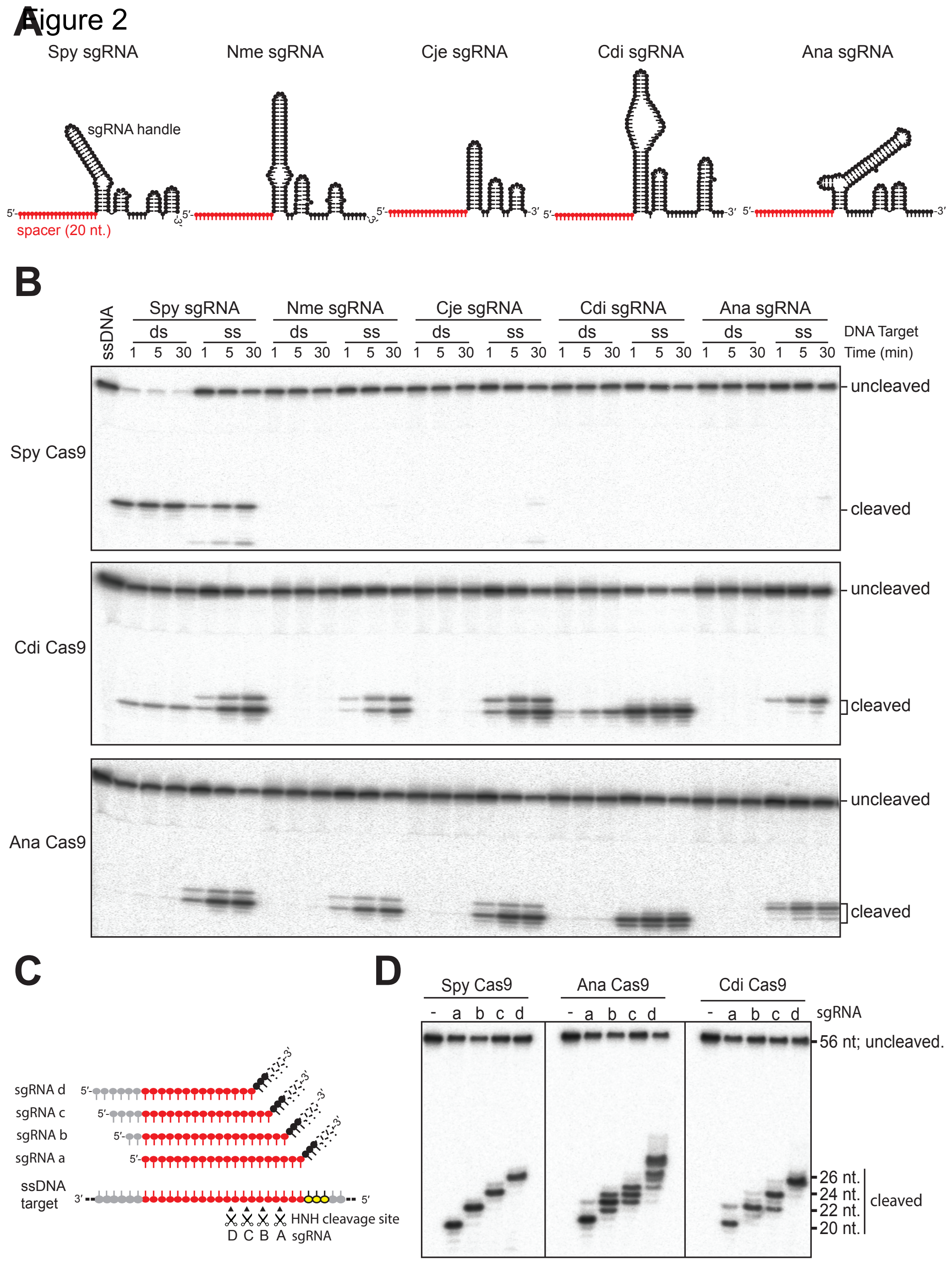


Figure 2.2. Programmable and PAM-independent Cas9-catalyzed DNA cutting. (A). Schematic representation of the five sgRNAs used in this study. All of the sgRNAs were drawn based on their mfold predictions (Zuker, 2003) and contain the same 20 nt. 'spacer' sequence (shown in red), which is complementary to the target sequence in the DNA substrates. (B). In vitro cleavage assay of ssDNA (ss) and dsDNA (ds) using various combinations of Cas9 proteins and sgRNAs. The 5′-end of the target strand is radiolabeled. Cas9 proteins are shown on the left. (C). Schematic presentation of the four sgRNAs (sgRNAs A-D) used in D. All four sgRNAs have a same Spy sgRNA handle and contain different 20-nt. spacer sequence that is able to base pair to the target sequence in the DNA substrate. (D). Programmable ssDNA cleavage by divergent Cas9 enzymes. The product sizes are labeled on the right; solid lines indicate borders between separate gels.

Figure 3

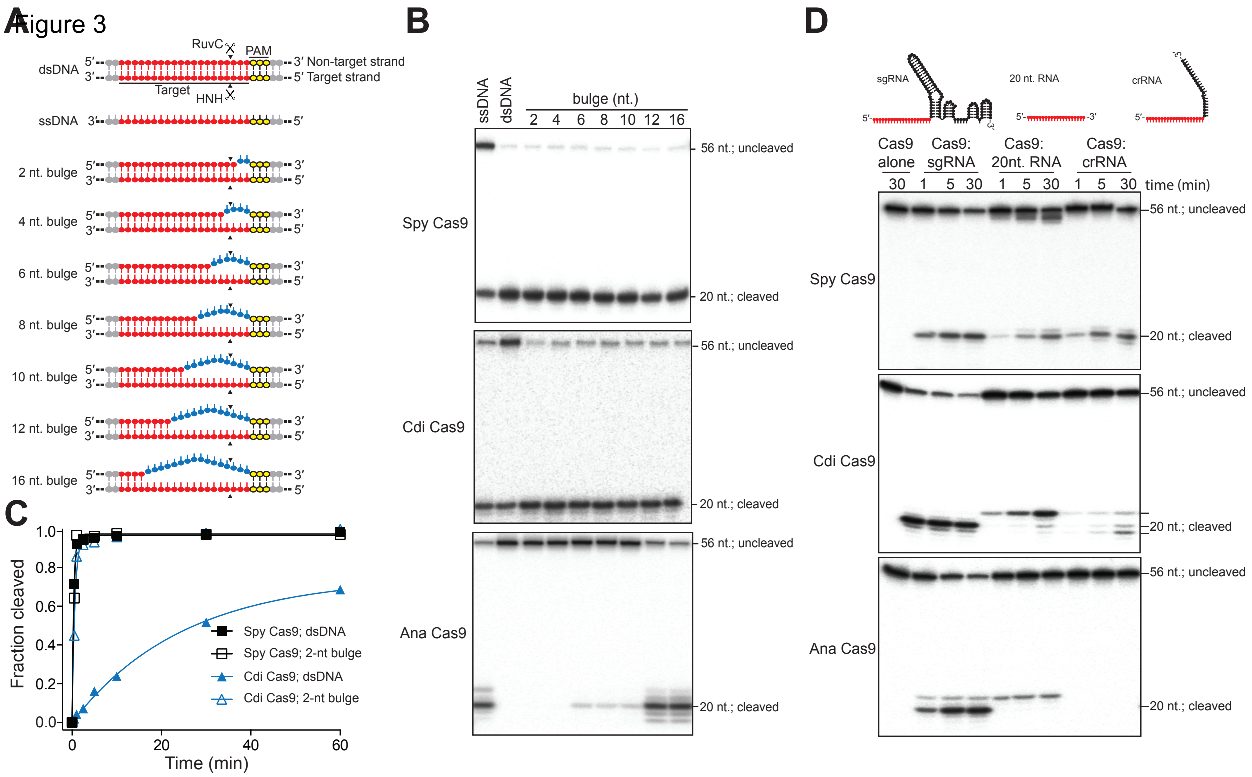


Figure 2.3. Substrate recognition varies among divergent Cas9 enzymes. (**A**). Schematic representation of DNA substrates used in this experiment. Target sequence is same in all of the substrates (red). For bulged dsDNAs, the bulge sizes are given on the left and mismatches are colored in blue. The PAM is indicated in yellow. (**B**). *In vitro* cleavage of bulged DNA substrates. Bulged substrates are indicated by the number of mismatches (2 to 16). Cas9 proteins are labeled on the left of each panel, and sizes of the substrate and cleaved products are labeled on the right. (**C**). Kinetic analysis of 2-nt. bulged substrate versus perfectly matched dsDNA. Cleavage assays were conducted in at least triplicate, as described in the Methods, and the quantified data were fitted with single-exponential decays to obtain pseudo first-order rate constants (k_{cleave}) for each reaction. k_{cleave} ± SD values for dsDNA cleavage by Spy and Cdi Cas9 are 2.9 ± 0.3 min^{-1} and 0.041 ± 0.002 min^{-1}, respectively**.** k_{cleave} ± SD values for 2-nt bulge dsDNA cleavage by Spy and Cdi Cas9 are 2.3 ± 0.1 min^{-1} and 0.99 ± 0.6 min^{-1}, respectively**.** (**D**). DNA cleavage by Cas9 proteins can be guided by short RNAs. Guide RNAs (sgRNA, 20-nt. and crRNA) used in this study are represented schematically above each lane. Cas9 proteins are labeled on the left of each panel, and sizes of the substrate and cleaved products are labeled on the right.

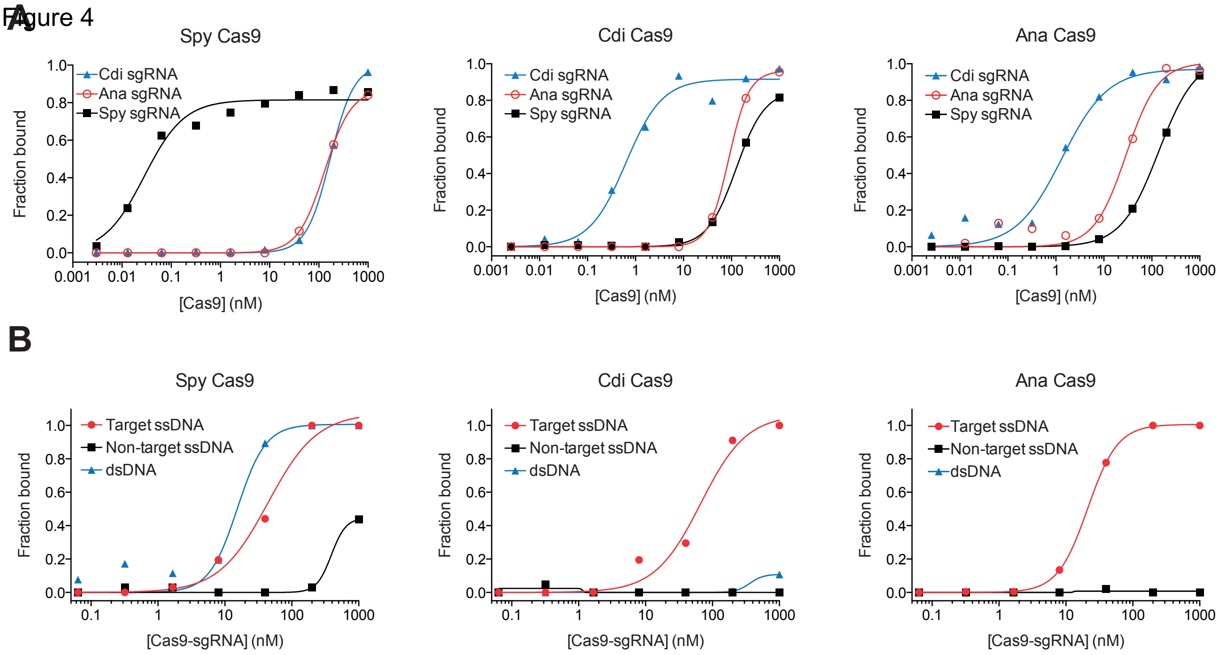


Figure 2.4 Binding affinity of Cas9 proteins for sgRNA and DNA. (**A**). Binding affinity of Cas9 proteins for cognate and non-cognate guides as determined by filter binding assays. Measurements were made in at least triplicate to determine K_D and a representative replicate is shown. Data were fit to a binding isotherm. (K_D ± SD for Spy Cas9 to Spy sgRNA, 28 pM ± 6 pM; Spy Cas9 to Cdi sgRNA, 146 ± 11nM; Spy Cas9 to Ana sgRNA, 152 ± 9nM; Cdi Cas9 to Spy sgRNA, 115 ± 51nM; Cdi Cas9 to Cdi sgRNA, 0.56 ± 0.13nM; Cdi Cas9 to Ana sgRNA, 46 ±17nM; Ana Cas9 to Spy sgRNA, 76 ± 11nM; Ana Cas9 to Cdi sgRNA, 0.35 ± 0.15nM; Ana Cas9 to Ana sgRNA, 6 ± 2nM) (**B**). Binding affinity of Cas9 proteins for dsDNA, target ssDNA and non-target ssDNA as measured by electrophoretic mobility shift assay (EMSA). Each Cas9 protein was incubated with its cognate guide except Ana, which was complexed with the Cdi sgRNA. Measurements were made in at least triplicate to determine K_D and a representative replicate is shown. Data were fit to a binding isotherm. (K_D ± SD [where appropriate] for Spy Cas9 to dsDNA, 25 ± 11 nM; Spy Cas9 to target ssDNA, 40 ± 11 nM; Spy Cas9 to non-target ssDNA, >1000 nM; Cdi Cas9 to dsDNA, >1000 nM; Cdi Cas9 to target ssDNA, 58 ± 32 nM; Cdi Cas9 for non-target ssDNA, >1000 nM; Ana Cas9 to dsDNA, >1000 nM; Ana Cas9 to target ssDNA, 12 ± 8 nM; Ana Cas9 for non-target ssDNA, >1000 nM)

Figure 5

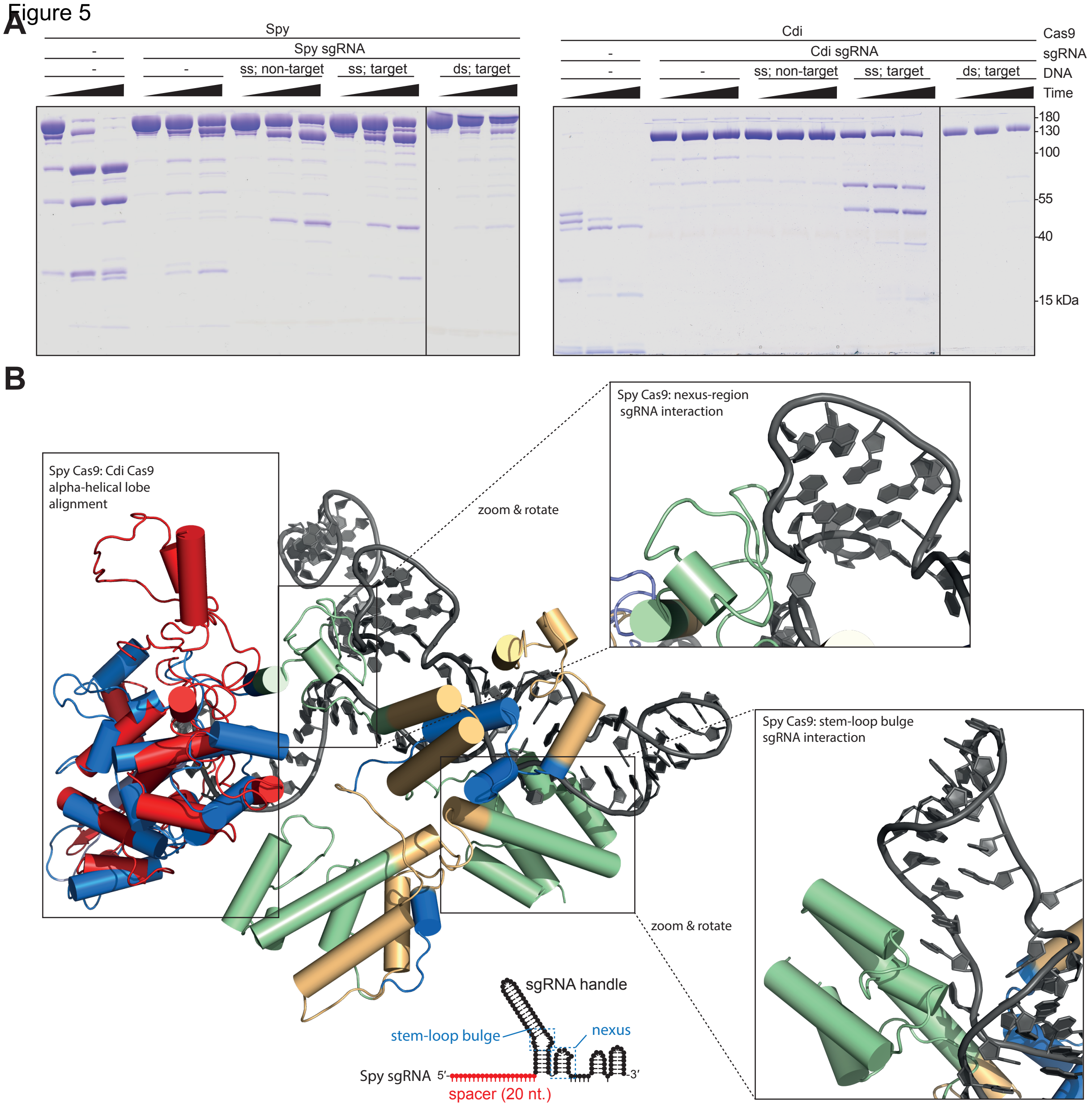


Figure 2.5. Limited proteolysis and structural comparison of Spy and Cdi Cas9 proteins. (**A**). Each Cas9 (5 µM) was incubated with its preferred guide RNA before the addition of DNA. Trypsin was added to the preformed complexes and allowed to digest the protein for 5, 15 or 30min. Degradation products were analyzed on 10% SDS-PAGE gels. ds: dsDNA; ss: ssDNA. Vertical lines indicate the border between two separate gels and approximate molecular weights are indicated on the right. (**B**). Structural alignment between Spy Cas9 alpha-helical (REC) lobe (PDB ID: 4ZT9) and the homology model of the alpha-helical lobe of Cdi Cas9 (based on the crystal structure of Ana Cas9 PDB ID: 4OGE); all atom RMSD: 3.9Å. The alpha-helical lobe of Cdi Cas9 is shown in red, sgRNA in shown in grey, and Spy Cas9 is colored according to primary sequence alignment with Cdi Cas9: green represents regions not present in the Cdi Cas9 protein sequence, blue represents regions present and conserved in the Cdi Cas9 sequence, and straw represents regions that are loosely conserved in the Cdi protein sequence. Spy sgRNA schematic is shown below with regions of interest boxed and labeled in blue.

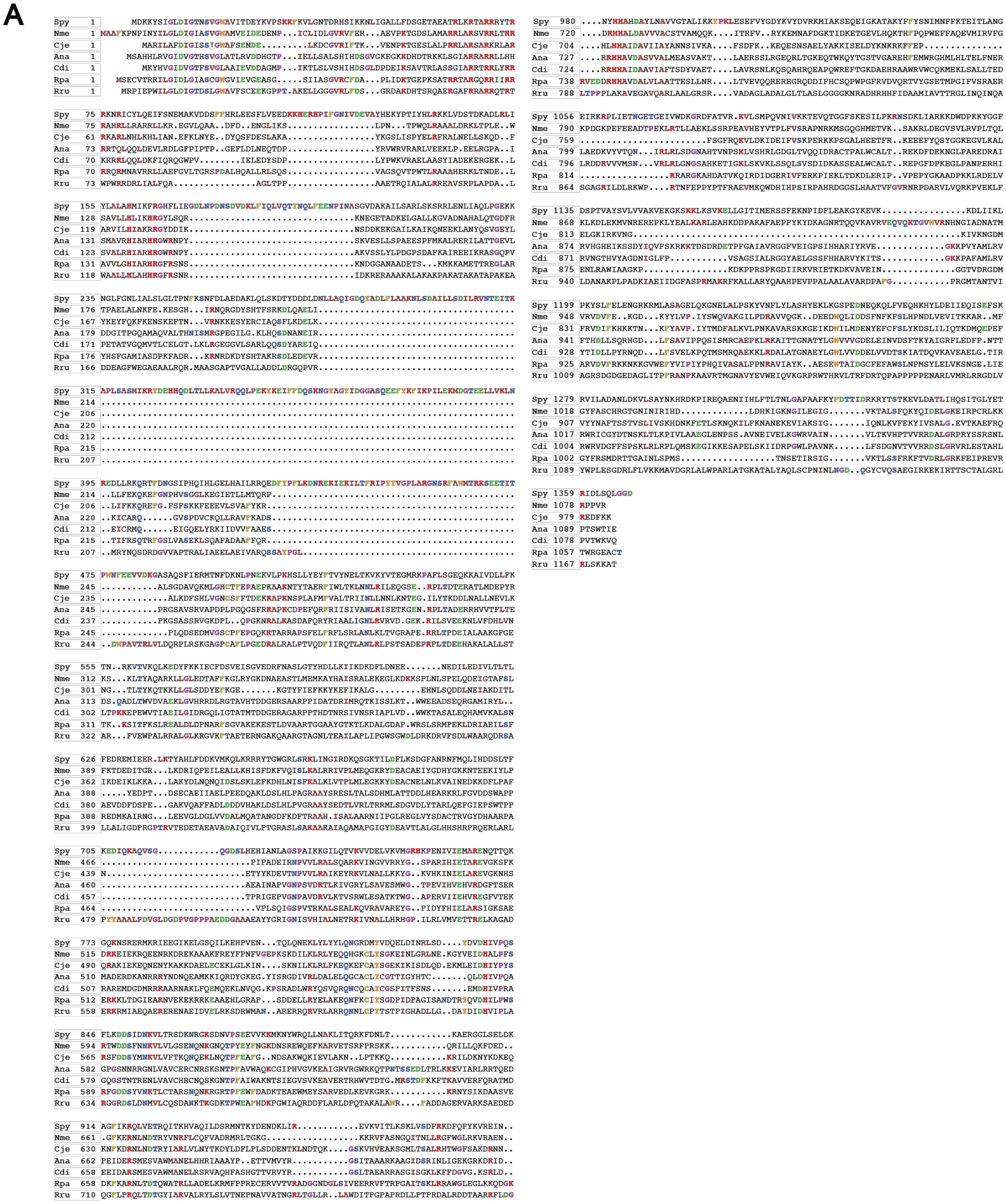


Figure 2.6. Cas9 proteins used in this work, Related to Figure 2.1. (A) Alignment of Cas9 proteins used in this work.

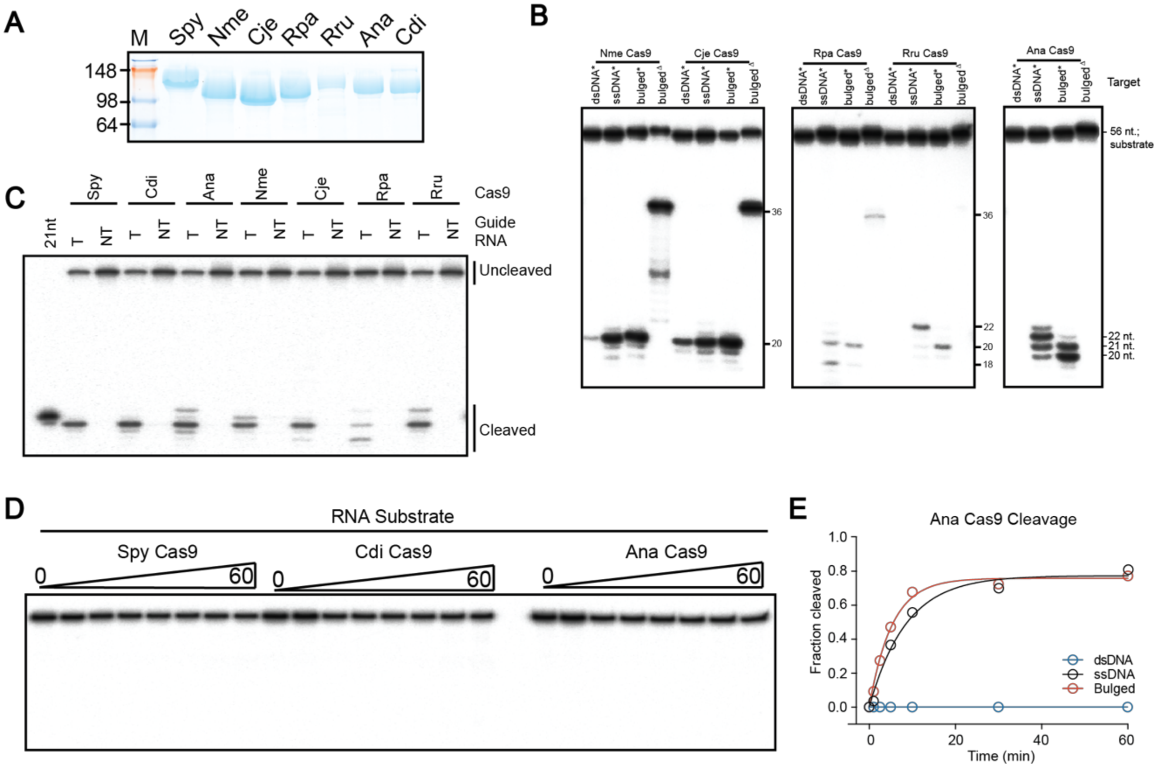


Figure 2.7. Cas9 proteins used in this work and their cleavage activity on various substrates, Related to Figure 2.1. (A) Superdex 200 purified Cas9 proteins analyzed on 12% SDS-PAGE gel. (B) *In vitro* cleavage assays of ssDNA, dsDNA and bulged DNAs mediated by Nme, Cje, Rpa and Rru Cas9s. The cleavage products were separated in 10% urea-PAGE gel. (C) ssDNA cleavage mediated by Cas9s using targeting (T) or non-targeting (NT) guide sequences. Guide RNAs are constructed from Spy sgRNA handle. (D) Cleavage assay of Cas9 enzymes using RNA substrates. The RNA sequence is the same as that used for ssDNA cleavage assays presented in this work. (E) Cleavage kinetics of Ana Cas9.

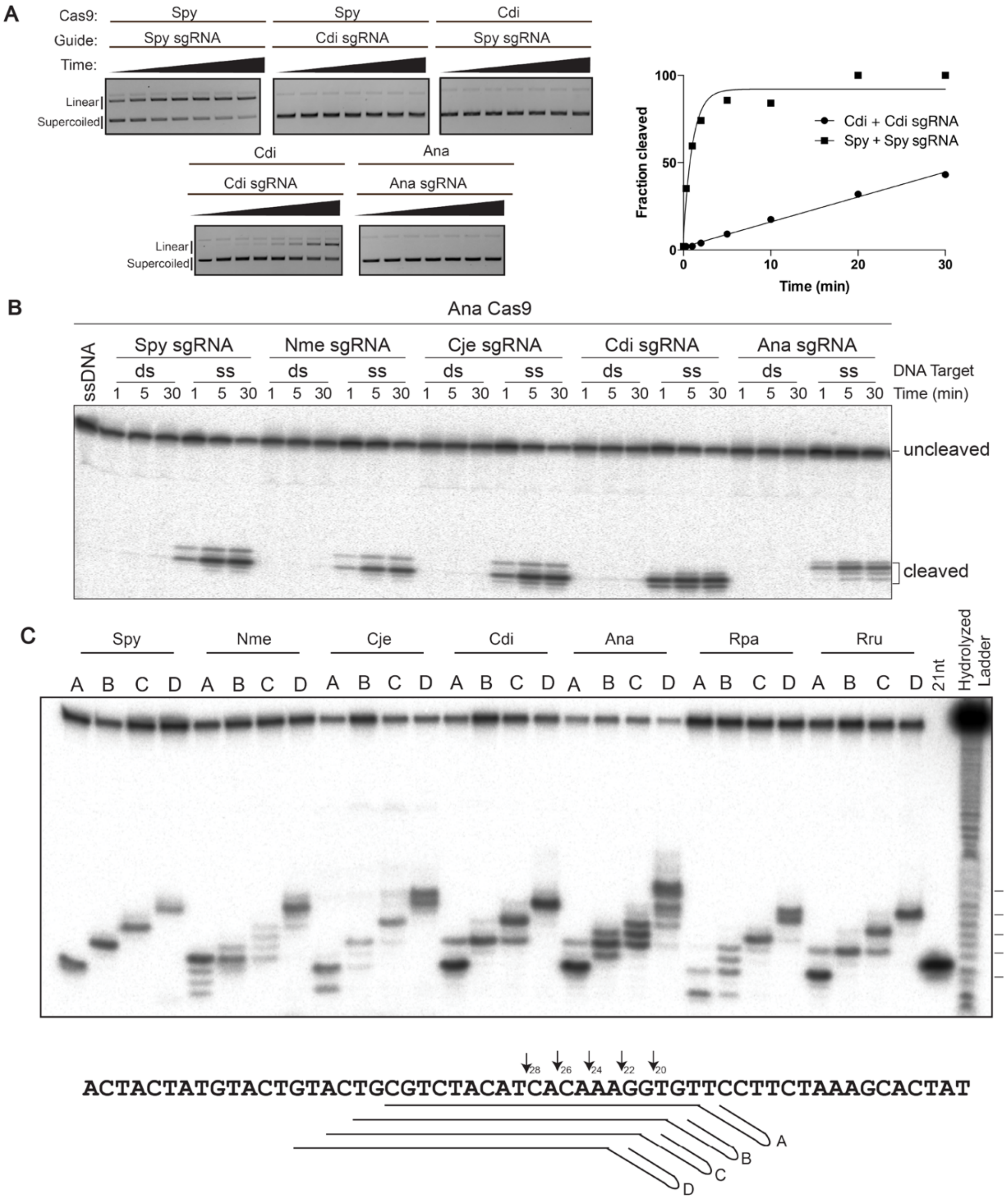


Figure 2.8. In vitro Cas9 cleavage activity on supercoiled plasmid and ssDNA substrates. Related to Figure 2.2. (A) Cleavage of supercoiled pUC19 plasmid by Cas9 enzymes. Products were analyzed on 1% agarose gel stained with ethidium bromide. (B) Cleavage of ssDNA by Ana Cas9 with various sgRNA sequences. “ss” and “ds” denote single stranded DNA and double stranded DNA respectively. (C) Cleavage of ssDNA substrate tiled with guide RNAs. Spy and Cdi Cas9 are shown from main text for reference. The cleavage products were separated by a 10% urea-PAGE gel.

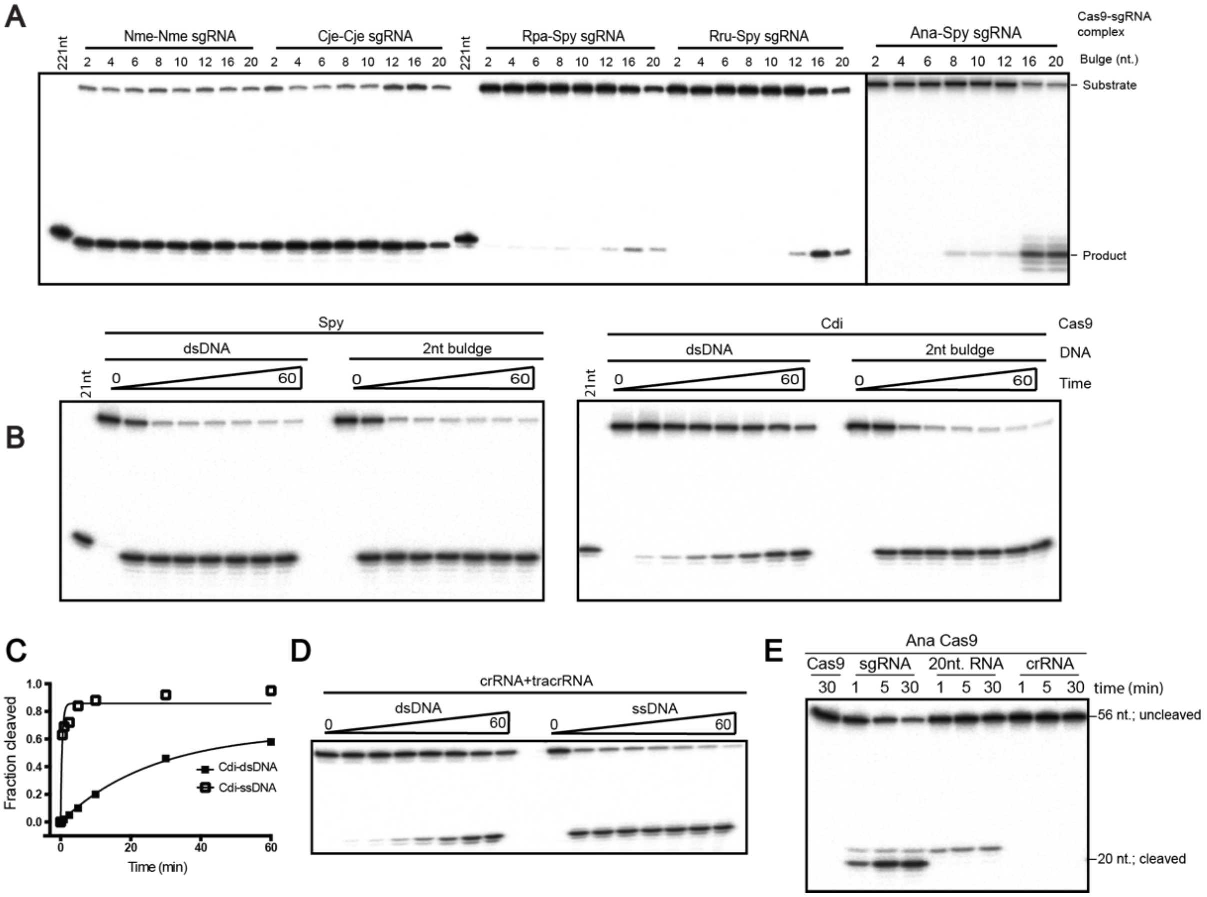


Figure 2.9. *In vitro* cleavage assays on budged substrates. Related to Figure 2.3. (A) *In vitro* cleavages of bulged DNA substrates mediated by Nme, Cje, Rpa and Rru Cas9. (B) Representative gels used for kinetic analysis of 2nt bulged substrate versus perfectly matched dsDNA. The cleavage products were separated in 10% urea-PAGE gel.(C) Cleavage kinetics of Cdi Cas9 using dual guide RNA. (D) Gels used for quantification shown in S4C. (E) Cleavage of Ana Cas9 using shortened guide constructs. See Figure 3D for diagrams of guide RNAs.

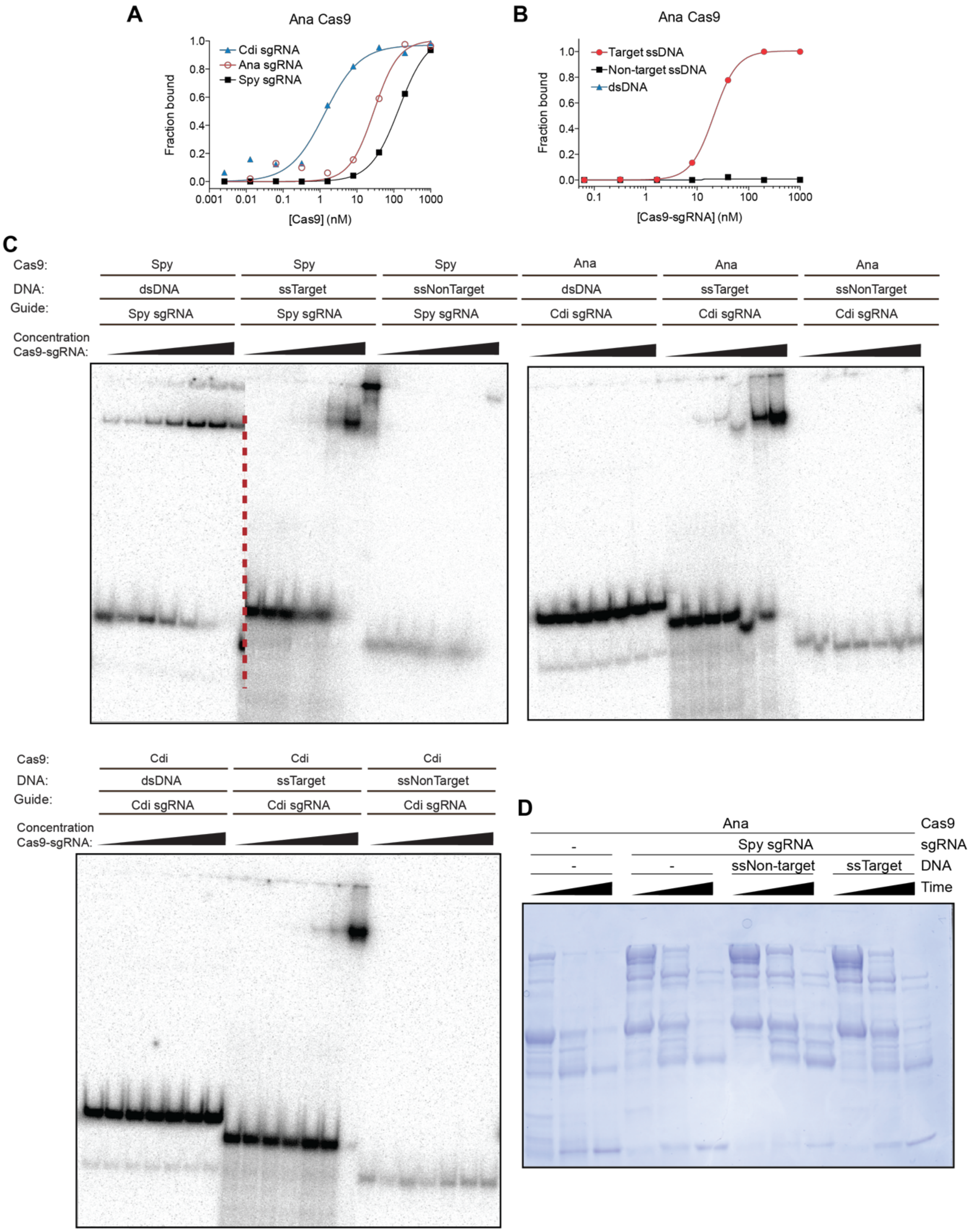


Figure 2.10. Electrophoresis mobility shift assay and limited proteolysis of Cas9 enzymes. Related to Figures 2.4 and 2.5. (A) Binding of Ana Cas9 to sgRNA. (K_D ± SD for Ana Cas9 to Spy sgRNA, 76 ± 11nM; Ana Cas9 to Cdi sgRNA, 0.35 ± 0.15nM; Ana Cas9 to Ana sgRNA, 6 ± 2nM) (B) Binding Ana Cas9 to DNA substrates. (K_D ± SD [where appropriate] for Ana Cas9 to target ssDNA, 12 ± 8 nM; Ana Cas9 for non-target ssDNA, >1000 nM) (C) Representative gels used for EMSA data, related to Figure 4. Red dotted line indicates the border between two separate gels. (D) Limited proteolysis of Ana Cas9, related to Figure 5.

Strains	Symbol	Subclass	# of a.a.	MW (Da)	Source
Streptococcus pyogenes SF370	Spy	IIA	1366	158445	ATCC 700294
Nesseria meningitidis Z2491	Nme	IIC	1082	124324	Pasteur Institute
Campylobacter jejuni NCTC 11168	Cje	IIC	984	114770	Pasteur Institute
Rhodopseudomonas palustris BisB5	Rpa	IIC	1064	123267	ATCC BAA-1123
Rhodospirillum rubrum ATCC 11170	Rru	IIC	1173	132088	ATCC 11170
Actinomyces naeslundii str. Howell 279	Ana	IIC	1101	127417	ATCC 12104
Corynebacterium diphtheriae NCTC 13129	Cdi	IIC	1084	125185	*ATCC* 700971

Table 2.1. Cas9 orthologues used in this study. Related to Figure 2.1.

DNA	**Sequence**
20-nt target strand DNA	5’-TGTGGAAACACTACATCTGC-3’
20-nt non-target strand DNA	5’-GCAGATGTAGTGTTTCCACA-3’
20-nt guide RNA	5’-GCAGAUGUAGUGUUUCCACA-3’
Perfectly matched dsDNA	5’-TGATGATACATGACATGACGCAGATGTAGTGTTTCCACAAGGAAGATTTCGTGATA-non-target 3’-ACTACTATGTACTGTACTGCGTCTACATCACAAAGGTGTTCCTTCTAAAGCACTAT-target
2-nt bulged dsDNA	5’-TGATGATACATGACATGACGCAGATGTAGTGTTTCCATTAGGAAGATTTCGTGATA-non-target 3’-ACTACTATGTACTGTACTGCGTCTACATCACAAAGGTGTTCCTTCTAAAGCACTAT-target
4-nt bulged dsDNA	5’-TGATGATACATGACATGACGCAGATGTAGTGTTTCTTTTAGGAAGATTTCGTGATA-non-target 3’-ACTACTATGTACTGTACTGCGTCTACATCACAAAGGTGTTCCTTCTAAAGCACTAT-target
6-nt bulged dsDNA	5’-TGATGATACATGACATGACGCAGATGTAGTGTTATTTTTAGGAAGATTTCGTGATA-non-target 3’-ACTACTATGTACTGTACTGCGTCTACATCACAAAGGTGTTCCTTCTAAAGCACTAT-target
8-nt bulged dsDNA	5’-TGATGATACATGACATGACGCAGATGTAGTGCCATTTTTAGGAAGATTTCGTGATA-non-target 3’-ACTACTATGTACTGTACTGCGTCTACATCACAAAGGTGTTCCTTCTAAAGCACTAT-target
10-nt bulged dsDNA	5’-TGATGATACATGACATGACGCAGATGTAGATCCATTTTTAGGAAGATTTCGTGATA-non-target 3’-ACTACTATGTACTGTACTGCGTCTACATCACAAAGGTGTTCCTTCTAAAGCACTAT-target
12-nt bulged dsDNA	5’-TGATGATACATGACATGACGCAGATGTTTATCAATTTTTAGGAAGATTTCGTGATA-non-target 3’-ACTACTATGTACTGTACTGCGTCTACATCACAAAGGTGTTCCTTCTAAAGCACTAT-target
16-nt bulged dsDNA	5’-TGATGATACATGACATGACGCAGTAATTTATCAATTTTTAGGAAGATTTCGTGATA-non-target 3’-ACTACTATGTACTGTACTGCGTCTACATCACAAAGGTGTTCCTTCTAAAGCACTAT-target
20-nt bulged dsDNA	5’-TGATGATACATGACATGACTTATTTTTTTTTTAATTTTTAGGAAGATTTCGTGATA-non-target 3’-ACTACTATGTACTGTACTGCGTCTACATCACAAAGGTGTTCCTTCTAAAGCACTAT-target
Spy sgRNA	GCAGAUGUAGUGUUUCCACAGUUUUAGAGCUAUGCUGAAAGCAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUG
Spy-crRNA	GCAGAUGUAGUGUUUCCACAGUUUUAGAGCUAUGCUGUUUUGAAUGGUCCCAAAAC
Nme-sgRNA	GCAGAUGUAGUGUUUCCACAGUUGUAGCUCCCUUUCUCAUUUCGCAGUGCGAAAGCACUGCGAAAUGAGAACCGUUGCUACAAUAAGGCCGUCUGAAAAGAUGUGCCGCAACGCUCUGCCCCUUAAAGCUUCUGC
Nme-crRNA	GCAGAUGUAGUGUUUCCACAGUUGUAGCUCCCUUUCUCAUUUCGCAGUGCG
Cje-sgRNA	GCAGAUGUAGUGUUUCCACAGUUUUAGUCCCUUUUUAAAUUUCUUUAUGGUAAAAUUAUAAUCUCAUAAGAAAUUUAAAAAGGGACUAAAAUAAAGAGUUUGCGGGACUCUGCGGGGUUACAAUCCCCUAAAACCGCUU
Cje-crRNA:	GCAGAUGUAGUGUUUCCACAGUUUUAGUCCCUUUUUAAAUUUCUUUAUGGU
Cdi-sgRNA:	GCAGAUGUAGUGUUUCCACAACUGGGGUUCAGUUCUCAAAAACCCUGAUAGACUUGAAAAGUCACUAACUUAAUUAAAUAGAACUGAACCUCAGUAAGCAUUGGCUCGUUUCCAAUGUUGAUUGCUCCGCCGGUGCUCCUUAUUUUUAAGGGCGCCGGCUUU
Cdi-crRNA	GCAGAUGUAGUGUUUCCACAACUGGGGUUCAGUUCUCAAAAACCCUGAUAGACUUC
Ana-sgRNA	GCAGAUGUAGUGUUUCCACAGCUGGGAUUCAGUCACCAGACCCCUUGAUAGACUUCAGAAACGUCUGUCAAGGGGGUCUGACCAGCCGUAAACACCUCGUCAGAGGUUCAGGAAGAUCAUGAGCUGUUGGGGC
Ana-crRNA	GCAGAUGUAGUGUUUCCACAGCUGGGAUUCAGUCACCAGACCCCUUGAUAGACUUC
Rpa-sgRNA	GCAGAUGUAGUGUUUCCACAGCCGUGGCUUCCCUACCGAUUUGAAAAAAUCGGUAGGAAAGCCACGGCAAGCAACGGAAACCUUGGUUUUCGUUGCGAAGGAUUACCCCCGUUGCGGAGACGUAACGGGGGG

Rru-sgRNA:	GCAGAUGUAGUGUUUCCACAACUCUACCAUGGCGGUGUGGGACGGGAAACCGUCCCACACGGCCAUGGUAGAGUGCGAUCACCCUAUCACACCUCUUUGAUUUGUAAAGGGGCACACCUCUUU
sgRNA A:	GCAGAUGUAGUGUUUCCACAGUUUUAGAGCUGUGCUGAAAAGCACAGCACGUUAAAAUAAGGCAGUGAUUUUUAAUCCAGUCCGUAUUCAGCUUGAAAAAGU
sgRNA B:	ACGCAGAUGUAGUGUUUCCAGUUUUAGAGCUGUGCUGAAAAGCACAGCACGUUAAAAUAAGGCAGUGAUUUUUAAUCCAGUCCGUAUUCAGCUUGAAAA AGU
sgRNA C:	UGACGCAGAUGUAGUGUUUCGUUUUAGAGCUGUGCUGAAAAGCACAGCACGUUAAAAUAAGGCAGUGAUUUUUAAUCCAGUCCGUAUUCAGCUUGAAAAAGU
sgRNA D:	CAUGACGCAGAUGUAGUGUUGUUUUAGAGCUGUGCUGAAAAGCACAGCACGUUAAAAUAAGGCAGUGAUUUUUAAUCCAGUCCGUAUUCAGCUUGAAAA AGU

Table 2.2. Sequences of DNA and guide RNA used in this paper. Related to Figure 2.1-5. Mismatches are shown in red and PAM sequences are identified in blue. Target strand of DNA and targeting sequence of RNA are underlined.

2.5 Discussion

CRISPR-Cas9 proteins share a common domain composition, but are highly diverse in overall size, primary sequence and guide RNA and PAM specificity (Jinek et al., 2012; Esvelt et al., 2013; Chylinski et al., 2014; Fonfara et al., 2014; Ran et al., 2015). We show here that evolutionarily divergent Cas9 enzymes from the type II-C Cas9 subclass preferentially cleave ssDNA, as opposed to the dsDNA activity typified by the type II-A Spy Cas9. Since dsDNA cleavage requires guide-RNA strand invasion and local DNA unwinding to form an R-loop (Szczelkun et al., 2014), these findings imply that the smaller and perhaps ancestral type II-C Cas9 proteins have limited DNA unwinding capability.

Distinct Cas9 cleavage activities correlate with protein size and architecture. Structural and biochemical studies indicate that the type II-A Spy Cas9 undergoes a large conformational rearrangement upon guide RNA binding, with additional changes upon dsDNA target recognition (F. Jiang et al., 2015; Martin Jinek et al., 2014). Similarly, we found that the type II-C Cdi Cas9 undergoes substantial structural change upon binding to guide RNA, as detected by partial proteolysis (Figure 2.5A). By contrast, the similar proteolytic digestion patterns detected for Ana Cas9 in the absence or presence of guide RNA or DNA suggest a similar or highly dynamic structure under all conditions tested, an observation likely related to the minimal dsDNA-cleavage activity of this enzyme without its cognate guide-RNA and PAM present.

We also found that the type II-C enzymes can utilize a variety of guide RNAs to recognize and cleave ssDNA substrates, and type II-A and II-C Cas9 proteins can employ the crRNA or even a simple 20-nt RNA for residual site-specific ssDNA cleavage. The ability of type II-C enzymes to use non-cognate full-length guide RNAs is markedly different from the activity of Spy Cas9, which is highly selective for its cognate guide RNA (Briner et al., 2014; Martin Jinek et al., 2012a). Structural differences between Spy and Ana Cas9 proteins, and likely similarity between the structures of the Ana and Cdi Cas9 proteins, suggest an interesting connection between recognition of the guide RNA in the region of the nexus (Briner et al., 2014; Figure 2.5B) and DNA binding capability. The lack of guide RNA binding specificity in the type II-C Cas9s may also relate to their limited ability to unwind dsDNA.

The observations that type II-C Cas9 enzymes are more promiscuous for guide RNA binding and have limited dsDNA unwinding and cleavage activity have important implications for genome editing applications. Because guide RNA binding is less specific, the potential for spurious guide RNA association may increase the likelihood of unintended editing events. In addition, the lack of robust dsDNA binding and cleavage activity imply inhibited ability to generate dsDNA breaks in cells. This finding may explain why type II-C Cas9 enzymes have reduced or in some cases undetectable genome editing activity relative to the type II-A Spy Cas9 (Hou et al., 2013; Ran et al., 2015b).

Based on previous phylogenetic analysis (Chylinski et al., 2014b; Fonfara et al., 2014), it is possible that the properties of the type II-C proteins represent ancestral activities of the Cas9 enzyme family. Although the physiological functions of ssDNA cleavage by type II-C Cas9 enzymes have not been determined, these enzymes may have evolved to recognize substrates such as ssDNA bacteriophage. However, bacteriophage ssDNA is generally stably bound and protected by ssDNA-binding proteins (Chase & Williams, 1986) and ssDNA cleavage by type II-C Cas9 enzymes would therefore require either transient exposure of naked ssDNA during replication or the ability to displace cellular ssDNA-binding proteins. In addition, Cas9 proteins could potentially silence active gene loci in a manner analogous to the ssDNA targeting that occurs during transcription *in vivo* by type III-A CRISPR/Cas systems (Deng, Garrett, Shah, Peng, & She, 2013; Samai et al., 2015b). In all cases, type II-C Cas9 enzymes might require additional cellular co-factors to promote efficient dsDNA cleavage. Persistent evolutionary pressure to protect bacteria from dsDNA phage or plasmids might have driven the emergence of larger variants of Cas9 with insertions in the alpha-helical (REC) lobe that confer both guide RNA recognition specificity and the ability to unwind dsDNA target sites. We also envision that despite limited efficiency for genome editing applications requiring dsDNA cleavage, it may be possible to harness the intrinsic type II-C Cas9 activity for ssDNA target recognition *in vitro* or possibly in cells.

2.6 Acknowledgments

We thank Doudna Lab members and Rodolphe Barrangou for helpful discussions. This work was supported by the Center for RNA Systems Biology (NIH P50, J. Cate, P.I.) and the Howard Hughes Medical Institute. L.H. is supported by a graduate fellowship from the National Science Foundation. J.A.D. is a Howard Hughes Medical Institute Investigator.

2.7 Author contributions

E.M., L.H., M.O. and J.A.D. designed the experiments; E.M., L.H. and K.Z. conducted the experiments; E.M., L.H., M.O. and J.A.D. analyzed the data; E.M., L.H., M.O. and J.A.D. wrote the manuscript.

Chapter 3

A thermostable Cas9 with increased lifetime in human plasma

*Part of the work presented in this chapter has previously been published in the following manuscript: Harrington, L. B., Paez-Espino, D., Staahl, B. T., Chen, J. S., Ma, E., Kyrpides, N. C., & Doudna, J. A. (2017). A thermostable Cas9 with increased lifetime in human plasma. *Nature Communications*, *8*(1), 1–7.

3.1 Abstract

CRISPR-Cas9 is a powerful technology that has enabled genome editing in a wide range of species. However, the currently developed Cas9 homologs all originate from mesophilic bacteria, making them susceptible to degradation and unsuitable for applications requiring cleavage at elevated temperatures. Here, we show that the Cas9 protein from the thermophilic bacterium *Geobacillus stearothermophilus* (GeoCas9) catalyzes RNA-guided DNA cleavage at elevated temperatures. GeoCas9 is active at temperatures up to 70°C, compared to 45°C for *Streptococcus pyogenes* Cas9 (SpyCas9), which expands the temperature range for CRISPR-Cas9 applications. We also found that GeoCas9 is an effective tool for editing mammalian genomes when delivered as a ribonucleoprotein (RNP) complex. Together with an increased lifetime in human plasma, the thermostable GeoCas9 provides the foundation for improved RNP delivery *in vivo* and expands the temperature range of CRISPR-Cas9.

3.2 Introduction

The use of CRISPR-Cas9 has rapidly transformed the ability to edit and modulate the genomes of a wide range of organisms(H. Wang, La Russa, & Qi, 2016). This technology, derived from adaptive immune systems found in thousands of bacterial species, relies on RNA-guided recognition and cleavage of invasive viral and plasmid DNA(Chylinski et al., 2014a). The Cas9 proteins from these species differ widely in their size and cleavage activities(Chylinski et al., 2014b; Ma, Harrington, O'Connell, Zhou, & Doudna, 2015; Ran et al., 2015b). Despite the abundance and diversity of these systems, the vast majority of applications have employed the first Cas9 homolog developed from *Streptococcus pyogenes* (SpyCas9)(Martin Jinek et al., 2012b). In addition to SpyCas9, several other Cas9 proteins have also been shown to edit mammalian genomes with varying efficiencies(Cong, Ran, Cox, Lin, & Barretto, 2013; Esvelt et al., 2013a; Hirano et al., 2016; Hou et al., 2013; Ran et al., 2015b). While these proteins together provide a robust set of tools, they all originate from mesophilic hosts, making them unsuitable for applications requiring cleavage at higher temperatures or extended protein stability.

This temperature restriction is particularly limiting for genome editing in obligate thermophiles(Xiang, Zhang, An, Cheng, & Wang, 2016). Recent efforts using SpyCas9 to edit a facultative thermophile have been possible by reducing the temperature within the organism(Mougiakos et al., 2017). While effective, this approach is not feasible for obligate thermophiles, and requires additional steps for moderate thermophiles. This is especially important for metabolic engineering for which thermophilic bacteria present enticing hosts for chemical synthesis due to decreased risk of contamination, continuous recovery of volatile products and the ability to conduct reactions that are thermodynamically unfavorable in mesophilic hosts(Zeldes et al., 2015). Developing a thermostable Cas9 system will enable facile genome editing in thermophilic organisms using technology that is currently restricted to mesophiles.

CRISPR-Cas9 has also emerged as a potential treatment for genetic diseases(Porteus, 2016). A promising method for the delivery of Cas9 into patients or organisms is the injection of preassembled Cas9 ribonucleoprotein complexes (RNP) into the target tissue or bloodstream(Staahl et al., 2017). One major challenge to this approach is that Cas9 must be stable enough to survive degradation by proteases and RNases in the blood or target tissue for efficient delivery. Limited protein lifetime will require delivery of higher doses of Cas9 into the patient or result in poor editing efficacy. In contrast, delivering a Cas9 with improved stability could greatly enhance genome-editing efficiency *in vivo*.

To address these challenges, we tested the thermostable Cas9 protein from *Geobacillus stearothermophilus* (GeoCas9). We find that GeoCas9 maintains activity over a wide temperature range. By harnessing the natural sequence variation of GeoCas9 from closely related species, we engineered a PAM variant that recognizes additional PAM sequences and thereby doubles the number of targets accessible to this system. We also engineered a highly efficient sgRNA using RNAseq data from the native organism and show that GeoCas9 can efficiently edit genomic DNA in mammalian cells. The functional temperature range of GeoCas9 complements that of previously developed Cas9 systems, greatly expanding the temperatures that Cas9 can be used for both *in vitro* cleavage and genome editing applications.

3.3 Materials and Methods

3.3.1 Identification of Cas9 homologs and generation of plasmids

We mined all isolate genomes from the public Integrated Microbial genomes (IMG) database(Markowitz et al., 2012) using the “Genome Search by Metadata Category tool.” We selected all the genomes annotated as “thermophile” (336) or “hyperthermophile” (94) and searched for the presence of Cas9-like candidates (hits to a TIGRfam model 01865 for Csn-like or 03031 for a Csx12-like) contained within a full CRISPR-Cas system (presence of Cas1, Cas2, and a Repeat-Spacers array). We initially selected the GeoCas9 variant due to its completeness, smaller gene size (shorter than the widely used SpyCas9), and growth in a large temperature range from 30-75 (optimal at 55C).The Cas9 from *Geobacillus stearothermophilus* was codon optimized for *E. coli,* ordered as Gblocks (IDT) and assembled using Gibson Assembly. For protein expression, a pET based plasmid containing an N terminal 10xHis-tag and MBP was used. For PAM depletion assays, a p15A plasmid was generated with the sgRNA constitutively expressed.

3.3.2 Cas9 purification

Cas9 was purified as previously described(Martin Jinek et al., 2012b) with modification. After induction, *E. coli* BL21(DE3) expressing Cas9 was grown in Terrific Broth overnight at 18°C. Cells were harvested, re-suspended in Lysis Buffer (50mM Tris-HCl, pH 7.5, 20mM imidazole, 0.5mM TCEP, 500mM NaCl, 1mM PMSF), broken by sonication, and purified on Ni-NTA resin. TEV was added to the elution and allowed to cleave overnight at 4°C. The resulting protein was loaded over tandem columns of an MBP affinity column onto a heparin column and eluted with a linear gradient from 300mM to 1250mM NaCl. The resulting fractions containing Cas9 were purified by gel filtration chromatography and flash frozen in Storage Buffer (20mM HEPES-NaOH pH 7.5, 5% Glycerol, 150mM NaCl, 1mM TCEP).

3.3.3 Differential Scanning Calorimetry

Cas9 proteins were dialyzed into degassed DSC Buffer (0.5mM TCEP, 50mM KH_2PO_4(pH 7.5), 150mM NaCl) overnight at 4°C. Samples were diluted to 0.3mg ml^{-1} and loaded a sample cell of a NanoDSC (TA instruments); buffer alone was used in the reference cell. The cell was pressurized to 3atm and the sample was heated from 20 to

90°C. Measurements made for buffer in both the sample and reference cells were subtracted from the sample measurements.

3.3.4 Biochemical cleavage assays

Radioactive cleavage assays were conducted as previously described(Wright et al., 2015). Reactions were carried out in 1× Reaction Buffer (20mM Tris-HCl, pH 7.5, 100mM KCl, 5mM $MgCl_2$, 1mM DTT and 5% glycerol (v v^{-1})). 100nM Cas9 and 125nM sgRNA were allowed to complex for 5min at 37°C. ~1nM radiolabeled probe was added to the RNP to initiate the reaction. Finally, the reaction was quenched with 2× Loading Buffer (90% formamide, 20mM EDTA, 0.02% bromophenol blue, 0.02% xylene cyanol and products were analyzed on 10% urea-PAGE gel containing 7M urea.

For thermostability measurements (Fig. 3.4a), 100nM Cas9 was complexed with 150nM sgRNA in 1× Reaction Buffer for 5min at 37°C. 100nM of a PCR product containing the targeted sequence was cleaved using dilutions of the estimated 100nM RNP complex to accurately determine a 1:1 ratio of Cas9 to target. Next, samples were challenged at the indicated temperature (40°C–75°C) for 10min and then returned to 37°C. 100nM PCR product containing the targeted sequence was added to the reaction and it was allowed to react for 30min at 37°C. The reaction was quenched with 6× Quench Buffer (15% glycerol (v v^{-1}), 1mg ml^{-1} Orange G, 100mM EDTA) and products were analyzed on a 1.25% agarose gel stained with ethidium bromide.

For thermophilicity measurements (Fig. 3.4b), 500nM Cas9 was complexed with 750nM sgRNA in 1x Reaction Buffer for 5min at 37°C. The samples were placed at the assayed temperature (20°C–80°C) and 100nM of PCR product was added to the reaction. Time points were quenched using 6× Quench Buffer and analyzed on a 1.25% agarose gel stained with SYBR Safe (Thermo Fisher Scientific).

To study the effect of human plasma on the stability of Cas9 proteins, preassembled Cas9-RNP was incubated for 8 hours either at 37°C or 4°C in Reaction Buffer with the specified amount of normal human plasma. Substrate was then added and cleavage products were analyzed as described for thermostability measurements. Uncropped gels related to Figure 4 and 5 can be found in Supplementary Figure 3.9.

3.3.5 Small RNA sequencing

Geobacillus stearothermophilus was obtained from ATCC and cultured at 55°C in Nutrient Broth (3g beef extract and 5g peptone per liter water, pH 6.8) to saturation. Cells were pelleted and RNA was extracted using a hot phenol extraction as previously described(Zhang et al., 2013). Total RNA was treated with TURBO DNase (Thermo Fisher Scientific), rSAP (NEB) and T4 PNK (NEB) according to manufactures instructions. Adapters were ligated onto the 3′ and 5′ ends of the small RNA, followed by reverse transcription with Superscript III. The library was amplified with limited cycles of PCR, gel-extracted on an 8% native PAGE gel and sequenced on an Illumina MiSeq. Adapters were trimmed using Cutadapt and sequences >10nt were mapped to the G. st. CRISPR locus using Bowtie 2(Langmead & Salzberg, 2012).

3.3.6 HEK293T EGFP disruption assay and indel analysis

HEK293T cells expressing a destabilized GFP were grown in Dulbecco's Modification of Eagle's Medium (DMEM) with 4.5g L^{-1} glucose L-glutamine and sodium pyruvate (Corning Cellgro), supplemented with 10% fetal bovine serum, penicillin and streptomycin at 37°C with 5% CO_2. ~24hrs before transfection, ~3 × 10^4 cells were seeded into each well of a 96-well plate. The next day, 20pmol (unless otherwise specified) of RNP was assembled as previously described(Lin et al., 2014) and mixed with 10µL OMEM. The RNP was added to 10µl of 1:10 dilution of Lipofectamine 2000 (Life Technologies) in OMEM and allowed to incubate at room temperature for 10min and added to the cells. Cells were analyzed for GFP fluorescence 48h later using Guava EasyCyte 6HT. Experiments were conducted in triplicate and the mean ±S.D. is shown. For analysis of indels, genomic DNA was extracted using Quick Extraction Solution (Epicentre), and the DNMT1 and AAVS1 loci were amplified by PCR. T7E1 reaction was conducted according to the manufacturer's instructions and products were analyzed on a 1.5% agarose gel stained with SYBR gold (Thermo Fisher Scientific).

3.4 Results

3.4.1 Identification of thermostable Cas9 homologs

Although thousands of *cas9* homologs have been sequenced, there have been no functionally validated Cas9 from archaea(Burstein, Harrington, Strutt, & Probst, 2017), restricting our search for a thermophilic Cas9 to thermophilic bacteria. We searched all the isolates in Integral Microbial Genomes database (IMG) from a thermophilic environment that contained a Cas9-like protein(Markowitz et al., 2012) (hits to a TIGRfam model 01865 for Csn1-like or 03031 for a Csx12-like). From them, the Cas9 from *Geobacillus stearothermophilus* (*G. st.;* formerly *Bacillus stearothermophilus*)(Donk, 1920) stood out because it was full-length and its sequence is shorter than the average Cas9. Most importantly, this candidate is from the organism that can grow in a reported temperature range from 30–75 (optimal at 55C). A BLASTn of GeoCas9 revealed several nearly identical homologs (from 93.19–99.91% identity over the full length) in 6 other *Geobacillus* species and 92.55% Identity over the full length in *Effusibacillus pohliae* DSM22757. *G. st.* has been a proven source of enzymes for thermophilic molecular cloning applications(Notomi et al., 2000), thermostable proteases(Fujii, Takagi, Imanaka, & Aiba, 1983) and enzymes for metabolic engineering(Ingram et al., 2010). Moreover, the wide temperature range that *G. st.* occupies(Cordova, Long, Venkataramanan, & Antoniewicz, 2015) holds promise that the Cas9 from this species (GeoCas9) may be able to maintain activity at both mesophilic and thermophilic temperatures (Fig. 3.1a). Notably, GeoCas9 is considerably smaller than SpyCas9 (GeoCas9, 1,087 amino acids; SpyCas9, 1,368 amino acids). A homology model of GeoCas9 based on available Cas9 crystal structures along with sequence alignments revealed that the small size of GeoCas9 is largely the result of a reduced REC lobe, as is the case with other compact Cas9 homologs from *Streptococcus aureus* Cas9 (SauCas9) and *Actinomyces naeslundii* Cas9 (AnaCas9) (Fig. 3.1b; Supplementary Fig. 3.6).

We purified GeoCas9 and performed initial thermostability tests using differential scanning calorimetry (DSC), which showed that in the absence of RNA or DNA, GeoCas9 has a denaturation temperature about 20°C higher than SpyCas9 (Fig. 3.1d).

Moreover, GeoCas9 denatures at 15°C higher than the slightly thermophilic *Streptococcus thermophilus* CRISPR III Cas9 (SthCas9) (Fig. 3.1e). Given these results, we selected GeoCas9 as a candidate for further development and optimization.

3.4.2 GeoCas9 PAM identification and engineering

CRISPR systems have evolved a preference for a protospacer adjacent motif (PAM) to avoid self-targeting of the host genome(Bolotin, Quinquis, Sorokin, & Dusko Ehrlich, 2005; Mojica, Diez-Villasenor, et al., 2009). These PAM sequences are divergent among Cas9 homologs and DNA targets are often mutated in this region to escape cleavage by Cas9(Paez-Espino et al., 2015). To identify the PAM for GeoCas9, we first searched for naturally targeted viral and plasmid sequences using CRISPRtarget(Biswas, Gagnon, Brouns, Fineran, & Brown, 2013). The three sequenced strains of *G. st*. provided 77 spacer sequences, and 3 of them had high-confidence viral and plasmid targets (Supplementary Fig. 3.7). Extracting the sequences 3′ of the targeted sequence revealed a consensus of 5′-NNNNCNAA-3′ (Fig. 3.2a, ED Fig. 3.2). Given the low number of viral targets, we next performed cleavage assays on substrates containing various PAM sequences, revealing a complete PAM sequence of 5′-NNNNCRAA-3′ (Fig. 3.2b).

In addition to the CRISPR loci found in *G. st.* strains, we also found a type II CRISPR locus in *Geobacillus* LC300 containing a Cas9 with ~97% amino acid identity to the *G. st.* Cas9. Despite having nearly identical sequences, alignment of these two homologs of GeoCas9 revealed a tight cluster of mutations in the PAM interacting domain (PI) (Fig. 3.2d). Furthermore, mapping these mutations onto the homology model of GeoCas9 showed that they are located near the PAM region of the target DNA (Fig. 3.2c). We hypothesized that this GeoCas9 variant might have evolved altered PAM specificity. By searching for viral targets using the spacers in the *G.* LC300 array, we identified a preference for GMAA in place of the CRAA PAM of *G. st.*, lending support to our hypothesis. We constructed and purified a hybrid Cas9 protein in which the PI domain of the *G.* LC300 Cas9 was substituted for the PI domain of *G. st.* Cas9 and tested cleavage activity on targets containing various PAM sequences (Fig. 3.2b). We found that, as predicted by protospacer sequences, the hybrid Cas9 preferred a GMAA PAM rather than the CRAA PAM utilized by GeoCas9. Moreover, *G.* LC300 appears to be more specific for its optimal PAM, which may result in lower off-target cleavage for genome editing applications(Kleinstiver, Prew, Tsai, Topkar, et al., 2015). By creating a hybrid Cas9 with this naturally occurring PAM-recognition variant, we double the sequence space that can be targeted by GeoCas9 without resorting to structure based protein engineering as has been done for other Cas9 homologs(Kleinstiver, Prew, Tsai, Topkar, et al., 2015).

3.4.3 Identification of tracrRNA and engineering of GeoCas9 sgRNA

CRISPR-Cas9 systems use a trans-activating crRNA (tracrRNA), which is required for maturation of the crRNA and activation of Cas9(Elitza Deltcheva et al., 2011; Martin Jinek et al., 2012b). To identify the tracrRNA for GeoCas9, we cultured *G. st.* and deep sequenced the small RNA it produced. We found that the CRISPR array was transcribed despite a lack of phage or plasmid challenge, and that the array was transcribed in the opposite direction of the Cas proteins (Fig. 3.3a). The crRNA was processed to 23nt (Fig. 3.3b) of the spacer sequence and 18nt of the repeat sequence *in vivo*, similar to other small type IIC Cas9 proteins(Fonfara et al., 2014; Hou et al., 2013). Mapping of the RNAseq reads to the CRISPR array also revealed a putative tracrRNA upstream of the Cas9 open reading frame (ORF).

We joined this putative tracrRNA to the processed crRNA using a GAAA-tetraloop to generate a single-guide RNA (sgRNA)(Savell & Day, 2017). Variations of this sgRNA were *in vitro* transcribed and tested for their ability to direct GeoCas9 to cleave a radiolabeled double-stranded DNA target at 37°C. We first varied the length of the crRNA:tracrRNA duplex and found that this modification had little impact on the DNA cleavage rate (left panel, Fig. 3. 3c), making it a valuable place for further sgRNA modifications(Konermann et al., 2014; Mali et al., 2013). Next, we tested the length of the tracrRNA, which here refers to the region after the tetraloop, choosing stopping points near predicted rho-independent terminators. In contrast to the crRNA:tracrRNA duplex length, the length of the tracrRNA had a dramatic effect on the cleavage rate, with sequences shorter than 91nt supporting only a small amount of cleavage (middle panel, Fig. 3.3c). Finally, we varied the length of the spacer sequence and found that 21–22nt resulted in a more than 5-fold increase in cleavage rate, compared to the 20nt spacer preferred by SpyCas9 (right panel, Fig. 3.3c). This finding contrasts with the most abundant spacer length of 23nt found in RNAseq. This difference may be due to inter- or intramolecular guide interactions in the *in vitro* transcribed sgRNA(Thyme, Akhmetova, Montague, Valen, & Schier, 2016). To test this we used an additional guide sequence with no predicted structure in the spacer region (Target 2). In contrast to Target 1, Target 2 had similar cleavage rates for 21, 22 and 23nt (right panel, Fig. 3.3c). In addition, testing cleavage of off-target substrates revealed that GeoCas9 has higher sensitivity of mismatched sequences proximal to the PAM then distal, similar to previously described Cas9 proteins (Supplementary Fig. 3.8A).

3.4.4 Genome editing by GeoCas9 RNP in mammalian cells

With evidence that GeoCas9 maintains cleavage activity at mesophilic temperatures, we assessed the ability of GeoCas9 to edit mammalian genomes. We tested GeoCas9 and SpyCas9 editing efficiency by delivering preassembled ribonucleoprotein complexes (RNPs) into cultured cells, circumventing differences between SpyCas9 and GeoCas9 protein expression. First, GeoCas9 RNPs targeting regions adjacent to various PAM sequences were delivered into HEK293T cells expressing a destabilized GFP (Fig. 3.4a). We found that when targeted to sequences adjacent to the preferred CRAA PAM, GeoCas9 decreased GFP fluorescence at levels comparable to those observed for SpyCas9 (Fig. 3.4a). Next, we targeted GeoCas9 to cleave the native genomic loci DNMT1 and AAVS1 (Fig. 3.4b,c). We varied the length of the targeting spacer sequence and found that at one site 21nt was a sufficient length to efficiently induce indels while at another site a 22nt spacer length was necessary. Given this variability and that extending the spacer length to 22nt had no detrimental effects, we conclude that a 22nt guide segment length is preferred for use in genome editing applications. Moreover, when we tested editing efficiency at a site containing an overlapping PAM for both GeoCas9 and SpyCas9, we observed similar editing efficiencies by both proteins (Fig. 3.4b). At the DNMT1 locus we titrated amounts of GeoCas9 and SpyCas9 RNPs to assess the effect on genome editing efficiency (Fig.

3.4c). Products analyzed by T7E1 assay again showed efficient production of indels by both GeoCas9 and SpyCas9. These results demonstrate that GeoCas9 is an effective alternative to SpyCas9 for genome editing in mammalian cells.

3.4.5 Stability of GeoCas9

Based on initial observations showing that GeoCas9 protein remains folded at elevated temperatures (Fig. 3.1d, e), we tested whether the GeoCas9 RNP maintains activity after exposure to high temperatures. We incubated SpyCas9 and GeoCas9 at a challenge temperature and added equimolar substrate to test the fraction of RNP that remained functional. After incubation for 10 min at 45°C, the fraction of active SpyCas9 was greatly reduced (Fig. 3.5a). In contrast, the fraction of GeoCas9 after incubation at 45°C remained at 100% and not until challenge at 70°C did we detect a decrease in activity (Fig. 3.5a).

Often thermostability comes at the cost of reduced activity at lower temperatures(Sawle & Ghosh, 2011). However, the wide range of natural growth temperatures for *G. st.* suggested that GeoCas9 might maintain activity at mesophilic temperatures. To examine this hypothesis, we measured the cleavage rate of SpyCas9 and GeoCas9 at various temperatures (Fig. 3.5b). SpyCas9 DNA cleavage rates increased between 20–35°C, reaching maximum levels from 35–45°C. Above these temperatures, SpyCas9 activity dropped sharply to undetectable levels, as predicted by thermostability measurements. In contrast, GeoCas9 activity increased to the maximum detection limit at 50°C and maintained maximum detectable activity up to 70°C, dropping to low levels at 75°C. These results make GeoCas9 a valuable candidate for editing obligate thermophilic organisms and for biochemical cleavage applications requiring Cas9 to operate at elevated temperatures.

The lifetime of proteins in blood is often limiting for their use as a therapeutic and many strategies have been employed to improve this, including fusion of other proteins and polymers to the protein of interest(Kontermann, 2011; Schellenberger et al., 2009). We investigated the lifetime of SpyCas9 RNP in mouse plasma and found that it had a half-life of ~1.5hrs (Fig. 3.5c). Further investigation revealed that a combination of RNA and protein degradation likely plays a role in this inactivation (Supplementary Fig. 3.8b, c). Although what determines the lifetime of a protein in blood is complex, it was shown previously that thermostabilization of a protein can increase its lifetime in blood(Narasimhan et al., 2010). To test if this is the case for GeoCas9, we incubated SpyCas9 and GeoCas9 in diluted human plasma at 37°C for 8 hrs and measured the amount of Cas9 activity remaining (Fig. 3.5d). Although SpyCas9 maintained activity when incubated in reaction buffer at 37°C, its activity was abolished even at the lowest concentration of plasma. In contrast, GeoCas9 maintained significant activity after incubation with human plasma, making it a promising candidate for *in vivo* RNP delivery.

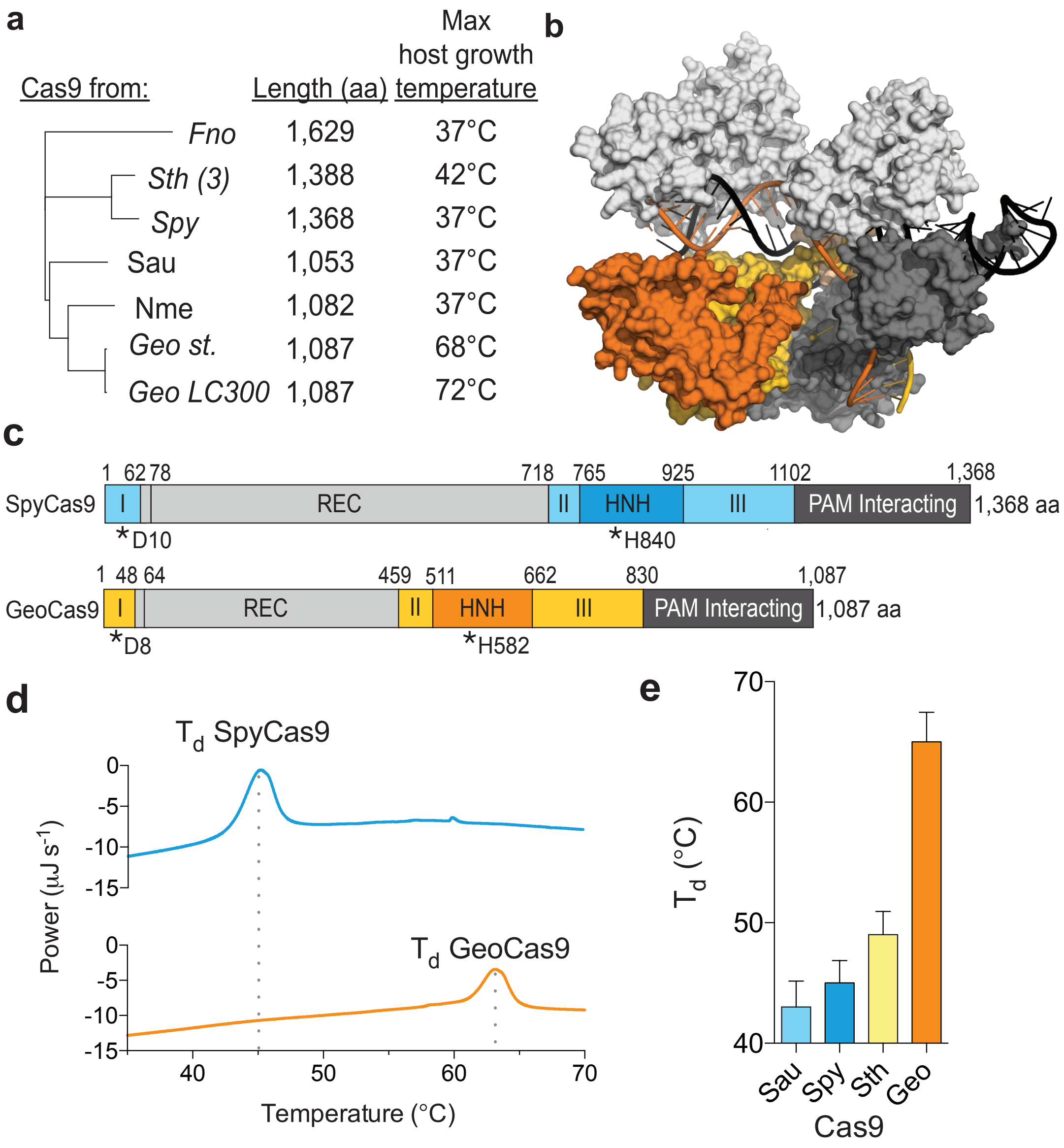


Figure 3.1 | GeoCas9 is a small, thermostable Cas9 homolog. **a,** Phylogeny of Cas9 proteins used for genome editing with their length (amino acids) and the maximum temperatures that supports growth of the host indicated to the right (Cordova et al., 2015) (Nme*, Neisseria meningitidis*; Geo, *Geobacillus stearothermophilus*; Geo LC300, *Geobacillus* LC300; Spy, *Streptococcus pyogenes;* Sau, *Streptococcus aureus*; Fno, *Francisella novicida*; Sth (3), *Streptococcus thermophilus* CRISPR III). **b**, Homology model of GeoCas9 generated using Phyre 2(Kelly, Mezulis, Yates, Wass, & Sternberg, 2015) with the DNA from PDB 5CZZ docked in. **c**, Schematic illustration of the domains of Spy Cas9 (blue) and GeoCas9 (orange) with active site residues indicated below with asterisks. **d**, Representative traces for Differential Scanning Calorimetry (DSC) of GeoCas9 and SpyCas9, T_d; Denaturation temperature. **e,** Denaturation temperature of various Cas9 proteins as measured by DSC, mean ± S.D. is shown.

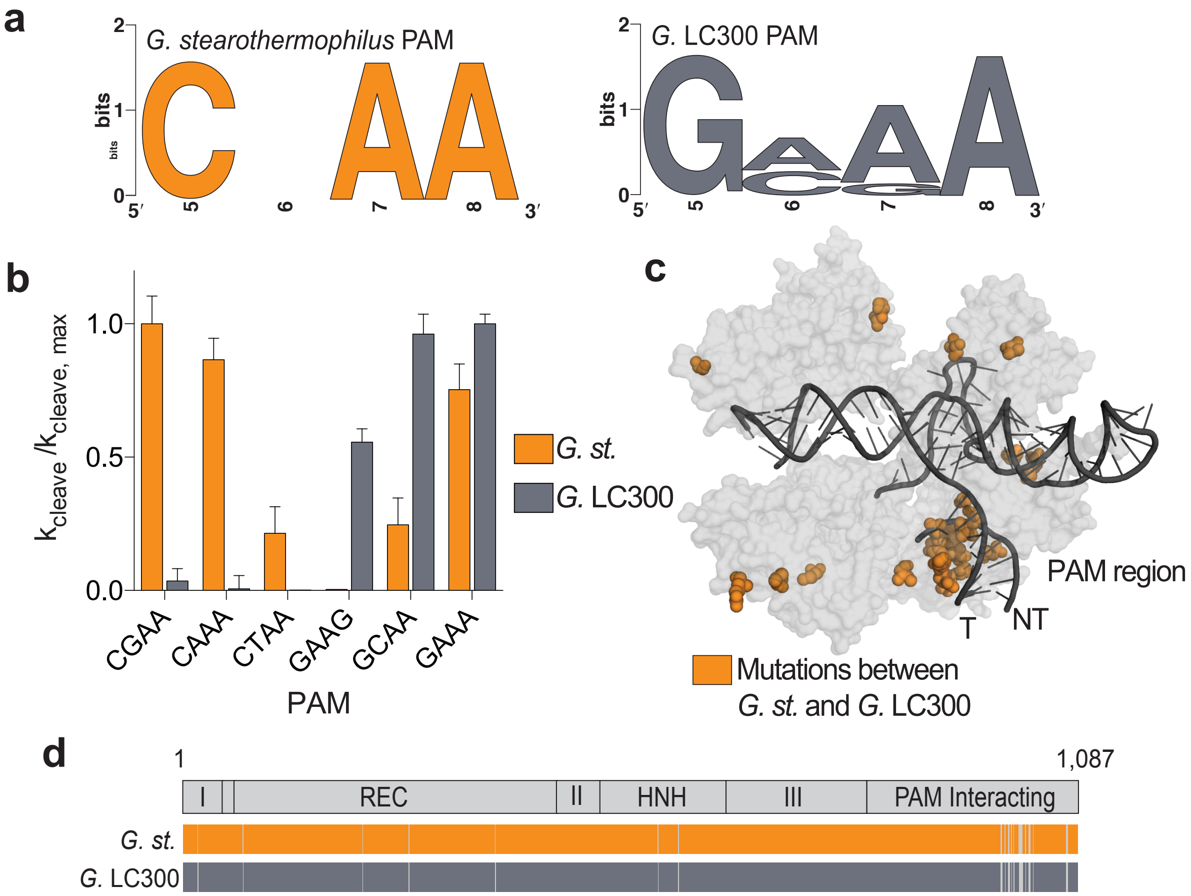


Figure 3.2 | PAM identification and engineering of GeoCas9. a, WebLogo for sequences found at the 3′ end of protospacer targets identified with CRISPRTarget for *Geobacillus stearothermophilus* (left panel) and *Geobacillus* LC300 (right panel). **b,** Cleavage assays conducted with the two homologs of GeoCas9. Substrates with various PAM sequences were P32-labelled and mean ± S.D. is shown. **c,** Mapping of mutated residues (orange spheres) between *G. st.* and *G.* LC300 onto the homology model of GeoCas9 showing high density in the PAM interacting domain near the PAM region of the target DNA. **c,** Alignment of the Cas9 proteins from *G. st.* and *G.* LC300 with the domain boundaries shown above. Solid colors represent identical residues and grey lines indicate residues that are mutated between the two Cas9 homologs.

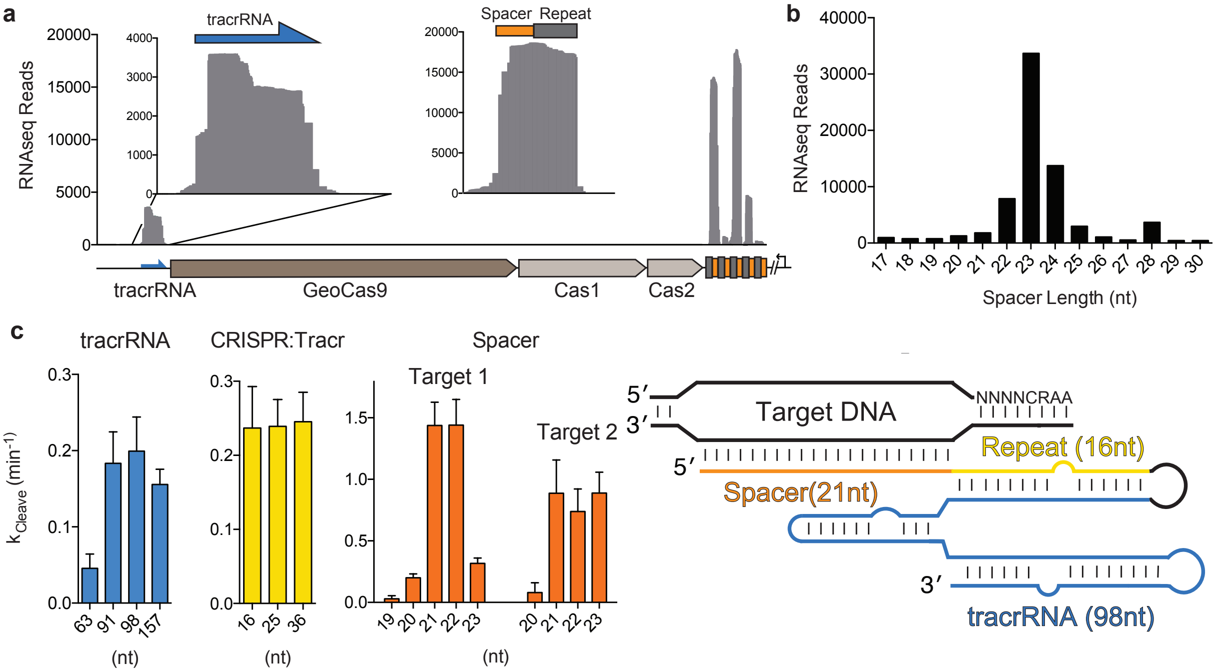


Figure 3.3 | Small RNA-seq and sgRNA engineering for GeoCas9. a, Small RNA sequenced from *G. stearothermophilus* mapped to the CRISPR locus. Inset shows enlargement of the region corresponding to the tracrRNA and the most highly transcribed repeat and spacer sequence. **b,** Distribution of the length of the spacer sequences extracted from the small RNA sequencing results. **c,** Length optimization of the tracrRNA and crRNA for GeoCas9 and the optimal guide RNA design (right panel). The length of the tracrRNA, crRNA:tracrRNA duplex and spacer was optimized sequentially by transcribing variations of the sgRNA and testing their ability to guide GeoCas9-mediated cleavage of a radiolabeled substrate. For spacer length, two different targets were used (Target 1 and Target 2). The mean k_{cleave} ± S.D. is shown and experiments were conducted in triplicate.

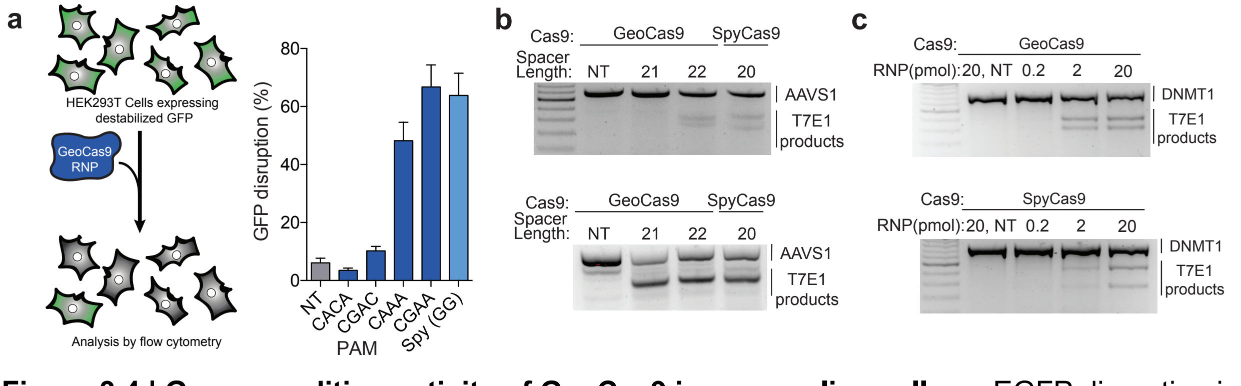


Figure 3.4 | Genome editing activity of GeoCas9 in mammalian cells. a, EGFP disruption in HEK293T cells by GeoCas9. HEK293FT cells expressing a destabilized GFP were transfected with GeoCas9 RNP preassembled with a targeting or non-targeting guide RNA. Cells were analyzed by flow cytometry and targets adjacent to the CRAA PAM resulted in efficient GFP disruption (NT, non-targeting; kb, kilobases). **b**, T7E1 analysis of indels produced at the AAVS1 locus when the guide length was varied from 21 to 22nt. The Cas9 used is indicated above each lane and the length of the spacer is shown below (NT; non-targeting). **c**, T7E1 analysis of indels produced using a titration

of GeoCas9 and SpyCas9 RNP targeting the DNMT1 locus in HEK293T cells. The Cas9 used is indicated above each lane and the amount of RNP delivered to each well of a 96-well plate is indicated below (NT, non-targeting; kb, kilobases).

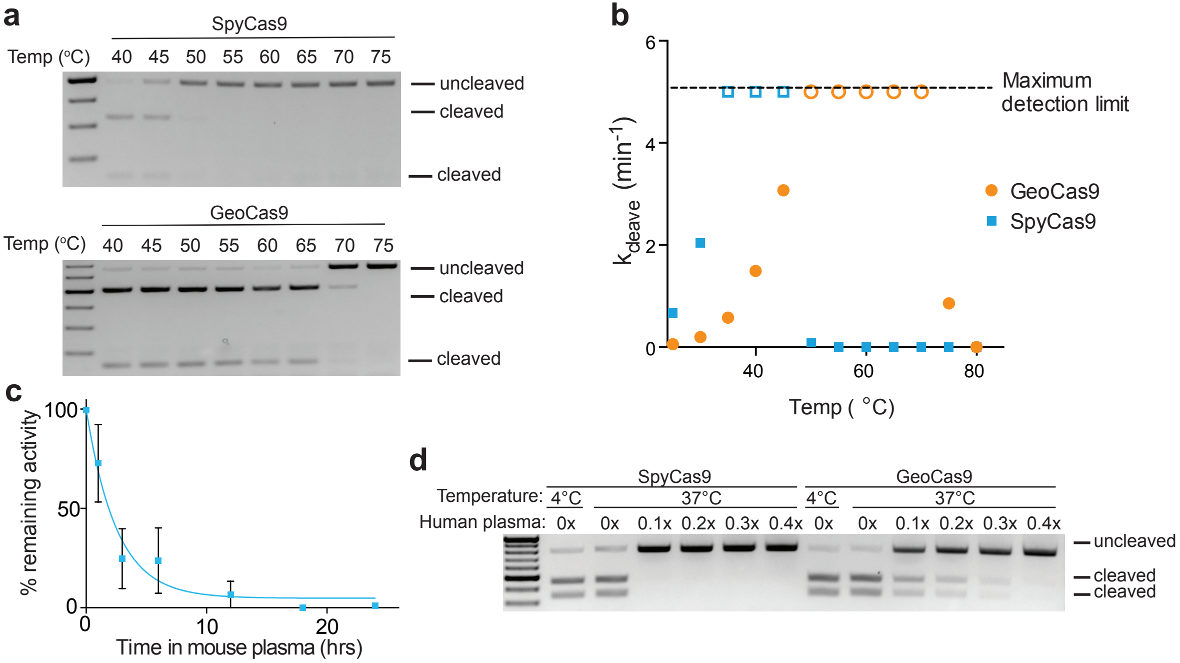


Figure 3.5 | Thermostability of GeoCas9 and longevity in human plasma. a, Activity of SpyCas9 and GeoCas9 after incubation at the indicated temperature. After challenging at the higher temperature, reactions were conducted at 37°C using a 1:1 ratio of substrate to RNP (kb, kilobases). **b,** Cleavage rate of SpyCas9 and GeoCas9 RNPs at various temperatures. Maximum detection limit is shown by the dashed line at k_{cleave}=5, indicating that the reaction completed in ≤30s. **c,** Lifetime of SpyCas9 after incubation in 50% mouse plasma. Incubation was done at 37°C for the specified amount of time after which DNA substrate was introduced at an equimolar ratio of substrate to RNP **d,** Effect of incubating GeoCas9 and SpyCas9 in human plasma. After incubation in varying concentrations of human plasma for 8hrs at 37°C, the reaction was carried out with 1:1 ratio of DNA substrate to RNP (kb, kilobases).

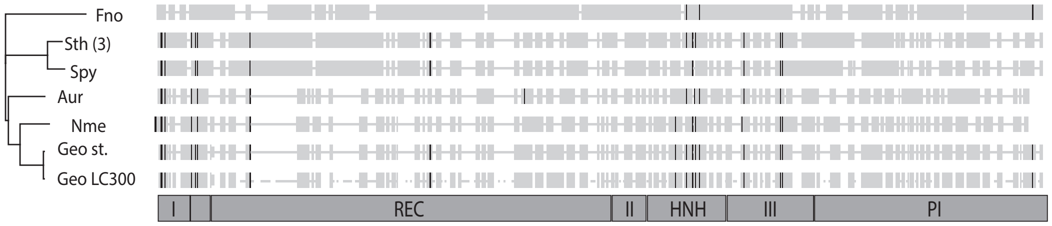
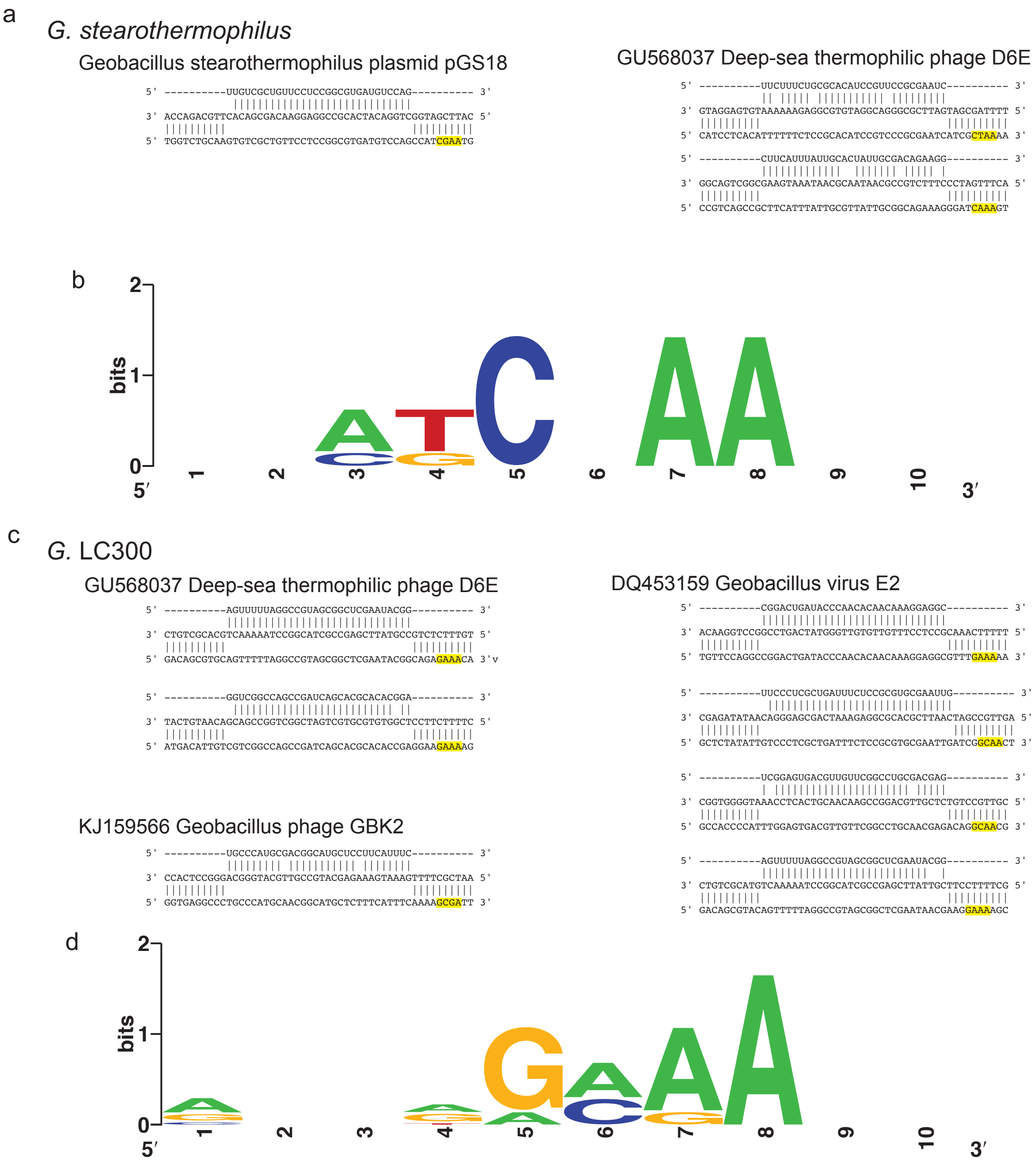


Figure 3.6 | Alignment of selected Cas9 homologs. Alignment was generated using MAFFT and the approximate domain boundies are shown below. Black lines indicate conserved residues.



a

G. stearothermophilus

Geobacillus stearothermophilus plasmid pGS18

```
5' ----------UUGUCGCUGUUCCUCCGGCGUGAUGUCCAG---------- 3'
             ||||||||||||||||||||||||||||||
3' ACCAGACGTTCACAGCGACAAGGAGGCCGCACTACAGGTCGGTAGCTTAC 5'
   ||||||||||                              ||||||||||
5' TGGTCTGCAAGTGTCGCTGTTCCTCCGGCGTGATGTCCAGCCATCGAATG
```

GU568037 Deep-sea thermophilic phage D6E

```
5' ----------UUCUUUCUGCGCACAUCCGUUCCGCGAAUC---------- 3'
             || ||||| ||||||||||| |||||||||
3' GTAGGAGTGTAAAAAAGAGGCGTGTAGGCAGGGCGCTTAGTAGCGATTTT 5'
   ||||||||||                              ||||||||||
5' CATCCTCACATTTTTTCTCCGCACATCCGTCCCGCGAATCATCGCTAAAA 3'

5' ----------CUUCAUUUAUUGCACUAUUGCGACAGAAGG---------- 3'
             |||||||||||||  ||||||| ||||| |
3' GGCAGTCGGCGAAGTAAATAACGCAATAACGCCGTCTTTCCCTAGTTTCA 5'
   ||||||||||                              ||||||||||
5' CCGTCAGCCGCTTCATTTATTGCGTTATTGCGGCAGAAAGGGATCAAAGT
```

b

c

G. LC300

GU568037 Deep-sea thermophilic phage D6E

```
5' ----------AGUUUUUAGGCCGUAGCGGCUCGAAUACGG---------- 3'
             ||||||||||||||||||||||||||||||
3' CTGTCGCACGTCAAAAATCCGGCATCGCCGAGCTTATGCCGTCTCTTTGT 5'
   ||||||||||                              ||||||||||
5' GACAGCGTGCAGTTTTTAGGCCGTAGCGGCTCGAATACGGCAGAGAAACA 3'v

5' ----------GGUCGGCCAGCCGAUCAGCACGCACACGGA---------- 3'
              |||||||||||||||||||||||||| ||
3' TACTGTAACAGCAGCCGGTCGGCTAGTCGTGCGTGTGGCTCCTTCTTTTC 5'
   ||||||||||                              ||||||||||
5' ATGACATTGTCGTCGGCCAGCCGATCAGCACGCACACCGAGGAAGAAAAG
```

KJ159566 Geobacillus phage GBK2

```
5' ----------UGCCCAUGCGACGGCAUGCUCCUUCAUUUC---------- 3'
             ||||||||| ||||||||||| ||||||||
3' CCACTCCGGGACGGGTACGTTGCCGTACGAGAAAGTAAAGTTTTCGCTAA 5'
   ||||||||||                              ||||||||||
5' GGTGAGGCCCTGCCCATGCAACGGCATGCTCTTTCATTTCAAAAGCGATT 3'
```

DQ453159 Geobacillus virus E2

```
5' ----------CGGACUGAUACCCAACACAACAAAGGAGGC---------- 3'
             ||||||||||||||||||||||||||||||
3' ACAAGGTCCGGCCTGACTATGGGTTGTGTTGTTTCCTCCGCAAACTTTTT 5'
   ||||||||||                              ||||||||||
5' TGTTCCAGGCCGGACTGATACCCAACACAACAAAGGAGGCGTTTGAAAAA 3'

5' ----------UUCCCUCGCUGAUUUCUCCGCGUGCGAAUUG---------- 3'
              ||||||||||||||||||||||||||||||
3' CGAGATATAACAGGGAGCGACTAAAGAGGCGCACGCTTAACTAGCCGTTGA 5'
   ||||||||||                               ||||||||||
5' GCTCTATATTGTCCCTCGCTGATTTCTCCGCGTGCGAATTGATCGGCAACT 3'

5' ----------UCGGAGUGACGUUGUUCGGCCUGCGACGAG---------- 3'
             | |||||||||||||||||||||| |||||
3' CGGTGGGGTAAACCTCACTGCAACAAGCCGGACGTTGCTCTGTCCGTTGC 5'
   ||||||||||                              ||||||||||
5' GCCACCCCATTTGGAGTGACGTTGTTCGGCCTGCAACGAGACAGGCAACG 3'

5' ----------AGUUUUUAGGCCGUAGCGGCUCGAAUACGG---------- 3'
             |||||||||||||||||||||||||||  |
3' CTGTCGCATGTCAAAAATCCGGCATCGCCGAGCTTATTGCTTCCTTTTCG 5'
   ||||||||||                              ||||||||||
5' GACAGCGTACAGTTTTTAGGCCGTAGCGGCTCGAATAACGAAGGAAAAGC
```

d

Figure 3.7 | Targets used to generate logos in figure 2b and 2c. a,c. Phage and plasmid targets matching the *G. stearothermophilus* and *G.* LC300 spacer sequences. PAM region is highlighted in yellow. **b,d,** Logo of the sequences 3′ of the protospacer targets identified in **a** and **c.**

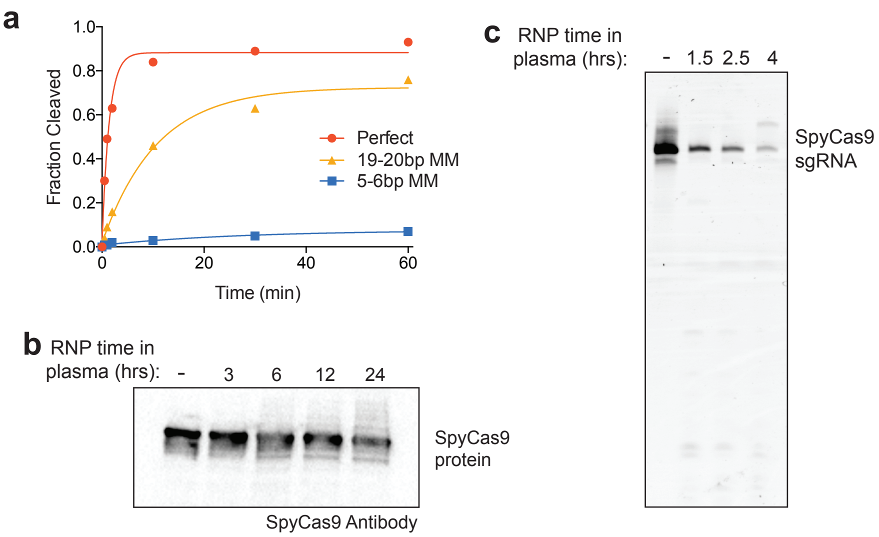


Figure 3.8 | Mismatch cleavage by GeoCas9 and degradation of SpyCas9 in mammalian plasma, GeoCas9 cleavage of targets containing mismatches in the PAM proximal (5-6bp MM) or PAM distal (19-20bp MM) regions of the target DNA showing higher specificity for the PAM proximal region. **b,** Western blot against SpyCas9 after incubated the RNP in human plasma for the specified period of time. **c,** SpyCas9 sgRNA after incubation in human plasma for the specified period of time.

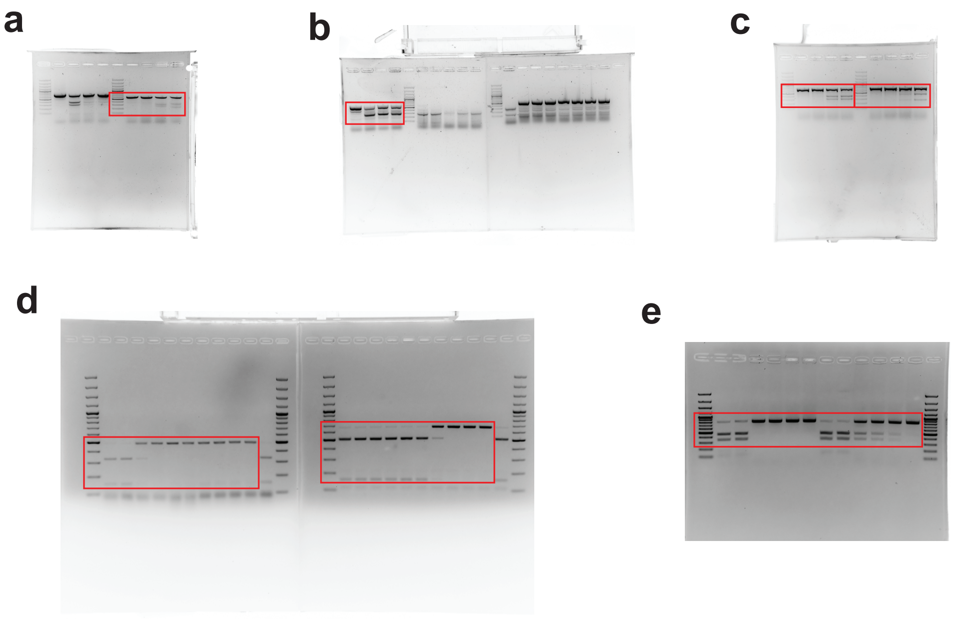


Figure 3.9 | Uncropped images of gels for Figure 4 and 5
a, Uncropped gel shown in the top panel of Figure 4b. **b,** Uncropped gel shown in the bottom panel of Figure 4b. **c,** Uncropped gel shown in Figure 4c. **d,** Uncropped gel shown in the top

panel of Figure 5a. **e,** Uncropped gel shown in Figure 5d. Red outline marks the region shown in the main text figures.

3.5 Discussion

Our results establish GeoCas9 as a thermostable Cas9 homolog and expand the temperatures at which Cas9 can be used. We anticipate that the development of GeoCas9 will enhance the utility of CRISPR-Cas9 technology at both mesophilic and thermophilic temperatures. The ability of Cas9 to function reliably in a wide range of species has been key to its rapid adoption as a technology, but the previously developed Cas9 homologs are limited for use in organisms that can grow below 42°C. The complementary temperature range of GeoCas9 with SpyCas9 (Fig. 3.5b) opens up Cas9 based genome editing to obligate thermophiles and facultative thermophiles, without the additional steps of altering the temperature of the organisms. We also anticipate that GeoCas9 will be useful for *in vitro* molecular biology applications requiring targeted cleavage at elevated temperatures. Furthermore, we predict that the extended lifetime of GeoCas9 in human plasma may enable more efficient delivery of Cas9 RNPs. GeoCas9 also has an extended PAM and spacer sequence compared to SpyCas9. These features are similar to SauCas9 and NmeCas9, both of which have been shown to be naturally higher fidelity than SpyCas9(Amrani et al., 2017; Friedland et al., 2015; Kleinstiver, Prew, Tsai, Nguyen, et al., 2015; Ran et al., 2015b) and it will be interesting for future work to investigate if this trend also holds true for GeoCas9.

We were interested to note that GeoCas9 and SpyCas9 induced similar levels of indels in HEK293T cells as SpyCas9 when delivered as an RNPs (Fig. 3.4b-d), despite GeoCas9's lower DNA cleavage rate at 37°C (Fig. 3.5b). We conclude that biochemical cleavage rates may not reflect the limiting step of target search in a human cell. It may be that GeoCas9 can persist longer in cells, which raises its effective concentration over time and compensates for its slower cleavage rate. Moreover, in applications requiring delivery of Cas9 into the bloodstream, the benefit of improved stability by GeoCas9 may become even more apparent. However, it remains to be seen if the efficient GeoCas9 mediated genome editing observed here will hold true in more challenging genome editing applications.

The development of GeoCas9 hinged upon utilizing the naturally occurring diversity of CRISPR systems. The sheer abundance and diversity of Cas9 makes it advantageous over newer type V systems, such as Cpf1, for developing specialized genome editing tools. It has previously been suggested that type II CRISPR systems are only found in mesophilic bacteria, and that protein engineering would be required to develop a thermophilic Cas9(Y. Li et al., 2015). The rarity of type II CRISPR systems in thermophiles is surprising given that CRISPR systems in general are enriched in thermophilic bacteria and archaea(Weinberger, Wolf, Lobkovsky, Gilmore, & Koonin, 2012). However, by searching the continually growing number of sequenced bacteria, we uncovered a naturally occurring thermophilic Cas9. Exploiting Cas9 sequence diversity, rather than engineering thermostability, revealed a protein that maintains activity over a broad temperature range, which is often difficult to select for using directed evolution. Using the natural context of this CRISPR system, including the transcribed RNA and targeted sequences, we further developed GeoCas9 with minimal experimental optimization. The strategy of mining the natural context and diversity of

CRISPR systems has proven successful for uncovering novel interference proteins(Burstein, Harrington, Strutt, & Probst, 2017; Shmakov et al., 2017b), and we anticipate that it can be applied more broadly to discover and develop new genome editing tools.

3.6 Acknowledgements

We thank A. Tambe, N. Ma and K. Zhou for technical assistance and discussions. L.B.H and J.S.C are supported by US National Science Foundation Graduate Research Fellowships. J.A.D. is an Investigator of the Howard Hughes Medical Institute. This research was supported in part by the Allen Distinguished Investigator Program, through The Paul G. Allen Frontiers Group and the National Science Foundation (MCB-1244557 to J.A.D.).

J.A.D. is an Investigator of the Howard Hughes Medical Institute. This research was supported in part by the Allen Distinguished Investigator Program, through The Paul G. Allen Frontiers Group and the National Science Foundation (MCB-1244557 to J.A.D.) J.A.D. is executive director of the Innovative Genomics Institute at the University of California, Berkeley (UC Berkeley) and the University of California, San Francisco (UCSF). J.A.D. is a co-founder of Editas Medicine, Intellia Therapeutics and Caribou Biosciences and a scientific adviser to Caribou, Intellia, eFFECTOR Therapeutics and Driver. UC Berkeley and HHMI have patents pending for CRISPR technologies on which the authors are inventors.

3.7 Author Contributions

L.B.H. and B.S. designed and executed experiments with help from J.S.C., E.M., and J.A.D.; The search for thermophilic Cas9 homologs was conducted by D.P.E. and N. C. K. All authors revised and agreed to the manuscript.

Chapter 4

Broad spectrum inhibitor of CRISPR-Cas9

Part of the work presented in this chapter has previously been published in the following manuscript: Harrington, L. B., Doxzen, K. W*., Ma, E., Liu, J. J., Knott, G. J., Edraki, A., … Doudna, J. A. (2017). A Broad-Spectrum Inhibitor of CRISPR-Cas9. Cell, 170(6), 1224–1233.

4.1 Abstract

CRISPR-Cas9 is a powerful technology that has enabled genome editing in a wide range of species. However, the currently developed Cas9 homologs all originate from mesophilic bacteria, making them susceptible to degradation and unsuitable for applications requiring cleavage at elevated temperatures. Here, we show that the Cas9 protein from the thermophilic bacterium *Geobacillus stearothermophilus* (GeoCas9) catalyzes RNA-guided DNA cleavage at elevated temperatures. GeoCas9 is active at temperatures up to 70°C, compared to 45°C for *Streptococcus pyogenes* Cas9 (SpyCas9), which expands the temperature range for CRISPR-Cas9 applications. We also found that GeoCas9 is an effective tool for editing mammalian genomes when delivered as a ribonucleoprotein (RNP) complex. Together with an increased lifetime in human plasma, the thermostable GeoCas9 provides the foundation for improved RNP delivery *in vivo* and expands the temperature range of CRISPR-Cas9.

4.2 Introduction

CRISPR systems provide bacteria and archaea with adaptive immunity against foreign DNA and RNA (R Barrangou et al., 2007; Brouns et al., 2008; Marraffini & Sontheimer, 2008). To initiate immunity, CRISPR-associated (Cas) proteins integrate fragments of invading DNA into the host genome at the CRISPR locus, where they serve as transcription templates for the synthesis of RNA that directs Cas nucleases to cleave infectious nucleic acids (Garneau et al., 2010; Hale et al., 2009). Class 2 CRISPR systems are streamlined versions that require only a single protein to target foreign DNA or RNA (K S Makarova et al., 2015; Shmakov et al., 2017b). CRISPR-Cas9, the most abundant and diverse of Class 2 CRISPR proteins, exists in three subtypes of which type IIA and IIC are more common compared to the relatively rare type IIB (Chylinski et al., 2014b; Shmakov et al., 2017b). The programmable nature of Cas9 has made it a powerful tool for gene editing and genomic modulation in a wide range of organisms.

In response to these robust prokaryotic immune systems, phages have evolved proteins that bind to and inactivate Cas proteins as they search for foreign nucleic acids (Joe Bondy-Denomy, Pawluk, Maxwell, & Davidson, 2013; Joseph Bondy-Denomy et al., 2015; Pawluk, Bondy-denomy, Cheung, Maxwell, & Davidson, 2014). Although only a small number of these anti-CRISPRs (Acrs) have been discovered to date, phylogenetic analysis suggests that Acrs are widespread and likely play a significant role in the evolution of Cas proteins (Houte et al., 2016; Pawluk, Staals, et al., 2016). In addition to their native functions, Acrs that inhibit Cas9 nucleases allow for control of Cas9 in genome editing applications (Pawluk, Amrani, et al., 2016; Rauch et al., 2017). Specifically, three unique Acrs that target the type IIC Cas9 from *Neisseria meningitidis* (NmeCas9) have been identified (AcrIIC1, 2 and 3) along with four that target select type IIA Cas9 orthologs (AcrIIA1, 2, 3 and 4). While some of these Acrs have been shown to inhibit NmeCas9 and SpyCas9 in mammalian cells (Pawluk, Amrani, et al., 2016; Rauch et al., 2017), their ability to inactivate other Cas9 orthologs used for genome editing remains unknown. Understanding this specificity as well as the mechanisms by which they disable Cas9 will be critical for their successful deployment as modulators of Cas9 in human and other cell types. Apart from applications, this mechanistic information is also fundamental to understanding how these Acrs have evolved to target distinct Cas9 orthologs and what evolutionary pressures they impose on CRISPR systems.

Here we investigated the inactivation of Cas9 by AcrIIC1 and AcrIIC3, uncovering unique mechanisms for both. We focused on these two Acrs because of their potent inhibition of NmeCas9 in human cells (Pawluk, Amrani, et al., 2016). Our data show that AcrIIC1 blocks DNA cleavage by multiple Cas9 orthologs without impacting DNA binding, effectively transforming catalytically active Cas9 into catalytically inactive dCas9. This mechanism is accomplished by AcrIIC1 binding directly to the HNH nuclease domain of Cas9, obscuring the active site and restricting conformational changes required for cleavage. AcrIIC3, by contrast, inhibits only a single Cas9 ortholog by blocking DNA binding. AcrIIC3 also causes Cas9 to dimerize, possibly contributing to its ability to interfere with target recognition and suggesting a mechanism distinct from that observed for AcrIIA4 (Dong et al., 2017; Shin et al., 2017). Together, AcrIIC1 and AcrIIC3 enable either broad-spectrum or selective inhibition of Cas9 orthologs respectively. The different mechanisms of these two Acrs allow separate control of binding to and cleavage of DNA by Cas9. Moreover, these mechanisms reveal vulnerabilities of Cas9 that are susceptible to inhibition, shedding light on the evolutionary arms race between bacteriophage and bacteria.

4.3 Materials and Methods

4.3.1 Phylogenetic analyses

Cas9 and Acr protein sequences were gathered from previous publications and additional Cas9 orthologs targeted by Acrs were added (Burstein, Harrington, Strutt, & Probst, 2017; Fonfara et al., 2014; Pawluk, Amrani, et al., 2016). A non-redundant set of proteins was compiled by clustering proteins with >90% identity. Proteins were aligned using MAFFT and maximum-likelihood phylogenies were constructed using RAxML (Stamatakis, 2014). Trees were visualized using FigTree 1.4.1.

4.3.2 Protein expression and purification

Proteins were purified as previously described (Martin Jinek et al., 2012a) with modification. For Cas9, *E. coli* BL21(DE3) was grown in Terrific Broth at 18°C for 14 h. Cells were harvested and resuspended in Lysis Buffer (50 mM Tris-HCl, pH 7.5, 20 mM imidazole, 0.5 mM TCEP-NaOH, 500 mM NaCl and 1 mM PMSF), disrupted by sonication, and purified on Ni-NTA resin. TEV protease was incubated with the elution overnight at 4°C. Next, the protein was run over an MBP affinity column onto a heparin column and eluted with a gradient from 300–1250 mM NaCl. The resulting fractions containing Cas9 were purified using a Superdex 200 10/300 gel filtration column and flash frozen in Storage Buffer (20 mM HEPES-NaOH, pH 7.5, 5% (v/v) glycerol, 150 mM NaCl and 1 mM TCEP-NaOH). The HNH domain and Acrs were purified using the same procedure, omitting the heparin and MBP column.

4.3.3 Cleavage assays

All reactions were carried out in 1× Reaction Buffer (20 mM Tris-HCl, pH 7.5, 100 mM KCl, 5 mM $MgCl_2$, 1 mM DTT and 5% glycerol (v/v)). 500 nM Cas9 was complexed with 625 nM sgRNA for 5 min at 37°C. The Acr was added at 2.5 µM and the reaction was incubated at 37°C for another 5 min. Next, DNA substrate was added to a final concentration of 25nM and the reaction was allowed to react for 15min (Fig 1A) or for 30 s, 1 min, 2 min, 5 min and 30 min (Figure 4.7B). Products were analyzed on 1.25% agarose, 0.5× TAE gels stained with ethidium bromide.

For P32-labeled DNA cleavage assays, 40 nM Cas9 was complexed with 40 nM sgRNA in 1× Reaction Buffer for 10 min at 37°C. The AcrIIC1 was added at 400 nM and the reaction was incubated at room temperature for another 20 min. Next probe was added to a final concentration of ~1nM and the reaction was allowed to react for 0, 1, 2.5, 5 and 10min (Figure 4.2B). Products were analyzed on 12% denature PAGE gel before being dried and visualized by phosphorimaging.

4.3.4 HEK293T transfection and indel analysis

Plasmids expressing NmeCas9, SpyCas9 and their respective sgRNAs targeting the DTS3 site, as well as plasmids expressing AcrE2 (Addgene #85677), AcrIIC1$_{Boe}$ (Addgene #85678) and AcrIIC1$_{Nme}$ (Addgene #85679), were previously described (Pawluk, Amrani, et al., 2016). AcrIIA4 expressing plasmid (Addgene #86842) was described previously (Rauch et al., 2017). The CjeCas9-expressing plasmid (PX404) was acquired from Addgene (#68338). For CjeCas9 sgRNA expression, the published sgRNA sequence (Kim et al., 2017) was synthesized as a gBlock (IDT), and was used to replace the NmeCas9 sgRNA cassette in pLKO.1-puro plasmid [(Pawluk, Amrani, et al., 2016); Addgene #85715] by Gibson Assembly. The resulting plasmid (pEJS676) contains the CjeCas9 sgRNA cassette with *Bfu*AI sites that can be used to insert any spacer of interest. Next, two previously validated guide sequences [targeting the *AAVS1* locus (TS2 and TS6; (Kim et al., 2017)] were inserted into the CjeCas9 sgRNA expression construct, yielding plasmids pEJS677 and pEJS678, respectively.

Plasmids were used to transfect HEK293T cells as previously described (Pawluk *et al.,* 2016). Briefly, 1.5 × 10^5 HEK293T cells [cultured at 37°C, 5% CO_2 in DMEM + 10% FBS + 1% Penicillin/Streptomycin (Gibco)] were transiently transfected with 100 ng Cas9-expressing plasmid, 100 ng sgRNA-expressing plasmid, and 100 ng Acr plasmid. 72 h after transfection, cells were harvested and genomic DNA was extracted with the DNeasy Blood and Tissue kit (Qiagen). 50 ng genomic DNA was used for PCR amplification [High Fidelity 2× PCR Master Mix (New England Biolabs)] with primers flanking the targeted site. The T7 Endonuclease I (T7E1, New England Biolabs) digestion was performed according to the manufacturer's instructions, and samples were fractionated in a 2.5% agarose/1×TAE gel, and visualized by ImageQuant LAS 4000.

For GeoCas9 RNP, 20 pmol of RNP was assembled as previously described (Lin et al., 2014) and mixed with 10 µl OMEM. The RNP was added to 10 µl of 1:10 dilution of Lipofectamine 2000 (Life Technologies) in OMEM and allowed to incubate at room temperature for 10 min before adding it to the cells. The same procedure was used to deliver the indicated molar ratio of AcrIIC1. For analysis of indels, genomic DNA was extracted using Quick Extraction Solution (Epicentre), and the *DNMT1* locus was amplified by PCR. T7E1 reactions were conducted according to the manufacturer's instructions and products were analyzed on a 1.5% agarose gel stained with SYBR gold (Thermo Fisher Scientific). Guide sequences and primers for amplification of targeted sites can be found in the Key Resources Table.

4.3.5 Filter binding

Filter binding assays were conducted as previously described (Ma et al., 2015). Assays were conducted in RNA Binding Buffer (20 mM Tris-HCl, pH 7.5, 150 mM KCl, 5 mM $MgCl_2$, 1 mM DTT, 5% (v/v) glycerol, 0.01% Igepal CA-630, 10 µg ml^{-1} yeast tRNA, and 10 µg ml^{-1} BSA). <0.02 nM radiolabeled sgRNA was incubated with Cas9 at the specified concentration for 20min and loaded onto a dot-blot apparatus through Tufryn, Protran and Hybond-N+ membranes, in that order. The membranes were washed with 50 µL Equilibration Buffer (20 mM Tris-HCl, pH 7.5, 150mM KCl, 5mM $MgCl_2$, 1mM DTT, 5% glycerol) before being dried and visualized by phosphorimaging. Data were fit to a binding isotherm using Prism (GraphPad Software).

4.3.6 Gel shift assays and Fluorescence polarization

Electrophoretic mobility shift assays (EMSA) were conducted as previously described (Ma et al., 2015). Binding reactions were conducted in 1× Binding Buffer (20 mM Tris-HCl, pH 7.5, 150 mM KCl, 5 mM EDTA, 5mM $MgCl_2$, 1 mM DTT, 5% (v/v) glycerol, 50 µg ml^{-1} heparin, 0.01% Tween 20, and 100 µg ml^{-1} BSA). Cas9 and sgRNA were incubated first for 5min at 37°C to allow for guide binding. Next the Cas9–sgRNA complex was diluted to the indicated concentration and a constant amount of 20 µM Acr was added to each sample and allowed to incubate for 10min at room temperature. Radiolabeled DNA target was then added (<0.05 nM). The binding reaction was incubated at 37°C for 30 min at room temperature. Samples were analyzed by 6% polyacrylamide/0.5× TBE gel electrophoresis. Gels were dried and imaged by phosphorimaging. Assays were conducted in triplicate with representative gels shown. The same procedure was used for fluorescence polarization except that 10nM FAM labeled probe was used in place of the radiolabeled DNA and binding was analyzed using fluoresce polarization.

4.3.7 Isothermal titration calorimetry

Isothermal titration calorimetry was conducted as previously described (Nuñez et al., 2014b). Proteins were dialyzed overnight into 20 mM HEPES-NaOH, pH 7.5, 300 mM NaCl and 1 mM TCEP-NaOH. 100μM AcrIIC1 was titrated into the cell containing 10μM NmeCas9. Origin software (OriginLab) was used for baseline correction, integration and curve fitting. The K_D reported is the mean ± standard deviation of three replicates.

4.3.8 Size exclusion binding assays

To test for binding, ~20 μM Cas9, Cas9 chimeras or Cas9 truncations were incubated with ~40 μM of Acr or Acr–GFP. Binding was conducted in Storage Buffer (above) omitting the glycerol. Complexes were resolved on either a Superdex 200 10/300 (full length Cas9 and chimeras) or Superdex 75 10/300 (HNH domains). Fractions were analyzed on a 4–20% PAGE gel and stained with coomassie G-250

4.3.9 *In vivo* assay of AcrIIC1 activity

A plasmid expressing GeoCas9 targeting *E. coli* phage Mu was constructed from pGeoCas9-sgRNA. This plasmid was linearized with *Bsa*I, and DNA encoding a crRNA targeting phage Mu was inserted (oligos 03Lys8F and 03Lys8R). AcrIIC1 was expressed by synthesizing its gene (GenScript) and ligating it to vector, pCDF-1 digested with *Nco*I and *Hin*dII.

The plasmid expressing the Mu-targeting GeoCas9 and the plasmid expressing WT or mutant AcrIIC1 were co-transformed into *E.coli* strain BB101 (a derivative of BL21(DE3) with a deletion of the *slyD* gene). Cells carrying both plasmids were subcultured (1:100) in LB containing chloramphenicol and streptomycin, grown for 2 hours, and the AcrIIC1 expression was induced with 0.01mM IPTG. After 3 h of induction, 200 μL of cells were mixed with top agar and poured onto LB agar plates containing both antibiotics, and supplemented with 200 ng/mL aTc (to induce GeoCas9), 0.2 % arabinose and $MgSO_4$ (10 mM). Phage Mu lysates were spotted in 10-fold serial dilutions onto these plates after the top agar had hardened.

To confirm AcrIIC1 mutant expression in *E. coli,* 500 μl of culture after IPTG induction was centrifuged, and then cells were resuspended in 100 μl SDS loading buffer. AcrIIC1 mutant expression was analyzed by SDS-PAGE on a 15% Tris-Tricine gel, followed by Coomassie Blue staining.

4.3.10 Generation of AcrIIC1 mutants

Mutations were introduced into the AcrIIC1 open reading frame, contained in the pCDF-1b-derived plasmid by site-directed mutagenesis. For each mutation, two 40bp complementary primers containing the desired mutation in the center with correct sequences on both sides were designed (see Key Resources Table). The PCR reaction was conducted using Pfu DNA polymerase (ThermoScientific), followed by DpnI digestion to eliminate the wild-type plasmid. The resulting DNA product was used to transform *E.coli* Stellar cells (Clontech). Plasmids were isolated from streptomycin resistant colonies and all mutations were verified by sequencing.

4.3.11 Crystallography

Crystals were obtained by hanging-drop vapor diffusion at 18°C. Purified NmeCas9 HNH domain was incubated with a 1.2× molar excess of AcrIIC1$_{\text{Nme}}$ on ice for 30 min. Complexed Nme HNH–AcrIIC1 was separated over a Superdex75 16/600 column in gel filtration buffer (150 mM NaCl, 20 mM HEPES-NaOH pH 7.5, 0.5 mM TCEP-NaOH). Preliminary crystallization conditions were identified by sparse-matrix screen using 400 nl drops set over 70 µl reservoir solutions in a 96-well format (Falcon). Optimized crystals were grown in Easy-Xtal 15-well trays (QIAGEN) in 2 µl drops with a 1:1 ratio of protein and reservoir solution with a final protein concentration ~12 mg ml^{-1} in 200 mM ammonium sulfate, 0.1 M sodium acetate pH 4.25, and 22% (w/v) PEG 4000.

Crystals were further optimized by micro-seeding to improve single crystal formation. Crystals were looped and crushed by vortexing and optimized in 2 µL hanging drops composed of 0.2 µl (1:10,000) micro-seed dilution, 1 µl of 7 mg ml^{-1} protein complex, and 0.8 µl reservoir solution. Single crystals were transferred with a nylon loop to a new drop containing reservoir solution supplemented with 15% (v/v) glycerol as a cryoprotectant and incubated for approximately 30 s before flash freezing in liquid nitrogen. Native and anomalous data were collected under cryogenic conditions at the Lawrence Berkeley National Laboratory Advanced Light Source (Beamline 8.3.1).

X-ray diffraction data were processed with *XDS* and *AIMLESS* (Kabsch, 2010) using the SSRL *autoxds* script (A. Gonzalez, Stanford SSRL). Nme HNH–AcrIIC1 crystals belonged to the orthorhombic space group $P\,2_1\,2_1\,2_1$, and contained one copy of each protein in the asymmetric unit. Sulfur single-wavelength anomalous dispersion (S-SAD) data was collected from a single native crystal for experimental phase determination. Briefly, iterative data sets were collected at ~6,000V and merged until a potential phase solution was obtained at ~90× multiplicity using HySS in *PHENIX* (Adams et al., 2010). Manual placement of a orthologous HNH domain from *Actinomyces naeslundii* was used to confirm the resulting map and allow extension of the correct solution to the native data processed to ~1.50Å using SOLVE/RESOLVE (Terwilliger, 1999). The correct phase solution contained 17 sites corresponding to 16 sulfur atoms and 1 SO_4 position. The Final model was completed by iterative model building in *COOT* (Emsley & Cowtan, 2004) and before refinement with PHENIX. X-ray data for refinement were extended according to an I/σ resolution cut-off of ~1.0, CC* correlation, R_{pim} parameters, and visual inspection of the resulting map.

4.3.12 Small-angle X-ray scattering

Small-angle X-ray scattering (SAXS) data were collected at the SIBYLS beamline at the Lawrence Berkeley National Laboratory Advanced Light Source. Data of NmeCas9, NmeCas+Acr1 and NmeCas9+Acr3 was collected over a dilution series from 1 to 4 mg ml^{-1}.

Data were background-corrected in *SCÅTTER* (BIOISIS) before primary data processing using the *ATSAS* software package (Konarev, Petoukhov, Volkov, & Svergun, 2006). The Guinier region, intensity at 0 [$I(0)$], and radius of gyration (R_g) were calculated using *PRIMUS* (Konarev, Volkov, Sokolova, Koch, & Svergun, 2003). The P(r) distribution and maximum dimension (D_{max}) were calculated using *GNOM* (Semenyuk & Svergun, 1991). Molecular weights were calculated from the mass parameter Q_R in *SCÅTTER* using the volume-of-correlation (V_c) and the power-law relationship for protein (Rambo & Tainer, 2013).

4.3.13 Electron microscopy

NmeCas9-sgRNA-AcrIIC3 was cross-linked by 0.02% glutaraldehyde at room temperature for 7 minutes. NmeCas9-sgRNA was used directly after gel filtration. All samples were diluted to a final concentration of ~100 nM and negatively stained in 2% (w/v) uranyl acetate solution following the standard deep-stain procedure on holey carbon-coated EM copper grids covered with a thin layer of continuous carbon (J.-J. Liu et al., 2014). Negetivly stained specimens were mounted on a transmission electron microscope holder and examined by a Tecnai Spirit electron microscope operated at 120-kV acceleration voltage. Magnified digital micrographs of the specimen were taken at a nominal magnification of 49,000 on a Gatan Ultrascan4000 CCD camera with a pixel size of 2.18 Å at the specimen level. The defocus values used were about -1.0 to -1.5µm, and the total accumulated dose at the specimen was about 50 electrons per Å2. The particle picking and 2D analysis were performed within Appion (Lander et al., 2009).

4.4 Results

4.4.1 AcrIIC1 inhibits diverse Cas9 orthologs

Phylogenetic analysis revealed that AcrIIC1 is part of an unusually diverse family of Acr proteins (Figure S1A). Mirroring this diversity, the bacterial genomes containing AcrIIC1 include Cas9 orthologs that span a large portion of the type IIC Cas9 tree (Figure 1A). Based on its phylogenetic distribution, we hypothesized that AcrIIC1 would be more promiscuous than other Acr proteins with respect to the Cas9 orthologs it can inhibit.

To test this idea, we conducted cleavage assays using various type IIC Cas9 orthologs previously shown to function in human cells (Esvelt et al., 2013b; Harrington, Paez-espino, et al., 2017; Hou et al., 2013; Kim et al., 2017). We found that in addition to NmeCas9, the AcrIIC1 from *Neisseria meningitidis* (AcrIIC1$_{\text{Nme}}$) exhibits robust inhibition of the Cas9 proteins from *Geobacillus stearothermophilus* (GeoCas9) and *Campylobacter jejuni* (CjeCas9) (Figure 1B and S1B). CjeCas9 and GeoCas9 are 36% and 42% identical to NmeCas9 respectively and represent diverse branches of the type IIC Cas9 phylogeny (Figure 1A). By contrast, AcrIIC2 and AcrIIC3 were both highly specific for NmeCas9, having no noticeable impact on CjeCas9- or GeoCas9-catalyzed DNA cleavage (Figure 1B and S1B).

To determine whether inhibition by AcrIIC1 can disable CjeCas9 in genome-editing applications, we transfected HEK293T cells with plasmids expressing NmeCas9, CjeCas9 or SpyCas9 and their respective single-guide RNAs (sgRNAs) in the presence or absence of a gene encoding AcrIIC1 (Figure 1C and S1C). Similar to the biochemical cleavage assays, we observed that in cells, CjeCas9 is inhibited by AcrIIC1 but not by AcrIIC3. Expressing AcrIIC1$_{\text{Nme}}$ or the AcrIIC1 from *Brackiella oedipodis* (AcrIIC1$_{\text{Boe}}$) resulted in efficient inhibition of CjeCas9, indicating that this promiscuity is not unique to the AcrIIC1$_{\text{Nme}}$ ortholog (Figure S1C). In similar cell-based assays, we found that AcrIIC1 is also a potent inhibitor of GeoCas9 ribonucleoprotein complexes (RNPs) in mammalian cells (Figure S1D), revealing that AcrIIC1 can also function when delivered as an expressed protein. The robust inhibition of both CjeCas9 and GeoCas9, in addition to NmeCas9, suggested that AcrIIC1 exploits a conserved feature of the Cas9 protein.

4.4.2 AcrIIC1 traps the DNA-bound Cas9 complex

Acrs can potentially inhibit Cas9 proteins at multiple distinct steps including guide RNA binding, target DNA binding or target cleavage. To determine the step at which AcrIIC1 inhibition occurs, we biochemically tested each of these possible mechanisms. First, we measured the binding affinity of NmeCas9 for its sgRNA in the presence and absence of AcrIIC1 (Figure S2A) and found that RNP assembly was unaffected by AcrIIC1. Next, we conducted equilibrium binding measurements of NmeCas9–sgRNA to its target DNA. Surprisingly, we found that NmeCas9 DNA binding was unimpeded by the presence of AcrIIC1, indicating that AcrIIC1 selectively blocks DNA cleavage (Figure

S2B-C). Titrating AcrIIC1 and AcrIIC3 in a cleavage assay revealed that both are capable of inhibiting NmeCas9 even at low concentrations (Figure S2D). We conducted end-labeled cleavage assays to determine if cutting of both the target and non-target DNA strands is inhibited to the same degree (Figure 2A and 2B). Here we found that AcrIIC1 strongly inhibits cleavage of both DNA strands but with a subtle difference in kinetics. Although slow cleavage of the non-target DNA strand catalyzed by the RuvC active site is observed, target-strand cleavage catalyzed by the HNH domain is undetectable. These results suggested that AcrIIC1 traps Cas9 in its DNA-bound state, while inhibiting DNA cleavage. We tested this hypothesis by conducting gel shift assays using catalytically active GeoCas9 with and without AcrIIC1. In the absence of AcrIIC1, GeoCas9 cleaved its DNA target substrate at concentrations above ~30 nM (Figure 2C). However, when AcrIIC1 was included in the reaction, Cas9 did not cleave the target DNA even though DNA binding was unaffected. This remarkable mechanism is distinct from the recently studied AcrIIA2 and AcrIIA4 anti-CRISPR proteins, which function as inhibitors of DNA binding by SpyCas9 (Dong et al., 2017; Shin et al., 2017). The unique ability of AcrIIC1 to trap Cas9 on its DNA target in a catalytically inactivate state effectively transforms the wild-type Cas9 into its catalytically inactive variant dCas9 (Martin Jinek et al., 2012a).

4.4.3 AcrIIC1 binds to the HNH domain of Cas9

The ability of AcrIIC1 to inhibit multiple Cas9 orthologs without preventing DNA binding suggested that it targets a conserved region of Cas9 involved in DNA cleavage. To determine which region of Cas9 interacts with AcrIIC1, we generated Cas9 truncations and tested their abilities to bind to AcrIIC1 using size exclusion chromatography (Figure 3A, 3B and S3A). Although many NmeCas9 truncations were insoluble, we took advantage of the thermostable GeoCas9 (Harrington, Paez-espino, et al., 2017) to generate soluble truncations. AcrIIC1 was able to associate with GeoCas9 without either the REC or PAM-interacting domains (Figure 3A), leaving the two nuclease domains as potential interacting partners. Truncating further to remove the RuvC domain allowed us to identify the HNH domain as being sufficient for AcrIIC1 binding to Cas9 (Figure 3A, 3B and S3A). Cas9 binding to RNA and DNA has been shown to be independent of the HNH domain(Sternberg, LaFrance, Kaplan, & Doudna, 2015; Yamada et al., 2017). In line with this, complexing of GeoCas9 with AcrIIC1 revealed that AcrIIC1 is able to interact with Cas9 irrespective of the presence of sgRNA or sgRNA and target DNA (Figure S3B). To determine if AcrIIC1 interacts specifically with the HNH domain, we exchanged the HNH domain of a Cas9 ortholog that does not interact with AcrIIC1 (*Actinomyces naeslundii,* AnaCas9) with an ortholog that does (GeoCas9). Here we found that the GeoCas9 chimera containing the AnaCas9 HNH domain no longer bound to AcrIIC1, while the AnaCas9 with the GeoCas9 HNH domain substitution was able to interact with AcrIIC1 (Figure 3A, 3B and S3C). These results indicated that the HNH domain is the primary site of interaction for AcrIIC1.

To examine whether this interaction also occurs with NmeCas9, we purified the HNH domain of NmeCas9 and tested its ability to bind to AcrIIC1 using size exclusion chromatography (Figure 3C). While AcrIIC1 and the HNH domain eluted at similar volumes in isolation, applying the two proteins to the size exclusion column together resulted in a large shift in elution volume, indicative of protein association. Importantly, the eluted HNH_{Nme}–AcrIIC1 complex remains in the included volume of the column, indicating that the large shift is not due to aggregation. Further analysis of AcrIIC1 binding to NmeCas9 by isothermal titration calorimetry (ITC) demonstrated an equilibrium binding affinity of AcrIIC1 6.3 ± 3.4 nM with a stoichiometry of one AcrIIC1 for each Cas9 (Figure S3D). The HNH nuclease domain is highly conserved across all Cas9 proteins (Figure S3E) and controls cleavage of both strands of the target DNA (Dagdas, Chen, Sternberg, Doudna, & Yildiz, 2017; Sternberg et al., 2015). Although the Cas9 HNH nuclease domain is directly responsible for cleavage of the target strand of the DNA (Gasiunas et al., 2012; Martin Jinek et al., 2012a), conformational activation of the HNH domain is a prerequisite for activating cleavage of the non-target strand by the RuvC nuclease domain (Sternberg et al., 2015). The ability of AcrIIC1 to bind to the most conserved domain of Cas9 explains its ability to robustly inhibit related Cas9 orthologs (Figure 1B) and its wide phylogenetic distribution (Figure 1A).

4.4.4 Structure of AcrIIC1 bound to the Cas9 HNH domain

To better understand how AcrIIC1 has evolved to bind to multiple Cas9 proteins, we determined a 1.5 Å resolution crystal structure of AcrIIC1$_{Nme}$ bound to the HNH domain of NmeCas9. The overall structure shows that AcrIIC1 binds directly to the HNH active site (Figure 4.4A), restricting it from accessing the target DNA. AcrIIC1 binds to the active site interface of the HNH domain through several ionic and hydrogen-bonding interactions. Critically, the HNH domain active site residues H588 and D587 hydrogen bond to AcrIIC1 residue S78 and to the backbone amine of C79, respectively (Figure 4.4B), possibly excluding the divalent cation necessary for target-strand DNA cleavage (Martin Jinek et al., 2012a). Mapping amino acid conservation onto the structure revealed that residues within the binding interface of both the HNH domain and AcrIIC1 are highly conserved (Figure 4.4D, S4A, S4B). In contrast to this observed conservation, antagonistic binding interfaces often evolve rapidly, leading to lower conservation (Franzosa & Xia, 2011), suggesting that AcrIIC1 is targeting a highly conserved surface in order to limit the chance for the host to escape inhibition.

Comparative structural homology searches of AcrIIC1 against protein structure databases using DALI and Vast revealed that AcrIIC1 adopts a novel protein fold (Holm & Laakso, 2016). The $\beta_1\beta_2\beta_3\alpha_1\alpha_2\beta_4\beta_5$ fold of AcrIIC1 comprises a five-stranded beta bundle interspaced by two alpha-helices. The beta bundle is the conserved core feature found in all AcrIIC1 orthologs while the internal loops connecting beta-strands and alpha-helices vary in length and composition across species (Figure 4.10B). All of the HNH-interacting residues occur within these variable loop regions, revealing how AcrIIC1 can evolve to target divergent HNH domains without compromising structural integrity.

For AcrIIC1 to effectively prevent Cas9 from cleaving the invading viral DNA, it must remain bound to the HNH domain for extended periods of time. This stable interaction is in part accomplished by multiple charged residues around the periphery of the active site that form an additional five hydrogen bonds with AcrIIC1 (Figure 4.10C). Interestingly, some interactions target conserved residues present in diverse Cas9 orthologs and other interactions appear to have evolved to target specific species. For example, S78 and E81 of AcrIIC1$_{Nme}$ interact with the highly conserved catalytic residues H588 and N616 of the HNH$_{Nme}$ domain, respectively. By contrast, AcrIIC1$_{Nme}$ residue D14 and the backbone carbonyl of P39 interact with K551 and K549 of

NmeCas9, which are mutated to a serine and glycine in AnaCas9. To assess the importance of individual amino acids for the biological function of AcrIIC1, we established an *in vivo* anti-CRISPR activity assay in *E. coli*. Plasmid-mediated expression of GeoCas9 and an sgRNA designed to target *E. coli* phage Mu (Morgan, Hatfull, Casjens, & Hendrix, 2002) led to a reduction in the plaquing efficiency of this phage by approximately 10^6-fold (Figure 4.4C). Co-expression of wild-type AcrIIC1 restored the full plaquing activity of phage Mu, implying that that GeoCas9 was completely inhibited by the anti-CRISPR. By contrast, the S78A mutant displayed very little anti-CRISPR activity in this assay as phage Mu plaquing in the presence of this mutant was barely above background (Figure 4.4C and 4.10D). Substitutions of other residues positioned in the HNH:AcrIIC1 interaction interface, such as M76 and E81, caused more modest reductions in anti-CRISPR activity (Figure 4.10D) while substitution of other interface residues caused no reduction in biological activity. Importantly, all mutant proteins were expressed at the same level as wild-type (Figure 4.10E). AcrIIC1 interacting residues on the active site interface of the NmeCas9 HNH domain closely align with those on the same interface of *S. aureus* Cas9 (SauCas9), but diverge from equivalent residues in SpyCas9 and AnaCas9 mainly near the N-terminus of the HNH domain (Figure 4.11A). Together with this structure, the high degree of structural similarity between the HNH domains of these species will enable rational engineering of AcrIIC1 to target specific Cas9 orthologs of interest.

Investigation of Cas9 target recognition and cleavage has uncovered several checkpoints along the interference pathway that ensure cleavage of the correct DNA sequence (Sternberg et al., 2015). The best understood of these checkpoints is mediated by the HNH domain, which undergoes a large rotation and translation to cleave the target DNA only when sufficient complementarity to the guide RNA is sensed (Dagdas et al., 2017; Sternberg et al., 2015). The structure presented here suggests that AcrIIC1 exploits one checkpoint in this process to ensure inhibited cleavage of both the target and non-target DNA strands. When modeled into a Cas9–sgRNA complex bound to dsDNA (Fuguo Jiang et al., 2016), AcrIIC1 sterically blocks the HNH domain from rotating into position above the scissile phosphate (Figure 4.4E and Figure 4.11B). We propose that the inability to correctly dock the HNH domain in the presence of AcrIIC1 inhibits RuvC cleavage of the non-target strand, explaining how a small protein that allows dsDNA engagement can still inhibit the two separate nucleases of Cas9.

4.4.5 AcrIIC3 blocks DNA binding and induces NmeCas9 dimerization

In contrast to AcrIIC1, AcrIIC3 has few natural orthologs and is found only in *Neisseria*. In HEK293T cells, expression of AcrIIC3 leads to the inability of dNmeCas9 to localize to a genomic target (Pawluk, Amrani, et al., 2016), suggesting that AcrIIC3 prevents NmeCas9 from binding to DNA. We tested this biochemically using fluorescence polarization, which detected a ~10 fold decrease in equilibrium DNA binding affinity of NmeCas9 in the presence of AcrIIC3 (71 ± 13.4 nM without AcrIIC3 versus 859 ± 149 nM with AcrIIC3) (Figure 4.5A). This reduced but not abolished binding affinity of NmeCas9 for DNA in the presence of AcrIIC3 may indicate that NmeCas9 can still interact with the PAM region of the DNA, but cannot achieve complete R-loop formation (Mekler, Minakhin, & Severinov, 2017).

After incubating NmeCas9 and AcrIIC3 together we noted a large shift in elution volume by size exclusion chromatography (SEC) compared to either AcrIIC1-bound NmeCas9 or NmeCas9 alone (Figure 4.5B). This large shift suggested either a substantial conformational change or oligomerization of Cas9. Analysis of the fractions by small-angle X-ray scattering (SAXS) revealed that the AcrIIC3-bound NmeCas9 increased in size relative to NmeCas9 alone or NmeCas9–AcrIIC1, as indicated by an elongated pair distance distribution, increased radius of gyration (R_g) and volume of correlation (V_c) (Figure 4.5C, Figure 4.12A). Using the power-law relationship for protein SAXS (Rambo & Tainer, 2013), we estimated the mass of NmeCas9 alone and AcrIIC1-bound NmeCas9 to be ~110kDa, a slight underestimate of the theoretical masses of 124kDa and 137kDa, respectively. In contrast, the estimated mass of AcrIIC3-bound NmeCas9 was ~210kDa. Together, our solution studies suggest that AcrIIC3 induces dimerization of NmeCas9, possibly contributing to its ability to block DNA binding. Analysis of AcrIIC3 alone by SEC and native mass spectrometry suggested that this Acr is monomeric in solution (Figure 4.12B and S6C), although dimerization of two AcrIIC3 monomers upon binding to NmeCas9 is possible.

To visualize the dimerization of NmeCas9, we examined NmeCas9 and AcrIIC3-bound NmeCas9 using electron microscopy. Although the protein interaction surfaces could not be identified due to limited resolution, an overall shape of the AcrIIC3-bound NmeCas9 complex can be observed in the 2D class averages (Figure 4.5D). To obtain better-resolution 2D class averages, the particles were cross-linked to reduce the flexibility of the NmeCas9–sgRNA complex and reduce dissociation of the dimer. These data reveal an overall symmetrical complex with dimensions consistent with tw o Cas9 proteins (Figure 4.5D). Cas9 inhibition by AcrIIC3-induced dimerization is consistent with the independent evolution of Acrs that act by diverse mechanisms.

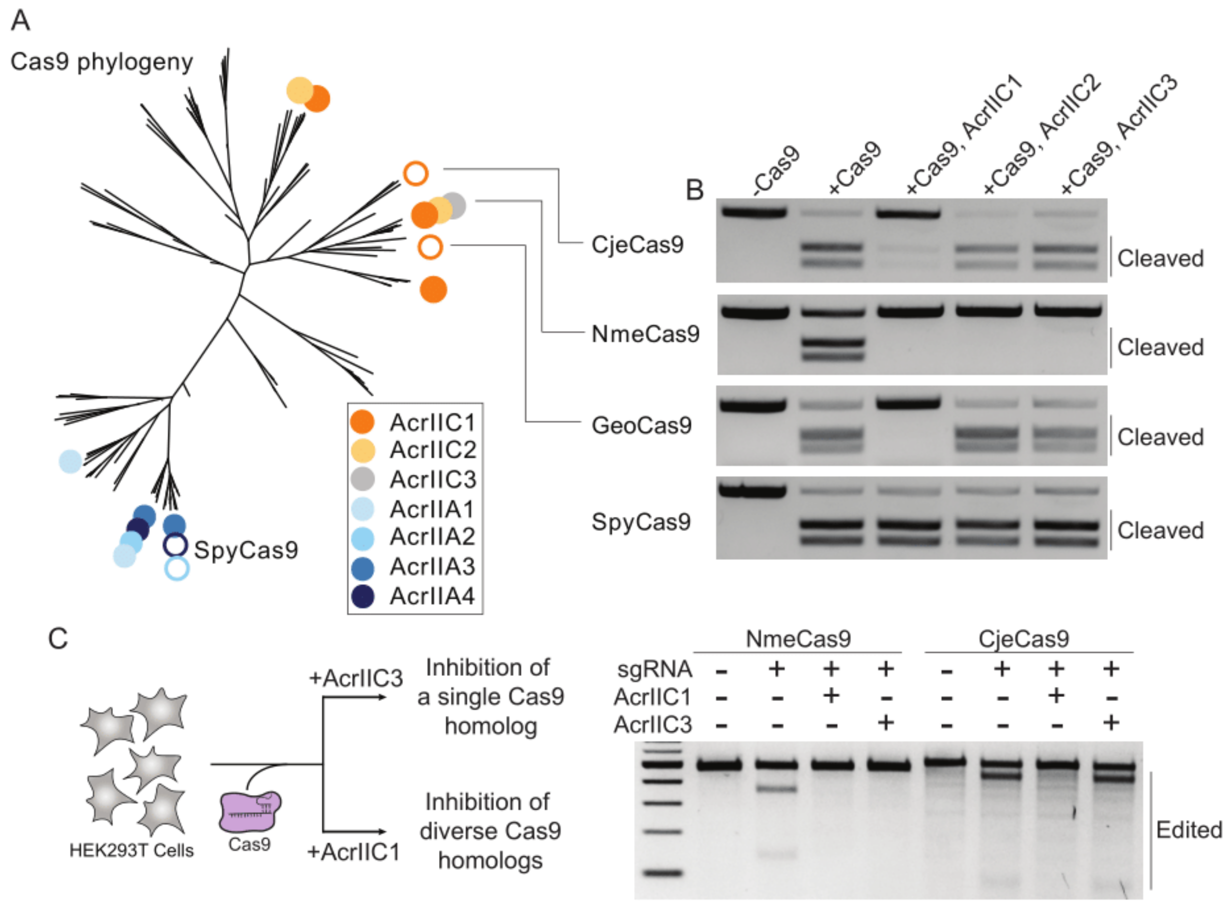


Figure 4.1 AcrIIC1 inhibits diverse Cas9 orthologs while AcrIIC2 and AcrIIC3 are highly specific

(A) Unrooted phylogenetic tree of Cas9. Cas9 orthologs targeted by Acrs are indicated with circles at ends of branches (closed circles, Cas9 orthologs naturally targeted by an Acr; open circles, Cas9 orthologs which have been shown experimentally to be inhibited by an Acr but without naturally occurring AcrIIC1 orthologs). For branches containing multiple Acrs of a given type only one circle is shown for simplicity (phylogeny adapted form (Burstein, Harrington, Strutt, & Probst, 2017).

(B) DNA cleavage assays conducted by various Cas9 orthologs in the presence of AcrIIC1, AcrIIC2 and AcrIIC3. (–Cas9, no Cas9 added; +Cas9, Cas9 and sgRNA added; Cje, *Campylobacter jejuni*; Nme*, Neisseria meningitidis*; Geo, *Geobacillus stearothermophilus*; Spy*, Streptococcus pyogenes*).

(C) (Left) Cartoon depicting experiment to test inhibition of Cas9 orthologs by AcrIIC1 in HEK293 cells. (Right) T7E1 assay analyzing indels produced by CjeCas9 and NmeCas9 shows that CjeCas9 genome editing is inhibited by AcrIIC1$_{Nme}$ but not AcrIIC3$_{Nme}$. See also Figure 4.7.

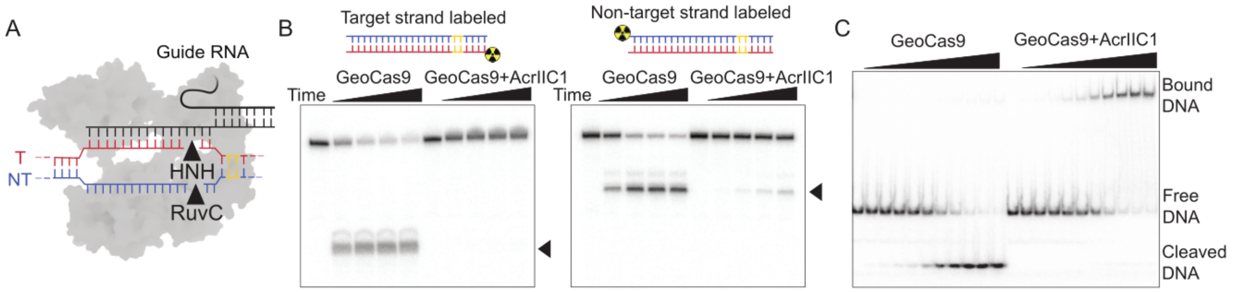


Figure 4.2 AcrIIC1 traps the DNA-bound Cas9 complex
(A) Cartoon of Cas9-mediated double-stranded DNA cleavage. Guide RNA (black) is duplexed to the DNA target strand (red), which is splayed from the DNA non-target strand (blue) adjacent to the PAM sequence (yellow). The HNH and RuvC nuclease domains (black triangles) cleave the target strand and non-target strand, respectively.
(B) Radiolabeled cleavage assays conducted using GeoCas9 to measure AcrIIC1 inhibition of cleavage on the target and non-target strands. Cas9–sgRNA RNP was complexed with or without AcrIIC1 and added to radiolabeled target DNA duplex with each strand labeled separately. The lanes for a given condition correspond to increasing time (0-30min) from left to right. Black triangles indicate cleavage products.
(C) Analysis of GeoCas9 binding and cleavage in the presence or absence of AcrIIC1 analyzed on a non-denaturing gel with the non-target strand labeled. GeoCas9 RNP concentration was varied in the absence or presence of excess AcrIIC1. The top band corresponds to GeoCas9 bound to the target DNA, the middle band is free DNA, and the lower band is cleaved DNA (concentration series correspond to 0, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512nM of GeoCas9 RNP). See also Figure 4.8.

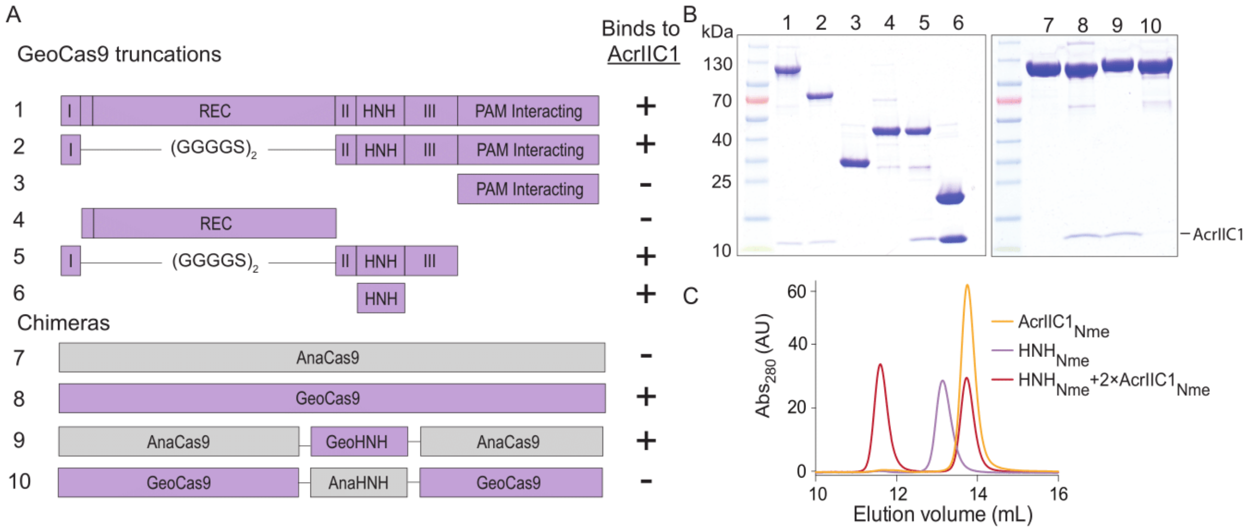


Figure 4.3 AcrIIC1 binds to Cas9 HNH domain
(A) Domain schematics of GeoCas9 truncations and Cas9 chimeras, designed to identify the Cas9 binding interface of AcrIIC1. Constructs 1-10 were incubated with AcrIIC1, fractionated over an S200 size-exclusion column and analyzed by SDS-PAGE. Constructs that bound to AcrIIC1 are indicated with a (+) and constructs that showed no interaction are indicated with a (-). The chimeric Cas9 proteins (7-10) were generated by switching the HNH domains of a Cas9 that is not inhibited by AcrIIC1 (AnaCas9) and a Cas9 that is inhibited (GeoCas9).
(B) Fractions from the S200 runs in Figure 4.9A were separated on a 4-20% SDS-PAGE gel. Numbers above the gel correspond to the construct or chimera numbers from Figure 4.2A.

(C) Elution from an S75 size exclusion column of NmeCas9 HNH domain (purple), AcrIIC1 (orange) or the two incubated together with 2-fold excess AcrIIC1(red). See also Figure 4.9.

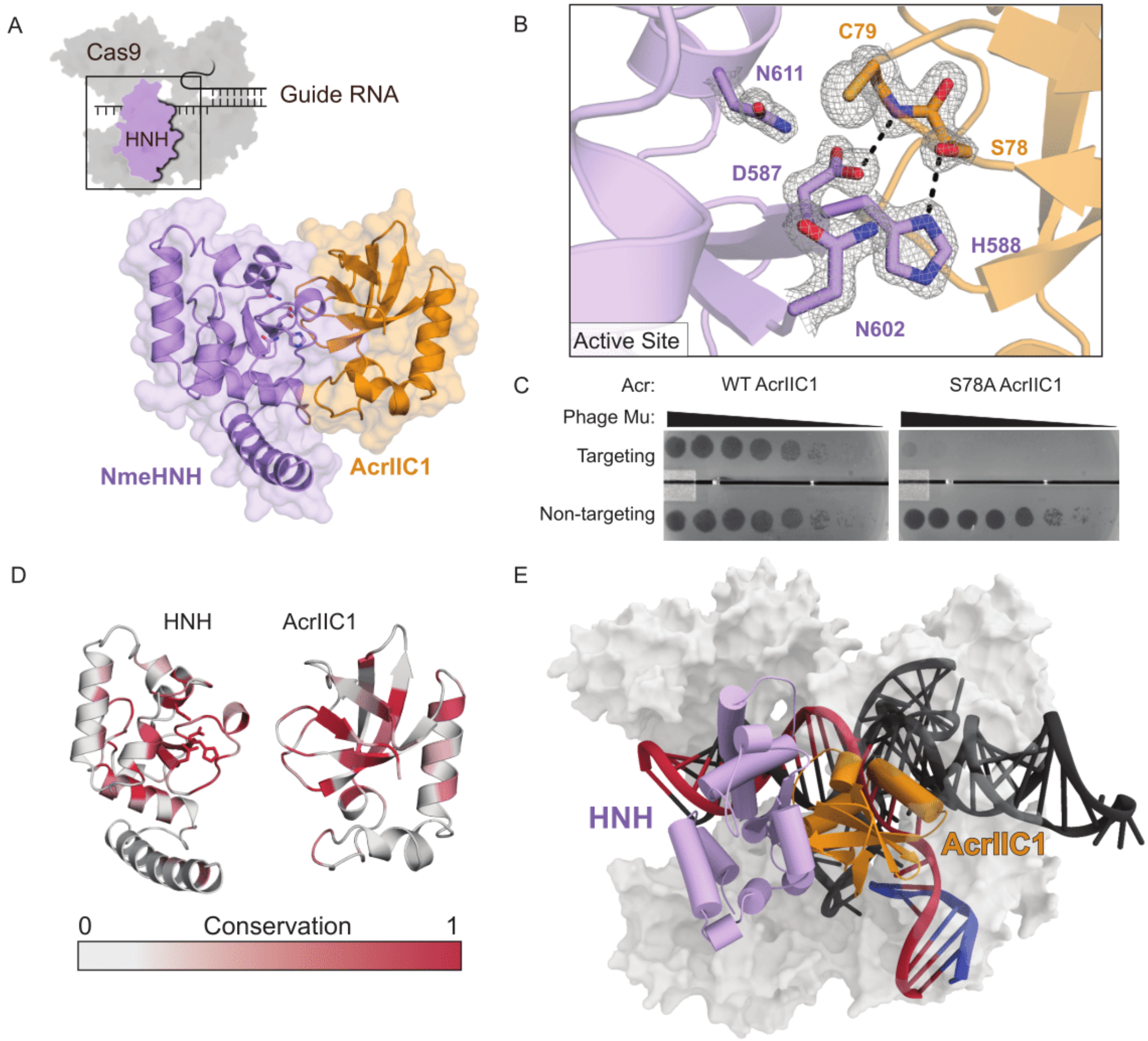


Figure 4.4 Structure of AcrIIC1 bound to the NmeCas9 HNH domain
(A) (Top) Cartoon depiction of NmeCas9 (grey) bound to a guide RNA (black.) The black outline of the HNH domain (purple) indicates the binding interface to AcrIIC1. (Bottom) Crystal structure of NmeCas9 HNH domain bound to AcrIIC1 (PDB:5VGB). Catalytic residues are depicted as sticks.
(B) Occlusion of HNH active site residues (purple) through hydrogen bonding with AcrIIC1 (orange). HNH catalytic residues H588 and D587 form hydrogen bonds (black dotted line) with S78 and the backbone amine of C79 of AcrIIC1, respectively. 2mF$_o$-DF$_c$ electron density map is shown for interacting residues and contoured at 1.8 σ.
(C) Plaquing of *E. coli* phage Mu targeted by GeoCas9 in the presence of wild-type AcrIIC1 or the S78A AcrIIC1 mutant. Mutation of S78A results in nearly complete inactivation of AcrIIC1's inhibitory effect on GeoCas9.
(D) Binding interfaces of NmeCas9 HNH domain and AcrIIC1 show residue conservation. Conservation was calculated using multiple sequence alignments of AcrIIC1 orthologs and Cas9

HNH domains. Conserved residues are colored red (1, 100% sequence identity) and non-conserved residues are colored white (0).

(E) Model of AcrIIC1 inhibiting cleavage of both target and non-target strands. NmeCas9 HNH domain (purple) was modeled into a “docked” position using dsDNA-bound SpyCas9 structure (PDB: 5F9R) as a reference for a homology model of NmeCas9. Placement of AcrIIC1 (orange) between the HNH domain and the target strand (red) prevents target cleavage and activation of the RuvC domain for non-target strand (dark blue) cleavage. See also Figure 4.10, 11 and Table 4.1.

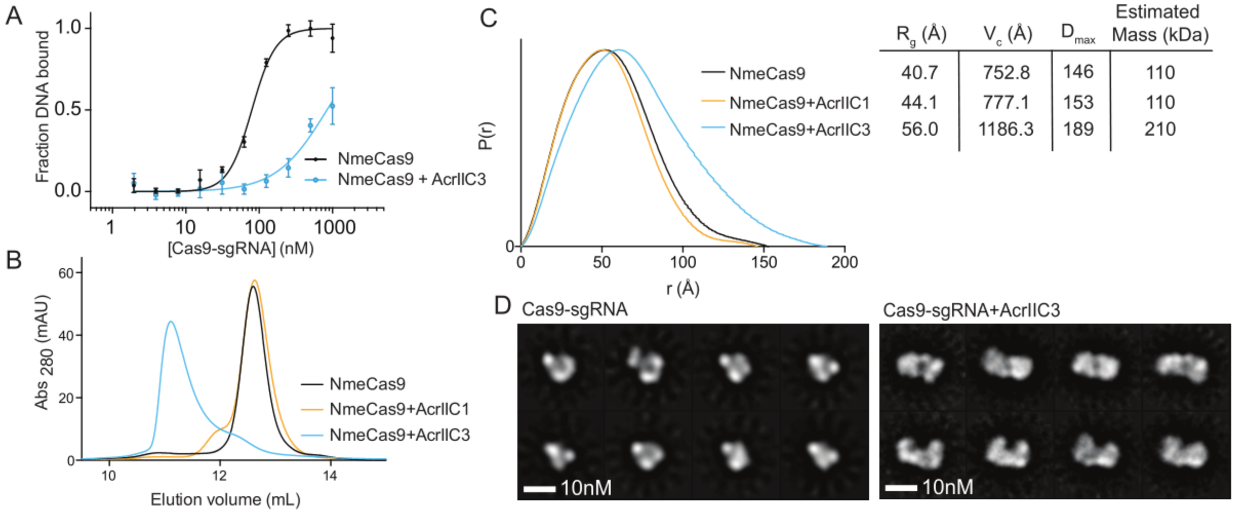


Figure 4.5 AcrIIC3 blocks DNA binding and dimerizes Cas9

(A) Equilibrium binding measurements of NmeCas9 to dsDNA using fluorescence polarization in the presence (blue) or absence (black) of AcrIIC3. Measurements were made in triplicate and the mean +/- S. D. is shown.

(B) Elution from a Superdex 200 10/300 size exclusion column for NmeCas9 (black), NmeCas9+AcrIIC1 (orange), and NmeCas9+AcrIIC3 (blue) showing a large shift in elution volume for NmeCas9-AcrIIC3, indicative of oligomerization.

(C) SAXs data for fractions collected from samples in (C). (Left) pair-distance distribution function for NmeCas9 alone (black), with AcrIIC1 (orange) or with AcrIIC3 (blue), indicating increased particle size upon AcrIIC3 binding. (R_g, radius of gyration; V_c, volume of correlation; D_{max}, maximum dimension.)

(D) 2D class averages of NmeCas9-sgRNA monomers (left) and NmeCas9-sgRNA bound to AcrIIC3 (right). Scale bar is 10nm. See also Figure 4.8 and 4.10.

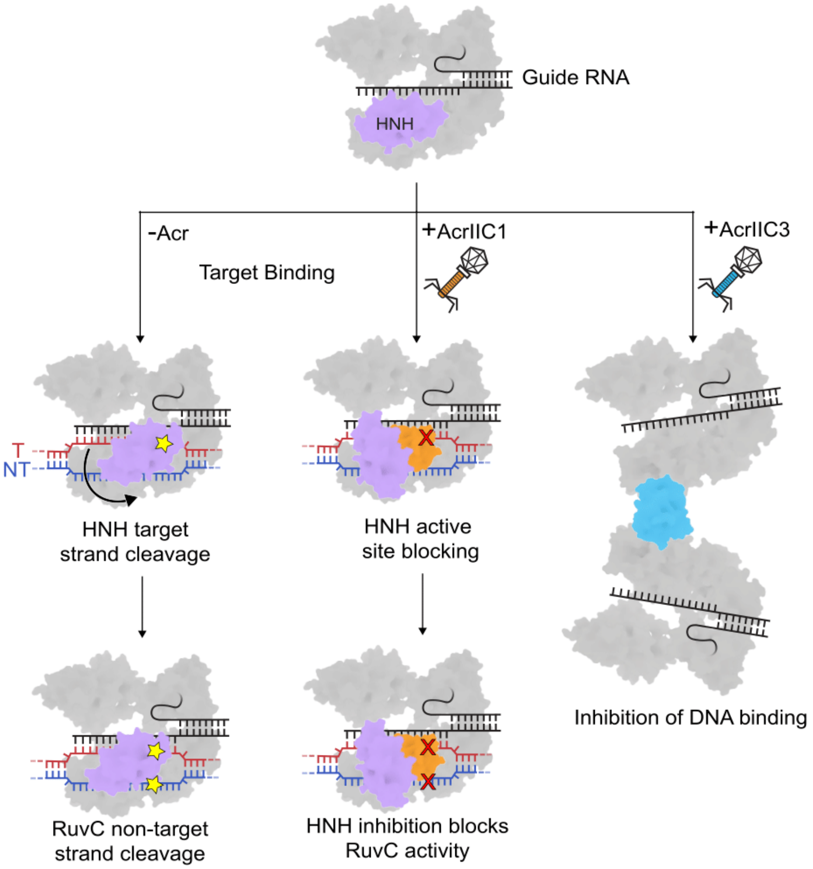


Figure 4.6 Model of AcrIIC1 and AcrIIC3 inhibition of Cas9

Cas9 assembles with its guide RNA to form the search complex. Phage encoded AcrIIC1 (orange) binds to Cas9, still allowing target dsDNA binding but occluding the HNH (purple) active site, and stopping cleavage of the target strand. AcrIIC1 also conformationally restricts HNH docking, stopping cleavage on the non-target strand. For AcrIIC3 (blue), Cas9's target DNA binding is inhibited and Cas9 is caused to dimerize.

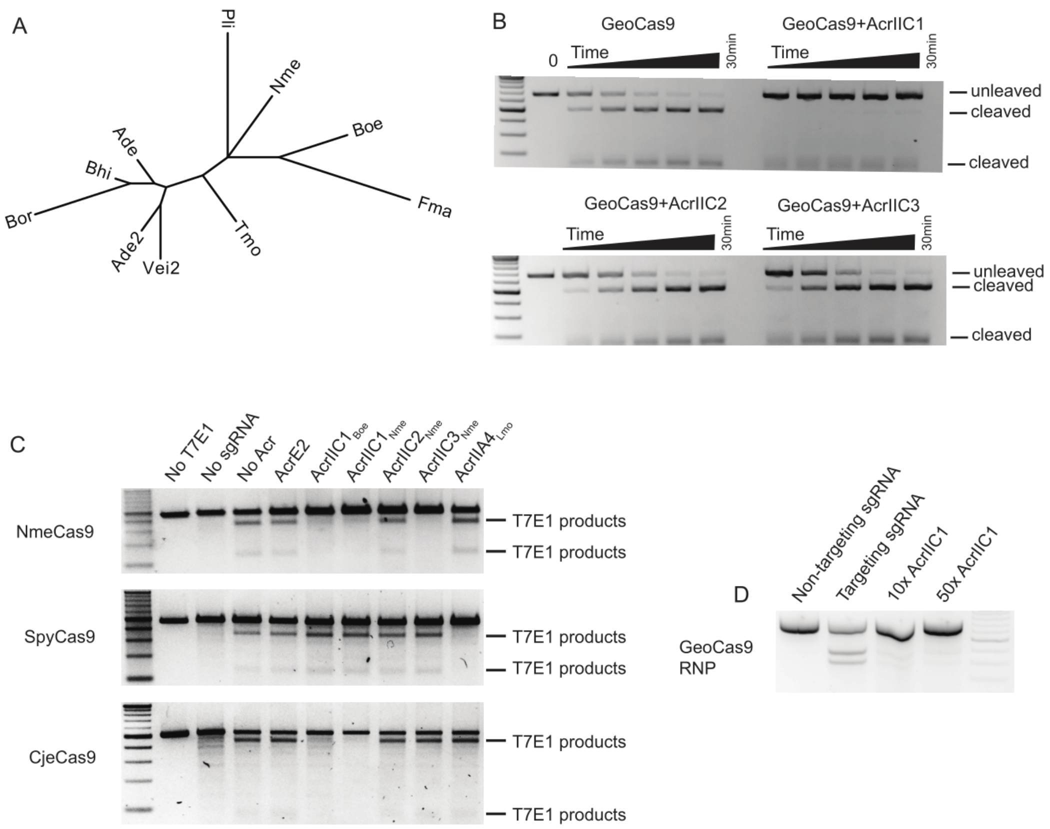


Figure 4.7 AcrIIC1 inhibits diverse Cas9 orthologs, while AcrIIC2 and AcrIIC3 are highly specific, related to Figure 4.1.

(A) Unrooted phylogenetic tree of AcrIIC1.

(B) Kinetic measurement of DNA cleavage mediated by GeoCas9 in the presence or absence of type IIC Acrs.

(C) Genome editing mediated by NmeCas9, SpyCas9 and CjeCas9 in the presence of various Acrs. Human (HEK293T) cells were transfected with plasmids expressing NmeCas9, SpyCas9, or CjeCas9, along with cognate, previously validated sgRNAs targeting genomic sites. T7E1 digestion was used to detect editing. A Type I anti-CRISPR (AcrE2) was used as a negative control for inhibition. As reported previously, NmeCas9 genome editing (upper panel) was inhibited by AcrIIC1$_{\mathrm{Boe}}$, AcrIIC1$_{Nme}$, and AcrIIC3$_{Nme}$; full inhibition by AcrIIC2$_{Nme}$ in human cells generally requires higher amounts of cotransfected expression plasmid. Inhibition of SpyCas9 genome editing (middle panel) was observed only with AcrIIA4$_{Lmo}$ (Rauch *et al.*, 2017). In contrast, CjeCas9 editing activity (bottom panel) was inhibited by AcrIIC1$_{Boe}$ and AcrIIC1$_{Nme}$, but not by any of the other anti-CRISPRs.

(D) GeoCas9 RNP mediated editing of HEK293T cells in the presence and absence of AcrIIC1. Indels were analyzed by T7E1 digestions.

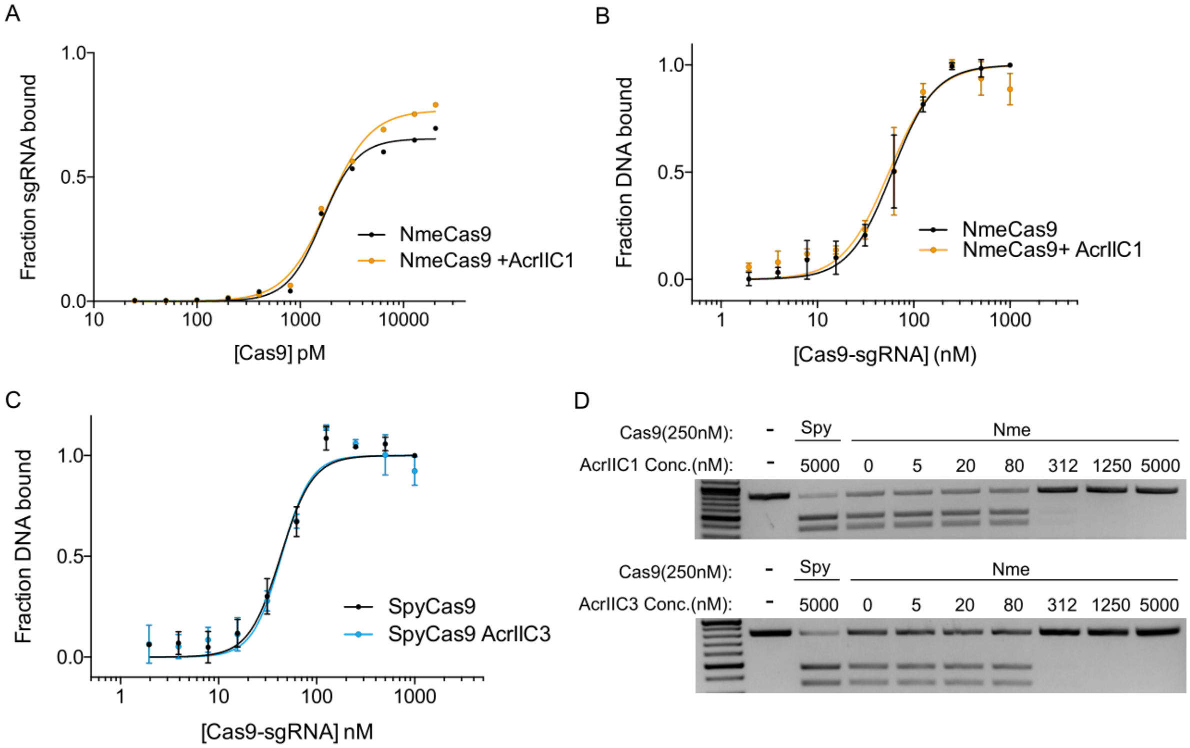


Figure 4.8 Impact of AcrIIC1 and AcrIIC3 on Cas9 RNA and DNA binding, related to Figures 4.2 and 4.5. (A) Filter-binding assays measuring the affinity of NmeCas9 to its guide in the presence and absence of AcrIIC1. Radiolabeled sgRNA was incubated with NmeCas9 in the presence (orange) or absence (black) of AcrIIC1. NmeCas9–sgRNA RNP formation was measured using a filter binding assay and fraction sgRNA bound was calculated and plotted against NmeCas9 concentration. When incubated with Cas9 prior to sgRNA binding, AcrIIC1 does not inhibit RNP formation.

(B) Equilibrium binding measurements of NmeCas9–sgRNA binding to dsDNA in the presence and absence of AcrIIC1, measured by fluorescence polarization.

(C) Equilibrium binding measurements of SpyCas9 in the presence and absence of AcrIIC3, related to Figure 4.5D.

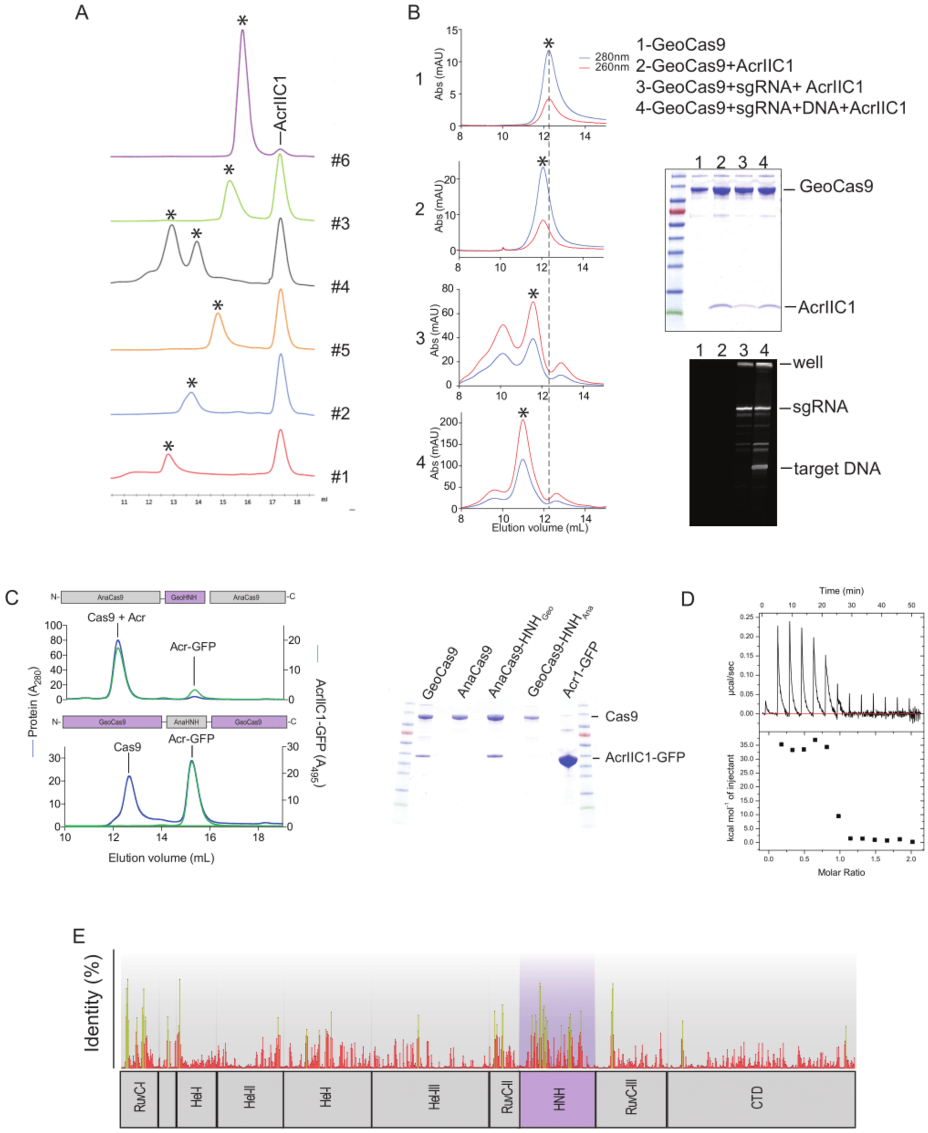


Figure 4.9 AcrIIC1 Binds to the Cas9 HNH domain, related to Figure 4.3.
(A) Superdex 200 10/300 traces of truncations used to identify the binding interface of AcrIIC1. Each trace included the indicated GeoCas9 truncation and excess AcrIIC1. Black asterisks indicate the fractions analyzed in Figure 4.3B. For construct #4 (REC lobe), multiple peaks resulted from bound contaminating nucleic acid species from the purification and both peaks were pooled and analyzed together.
(B) Superdex 200 10/300 traces for GeoCas9 complexed with the components indicated in the top right. Fractions indicated with asterisk were analyzed on SDS-PAGE (upper gel) gel and denaturing urea PAGE gel (lower gel) and components were added in the order listed. (1, Apo GeoCas9; 2, GeoCas9 +AcrIIC1; 3, GeoCas9+sgRNA+AcrIIC1; 4, GeoCas9+sgRNA+DNA+AcrIIC1).
(C) (top) HNH domains between GeoCas9 (orange) and AnaCas9 (yellow) were swapped to create chimeric Cas9 proteins. AcrIIC1 was fused with GFP (to more easily visualize a change in elution volume once bound to Cas9) and run over a S200 size-exclusion column. Additionally, AcrIIC1–GFP detection at a wavelength of 495 nm offers another indication of change in

AcrIIC1 elution volume. (bottom) SDS-PAGE gradient gel (4–20%) with protein samples from S200 elution peaks of chimeric Cas9 proteins incubated with AcrIIC1.
(D) Representative ITC trace for AcrIIC1 binding to the NmeHNH domain.
(E) Conservation of Cas9 mapped onto the approximate domain boundaries below using a non-redundant list of Cas9 orthologs from all Cas9 subtypes. HNH is highlighted in purple. Bar heights are proportional to protein identity, with yellow bars indicating highly conserved residues.

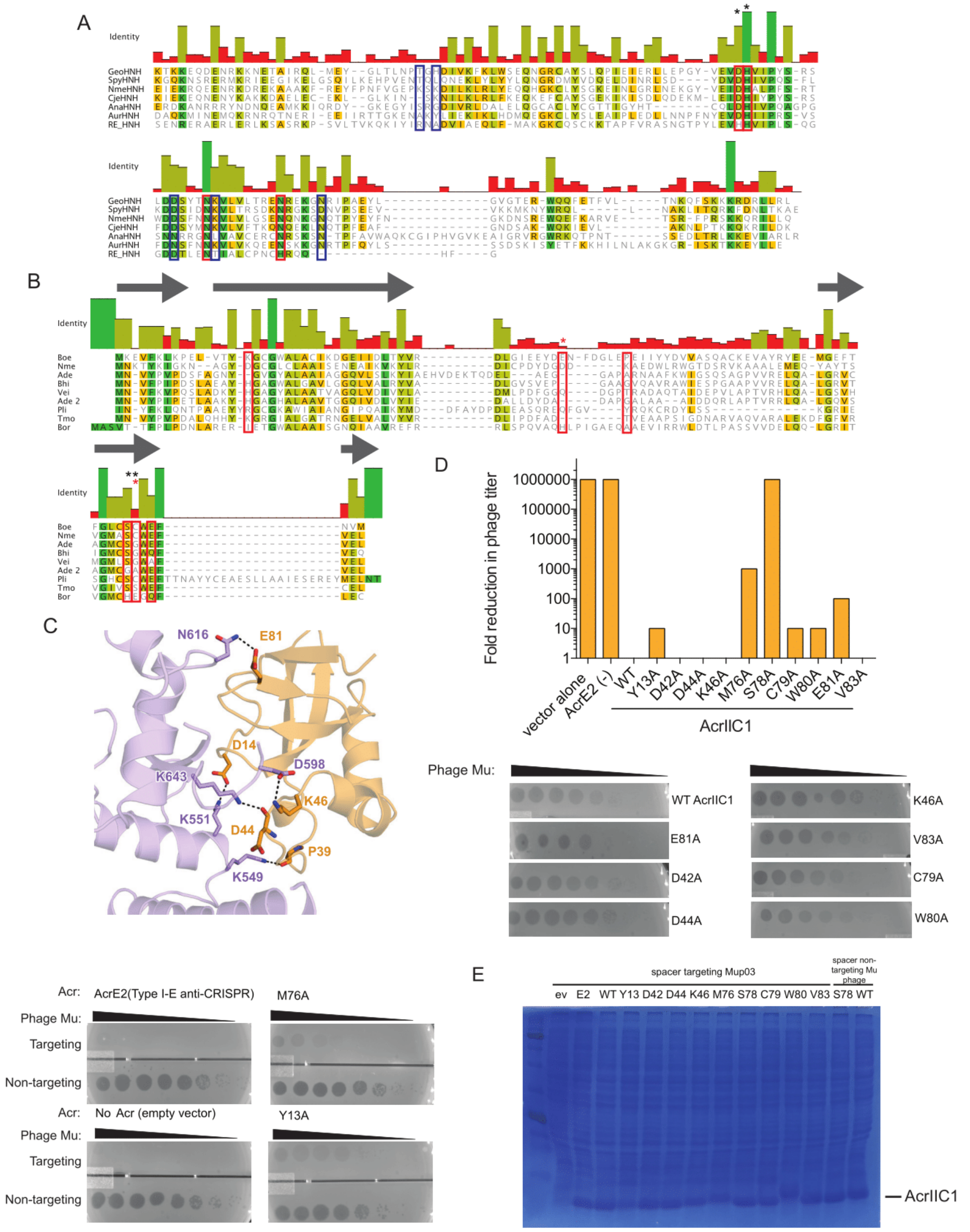


Figure 4.10 Alignment of HNH domains and AcrIIC1 orthologs and structure of AcrIIC1–HNH, related to Figure 4.4
(A) A multiple sequence alignment of selected HNH-domains. The multiple sequence alignment was generated using the extracted HNH domains of 6 Cas9 orthologs and the restriction enzyme from *Pseudoaltermonas lipolytica* (RE_HNH). Red boxes surround catalytic residues

and black asterisks indicate catalytic residues involved in AcrIIC1 binding. Blue boxes surround other residues involved in AcrIIC1 interaction.

(B) A multiple sequence alignment of AcrIIC1 using 11 AcrIIC1 orthologs. Highest degree of conservation occurs within a beta barrel (marked by a grey arrow). Red boxes surround selected residues involved in the binding between the NmeCas9 HNH domain and AcrIIC1$_{\mathrm{Nme}}$. Red asterisks indicate AcrIIC1 residues that interact with the backbone of HNH. Black asterisks indicate AcrIIC1 residues that interact with catalytic residues of the HNH domain.

(C) Charged residues (depicted as sticks) surrounding the active site of the HNH domain (purple) form ionic and hydrogen binding interactions (depicted as dotted black lines) with AcrIIC1 (orange).

(D) Fold reduction in phage titer in response to GeoCas9 targeting of phage Mu in the presence of AcrIIC1 mutants. One representative plate is shown for each mutant tested. 10-fold serial dilutions of phage Mu lysate were spotted lawns of bacteria expressing the indicated AcrIIC1 mutant. The fold reductions shown in the bar graph were qualitatively evaluated from inspecting three replicates of each experiment.

(E) SDS-PAGE gel showing the expression levels of WT and mutant AcrIIC1s used for experiments in panel D. The approximate mass of AcrIIC1 is indicated on the right (ev, empty vector; E2, AcrIE2).

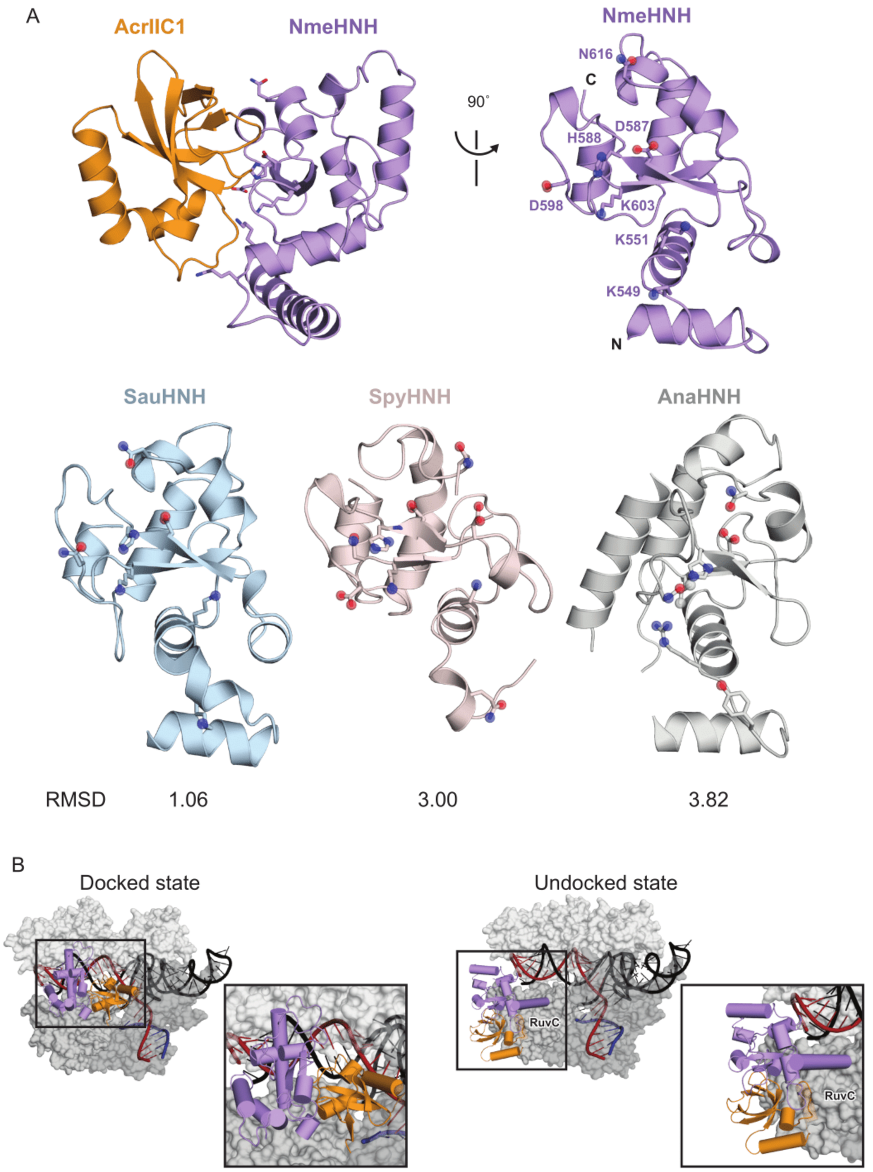


Figure 4.11 Conservation of the Cas9 HNH domain in the context of full length Cas9, related to Figure 4.4.

(A) Crystal structure of NmeCas9 HNH domain bound to AcrIIC1 (PDB: 5VGB). The HNH is rotated 90˚C to show the active site interface with labeled residues (depicted as sticks) involved in AcrIIC1 binding. (Bottom) Crystal structures of the HNH domains from three Cas9 orthologs (*Staphylococcus aureus*, PDB: 5CZZ; *Streptococcus pyogenes*, PDB: 4CMP; *Actinomyces naeslundii*, PDB: 4OGE). RMSD values were generated using super alignment in PyMol. Comparison of crystallized Cas9 orthologs reveals strong similarity between AcrIIC1 interacting residues of NmeCas9 and SauCas9 HNH domains.

(B) (Left) Model of AcrIIC1 inhibiting cleavage of both target and non-target strands. NmeCas9 HNH domain (purple) was modeled into a “docked” position using the dsDNA-bound SpyCas9 structure (PDB: 5F9R) as a reference. Placement of AcrIIC1 (orange) between the HNH domain and the target strand (red) prevents target cleavage and activation of the RuvC domain for non-target strand (dark blue) cleavage. The black box shows a zoomed-in view of the NmeHNH-AcrIIC1 complex. (Right) Model of AcrIIC1 clashing with the RuvC domain when the HNH domain is in an “undocked” conformation. NmeCas9 HNH was placed in a undocked conformation using a GeoCas9 Phyre model as a reference. Placement of AcrIIC1 indicates steric clashing with the RuvC domain, indicating that binding of AcrIIC1 to the HNH domain must position the HNH domain between the docked and undocked position. The black box shows a zoomed-in view of the HNH_{Nme}-$AcrIIC1_{Nme}$ complex.

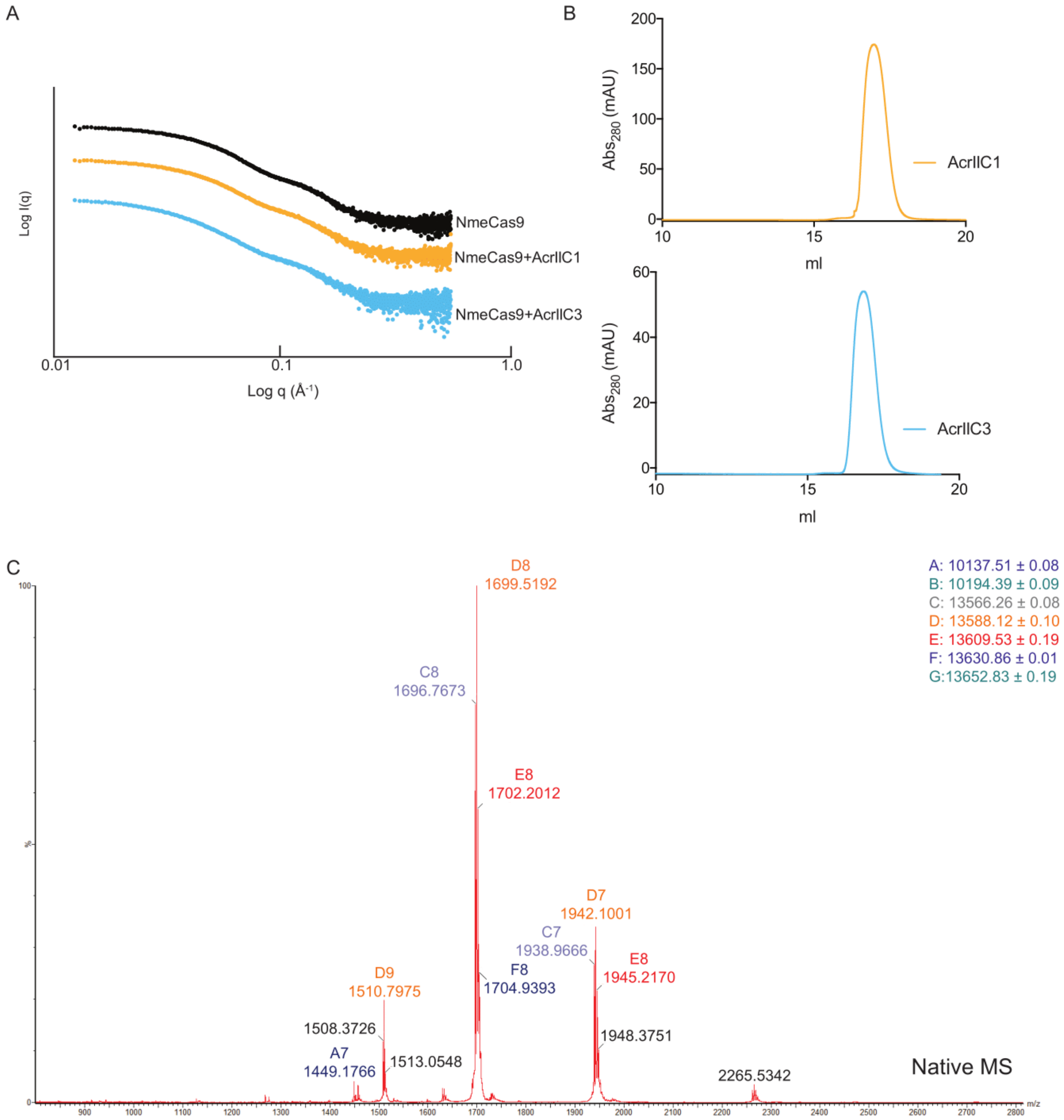


Figure 4.12 AcrIIC3 is monomeric and dimerizes NmeCas9, related to Figure 4.5.
(A) Small-angle X-ray scattering (SAXS) curves of NmeCas9, NmeCas9+AcrIIC1 and AcrIIC3.
(B) Analysis of AcrIIC1 (yellow, top) and AcrIIC3 (blue, below) on a Superdex 200 10/300 size exclusion column.
(C) Native mass spectrometry of AcrIIC3. The estimated masses from deconvoluting the charge series are identified in the top right corner.

	Nme HNH : Acr1	**Nme HNH: Acr1** (S-SAD)
Data collection		
Space group	$P\ 2_12_12_1$	$P\ 2_12_12_1$
Cell dimensions		
a, b, c (Å)	46.16, 60.49, 77.32	46.25, 60.45, 77.32
α, β, γ (°)	90, 90, 90	90, 90, 90
Wavelength	1.11583	2.06633
Resolution (Å) [a]	39.64–1.50 (1.55–1.50)	39.69–2.76 (2.92-2.76)
R_{merge} (%)	8.1 (195.2)	15.5 (52.6)
R_{pim} (%)	4.4 (1.2)	1.6 (12.4)
I/σ	13.25 (0.97)	34.5 (5.2)
$CC_{1/2}$	1.0 (0.48)	1.0 (0.81)
Completeness (%)	99.0 (100.0)	98.1 (87.1)
Redundancy	6.3 (6.0)	90.1 (18.3)
Refinement		
Resolution (Å)	39.64–1.50	
No. reflections		
Total	228,832	
Unique	35,654	
Free (%)	5	
$R_{\mathrm{work}}/R_{\mathrm{free}}$	16.4 / 20.9	
No. atoms		
Protein	1930	
Ligands/ion	23 (Glycerol, SO_4)	
Water	145	
Average B-factors (Å^2)		
Protein	29.6	
Ligand/ion	51.6	
Water	37.6	
R.m.s deviations		
Bond lengths (Å)	0.006	
Bond angles (º)	0.82	
Ramachandran		
Favored (%)	98.3	
Allowed (%)	1.3	

Outliers (%)	0.4

A single crystal was used for both the Native and S-SAD data sets

[a]Values in parentheses are for the highest resolution shell.

Table 4.1. Data collection and refinement statistics, related to Figure 4.4.

4.5 Discussion

We investigated the functions of two anti-CRISPR proteins, AcrIIC1 and AcrIIC3, and found that they block Cas9 activity by distinct mechanisms (Figure 4.6). AcrIIC1, an 85 amino acid protein, inactivates a wide range of type IIC Cas9 orthologs by binding to and conformationally restraining the conserved HNH domain. The direct interaction with essential catalytic residues of the HNH domain limits the opportunity for Cas9 to mutate and escape inhibition by AcrIIC1, explaining the phylogenetic propagation of this inhibitor to target multiple Cas9 orthologs. Intriguingly, AcrIIC1 traps Cas9 in an inactive but DNA-bound state, effectively converting wild-type Cas9 into a catalytically inactive dCas9. In contrast, the 116-amino acid AcrIIC3 binds specifically to the NmeCas9 enzyme to trigger dimerization and prevent DNA binding. Both of these mechanisms are different from that of the anti-CRISPR protein AcrIIA4, which acts as a DNA mimetic that prevents DNA binding by occupying the PAM-recognition site within a small subset of related type IIA Cas9 orthologs (Dong et al., 2017; Shin et al., 2017).

The CRISPR inhibition mechanisms determined in this study concur with two general strategies observed previously for blocking interference proteins in both type I and type II CRISPR systems. The first and most common mechanism is to target the CRISPR surveillance complex by disrupting DNA binding (AcrIIC3, AcrIIA4, AcrF1, AcrF2; (Bondy-Denomy et al., 2015; Chowdhury et al., 2017; Dong et al., 2017; Pawluk et al., 2016b; Rauch et al., 2017; Shin et al., 2017). The second is to target nucleases or nuclease domains, thereby allowing DNA binding but not cleavage (AcrIIC1, AcrF3; (Bondy-Denomy et al., 2015; Wang et al., 2016). Of the currently studied Acrs, the strategy of inhibiting crRNA assembly with Cas proteins has yet to be found. The absence of this mechanism is possibly because this strategy would be unable to interfere with CRISPR nucleases that were assembled prior to infection. Nonetheless, many CRISPR systems are tightly regulated and are often activated in response to cell density and other factors (Høyland-Kroghsbo et al., 2016; Patterson et al., 2016, 2017). Inhibition of crRNA binding could be an effective Acr method to inhibit those systems where RNP assembly coincides with phage infection, and such inhibitors may yet be discovered. Moreover, only Acrs that target CRISPR interference proteins have been found, despite the fact that the methods currently used to identify Acrs are capable of finding inhibitors of spacer acquisition and crRNA processing. Nonetheless, it is likely that in cases where interference proteins are linked to other steps in CRISPR adaptive immunity, such as acquisition for type IIA Cas9 or RNA processing for Cas12a, such Acrs exist (Fonfara et al., 2016; Heler et al., 2015; Wei et al., 2015).

The mechanism of HNH domain binding by AcrIIC1 is particularly interesting for several reasons. First, the high structural similarity of the Cas9 HNH domain across Cas9 orthologs implies that AcrIIC1-type inhibitors may be more widespread than current analysis has identified. There also may be other Acrs that have converged on this mechanism. Second, type IIC Cas9 orthologs are capable of tracrRNA- and PAM-independent DNA cleavage of single stranded DNA catalyzed by the HNH-domain (Ma et al., 2015; Zhang et al., 2015). In addition to inhibition of double-stranded DNA cleavage, AcrIIC1 would also be able to inhibit this single-stranded cleavage activity, whereas inhibitors of PAM binding such as AcrIIA4 may not be able to. Third, the fact that AcrIIC1 traps Cas9-guide RNA complexes in a catalytically inactive but DNA-bound state is consistent with additional roles for this inhibitor that include gene regulation rather than genome protection. Notably, efforts to engineer regulatory forms of Cas9 have utilized dCas9, a catalytically inactive mutant of the enzyme that retains RNA-programmed DNA binding activity. It would be exciting to determine whether bacteria natively employ AcrIIC1 to repurpose Cas9 as a gene regulator in cells. Whether or not this occurs in nature, it is an enticing possibility that AcrIIC1 can be employed in gene editing applications to obviate the need to generate separate dCas9 enzymes for gene regulatory purposes (Gilbert et al., 2014; Mali et al., 2013; Qi et al., 2013). Finally, we note that the HNH fold is not unique to Cas9 but is in fact common to many bacterial restriction enzymes (Vasu and Nagaraja, 2013). This raises the possibility that in addition to targeting CRISPR-based adaptive immunity, AcrIIC1 also inhibits restriction enzymes. In line with this hypothesis, we observed that no Cas9 ortholog is present in Pseudoaltermonas lipolytica despite the presence of an AcrIIC1. A blastp search of the P. lipolytica genome for Cas9 revealed an HNH restriction enzyme with homology to the Cas9 HNH domain (20% sequence identity of HNH domains). It may be that AcrIIC1 has evolved to inhibit both adaptive (CRISPR) and innate (restriction) immune systems by targeting this conserved protein domain.

Given the rapid evolution and resourcefulness of phage, it is likely that Acrs are much more widespread than is currently known. As the toolbox of proteins used to edit genomes continues to expand to include other Class 2 CRISPR systems, discovery of new Acrs can serve as potent tools to control these new systems. Continued analysis of the abundance of Acrs as well as their mechanisms will provide unique opportunities to regulate and disable CRISPR systems, and in the process illuminate the influence of Acrs on CRISPR diversity.

4.6 Acknowledgements

We thank G. Meigs and the 8.3.1 beamline staff at the Advanced Light Source for assistance with data collection and members of the Doudna laboratory for comments and discussions. The 8 .3.1 beamline is supported by UC Office of the President, Multicampus Research Programs and Initiatives grant MR-15-328599 and Program for Breakthrough Biomedical Research, which is partially funded by the Sandler Foundation. J.A.D. is an Investigator of the Howard Hughes Medical Institute. This research was supported in part by the Allen Distinguished Investigator Program, through The Paul G. Allen Frontiers Group, the National Science Foundation (MCB-1244557 to J.A.D.), the National Institutes of Health (GM115911 to E.J.S.) and the Canadian Institutes of Health Research (MOP-130482 to A.R.D and MOP-136845 to K.L.M.). L.B.H., K.W.D. and J.S.C. are supported by US National Science Foundation Graduate Research

4.7 Author Contributions

L.B.H., K.W.D., E.M., K.L.M, J.A.D. and A.R.D. designed experiments. L.B.H, A.E., N. A., and E.J.S. designed and performed cell-based assays. L.B.H, K.W.D., J.S.C. and J.C.C. purified proteins. L.B.H. designed and purified RNA and DNA substrates. L.B.H and E.M conducted in vitro cleavage and binding experiments. K.W.D and E.M. assembled complex and set trays for protein crystallization. K.W.D and L.B.H acquired diffraction data. P.J.K. and K.W.D. determined the experimental phases of crystallographic data and traced the initial models and K.W.D. completed model building. L.B.H. and G.J.K. conducted and analyzed SAXS experiments. J. L. conducted electron microscopy experiments. A.R.D, K.L.M and B.G. designed and executed phage plaquing experiments. L.B.H, K.W.D, and J.A.D. wrote the manuscript. All authors revised and agreed to the final manuscript.

Chapter 5

New CRISPR-Cas Systems from Uncultivated Microbes

Part of the work presented in this chapter has previously been published in the following manuscript: Burstein, D., Harrington, L. B.*, Strutt, S. C.*, & Probst, A. J. (2017). New CRISPR-Cas systems from uncultivated microbes. *Nature Publishing Group*, *542*(7640), 237–241.

5.1 Abstract

CRISPR-Cas adaptive immune systems have revolutionized genome engineering by providing programmable enzymes capable of site-specific DNA cleavage. However, current CRISPR-Cas technologies are based solely on systems from cultured bacteria, leaving untapped the vast majority of enzymes from organisms that have not been isolated. Here, using cultivation-independent genome-resolved metagenomics, we identified new CRISPR-Cas systems, including the first reported Cas9 in the archaeal domain of life. This divergent Cas9 enzyme was found in little-studied nanoarchaea as part of an active CRISPR-Cas system. In bacteria, we discovered two previously unknown systems, CRISPR-CasX and CRISPR-CasY, which are among the most streamlined systems yet identified. Notably, all required functional components were identified by metagenomics, enabling validation of robust in vivo RNA-guided DNA interference activity in E. coli. Interrogation of environmental microbial communities combined with in vivo experiments allows access to an unprecedented diversity of genomes whose content will expand the repertoire of microbe-based biotechnologies.

5.2 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 viruses1,2. CRISPR-Cas systems consist of Cas proteins that 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 sequence3,4. The programmable nature of these minimal systems has enabled their use as a versatile technology that is broadly revolutionizing biology and clinical medicine5.

Metagenomics, the sequencing of DNA extracted from natural microbial communities, provides access to the genetic material of organisms that have never been studied in the laboratory. Metagenomic reads are assembled into contiguous DNA sequences (contigs) and the contigs clustered into genome bins6. The resulting genomes encode an enormous variety of proteins for which functional insights are lacking7,8. Since this approach offers an opportunity to investigate both microbial defense systems and the entities they target (e.g., phage and plasmids) it is ideal to explore both diversity and evolution of CRISPR-Cas systems. Here, we analyzed terabase-scale metagenomic datasets, seeking class 2 CRISPR-Cas systems that are not represented among cultured organisms. We searched for large uncharacterized genes in proximity to a CRISPR array and cas1, the universal CRISPR integrase9–11, in a dataset of more than 155 million proteins from groundwater, sediment, acid mine drainage biofilms, soil, infant gut, and other microbial communities. We identified the first Cas9 proteins in domain Archaea, and discovered in uncultivated bacteria two new CRISPR-Cas systems, which we refer to as CRISPR-CasX and CRISPR-CasY (Fig. 5.1). Notably, both the archaeal Cas9 and CasY are encoded exclusively in the genomes of organisms from lineages with no known isolated representatives.

5.3 Materials and Methods

3.3.1 Metagenomics and metatranscriptomics

Metagenomic samples from three different sites were analyzed: (1) Acid mine drainage (AMD) samples collected between 2006 and 2010 from the Richmond Mine, Iron Mountain, California42,43 (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, Colorado7,29. (3) Groundwater collected in 2009 and 2014 from Crystal Geyser, a cold, CO2-driven geyser on the Colorado Plateau in Utah44.

For the AMD data, DNA extraction methods and short read sequencing were reported by Denef and Banfield (2012)42 and Miller et al. (2011)43. For the Rifle data, DNA extraction, sequencing, assembly, and genome reconstruction were described by Anantharaman et al. (2016)29 and Brown et al. (2015)7. For samples from Crystal Geyser, methods follow those described by Probst et al. (2016)44 and Emerson et al. (2016)45. Rifle metatranscriptomic data was used from the data reported by Brown et al. (2015)7.

Briefly, DNA was extracted from samples using the PowerSoil DNA Isolation Kit (MoBio Laboratories Inc., Carlsbad, CA, USA). RNA was extracted from 0.2 μm filters collected from six 2011 Rifle groundwater samples. Following RNA extraction using the Invitrogen TRIzol reagent, DNA removal was done 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 Metatrancriptomic cDNA on 5500XL SOLiD platform after emulsion clonal bead amplification using the SOLiD EZ Bead system (Life Technologies). For the newly reported Crystal Geyser data and reanalysis of the AMD data, sequences were assembled using IDBA-UD46. DNA and RNA (cDNA) read-mapping used to determine sequencing coverage and gene expression, respectively, was performed using Bowtie247. Open reading frames (ORFs) were predicted on assembled scaffolds using Prodigal48. Scaffolds from the Crystal Geyser dataset were binned on the basis of differential coverage abundance patterns using a combination of ABAWACA7, ABAWACA2 (https://github.com/CK7) Maxbin249, and tetranucleotide frequency using Emergent Self-Organizing Maps (ESOM)50. Genomes were manually curated using % GC content, taxonomic affiliation, and genome completeness. Scaffolding errors were corrected using ra2.py (https://github.com/christophertbrown).

5.3.2 CRISPR-Cas computation analysis

The assembled contigs from the various samples were scanned for known Cas proteins using Hidden Markov Model (HMMs) profiles, which were built using the

HMMer suite39, based on alignments from Makarova et al. (2015)3 and Shmakov et al. (2015)4. CRISPR arrays were identified using a local version of the CrisprFinder software51. Loci that contained both Cas1 and a CRISPR array were further analyzed if one of the ten ORFs adjacent to the cas1 gene encoded for an uncharacterized protein larger than 800 aa, and no known cas interference genes were identified on the same contig. These large proteins were further analyzed as potential class 2 Cas effectors. The potential effectors were clustered to protein families based on sequence similarities using MCL52. 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 make sure that the protein families are indeed new, known homologs were searched using BLAST53 against NCBI's non-redundant (nr) and metagenomic (env_nr) protein databases, as well as HMM searches against the UniProt KnowledegeBase39,54. 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-suite32. High scoring HHpred hits were used to infer domain architecture based on comparison to solved crystal structures55,56, and secondary structure that was predicted by JPred457. The HMM database, including the newly discovered Cas proteins are available in Supplementary Data 6.

Spacer sequences were determined from the assembled data using CrisprFinder51. CRASS58 was used to locate additional spacers in short DNA reads of the relevant samples. Spacer targets (protospacers) were then identified by BLAST53 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). Protospacer adjacent motifs (PAMs) were identified by aligning regions flanking the protospacers and visualized using WebLogo59. In cases that one spacer had multiple putative protospacers with different composition of flanking nucleotides, each distinct combination of protospacer and downstream nucleotides was taken into account for the logo calculation. RNA structures were predicted using mFold60. Average nucleotide identity was computed with the pyani Python module (https://github.com/widdowquinn/pyani), using the Mummer61 method. CRISPR array diversity was analyzed 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 Makarova et al. (2015)3 and Shmakov et al. (2015)4. A non-redundant set was compiled by clustering together proteins with ≥ 90% identity using CD-HIT62. Alignments were produced with MAFFT63, and maximum-likelihood phylogenies were constructed using RAxML64 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 v365.

5.3.3 Generation of heterologous plasmids

Metagenomic contigs were made into minimal CRISPR interference plasmids by removing proteins associated with acquisition for CRISPR-CasX and reducing the size of the CRISPR array for both CRISPR-CasX and CRISPR-CasY. The minimal locus was synthesized as Gblocks (Integrated DNA Technology). Native promoters were used, with the exception of the overexpression of CasY.1 and expression of the crRNA alone or sgRNA for CasX in figure 3c where the J23119 constitutive promoter was used. The minimal CRISPR loci were assembled using Gibson Assembly66 into a plasmid with a p15A origin of replication and chloramphenicol resistance gene. Detailed plasmid maps are available at the links provided in Supplementary Table 2.

5.3.4 PAM depletion assay

PAM depletion assays were conducted as previously described with modification67. Plasmid libraries containing randomized PAM sequences were assembled by annealing a DNA oligonucleotide containing a target with a 7 nt randomized PAM region with a primer (Supplementary Table 2) and extended with Klenow Fragment (NEB). The double stranded DNA was digested with EcoRI and NcoI and ligated into a pUC19 backbone. The ligated library was transformed into E. coli DH5α and >108 cells were harvested and the plasmids extracted and purified. 200 ng of the pooled library was transformed into electrocompetent E. coli harboring a CRISPR locus or a control plasmid with no locus. The transformed cells were plated on selective media containing carbenicillin (100 mg L-1) and chloramphenicol (30 mg L-1) for 30 hours at 25°C. Plasmid DNA was extracted and the PAM sequence was amplified with adapters for Illumina sequencing. The 7 nt PAM region was extracted and PAM frequencies calculated for each 7 nt sequence. PAM sequences depleted above the specified threshold were used to generate a sequence logo with WebLogo59.

5.3.5 Plasmid Interference

Putative targets identified from metagenomic sequence analysis or PAM depletion assays were cloned into a pUC19 plasmid. 10 ng of target plasmid were transformed into electrocompetent E. coli (NEB Stable) containing the CRISPR loci plasmid. CasX.1 was used for the plasmid interference assays under control of native promoters or using a strong heterologous promoter (J23119) for sgRNA and crRNA expression. CasY.1 was put under the control of a heterologous promoter (J23119) for these assays. Cells were recovered for 2 hrs at 25°C in Super Optimal Broth (SOB) and an appropriate dilution was plated on selective media. Plates were incubated at 25°C and colony forming units were counted. All plasmid interference experiments were performed in triplicate and electrocompetent cells were prepared independently for each replicate.

5.3.6 Northern Blots

E. coli containing the deltaproteobacteria CasX CRISPR locus was grown to OD_{600}=1 at 25°C in SOB media. RNA was extracted by warm phenol extraction, separated on 10% denaturing polyacrylamide gel and blotted as previously described by Zhang *et al.* (2013)(Zhang et al., 2013).

5.4 Results and Discussion

5.4.1 CRISPR-CasX is a novel dual-RNA-guided CRISPR system

In addition to Cas9, only three families of class 2 Cas effector proteins have been discovered and experimentally validated: Cpf1, C2c1, and C2c2(Abudayyeh et al., 2016a; Shmakov et al., 2015; Zetsche et al., n.d.). Another gene, *c2c3*, which was identified only on small DNA fragments, has been suggested to also encode such a protein family(Shmakov et al., 2015). We hypothesized that other distinct types of effector proteins might exist within uncultivated microbes whose genomes were reconstructed from our metagenomic datasets. Indeed, a new type of class 2 CRISPR-Cas system was found in the genomes of two bacteria recovered repeatedly from groundwater and sediment samples(Anantharaman et al., 2016). This newly described system includes Cas1, Cas2, Cas4 and an uncharacterized ~980 aa protein that we refer to as CasX. The high conservation (68% protein sequence identity, Supplementary Data 1) of this protein in two organisms belonging to different phyla, Deltaproteobacteria and Planctomycetes, suggests a recent cross-phyla transfer(Burstein et al., 2016; Godde & Bickerton, 2006). The CRISPR arrays associated with each CasX had highly similar repeats (86% identity) of 37 nt, spacers of 33–34 nt, and a putative tracrRNA between the Cas operon and the CRISPR array (Fig. 1b, Extended Data Table 1). BLAST searches revealed only weak similarity (e-value > 1×10^{-4}) to transposases, with similarity restricted to specific regions of the CasX C-terminus. Distant homology detection(Remmert, Biegert, Hauser, & Söding, 2012) and protein modeling(Kelley, Mezulis, Yates, Wass, & Sternberg, 2015) identified a RuvC domain near the CasX C-terminal end, with organization reminiscent of that found in type V CRISPR-Cas systems (Extended Data Fig. 6). The rest of the CasX protein (630 N-terminal amino acids) showed no detectable similarity to any known protein, suggesting this is a novel class 2 effector. The combination of tracrRNA and separate Cas1, Cas2 and Cas4 proteins is unique among type V systems. Further, CasX is considerably smaller than any known type V proteins: 980 aa compared to a typical size of more than 1,200 aa for Cpf1, C2c1 and C2c3.

We wondered whether CasX would be capable of RNA-guided DNA targeting analogous to Cas9 and Cpf1 proteins. To test this possibility, we synthesized a plasmid encoding a minimal CRISPR-CasX locus including *casX*, a short repeat-spacer array and intervening noncoding regions. We found that when expressed in *E. coli*, this minimal locus blocked transformation by a plasmid bearing a target sequence identified by metagenomic analysis (Fig. 3a–c, Extended Data Fig. 7). Furthermore, interference with transformation occurred only when the spacer sequence in the mini-locus matched the protospacer sequence in the plasmid target.

To identify a PAM sequence for CasX, we repeated the transformation assay in *E. coli* using a plasmid containing either a 5′ or 3′ randomized sequence adjacent to the target site. This analysis revealed a stringent preference for the sequence ‘TTCN’ located 5′ of the protospacer sequence (Fig. 3d). No 3′ PAM preference was observed (Extended

Data Fig. 7). Consistent with this finding, we observed that ‘TTCA’ is the sequence found upstream of the putative Deltaproteobacteria CRISPR-CasX protospacer that was identified in the environmental samples. Notably, both CRISPR-CasX loci share the same PAM sequence, in line with their high degree of protein sequence homology.

Examples of both single-RNA and dual-RNA guided systems exist among type V CRISPR loci. We used environmental RNA (metatranscriptomic) data to determine whether CasX requires a tracrRNA for DNA targeting activity. This analysis revealed a non-coding RNA transcript with a sequence complementary to the CRISPR repeat encoded between the Cas2 open reading frame and the CRISPR array (Fig. 4a). To check for expression of this non-coding RNA in *E. coli* expressing the CasX locus, Northern blots were conducted against this transcript in both directions (Extended Data Fig. 7). The results showed expression of a transcript on the same stand as the CasX ORF of about 110 nt with a more heterogeneous band at about 60–70 nt, suggesting that the leader sequence for the CRISPR array lies between the tracrRNA and the array. Transcriptomic mapping further suggests that the CRISPR RNA (crRNA) is processed to include ~23 nt of the repeat and 20 nt of the adjacent spacer, similar to the crRNA processing that occurs in CRISPR-Cas9 systems(Elitza Deltcheva et al., 2011; Martin Jinek et al., 2012a) (Fig. 4a). To determine the dependence of CasX activity on the putative tracrRNA, we deleted this region from the minimal CRISPR-CasX locus described above, and repeated the plasmid interference assays. Deletion of the putative tracrRNA-encoding sequence from the CasX plasmid abolished the robust transformation interference observed in its presence (Fig. 4c). This putative tracrRNA was joined with the processed crRNA using a tetraloop to form a single-guide RNA (sgRNA)(Martin Jinek et al., 2012a). While expression using a heterologous promoter of the crRNA alone or a shortened version of the sgRNA did not have any significant plasmid interference, expression of the full-length sgRNA conferred resistance to plasmid transformation (Fig. 4c). Together, these results establish CasX as a new functional DNA-targeting, dual-RNA guided CRISPR enzyme.

5.4.2 CRISPR-CasY, a system found exclusively in bacterial lineages lacking isolates

We identified another new class 2 Cas protein encoded in the genomes of certain candidate phyla radiation (CPR) bacteria(Brown et al., 2015a; Hug et al., 2016) (Fig. 1, Extended Data Table 1). These bacteria typically have small cell sizes (based on cryo-TEM data and enrichment via filtration), very small genomes and a limited biosynthetic capacity, indicating they are most likely symbionts(Brown et al., 2015a; Kantor et al., 2013; Luef et al., 2015; Nelson & Stegen, 2015; Rinke et al., 2013). The new ~1,200 aa Cas protein, which we named CasY, appears to be part of a minimal CRISPR-Cas system that includes Cas1 and a CRISPR array (Fig. 5a). Most of the CRISPR arrays have unusually short spacers of 17–19 nt, but one system, which lacks Cas1 (CasY.5), has longer spacers (27–29 nt). No predicted tracrRNA was detected in the vicinity of CRISPR-CasY, based on partial complementarity to the repeat sequences; however, we had insufficient metatranscriptomic data mapped to the CasY loci to detect potential tracrRNA sequences. Thus, we cannot exclude the dependence of CasY on a tracrRNA for robust interference from the available data.

The six examples of CasY proteins we identified had no significant sequence similarity to any protein in public databases. A sensitive search using profile models (HMMs(Finn, Clements, & Eddy, 2011)) built from published Cas proteins(Kira S.

Makarova et al., 2015; Shmakov et al., 2015) indicated that four of the six CasY proteins had local similarities (e-values $4×10^{-11}$– $3×10^{-18}$) to C2c3 in the C-terminal region overlapping the RuvC domains and a small region (~45 aa) of the N-terminal region (see Extended Data Fig. 6). C2c3 proteins are putative type V Cas effectors(Shmakov et al., 2015) that were identified on short contigs with no taxonomic affiliation, and have not been validated experimentally. Like CasY, the C2c3s were found next to arrays with short spacers and Cas1, but with no other Cas proteins. Notably, two of the CasY proteins identified in the current study had no significant similarity to C2c3s, despite sharing significant sequence similarity (best Blast hits: e-values $6×10^{-85}$, $7×10^{-75}$) with the other CasY proteins (Supplementary Data 2).

Given the low homology of CRISPR-CasY to any experimentally validated CRISPR loci, we wondered whether this system confers RNA-guided DNA interference, but due to the short spacer length we did not have reliable information about a possible PAM motif that might be required for such activity. To work around this, the entire CRISPR-CasY.1 locus was synthesized with a shortened CRISPR array and introduced into *E. coli* on a plasmid vector. These cells were then challenged in a transformation assay using a target plasmid with a sequence matching a spacer in the array and containing an adjacent randomized 5′ or 3′ region to identify a possible PAM. Analysis of transformants revealed depletion of sequences containing a 5′ TA directly adjacent to the targeted sequence (Fig. 5b). Using this identified PAM sequence, the CasY.1 locus was overexpressed using a heterologous promoter and was tested against plasmids containing single PAMs. Plasmid interference was strongest in the presence of a target containing the identified 5′ TA PAM sequence (Fig. 5c). Thus, we conclude that CRISPR-CasY has DNA interference activity.

5.4.3 Discussion

We discovered new class 2 CRISPR-Cas adaptive immune systems in genomes from uncultivated bacteria and archaea. Evolutionary analysis of Cas1 (Fig. 6a), which is universal to active CRISPR loci, suggests that the archaeal Cas9 system described here does not clearly fall into any existing type II subtype. The Cas1 phylogeny (as well as the existence of *cas4*) affiliate it as a type II-B system(Chylinski et al., 2014b; Kira S. Makarova et al., 2015), yet the Cas9 sequence is more similar to type II-C proteins (Extended Data Fig. 8, Supplementary Data 3). Thus, the archaeal type II system may have arisen as a fusion of type II-C and II-B systems (Fig. 6b). Cas1 phylogenetic analyses also indicate that the Cas1 from the CRISPR-CasX system is distant from any other known type V system. Type V systems have been suggested to be the result of the fusion of a transposon with the adaptation module (Cas1–Cas2) from an ancestral type I system(Shmakov et al., 2015). We therefore hypothesize that the CRISPR-CasX system emerged following a fusion event different from those that gave rise to the previously described type V systems. Strikingly, both CRISPR-CasY and the putative C2c3 systems seem to lack Cas2, a protein that is considered essential for integrating DNA into the CRISPR locus(Nuñez et al., 2014a; Yosef et al., 2012). Given that all CRISPR-Cas systems are thought to be descendants of an ancestral type I system that contained both Cas1 and Cas2(Shmakov et al., 2015), CRISPR-CasY and C2c3 systems may either have different ancestry than the rest of the CRISPR-Cas systems, or alternatively, Cas2 might have been lost during their evolutionary history. It remains to be seen whether these new type V systems are functional for acquisition.

The discovery of Cas9 in archaea and two previously unknown bacterial CRISPR-Cas systems was enabled by access to extensive DNA and RNA sequence datasets obtained from complex natural microbial communities. In the case of CasX and CasY, genome context was critical to prediction of functions that would not have been evident from unassembled sequence information. Further, the identification of a putative tracrRNA, as well as targeted sequences uncovered through analysis of the metagenomic data, guided the functional testing. Interestingly, we discovered some of the most compact CRISPR-Cas loci identified to date in organisms with very small genomes. A consequence of small genome size is that these organisms likely depend on other community members for basic metabolic requirements, and thus they have remained largely outside the scope of traditional cultivation-based methods. The small number of proteins that are required for interference, and their relatively short length make these minimal systems especially valuable for the development of new genome editing tools. Importantly, we show that metagenomic discoveries related to CRISPR-Cas systems are not restricted to *in silico* observations, but can be introduced into an experimental setting where their function can be tested. Given that virtually all environments where life exists can now be probed by genome-resolved metagenomic methods, we anticipate that the combined computational-experimental approach will greatly expand the diversity of known CRISPR-Cas systems, enabling new technologies for biological research and clinical applications.

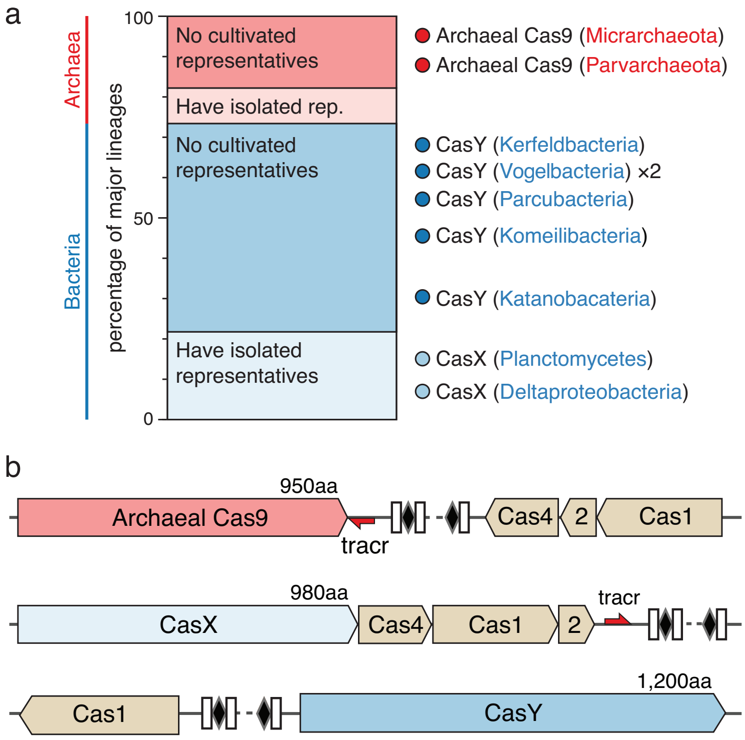

a
Archaea
Bacteria
percentage of major lineages
100
50
0
No cultivated representatives
Have isolated rep.
No cultivated representatives
Have isolated representatives
● Archaeal Cas9 (Micrarchaeota)
● Archaeal Cas9 (Parvarchaeota)
● CasY (Kerfeldbacteria)
● CasY (Vogelbacteria) ×2
● CasY (Parcubacteria)
● CasY (Komeilibacteria)
● CasY (Katanobacateria)
○ CasX (Planctomycetes)
○ CasX (Deltaproteobacteria)
b
950aa
Archaeal Cas9
tracr
Cas4
2
Cas1
980aa
CasX
Cas4
Cas1
2
tracr
Cas1
1,200aa
CasY

Figure 5.1 I Novel identified CRISPR-Cas systems from uncultivated organisms. a, Ratio of major lineages with and without isolated representatives in all bacteria and archaea, based on data of Hug *et al*. (2016)(Hug et al., 2016). The results highlight the massive scale of as-yet little

investigated biology in these domains. Archaeal Cas9 and the novel CRISPR-CasY were found exclusively in lineages with no isolated representatives. **b,** Locus organization of the newly discovered CRISPR-Cas systems.

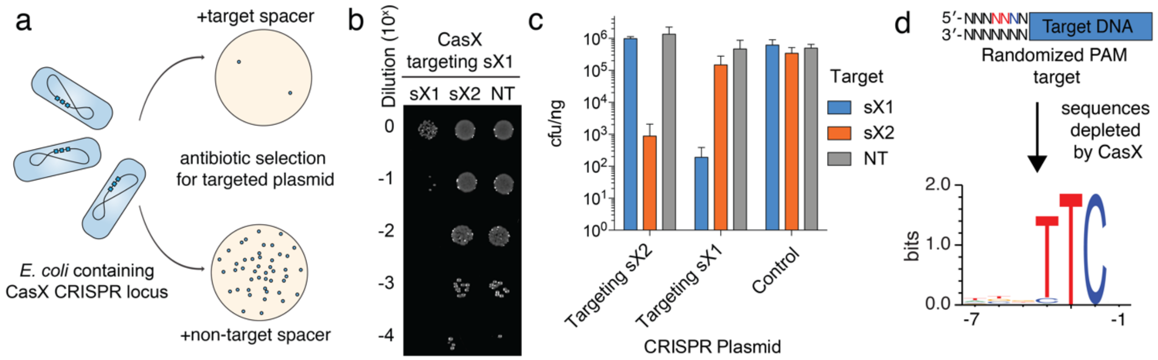


Figure 5.2 | CasX mediates programmable DNA interference in *E. coli*. a, Diagram of CasX plasmid interference assays. *E. coli* expressing a minimal CasX locus is transformed with a plasmid containing a protospacer matching the sequence in the CRISPR array (target) or plasmid containing a non-matching sequence (non-target). After being transformed, cultures are plated and colony-forming units (cfu) are quantified. **b,** Serial dilution of *E. coli* expressing the Planctomycetes CasX locus with spacer 1 (sX1) and transformed with the specified target (sX1, CasX protospacer 1; sX2, CasX protospacer 2; NT, non-target). **c,** Plasmid interference by Deltaproteobacteria CasX, using the same spacers and targets as in (**b**). Experiments were conducted in triplicate and mean ± s.d. is shown. **d,** PAM depletion assays for the Planctomycetes CasX locus expressed in *E. coli.* PAM sequences depleted greater than 30-fold compared to a control library were used to generate the sequence logo (see also Extended Data Fig. 7).

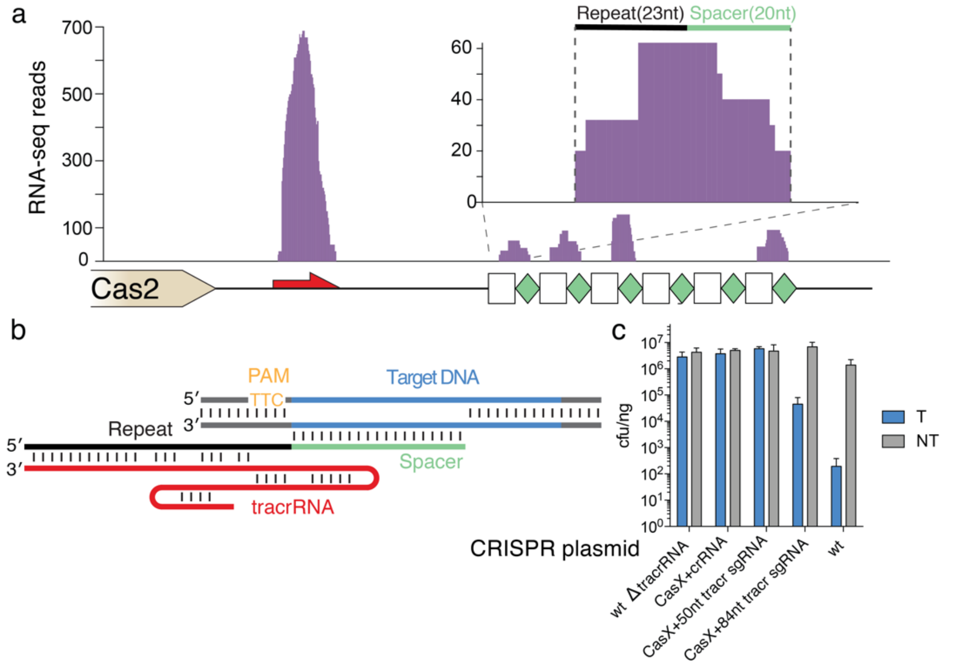


Figure 5.3 | CasX is a dual-guided CRISPR complex. a, Mapping of environmental RNA sequences (metatranscriptomic data) to the CasX CRISPR locus diagramed below

(red arrow, putative tracrRNA; white boxes, repeat sequences; green diamonds, spacer sequences). Inset shows detailed view of the first repeat and spacer. **b,** Diagram of CasX DNA interference. **c,** Results of plasmid interference assays with the putative tracrRNA knocked out of the CasX locus and CasX coexpressed with a crRNA alone, a truncated sgRNA or a full length sgRNA (T, target; NT, non-target). Experiments were conducted in triplicate and mean ± s.d. is shown.

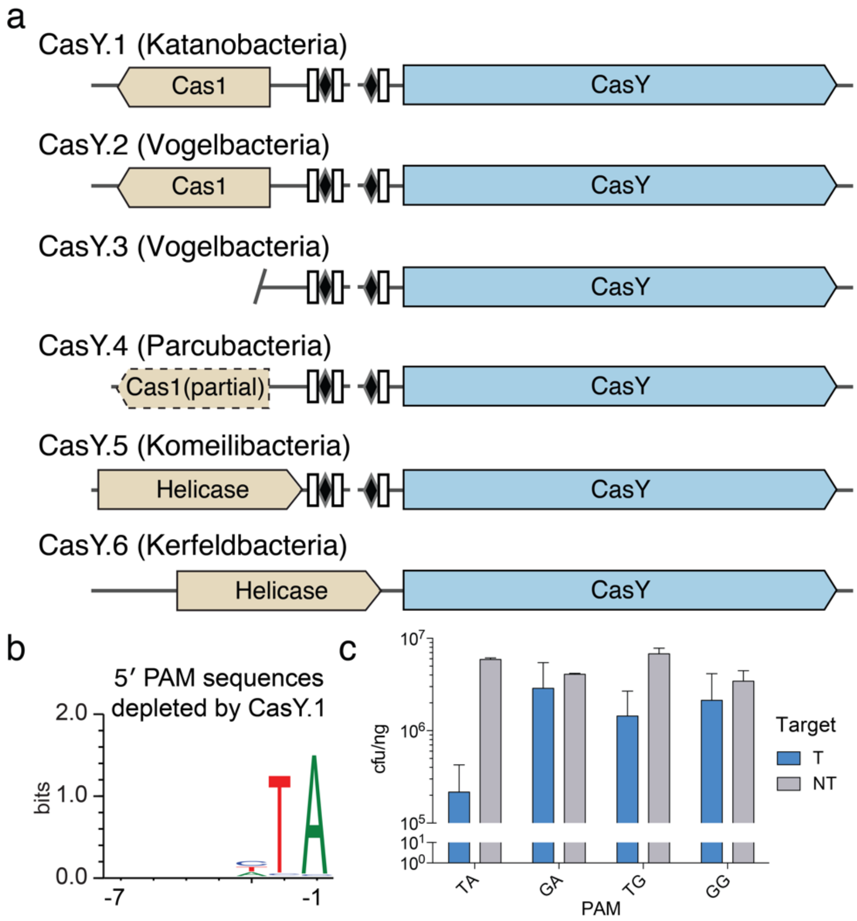


Figure 5.4 | Expression of a CasY locus in *E. coli* is sufficient for DNA interference. **a,** Diagrams of CasY loci and neighboring proteins (see also Extended Data Table 1). **b,** Sequence logo of the 658 5′ PAM sequences depleted greater than 3-fold by CasY relative to a control library. **c,** Plasmid interference by *E. coli* expressing CasY.1 and CRISPR array expressed with a heterologous promoter and transformed with targets containing the indicated PAM. Experiments were conducted in triplicate and mean ± s.d. is shown.

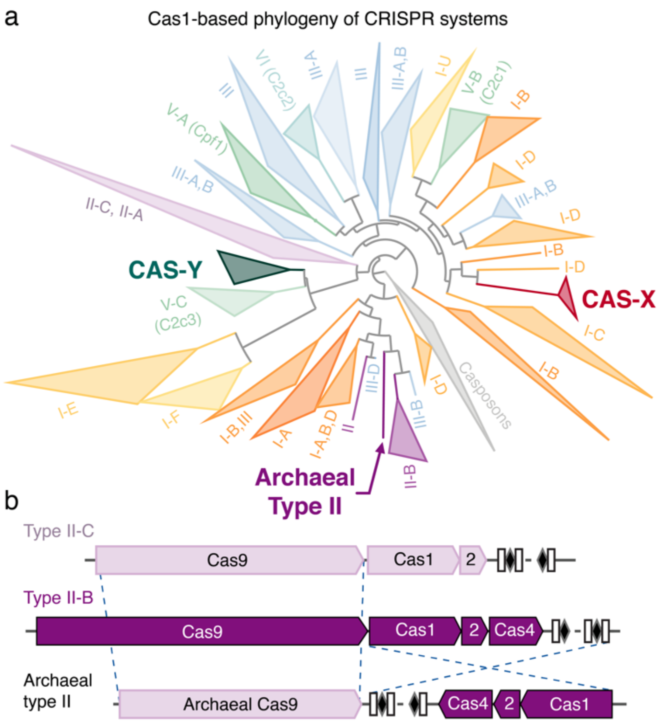


Figure 5.5 | The newly identified CRISPR-Cas in context of known systems. a, Simplified phylogenetic tree of the universal Cas1 protein. CRISPR types of known systems are noted on the wedges and branches; the newly described systems are in bold. Detailed Cas1 phylogeny is provided in Supplementary Data 4. **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.

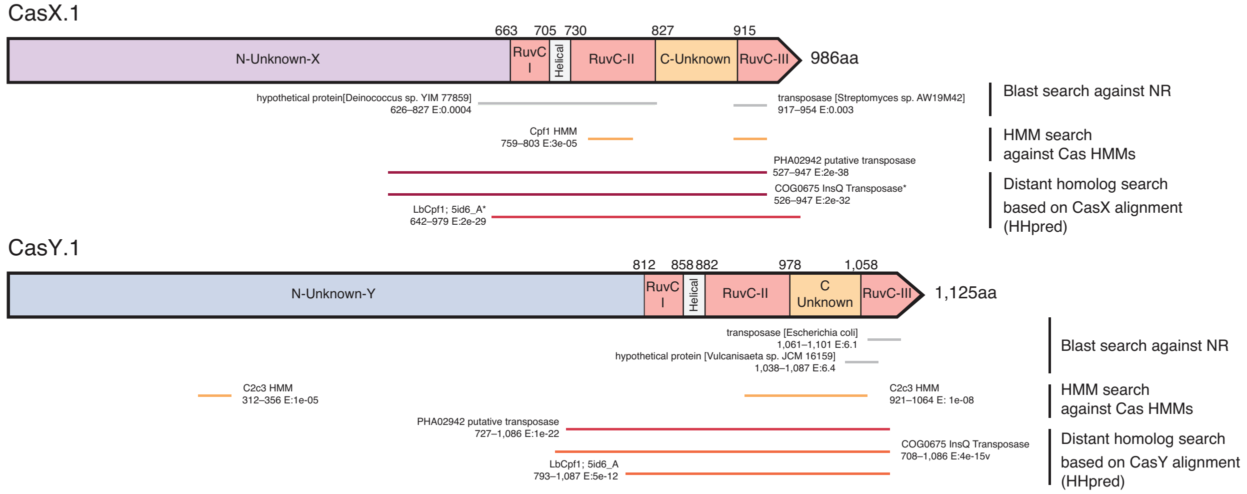


Figure 5.6 | Newly identified CRISPR-Cas systems compared to known proteins. Similarity of CasX and CasY to known proteins based on the following searches: (1) Blast search against

the non-redundant (NR) protein database of NCBI, (2) Hidden Markov model (HMM) search against an HMM database of all known proteins and (3) distant homology search using HHpred(Remmert et al., 2012, p.) (E, e-value).

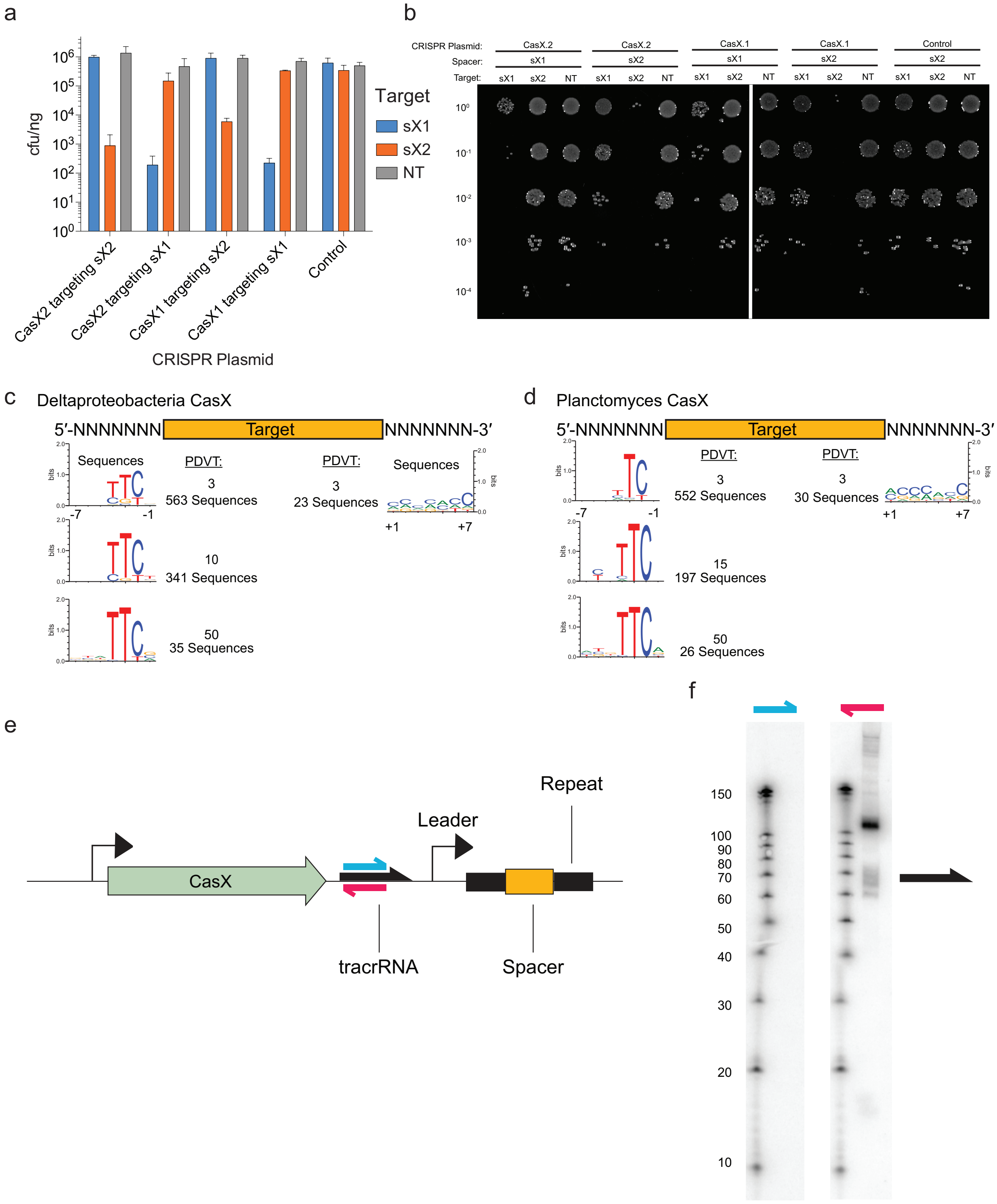


Figure 5.7 | Programmed DNA interference by CasX. a, Plasmid interference assays for CasX.1 (Deltaproteobacteria) and CasX.2 (Planctomycetes), continued from Figure 3c (sX1, CasX spacer 1; sX2, CasX spacer 2; NT, non-target). Experiments were conducted in triplicate

and mean ± s.d. is shown. **b,** Serial dilution of *E. coli* expressing a CasX locus and transformed with the specified target, continued from Figure 5.2b. **c,** PAM depletion assays for the Deltaproteobacteria CasX and **d,** Planctomycetes CasX expressed in *E. coli.* PAM sequences depleted greater than the indicated PAM depletion value threshold (PDVT) compared to a control library were used to generate the sequence logo. **e,** Diagram depicting the location of Northern blot probes for CasX.1. **f,** Northern blots for CasX.1 tracrRNA in total RNA extracted from *E. coli* expressing the CasX.1 locus. The sequences of the probes used are provided in Supplementary Table 2.

5.6 Acknowledgements

We thank Nan Ma, Kaihong Zhou and David McGrath for technical assistance; Chris Brown, Matt Olm, Mitchell O'Connell, Janice Chen and Stephen Floor for critical reading of the manuscript and useful discussions. We thank V. Yu for gift of S. cerevisiae expression strain. D.B. was supported by a long-term EMBO fellowship, L.B.H. by a US National Science Foundation Graduate Research Fellowship, and A.J.P. by a fellowship of the German Science Foundation (DFG PR 1603/1-1). J.A.D. is an Investigator of the Howard Hughes Medical Institute. This work was supported in part by a Frontiers Science award from the Paul Allen Institute to J.A.D. and J.F.B., the National Science Foundation (MCB-1244557 to J.A.D.) and the Lawrence Berkeley National Laboratory’s Sustainable Systems Scientific Focus Area funded by the U.S. Department of Energy, (DE-AC02-05CH11231 to J.F.B.). DNA sequencing was conducted at the DOE Joint Genome Institute, a DOE Office of Science User Facility, via the Community Science Program.

5.7 Author Contributions

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 analyzed the CRISPR-Cas systems. L.B.H. and D.B. designed and executed experimental work with CRISPR-CasX and CRISPR-CasY. S.C.S. designed and executed the experimental work with ARMAN Cas9. The manuscript that was read, edited, and approved by all authors.

Chapter 6

Programmed DNA destruction by miniature CRISPR-Cas14 enzymes

Part of the work presented in this chapter has previously been published in the following manuscript: Harrington, L. B., Burstein, D.*, Chen, J. S., Paez-Espino, D., Ma, E., Witte, I. P., … Doudna, J. A. (2018). Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science (New York, N.Y.)*.

6.1 Abstract

CRISPR-Cas systems provide microbes with adaptive immunity to infectious nucleic acids and are widely employed as genome editing tools. These tools utilize RNA-guided Cas proteins whose large size (950—1400 amino acids) has been considered essential to their specific DNA- or RNA-targeting activities. Here we present a set of CRISPR-Cas systems from uncultivated archaea that contain Cas14, a family of exceptionally compact RNA-guided nucleases (400—700 amino acids). Despite their small size, Cas14 proteins are capable of targeted single-stranded DNA (ssDNA) cleavage without restrictive sequence requirements. Moreover, target recognition by Cas14 triggers non-specific cutting of ssDNA molecules, an activity that enables high-fidelity SNP genotyping (Cas14-DETECTR). Metagenomic data show that multiple CRISPR-Cas14 systems evolved independently and suggest a potential evolutionary origin of single-effector CRISPR-based adaptive immunity.

6.2 Introduction

Competition between microbes and viruses stimulated the evolution of CRISPR-based adaptive immunity to provide protection against infectious agents (Rodolphe Barrangou et al., 2007; Jackson et al., 2017). In class 2 CRISPR-Cas systems, a single 100–200 kilodalton (kDa) CRISPR-associated (Cas) protein with multiple functional domains carries out RNA-guided binding and cutting of DNA or RNA substrates (J. S. Chen & Doudna, 2017; Shmakov et al., 2017a). To determine whether simpler, smaller RNA-guided proteins occur in nature, we queried terabase-scale metagenomic datasets (Anantharaman et al., 2016; Brown et al., 2015b; I. M. A. Chen et al., 2017; Markowitz et al., 2014; Probst et al., 2017) for uncharacterized genes proximal to both a CRISPR array and *cas1*, the gene that encodes the universal CRISPR integrase (Nuñez, Lee, et al., 2015; Yosef et al., 2012). This analysis identified a diverse family of CRISPR-Cas systems that contain *cas1*, *cas2, cas4,* and a new gene, *cas14*, encoding a 40–70 kDa polypeptide (Fig. 6.6.1A). We initially identified 24 different *cas14* gene variants that cluster into three subgroups (Cas14a–c) based on comparative sequence analysis (6.6.1A–B, fig. 6.5, 6.6). Cas14 proteins are ~400–700 amino acids (aa), about half the size of previously known class 2 CRISPR RNA-guided enzymes (950—1400 aa) (Fig. 6.1C–D). While the identified Cas14 proteins exhibit considerable sequence diversity, all are united by the presence of a predicted RuvC nuclease domain, whose organization is characteristic of Type V CRISPR-Cas DNA-targeting enzymes (Fig. 6.6.1D) (Burstein, Harrington, Strutt, Probst, et al., 2017; Shmakov et al., 2015, 2017a).

6.3 Materials and Methods

6.3.1 Metagenomics and metatranscriptomics

The initial analysis was performed on previously assembled and binned metagenomes from two sites: the Rifle Integrated Field Research (IFRC) site, adjacent to the Colorado River near Rifle, Colorado(Anantharaman et al., 2016; Brown et al., 2015b) and Crystal Geyser, a cold, CO_2-driven geyser on the Colorado Plateau in Utah(Probst et al., 2017). Metatranscriptomic data from IFRC site (Anantharaman et al., 2016; Brown et al., 2015b) was used to detect transcription of non-coding elements in nature. Further mining of CRISPR-Cas14 systems was then performed on public metagenomes from IMG/M (Markowitz et al., 2014, 2012).

6.3.2 CRISPR-Cas computation analysis

The assembled contigs from the various samples were scanned with the HMMer suite(Finn et al., 2011) for known Cas proteins using Hidden Markov Model (HMMs) profiles(Burstein, Harrington, Strutt, Probst, et al., 2017). Additional HMMs were constructed for Cas14 proteins based on the MAFFT alignments of putative type V effectors that contained less than 800 aa, and were adjacent to acquisition *cas* genes and CRISPR arrays. These HMMs were iteratively refined by augmenting them with manually selected novel putative Cas14 sequences that were found using the existing Cas14 HMM models. The sequence of newly identified Cas14 are provided in Data S1. CRISPR arrays were identified using a local version of the CrisprFinder software(Grissa, Vergnaud, & Pourcel, 2007) and CRISPRDetect(Biswas, Staals, Morales, Fineran, & Brown, 2016). Phylogenetic trees of Cas1 and type V effector proteins were constructed using RAxML(Stamatakis, 2014) with PROTGAMMALG as the substitution model and 100 bootstrap samplings. Trees were visualized using FigTree 1.4.1 (http://tree.bio.ed.ac.uk/software/figtree/). Metatranscriptomic reads were mapped to assembled contigs using Bowtie2(Langmead & Salzberg, 2012). RNase presence analysis was based on HMMs that were built from alignment of KEGG orthologous groups (KOs) downloaded from KEGG database (Ogata et al., 1999).

6.3.3 Generation of expression plasmids, RNA and DNA substrates

Minimal CRISPR loci for putative systems were designed by removing acquisition proteins and generating minimal arrays with a single spacer. These minimal loci were ordered as gBlocks (IDT) and assembled into a plasmid with a tetracycline inducible promoter driving expression of the locus. Plasmid maps are available on Addgene and in supplementary materials. All RNA was in vitro transcribed using T7 polymerase and PCR products as dsDNA template. Resulting IVTs were gel extracted and ethanol precipitated. DNA substrates were obtained from IDT and their sequences are available in Supplementary Table 1. For radiolabeled cleavage assays DNA oligos were gel extracted from a PAGE gel before radiolabeling. For FQ assays, DNA substrates were used without further purification.

6.3.4 *E. coli* RNAseq

Small RNA sequencing was conducted as described previously with modification(Harrington, Paez-Espino, et al., 2017). E. coli NEB Stable3 was transformed with a plasmid expressing Cas14a1 system with a tetracycline inducible promoter upstream of the Cas14a1 ORF or the same plasmid with an N-terminal 10x-histidine tag fused to Cas14. Starters were grown up overnight in SOB, diluted 1:100 in 5mL fresh SOB containing 214nM anhydrotetracycline and grown up overnight at 25°C. For sequencing of RNA pulled down with Cas14a, the plasmid containing an N-terminal His-tag fused to Cas14a1 was grown up at 18°C before lysis and purification as described in “Protein purification”, stopping after the Ni-NTA elution. Cells were pelleted and RNA was extracted using hot phenol as previously described. Total nucleic acids were treated with TURBO DNase and phenol extracted. The resulting RNA was treated with rSAP which was heat inactivated before addition of T4 PNK. Adapters were ligated onto the small RNA using the NEBnext small RNA kit and gel-extracted on an 8% native PAGE gel. RNA was sequenced on a MiSeq with single end 300bp reads. For analysis, the resulting reads were trimmed using Cutadapt, discarding sequences <8nt and mapped to the plasmid reference using Bowtie2 (Langmead & Salzberg, 2012).

6.3.5 PAM depletion assays

PAM depletion assays were conducted as previously described (Burstein, Harrington, Strutt, Probst, et al., 2017). Randomized plasmid libraries were generated using a primer containing a randomized PAM region adjacent to the target sequence. The randomized primers were hybridized with a primer that was complementary to the 3' end of the primer and the duplex was extended using Klenow Fragment (NEB). The dsDNA containing the target and were digested with EcoRI and NcoI, ligated into pUC19 backbone and transformed into *E. coli* DH5α and >10^7 cells were harvested. Next *E. coli* NEBstable was transformed with either a CRISPR plasmid or an empty vector control and these transformed *E. coli* were made electrocompetent by repeated washing with 10% glycerol. These electrocompetent cells were transformed with 200ng of the target library and plated on bioassay dishes containing selection for the target (carbenicillin, 100mg l^{-1}) and CRISPR plasmid (chloramphenicol, 30mg l^{-1}). Cells were harvested and prepared for amplicon sequencing on an Illumina MiSeq. The PAM region was extracted using Cutadapt and depletion values were calculated in python. PAMs were visualized using WebLogo(Crooks, Hon, Chandonia, & Brenner, 2004).

6.3.6 Protein purification

Cas14a1 was purified as described previously with modification(Harrington, Doxzen, et al., 2017). *E. coli* BL21(DE3) RIL were transformed with 10xHis-MBP-Cas14a1 expression plasmid and grown up to OD_{600} =0.5 in Terrific Broth (TB) and induced with 0.5mM IPTG. Cells were grown overnight at 18°C, collected by centrifugation, resuspended in Lysis Buffer (50 mM Tris-HCl, pH 7.5, 20 mM imidazole, 0.5 mM TCEP, 500 mM NaCl) and broken by sonication. Lysate was batch loaded on to Ni-NTA resin, washed with the above buffer before elution with Elution Buffer (50 mM Tris-HCl, pH 7.5, 300 mM imidazole, 0.5 mM TCEP, 500 mM NaCl). The MBP and His-tag were removed by overnight incubation with TEV at 4°C. The resulting protein exchanged into Buffer A (20 mM HEPES, pH 7.5, 0.5 mM TCEP, 150 mM NaCl) and loaded over tandem MBP, heparin columns (GE, Hi-Trap) and eluted with a linear gradient from Buffer A to Buffer B (20 mM HEPES, pH 7.5, 0.5 mM TCEP, 1250 mM NaCl). The resulting fractions containing Cas14a1 were loaded onto an S200 gel filtration column, flash frozen and stored at -80°C until use.

6.3.7 *In vitro* cleavage assays

Radiolabeled

Radiolabeled cleavage assays were conducted in 1× Cleavage Buffer (25 mM NaCl, 20 mM HEPES, pH 7.5, 1 mM DTT, 5% glycerol and 5mM $MgCl_2$). For cleavage assays with different divalent cations, 5mM $MgCl_2$ was replaced with 5mM of the indicated metal or EDTA. 100nM Cas14a1 was complexed with 125nM crRNA and 125nM tracrRNA for 10min at RT. ~1nM radiolabeled DNA or RNA substrate was added and allowed to react for 30min at 37°C. The reaction was stopped by adding 2x Quench Buffer (90% formamide, 25mM EDTA and trace bromophenol blue), heated to 95°C for 2min and run on a 10% polyacrylamide gel containing 7M Urea and 0.5×TBE. Products were visualized by phosphorimaging.

M13 DNA cleavage

M13 DNA cleavage assays were conducted in 100 mM NaCl, 20 mM HEPES, pH 7.5, 1 mM DTT, 5% glycerol, 5mM $MgCl_2$. 250nM Cas14a1 was complexed with 250nM crRNA and 250nM tracrRNA and 250nM ssDNA activator. The reaction was initiated by addition of 5 nM M13 ssDNA plasmid and was quenched by addition of loading buffer supplemented with 10mM EDTA. Products were separated on a 1.5% agarose TAE gel prestained with SYBR gold (Thermofisher).

FQ detection of trans-cleavage

FQ detection was conducted as previously described with modification(J. S. Chen et al., 2018). 100nM Cas14a1 was complexed with 125nM crRNA, 125nM tracrRNA, 50nM FQ probe and 2nM ssDNA activator in 1× Cleavage Buffer at 37°C for 10 min. The reaction was then initiated by addition of activator DNA when for all reactions except for the RNA optimization experiments where the variable RNA component was used to initiate. The reaction was monitored in a fluorescence plate reader for up to 120 minutes at 37°C with fluorescence measurements taken every 1 min (λ_{ex}: 485 nm; λ_{em}: 535 nm). The resulting data were background subtracted using the readings taken in the absence of activator and fit using a single exponential decay curve.

6.3.8 Cas14 and Cas12a-DETECTR assays

For Cas14a1 or Cas12a detection of the A/G SNP in the HERC2 gene, saliva samples were taken at three independent times by brown and blue-eyed individuals. For crude DNA extraction, saliva was pelleted and washed twice in phosphate buffered saline (1 × PBS), incubated for 5 min at 100°C, and centrifuged for 5 min at 10000×g.

DETECTR assays involved an initial PCR amplification followed by detection by Cas14a1 or LbCas12a. 50 uL PCRs consisting of 1 uL template DNA, 10uL 5X Q5 Buffer, 0.48uM forward/reverse primers, 200uM (each) dNTPs, and 1 U Q5 DNA Polymerase underwent 25 cycles of amplification. The first four 5′ nucleotides of the forward primer were phosphorothioated to protect from degradation by T7 exonuclease in the subsequent detection step, while the 5′ end of the reverse primer was unmodified. 2 uL of the PCR product was transferred to a 384 well plate and combined directly in the plate with the DETECTR reaction mix. The Cas14a1 DETECTR reaction consisted of a final concentration of 100nM Cas14a1, 125nM sgRNA, 50 nM ssDNA-FQ reporter, and 5 U T7 exonuclease in a total reaction volume of 20 uL. The LbCas12a DETECTR reaction consisted of a final concentration of 50 nM LbCas12a, 50 nM sgRNA, 50 nM ssDNA-FQ reporter, and 2.5 U T7 exonuclease in a total reaction volume of 20 uL. Reactions were incubated in a fluorescence plate reader (Tecan Infinite Pro 200 M Plex) for 2 hours at 37°C with fluorescence measurements taken every 30 seconds (λex: 485 nm; λem: 535 nm).

6.4 Results and Discussion

The Cas14 proteins we identified occur almost exclusively within DPANN, a super-phylum of symbiotic archaea characterized by small cell and genome sizes (Castelle et al., 2015; Rinke et al., 2013). Phylogenetic comparisons showed that Cas14 proteins are widely diverse with similarities to C2c10 and C2c9, families of bacterial RuvC-domain-containing proteins that are sometimes found near a CRISPR array but not together with other *cas* genes (Fig. 6.1B, fig. 6.5) (Shmakov et al., 2017a). This observation and the small size of *c2c10, c2c9* and *cas14* genes made it improbable that these systems could function as standalone CRISPR effectors (Shmakov et al., 2017a).

Based on their proximity to conserved genes responsible for creating genetic memory of infection (*cas1*, *cas2*, *cas4*) (fig. 6.7A), we explored whether CRISPR-Cas14 systems can actively acquire DNA sequences into their CRISPR arrays. Assembled metagenomic contiguous DNA sequences (contigs) for multiple CRISPR-Cas14 loci revealed that otherwise identical CRISPR systems showed diversity in their CRISPR arrays. These results are consistent with active adaptation to new infections although without longitudinal sampling this data could also be explained by alternative biological mechanisms (Fig. 6.2A, fig. 6.7B) (Burstein, Harrington, Strutt, Probst, et al., 2017). The evidence suggesting acquisition of new DNA sequences led us to hypothesize that these CRISPR-Cas14 loci encode functional enzymes with nucleic acid targeting activity despite their small size. To test this possibility, we first investigated whether RNA components are produced from CRISPR-Cas14 loci. Environmental metatranscriptomic sequencing data were analyzed for the presence of RNA from the native archaeal host that contains CRISPR-Cas14a (Fig. 6.6B, fig. 6.8A). In addition to CRISPR RNAs (crRNAs), a highly abundant non-coding RNA was mapped to a ~130-base pair sequence located between *cas14a* and the adjacent CRISPR array. Notably, the 3′ end of this transcript was mostly complementary to the repeat segment of the crRNA (Fig. 6.2C, fig. 6.8B), as observed for trans-activating CRISPR RNAs (tracrRNAs) found in association with Cas9, Cas12b and Cas12e CRISPR systems (Burstein, Harrington, Strutt, Probst, et al., 2017; E Deltcheva et al., 2011; Shmakov et al., 2015). In these previously studied systems, the double-stranded-RNA-cutting enzyme Ribonuclease III (RNase III) generates mature tracrRNAs and crRNAs, but no genes encoding RNase III were present in *cas14*-containing reconstructed genomes (fig. 6.9A), nor did Cas14a cleave its own pre-crRNA when tested biochemically (fig. 6.9B). These observations imply that an alternative mechanism for CRISPR-associated RNA processing exists in these hosts.

To test whether the Cas14a proteins and associated RNA components can assemble together in a heterologous organism, we introduced a plasmid into *E. coli* containing a minimal CRISPR-Cas14a locus that includes the Cas14 gene, the CRISPR array and intergenic regions containing the putative tracrRNA. Affinity purification of the Cas14a protein from cell lysate and sequencing of co-purifying RNA revealed a highly abundant mature crRNA as well as the putative tracrRNA, albeit in lower relative abundance than environmental metatranscriptomics, suggesting that Cas14 associates with both crRNA and tracrRNA (fig. 6.9B). The calculated mass of the assembled Cas14a protein-tracrRNA-crRNA particle is 48% RNA by weight compared to just 17%

for *S. pyogenes* Cas9 (SpCas9) and 8% for *F. novicida* Cas12a (FnCas12a) (Fig. 6.2D), hinting at a central role of the RNA in the architecture of the Cas14a complex. Known class 2 CRISPR systems require a short sequence called a protospacer adjacent motif (PAM) to target double-stranded DNA (dsDNA) (Mojica, Díez-Villaseñor, García-Martínez, & Almendros, 2009). To test whether Cas14a requires a PAM and can conduct dsDNA interference, we transformed *E. coli* expressing a minimal Cas14a locus with a dsDNA plasmid containing a randomized PAM region next to a sequence matching the target-encoding sequence (spacer) in the Cas14 array. Notably, no depletion of a PAM sequence was detected among *E. coli* transformants, suggesting that the CRISPR-Cas14a system is either unable to target dsDNA, can do so without requiring a PAM, or is inactive in this heterologous host (fig. S6A, B).

We next tested whether purified Cas14a-tracrRNA-crRNA complexes are capable of RNA-guided nucleic acid cleavage *in vitro*. All currently reconstituted DNA-targeting class 2 interference complexes are able to recognize both dsDNA and ssDNA substrates (J. S. Chen et al., 2018; Ma et al., 2015; Zhang, Rajan, Seifert, Mondragón, & Sontheimer, 2015). We incubated purified Cas14a-tracrRNA-crRNA complexes with radiolabeled target oligonucleotides (ssDNA, dsDNA, and ssRNA) bearing 20-nucleotide sequence complementary to the crRNA guide sequence, or a non-complementary ssDNA, and we analyzed these substrates for Cas14a-mediated cleavage. Only in the presence of a complementary ssDNA substrate was any cleavage product detected (Fig. 6.3A, fig. 6.11A-C), and cleavage was dependent on the presence of both tracrRNA and crRNA, which could also be combined into a single-guide RNA (sgRNA) (Fig. 6.3B, fig. 6.12). The lack of detectable dsDNA cleavage suggests that Cas14a targets ssDNA selectively, although it is possible that some other host factor or sequence requirement could enable dsDNA recognition in the native host. Mutation of the conserved active site residues in the Cas14a RuvC domain eliminated cleavage activity (fig. 6.11D-E), implicating RuvC as the domain responsible for DNA cutting. Moreover, Cas14a DNA cleavage was sensitive to truncation of the RNA components to lengths shorter than the naturally produced sequences (fig. 6.13A-D). These results establish Cas14a as the smallest class 2 CRISPR effector demonstrated to conduct programmable RNA-guided DNA cleavage thus far.

Although we were unable to identify a dsDNA PAM *in vivo,* we tested whether Cas14a requires a PAM for ssDNA cleavage *in vitro* by tiling Cas14a guides across a ssDNA substrate (Fig. 6.3C). Despite sequence variation adjacent to the targets of these different guides, we observed cleavage for all four sequences. Notably, the cleavage sites occur beyond the guide-complementary region of the ssDNA and shift in response to guide binding position (Fig. 6.3C). These data demonstrate Cas14a is a ssDNA-targeting CRISPR endonuclease that does not require a PAM for activation.

Based on the observation that Cas14a cuts outside of the crRNA/DNA targeting heteroduplex, we hypothesized that Cas14a might possess target-activated non-specific ssDNA cleavage activity, similar to the RuvC-containing enzyme Cas12a (J. S. Chen et al., 2018; Zetsche et al., 2015). To test this possibility, we incubated Cas14a-tracrRNA-crRNA with a complementary activator DNA and an aliquot of M13 bacteriophage ssDNA bearing no sequence complementarity to the Cas14a crRNA or activator (Fig. 6.3D). The M13 ssDNA was rapidly degraded to small fragments, an activity that was eliminated by mutation of the conserved Cas14a RuvC active site, suggesting that activation of Cas14a results in non-specific ssDNA degradation. However, we were unable to observe Cas14a-mediated interference against the ssDNA bacteriophage ΦX174 when we expressed Cas14a heterologously in *E. coli* (Fig. 6.14A-C), possibly due to the dissimilarity between *E. coli* and Cas14a's native archaeal host. To investigate the specificity of target-dependent non-specific DNA cutting activity by Cas14a, we adapted a fluorophore-quencher (FQ) assay in which cleavage of dye-labeled ssDNA generates a fluorescent signal (Fig. 6.4A) (East-Seletsky et al., 2016). When Cas14a was incubated with various guide RNA-target ssDNA pairs, a fluorescent signal was observed only in the presence of the cognate target and showed strong preference for longer FQ-containing substrates (fig. 6.14D, Fig. 6.4A). We next tested Cas14a mismatch tolerance by tiling 2-nt mismatches across the targeted region in various ssDNA substrates. Surprisingly, mismatches near the middle of the ssDNA target strongly inhibited Cas14a activity, revealing an internal seed sequence that is distinct from the PAM-proximal seed region observed for dsDNA-targeting CRISPR-Cas systems (Fig. 6.4B, fig. 6.15A-D). Moreover, DNA substrates containing strong secondary structure resulted in reduced activation of Cas14a (fig. 6.15E). Truncation of ssDNA substrates also resulted in reduced or undetectable *trans* cleavage (fig. 6.15F). Together, these results suggest a mechanism of fidelity distinct from dsDNA-targeting class 2 CRISPR systems, possibly utilizing a mechanism similar to the ssRNA-targeting Cas13a enzymes (Abudayyeh et al., 2016b; Knott et al., 2017; L. Liu et al., 2017).

The target-dependent, non-specific DNase activity of Cas12a serves as a DNA detection platform (DNA endonuclease-targeted CRISPR trans reporter; DETECTR) for diagnostic uses (J. S. Chen et al., 2018; S. Y. Li et al., 2018). While Cas12a exhibits low fidelity in discriminating against ssDNA substrates (J. S. Chen et al., 2018), Cas14a requires complementarity in the seed region for ssDNA substrate recognition. This improved specificity raised the possibility of using Cas14a for high-fidelity detection of DNA single nucleotide polymorphisms (SNPs) without the constraint of a PAM sequence. To test this idea, DNA substrates were amplified using a phosphorothioate (PT)-containing primer to protect one strand from degradation by exonucleases. Upon addition of T7 exonuclease, the unmodified strand was degraded, leaving ssDNA substrates that can be detected by Cas14a (Fig. 6.4C–D). As a proof of principle, we aimed to detect the human HERC2 gene, which contains a SNP responsible for eye color (Eiberg et al., 2008). We amplified the HERC2 gene from DNA in human saliva from both blue-eyed and brown-eyed individuals, using the PT amplification approach described above. When programmed with a guide RNA targeting the blue-eyed SNP, Cas12a failed to discriminate between the two ssDNA targets, exhibiting robust *trans* activity in both cases, while Cas14a exhibited strong activation in recognition of the blue-eyed SNP with near-background signal for the brown-eyed sample (Fig. 6.4E). The development of Cas14-DETECTR now allows for CRISPR-based detection of medically and ecologically important ssDNA pathogens as well as high-fidelity detection of SNPs without the constraint of a PAM sequence.

Further investigation of compact Type V systems in metagenomic data revealed a large diversity of systems that, like Cas14a–c, include a gene encoding a short RuvC-containing protein adjacent to acquisition-associated *cas* genes and a CRISPR array. We found 20 additional such systems in various uncultivated microbes that cluster into five main families (Cas14d–h). Excluding *cas*14g, which is related to *cas*12b, the *cas*14-

like genes form separate clades on the type V effector phylogeny (fig. 6.16A, B), suggesting these families evolved from independent domestication events of TnpB, the transposase-associated protein implicated as the evolutionary ancestor of type V CRISPR effectors (Shmakov et al., 2017b). Phylogenetic reconstruction of their associated *cas1* genes indicated that they too have different origins for the *cas14* subtypes (fig. 6.6A). Altogether we identified 38 CRISPR-Cas14 systems belonging to eight families (Cas14a–h) and eight additional systems that could not be clustered with our analysis (termed Cas14u, data S1).

The small size of the Cas14 proteins described here and their resemblance to type V effector proteins suggest that RNA-guided ssDNA cleavage may have existed as an ancestral class 2 CRISPR system (Koonin, Makarova, & Zhang, 2017; K S Makarova et al., 2015). In this scenario, a small, domesticated TnpB-like ssDNA interference complex may have gained additional domains over time, gradually improving dsDNA recognition and cleavage. Related to this hypothesis, smaller Cas9 orthologs exhibit weaker dsDNA-targeting activity than their larger counterparts but retain the ability to robustly cleave ssDNA (Ma et al., 2015). Aside from the evolutionary implications, the ability of Cas14 to specifically target ssDNA suggests a role in defense against ssDNA viruses or mobile genetic elements (MGEs) that propagate through ssDNA intermediates (Barabas et al., 2008). A ssDNA-targeting CRISPR system would be particularly advantageous in certain ecosystems where ssDNA viruses comprise the vast majority of viral abundance (Yoshida et al., 2018). The unexpected finding that these miniature CRISPR proteins can conduct targeted DNA cleavage highlights the diversity of CRISPR systems hidden in uncultivated organisms. Ongoing exploration of these underrepresented microbial lineages will likely continue to reveal new, unexpected insights into this microscopic arms race and lead to continued development of valuable CRISPR-based technologies.

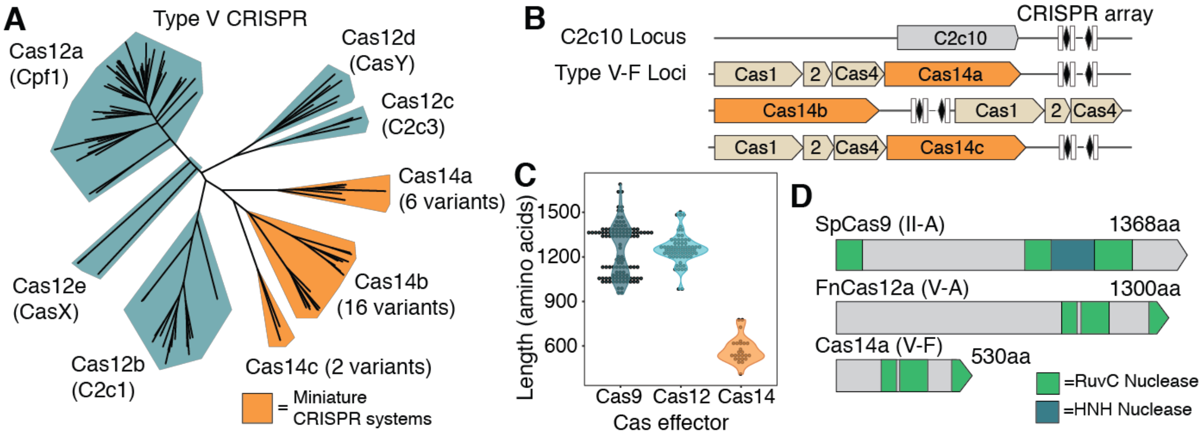


Figure 6.1 Novel identified CRISPR-Cas systems from uncultivated organisms. (A) Phylogenetic tree of Type V CRISPR systems. Newly identified miniature CRISPR systems are highlighted in orange. (B) Representative loci architectures for C2c10 and CRISPR-Cas14 systems. (C) Length distribution of Cas14a–c systems compared to Cas12a-e and Cas9. (D) Domain organization of Cas14a compared to Cas9 and Cas12a with the nuclease domains (RuvC and HNH) indicated. Protein lengths are drawn to scale.

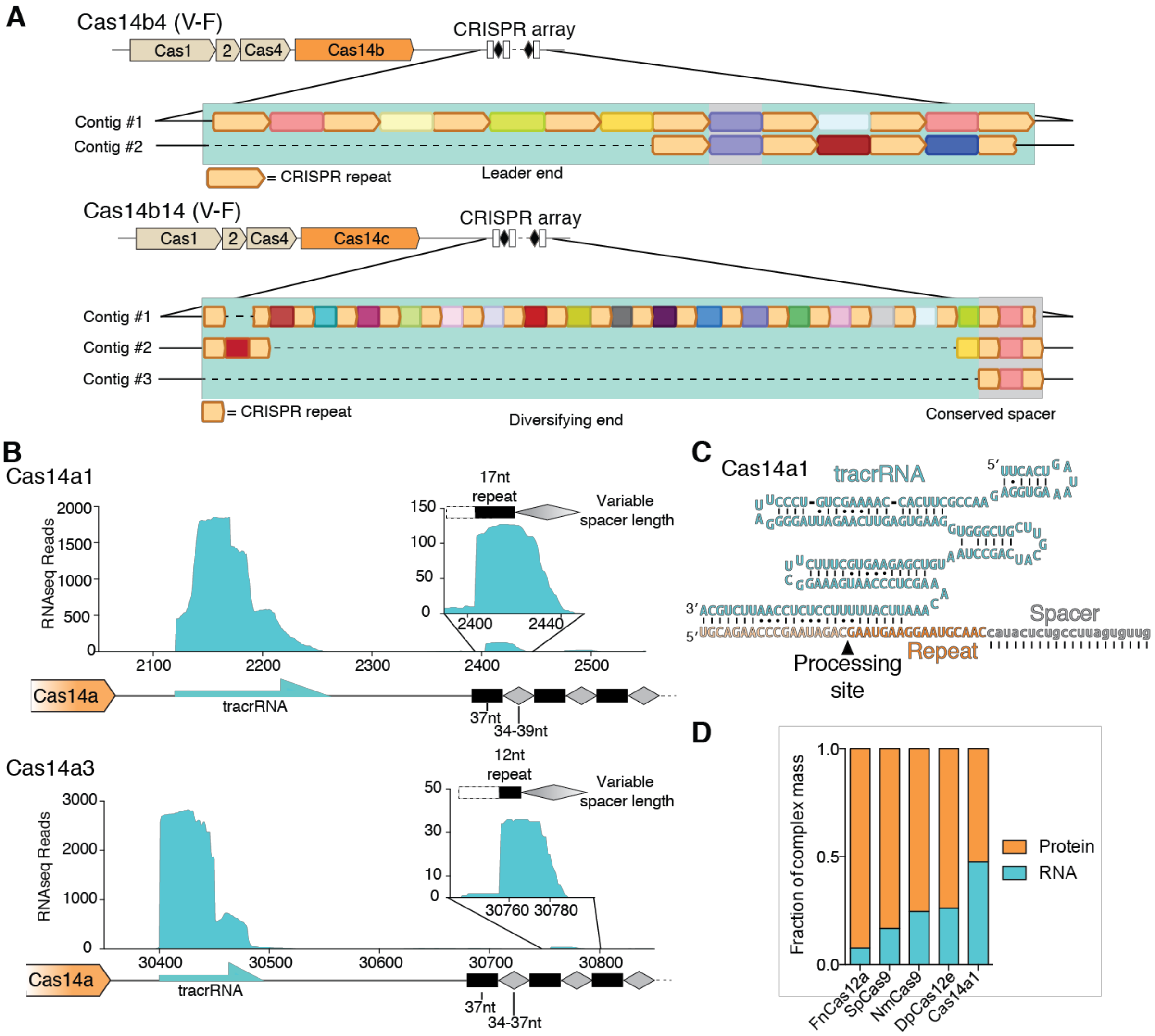


Figure 6.2 CRISPR-Cas14a actively adapts and encodes a tracrRNA. (A) Spacer diversity for Cas14b4 and Cas14b14 with CRISPR repeats diagramed in tan and unique spacers shown in different colors. (B) Metatranscriptomics reads mapped to Cas14a1 and Cas14a3. Inset shows expansion of most abundant repeat and spacer sequence. (C) *In silico* predicted structure of Cas14a1 crRNA and tracrRNA. Notably, RNase III orthologs were not identified in host genomes (fig. 6.9A). (D) Fraction of various CRISPR complexes mass made up of by RNA and protein.

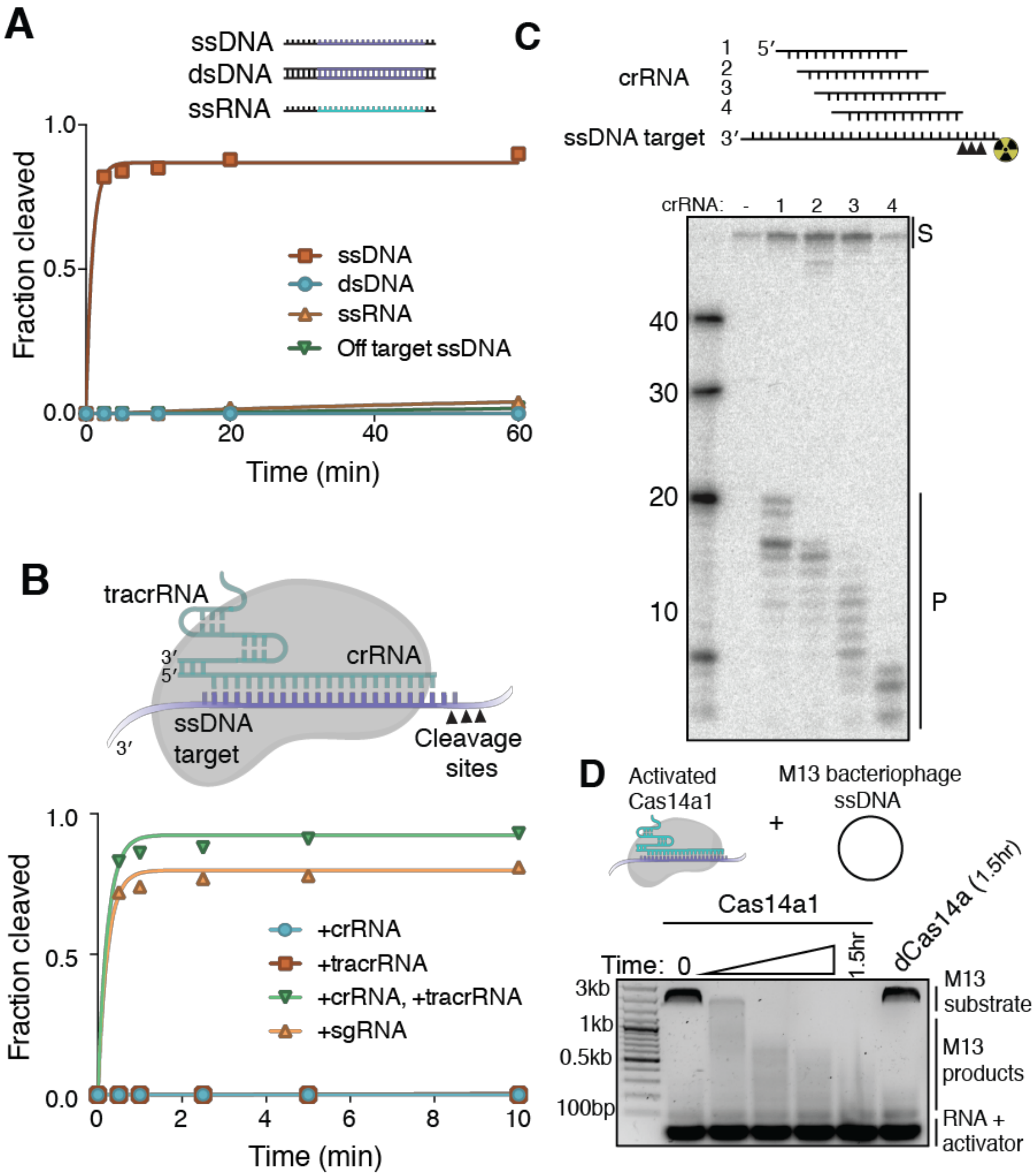


Fig. 6.3 CRISPR-Cas14a is an RNA-guided DNA-endonuclease. (A) Cleavage kinetics of Cas14a1 targeting ssDNA, dsDNA, ssRNA and off-target ssDNA. (B) Diagram of Cas14a RNP bound to target ssDNA and Cas14a1 cleavage kinetics of radiolabeled ssDNA in the presence of various RNA components. (C) Tiling of a ssDNA substrate by Cas14a1 guide sequences. (D) Cleavage of the ssDNA viral M13 genome with activated Cas14a1.

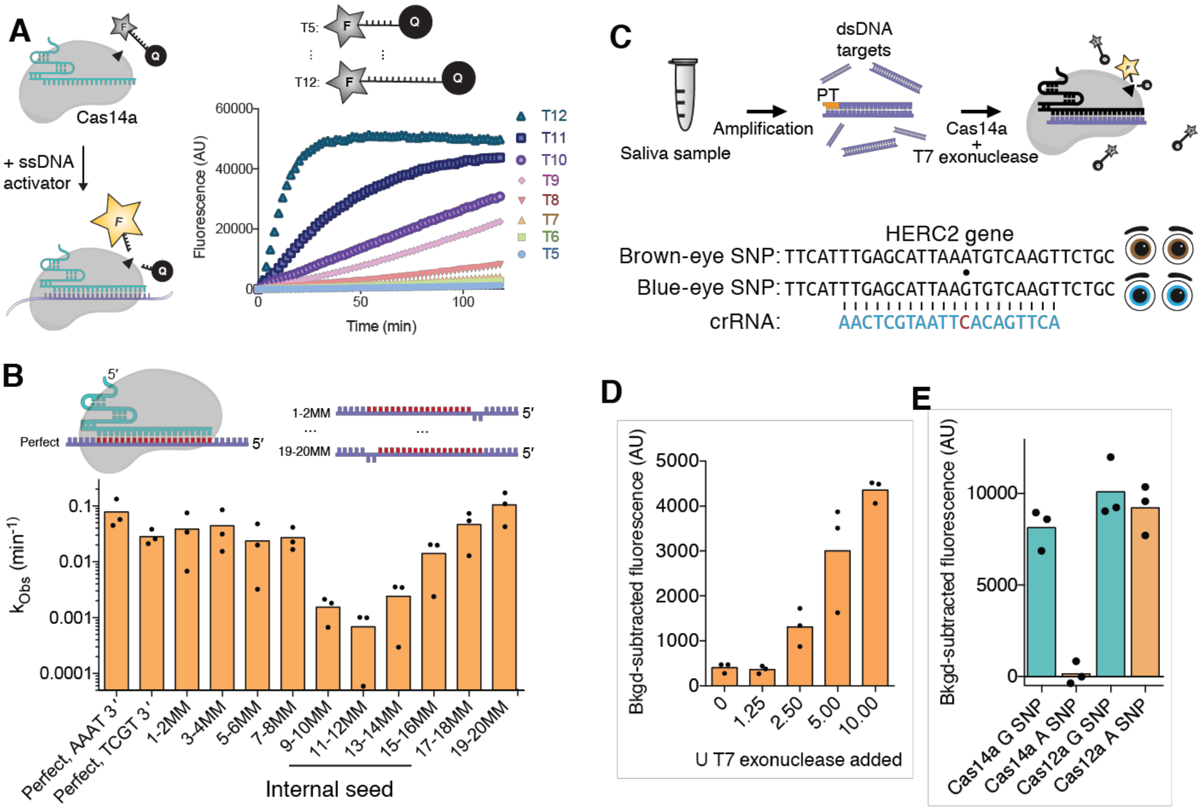


Fig. 6.4 High fidelity ssDNA SNP detection by CRISPR-Cas14a. (A) Fluorophore-quencher (FQ) assay for detection of ssDNA by Cas14a1 and the cleavage kinetics for various length FQ substrates. (B) Cleavage kinetics for Cas14a1 with mismatches tiled across the substrate (individual points represent replicate measurements). (C) Diagram of Cas14-DETECTR strategy and HERC2 eye color SNP. (D) Titration of T7 exonuclease and impact on Cas14a-DETECTR. (E) SNP detection using Cas14a-DETECTR with a blue-eye targeting guide for a blue-eyed and brown-eyed saliva sample compared to ssDNA detection using Cas12a.

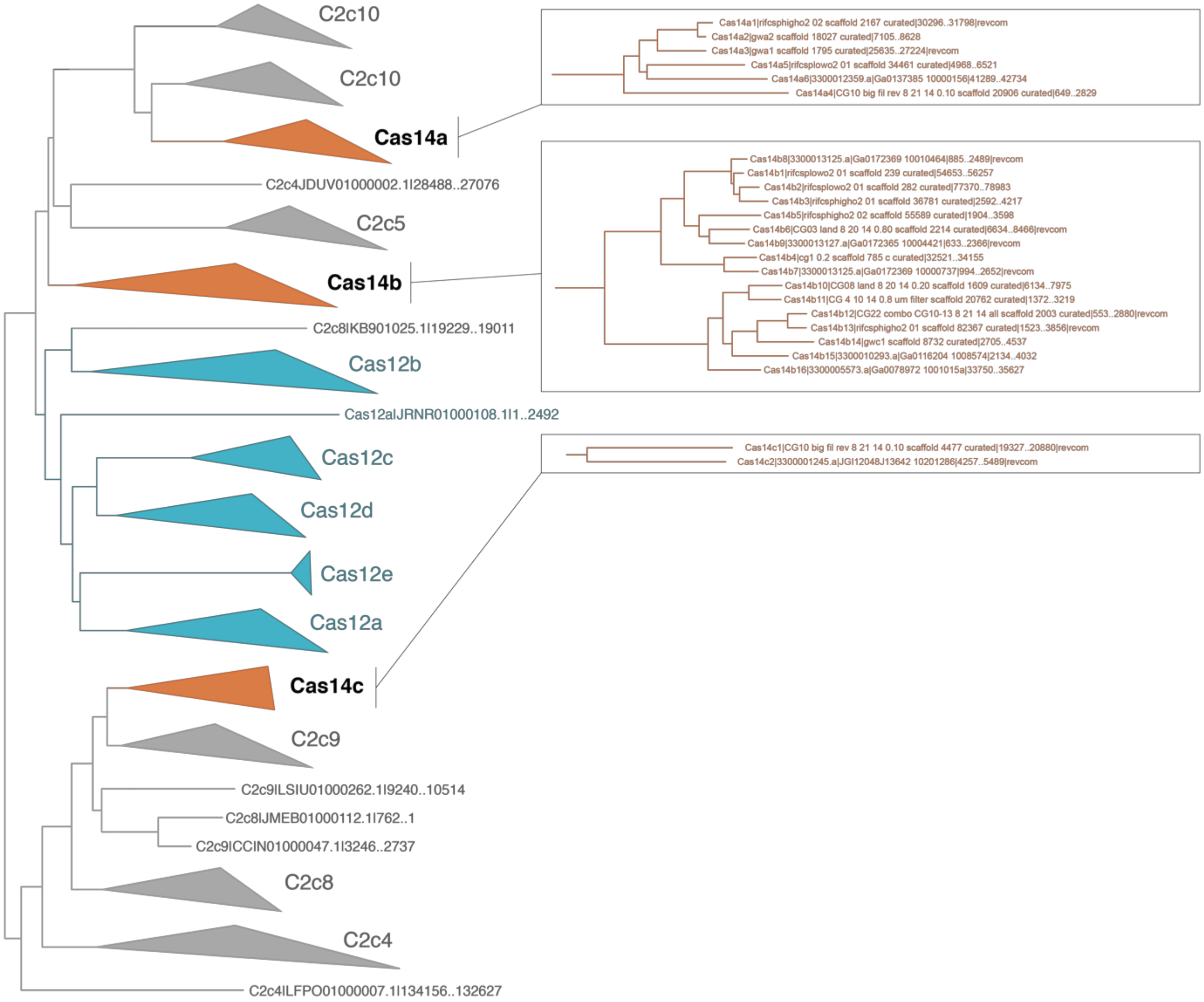

C2c10
C2c10
Cas14a
C2c4JDUV01000002.1I28488..27076
C2c5
Cas14b
C2c8IKB901025.1I19229..19011
Cas12b
Cas12aIJRNR01000108.1I1..2492
Cas12c
Cas12d
Cas12e
Cas12a
Cas14c
C2c9
C2c9ILSIU01000262.1I9240..10514
C2c8IJMEB01000112.1I762..1
C2c9ICCIN01000047.1I3246..2737
C2c8
C2c4
C2c4ILFPO01000007.1I134156..132627
Cas14a1|rifcsphigho2 02 scaffold 2167 curated|30296..31798|revcom
Cas14a2|gwa2 scaffold 18027 curated|7105..8628
Cas14a3|gwa1 scaffold 1795 curated|25635..27224|revcom
Cas14a5|rifcsplowo2 01 scaffold 34461 curated|4968..6521
Cas14a6|3300012359.a|Ga0137385 10000156|41289..42734
Cas14a4|CG10 big fil rev 8 21 14 0.10 scaffold 20906 curated|649..2829
Cas14b8|3300013125.a|Ga0172369 10010464|885..2489|revcom
Cas14b1|rifcsplowo2 01 scaffold 239 curated|54653..56257
Cas14b2|rifcsplowo2 01 scaffold 282 curated|77370..78983
Cas14b3|rifcsphigho2 01 scaffold 36781 curated|2592..4217
Cas14b5|rifcsphigho2 02 scaffold 55589 curated|1904..3598
Cas14b6|CG03 land 8 20 14 0.80 scaffold 2214 curated|6634..8466|revcom
Cas14b9|3300013127.a|Ga0172365 10004421|633..2366|revcom
Cas14b4|cg1 0.2 scaffold 785 c curated|32521..34155
Cas14b7|3300013125.a|Ga0172369 10000737|994..2652|revcom
Cas14b10|CG08 land 8 20 14 0.20 scaffold 1609 curated|6134..7975
Cas14b11|CG 4 10 14 0.8 um filter scaffold 20762 curated|1372..3219
Cas14b12|CG22 combo CG10-13 8 21 14 all scaffold 2003 curated|553..2880|revcom
Cas14b13|rifcsphigho2 01 scaffold 82367 curated|1523..3856|revcom
Cas14b14|gwc1 scaffold 8732 curated|2705..4537
Cas14b15|3300010293.a|Ga0116204 1008574|2134..4032
Cas14b16|3300005573.a|Ga0078972 1001015a|33750..35627
Cas14c1|CG10 big fil rev 8 21 14 0.10 scaffold 4477 curated|19327..20880|revcom
Cas14c2|3300001245.a|JGI12048J13642 10201286|4257..5489|revcom

Fig. 6.5 Phylogenetic analysis of Cas14 orthologs

Maximum likelihood tree for known Type V CRISPR effectors and class 2 candidates containing a RuvC domain. Inset shows individual orthologs for each newly identified subtype.

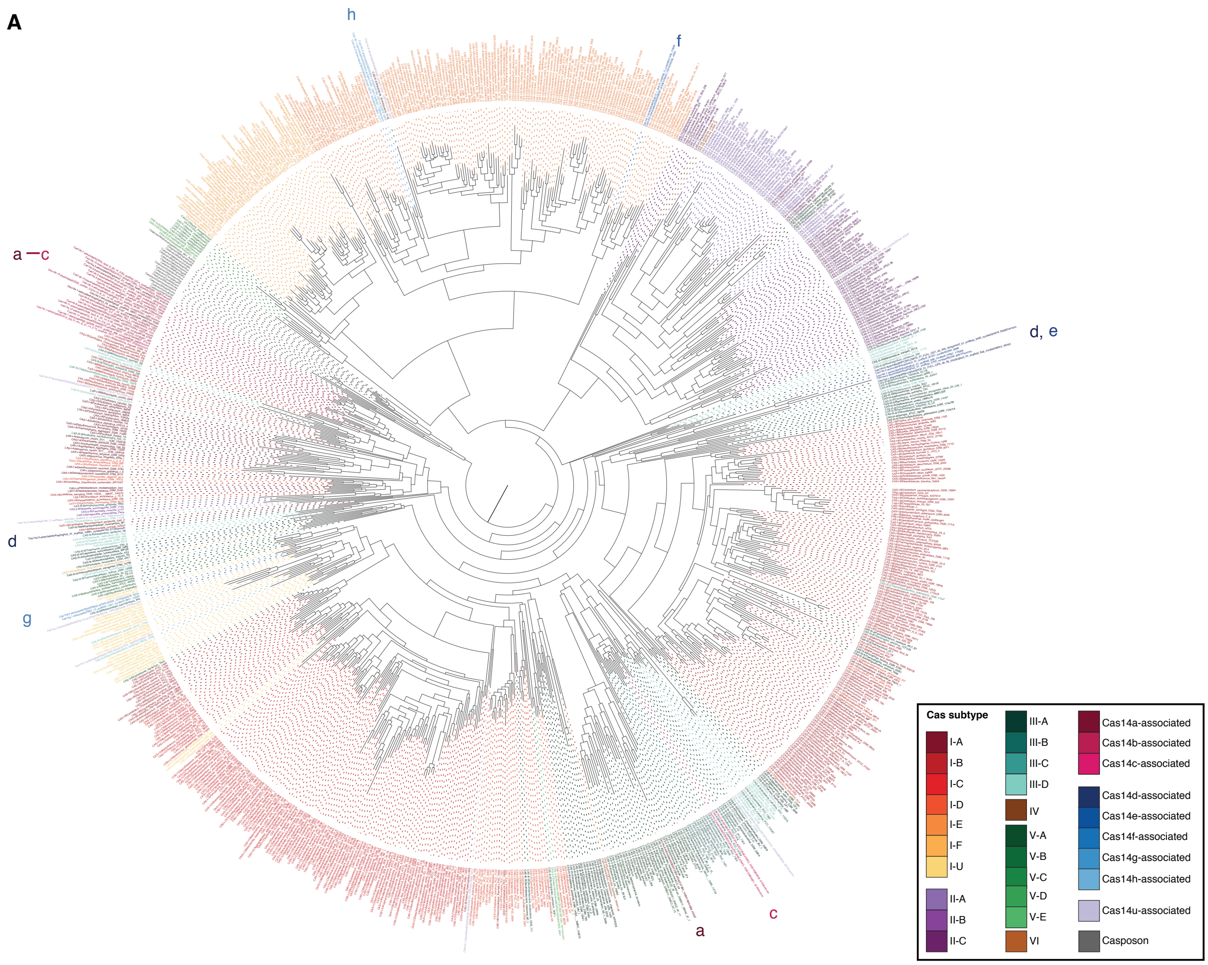


Fig. 6.6 Maximum likelihood tree for Cas1 from known CRISPR systems. Newick format of the Cas1 tree is provided in Data S4.

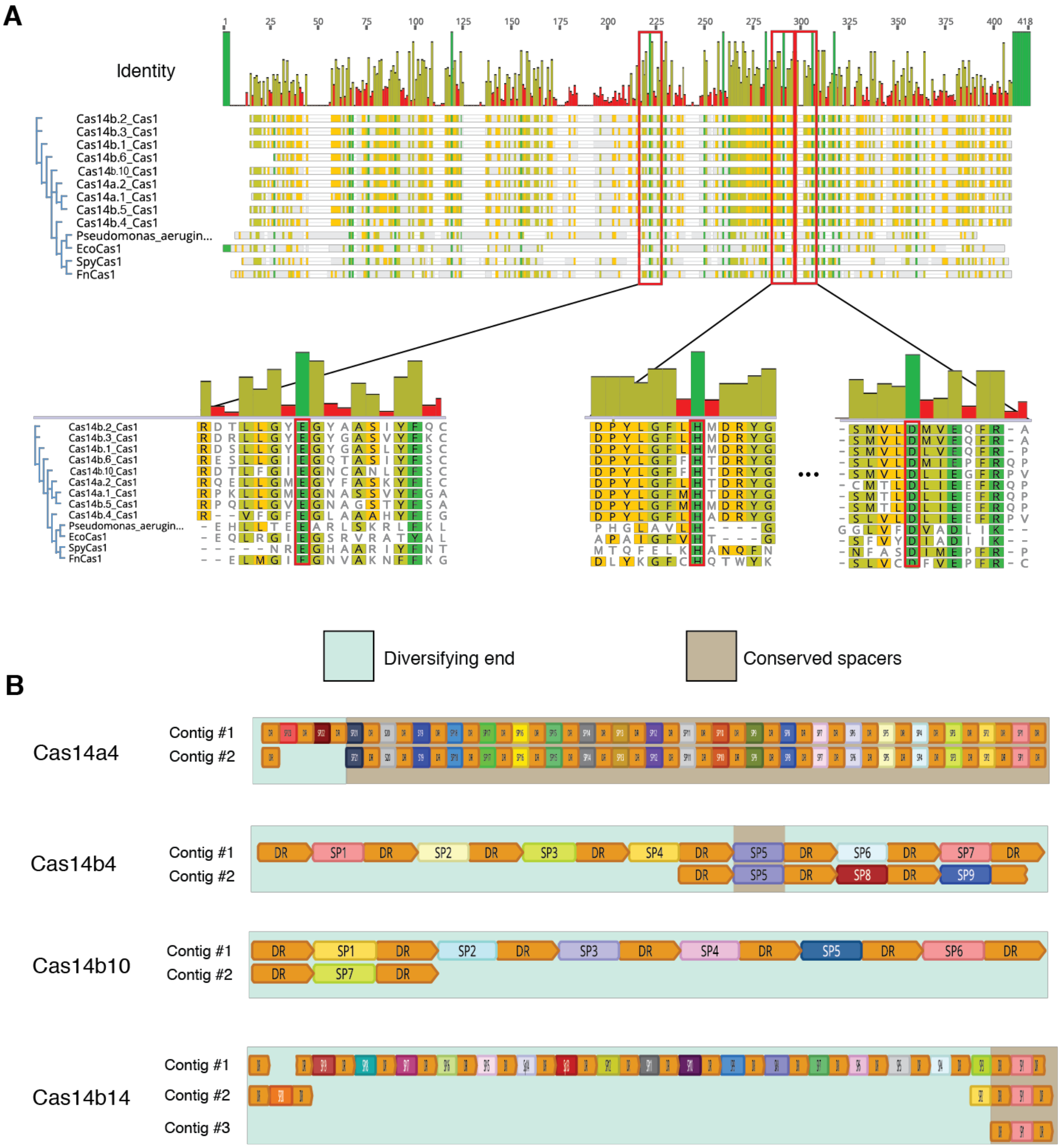


Fig. 6.7 Acquisition of new spacers by CRISPR-Cas14 systems

(A) Alignment of Cas14 Cas1 orthologs. Expansion shows conservation of previously implicated active site residues highlighted in red boxes. (B) Multiple CRISPR arrays assembled for various CRISPR-Cas14 systems revealing spacer diversity for these CRISPR systems. Orange arrows indicate repeats while variously colored boxes indicate unique spacers.

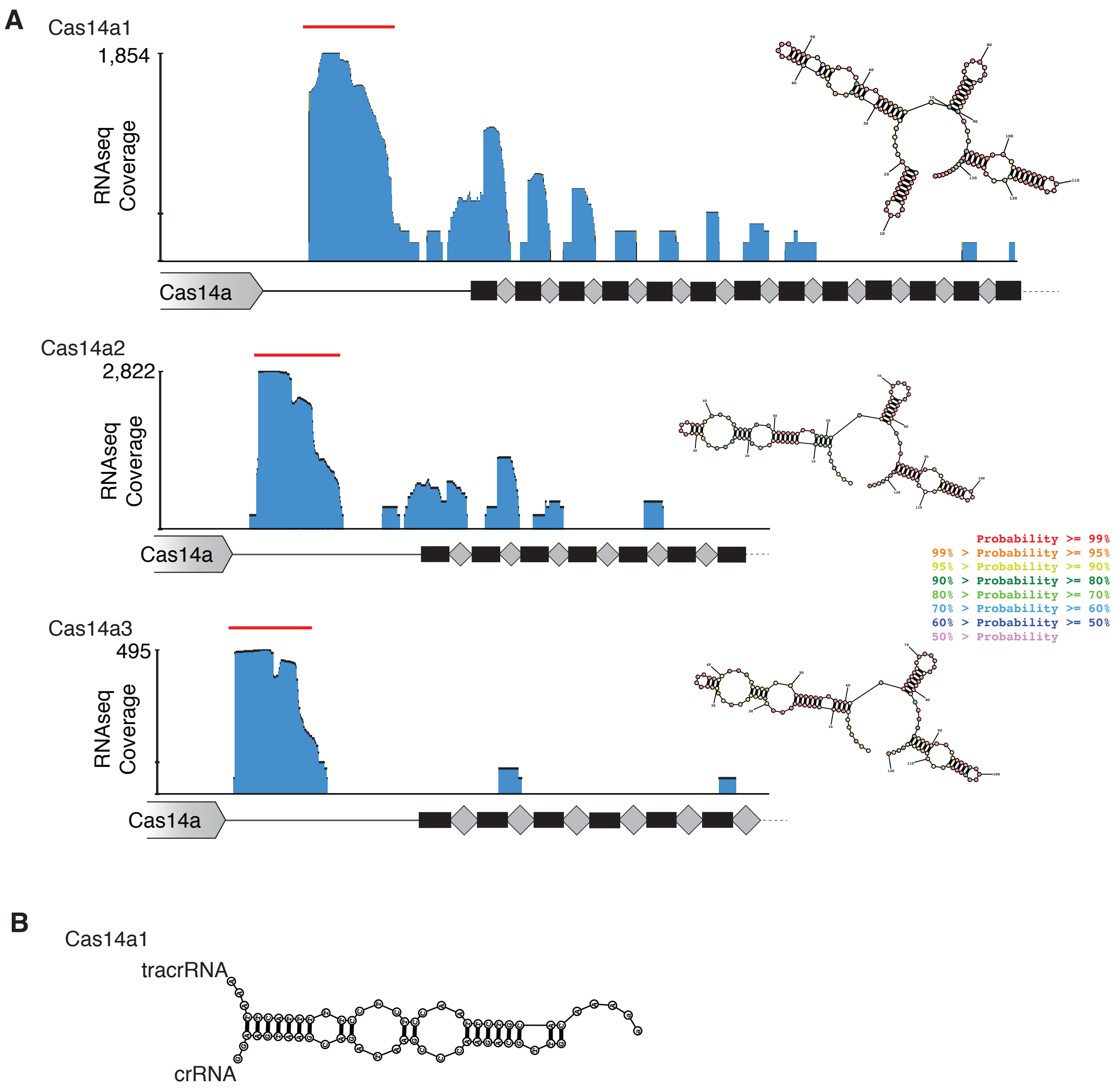


Fig. 6.8 Metatranscriptomics for CRISPR-Cas14 loci

(A)Environmental RNA sequencing reads for Cas14a orthologs. Location of Cas14 and the CRISPR array indicated below. RNA structures to the right show the *in silico* predicted structure of the tracrRNA identified from metatranscriptomics. (B) Predicted hybridization for Cas14a1 crRNA:tracrRNA duplex.

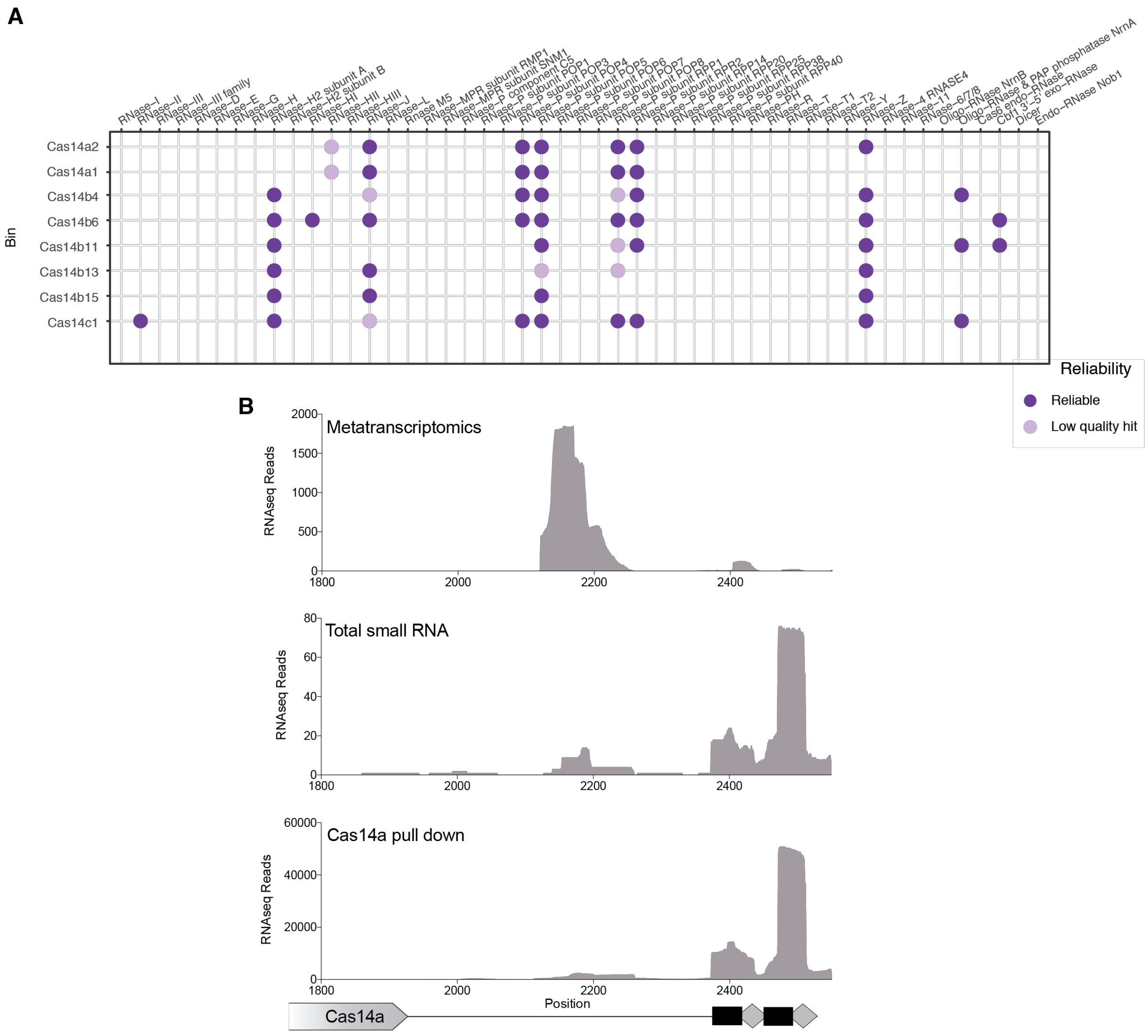


Fig. 6.9 RNA processing and heterologous expression by CRISPR-Cas14
(A) Presence of common RNase orthologs in Cas14 containing genomes. Light purple represents hits that were significantly shorter than the expected length for the given RNase. Note that RNase III is absent in all investigated genomes. (B) Small RNAseq reads from heterologous expression of Cas14a1 locus in *E. coli* (bottom two) compared to metatranscriptomic reads (top panel). Pull down refers to RNA that copurified with Ni-NTA affinity purified Cas14a1.

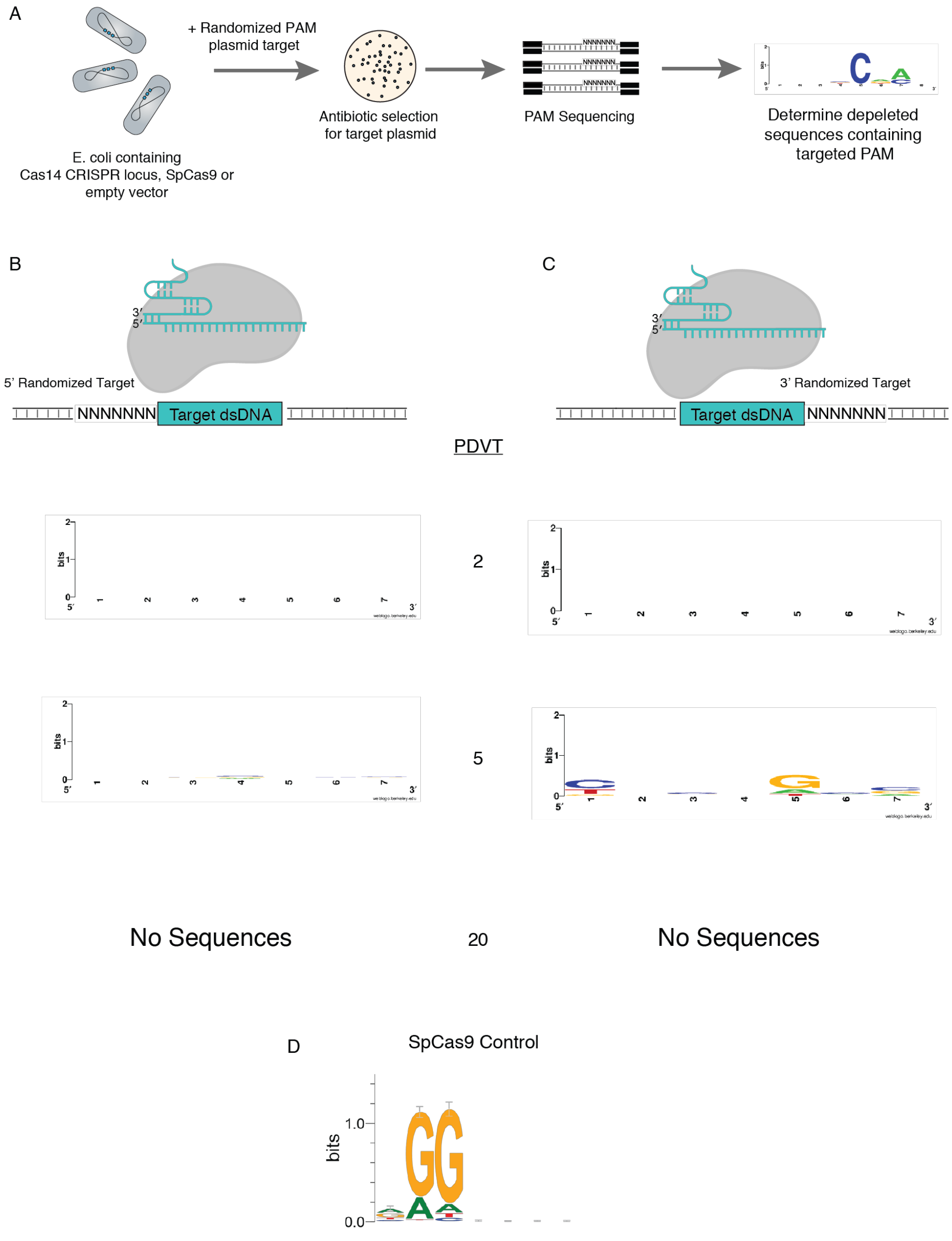


Fig. 6.10 Plasmid depletion by Cas14a1 and SpCas9

(A) Diagram outlining PAM discovery experiment. *E. coli* expressing the CRISPR system of interest is challenged with a plasmid containing a randomized PAM sequence flanking the target. The surviving (transformed) cells are harvest and sequenced along with a control harboring an empty vector. The depleted sequences are then sequenced and PAMs depleted more than the PAM Depletion Value Threshold (PDVT) are used to generate a Weblogo. (B-C) PAM sequences depleted by heterologously expressed Cas14a1 transformed with a target plasmid containing a randomized PAM sequence 5' (B) or 3' (C) of the target. “No sequences” indicates that no sequences were found to be depleted at or above the given PDVT.

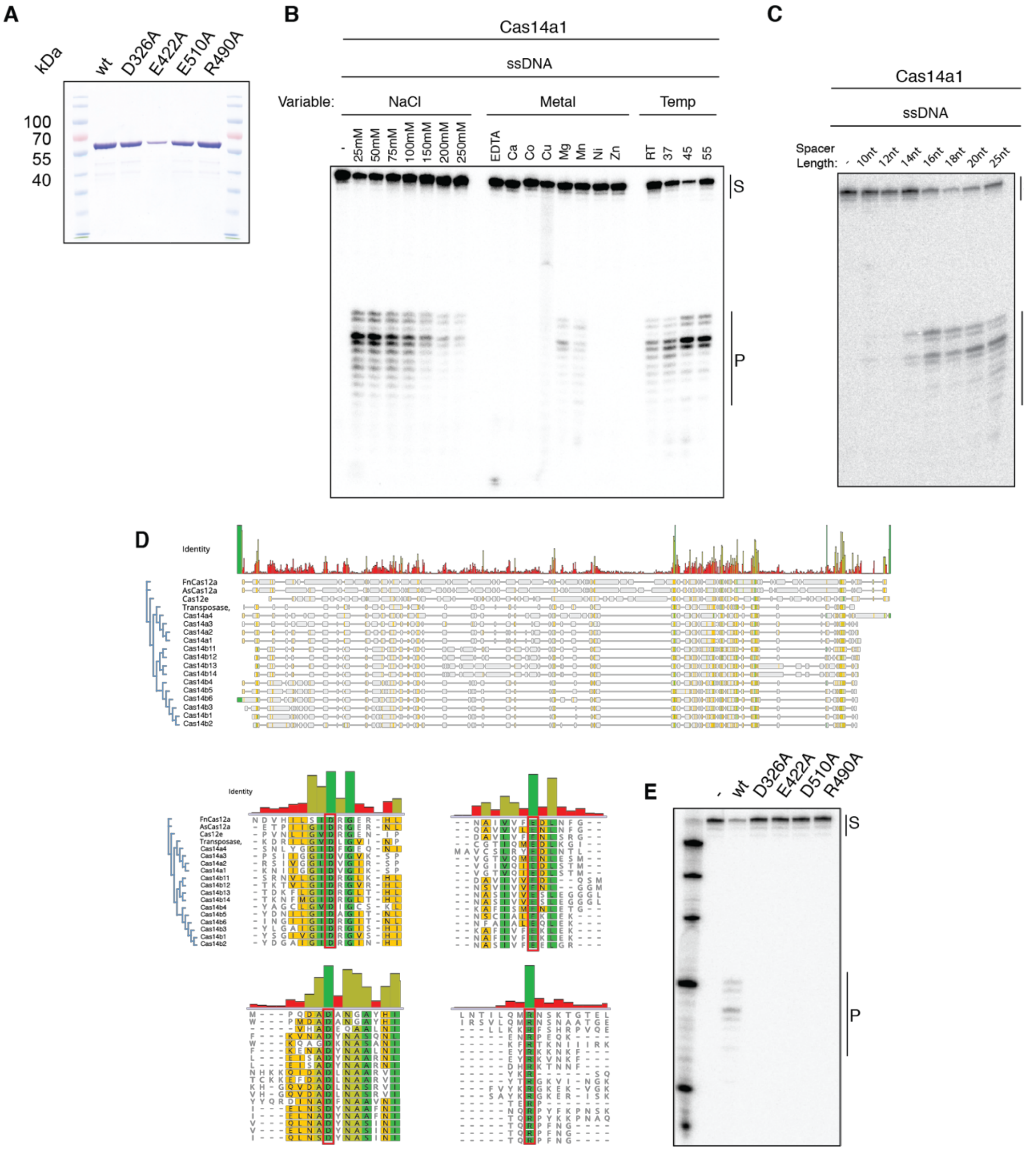


Fig. 6.11 Degradation of ssDNA by Cas14a1

(A) SDS-PAGE of purified Cas14a1 and Cas14a1 point mutants. (B) Optimization of salt, cation and temperature for Cas14a1 cleavage of ssDNA targets. (C) Radiolabled cleavage of ssDNA by Cas14a1 with spacer sequences of various lengths. (D) Alignment of Cas14 with previously studied Cas12 proteins to identify RuvC active site residues and (E) cleavage of ssDNA by purified Cas14a1 RuvC point mutants.

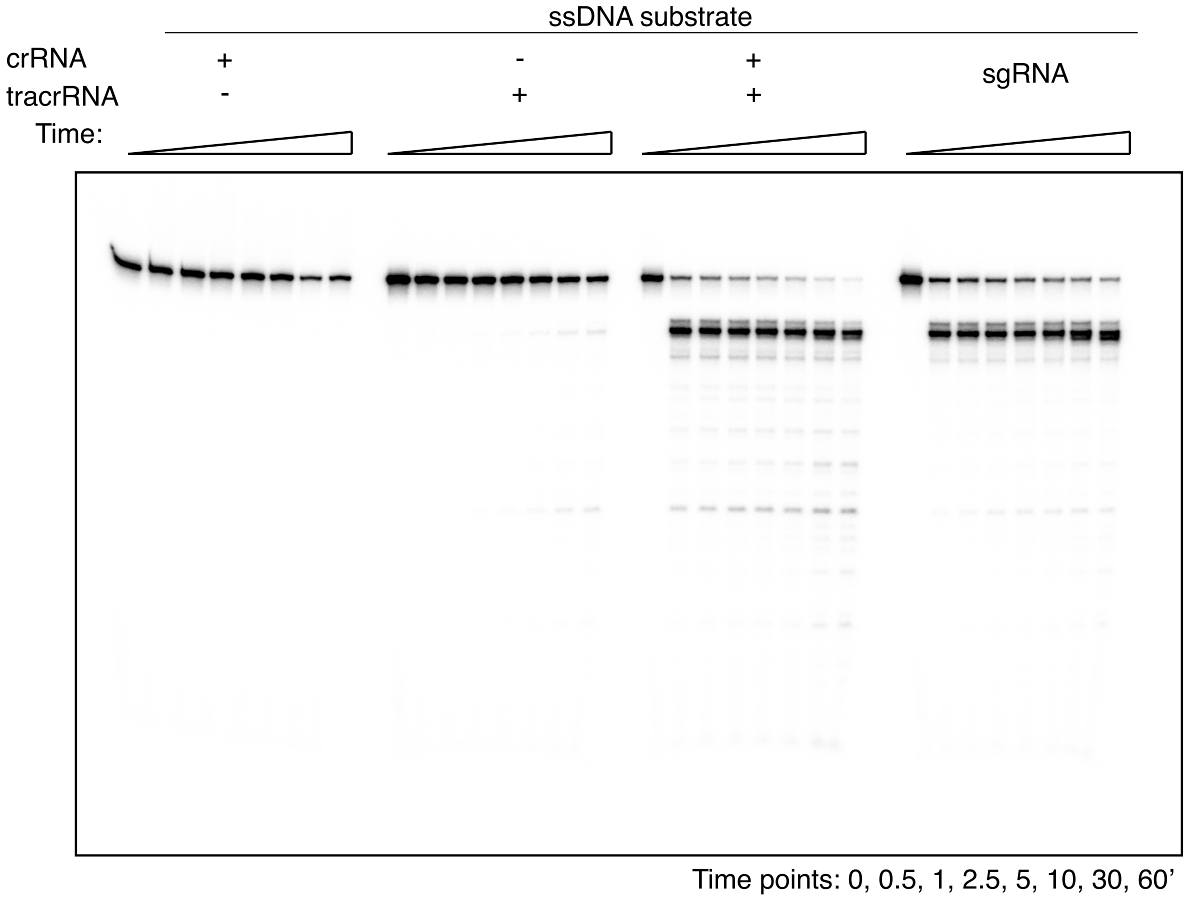


Fig. 6.12 Kinetics of Cas14a1 cleavage of ssDNA with various guide RNA components.

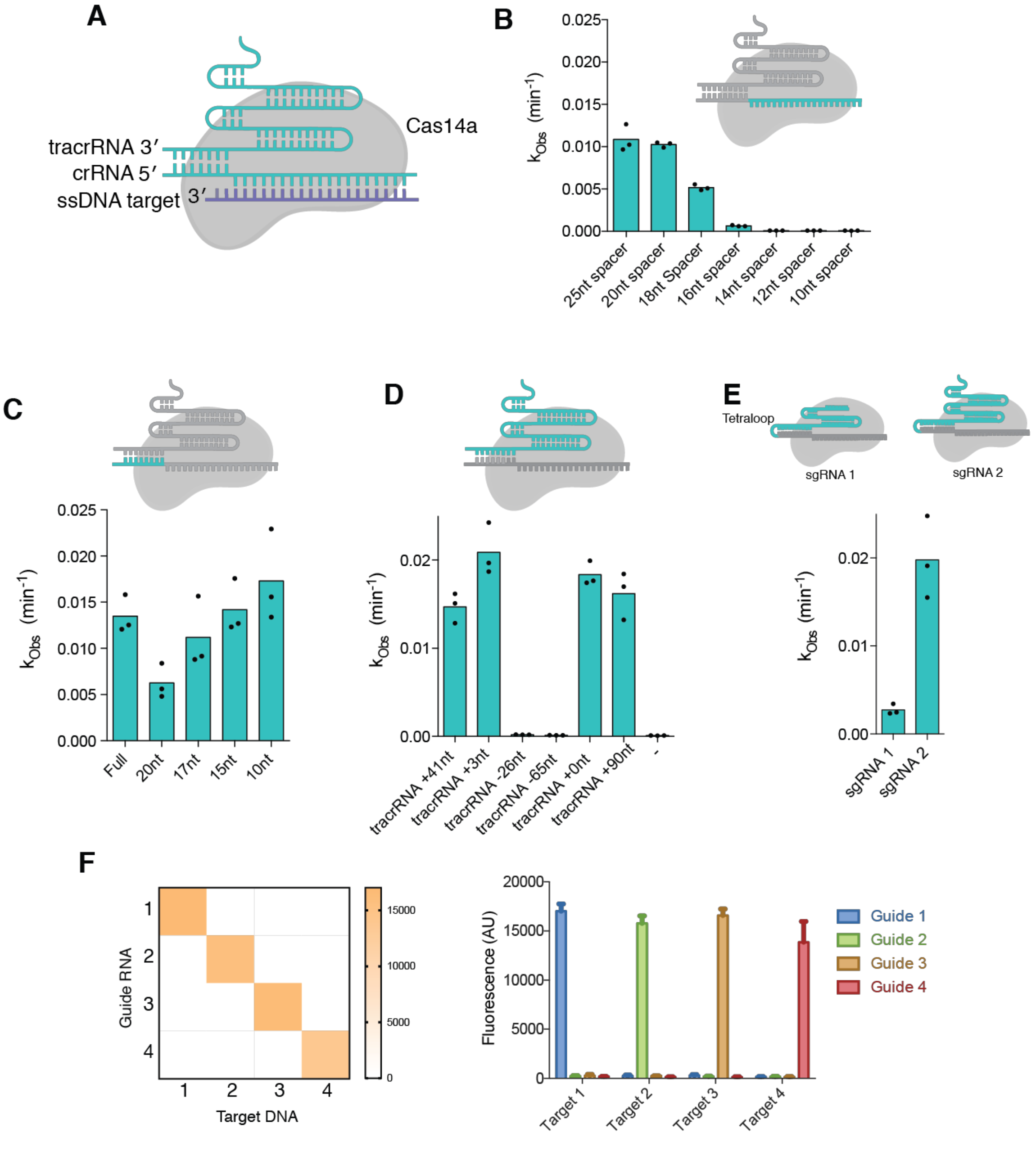


Fig. 6.13 Optimization of Cas14a1 guide RNA components

(A) Diagram of Cas14a1 targeting ssDNA. Impact on Cas14a1 cleavage of an FQ ssDNA substrate by varying the spacer length (B), repeat length (C), tracrRNA (D), and fusing the crRNA and tracrRNA together(E). For the tracrRNA variants +nt label refers to extensions at the 3’ end of the RNA and for the singe guide RNA (sgRNA) variants sgRNA 1 contains a truncated tracrRNA portion of the sgRNA. (F) Heat map showing the background subtracted fluorescence resulting from cleavage of a ssDNA FQ reporter in the presence of various guide and target combinations.

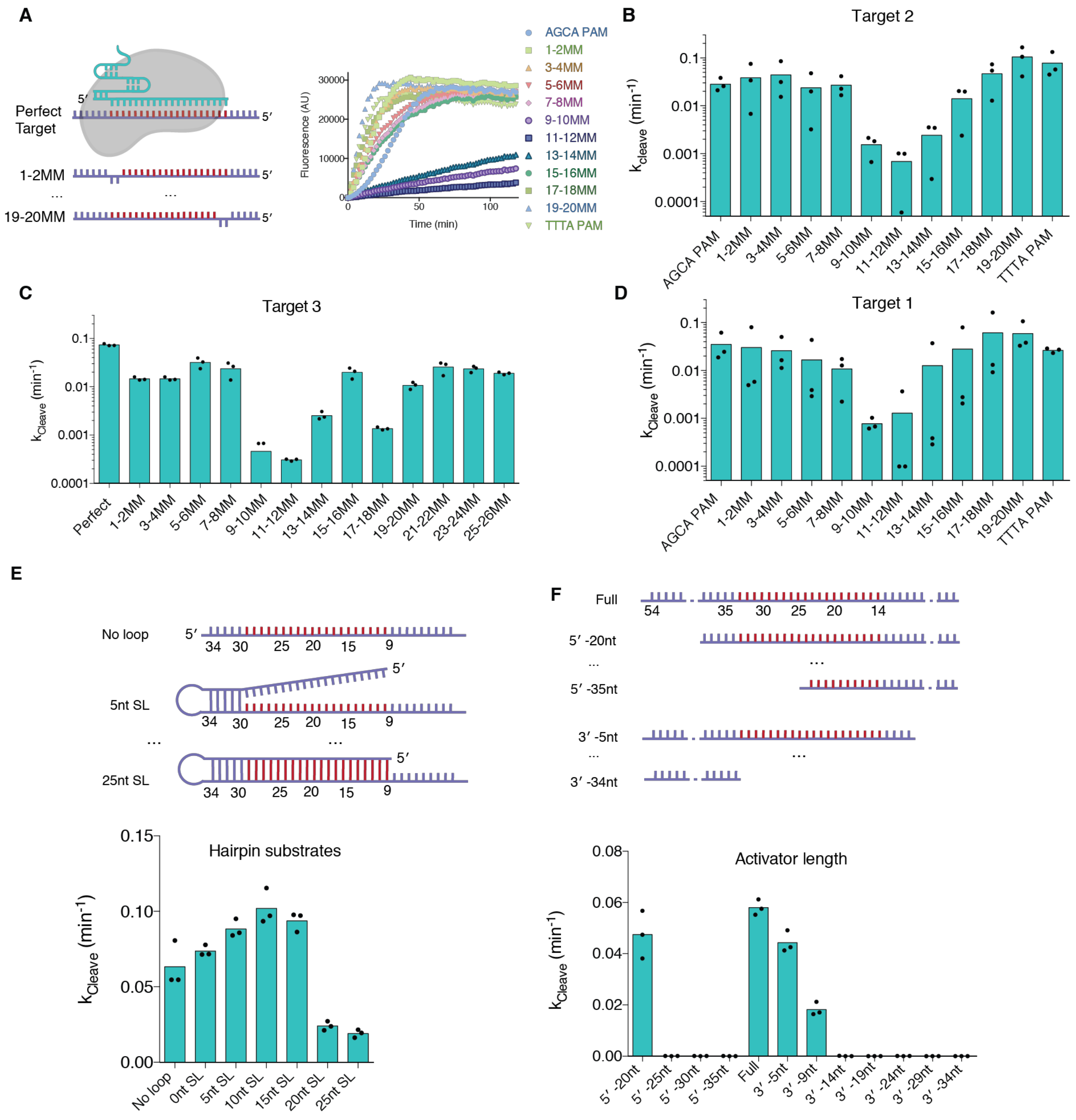


Fig. 6.14 Impact of various activators on Cas14a1 cleavage rate

(A) Diagram of Cas14a1 targeting of ssDNA with position of mismatches used in panels A-D and raw rates for representative replicates of mismatch (MM) position for Target 1. Cleavage rates for Cas14a targeting substrates with mutations tiled across three different substrates (B-D). (E) Trans cleavage rates for substrates with increasing amounts of secondary structure. (F) Trans leavage rates with truncated substrates. Points represent individual measurements.

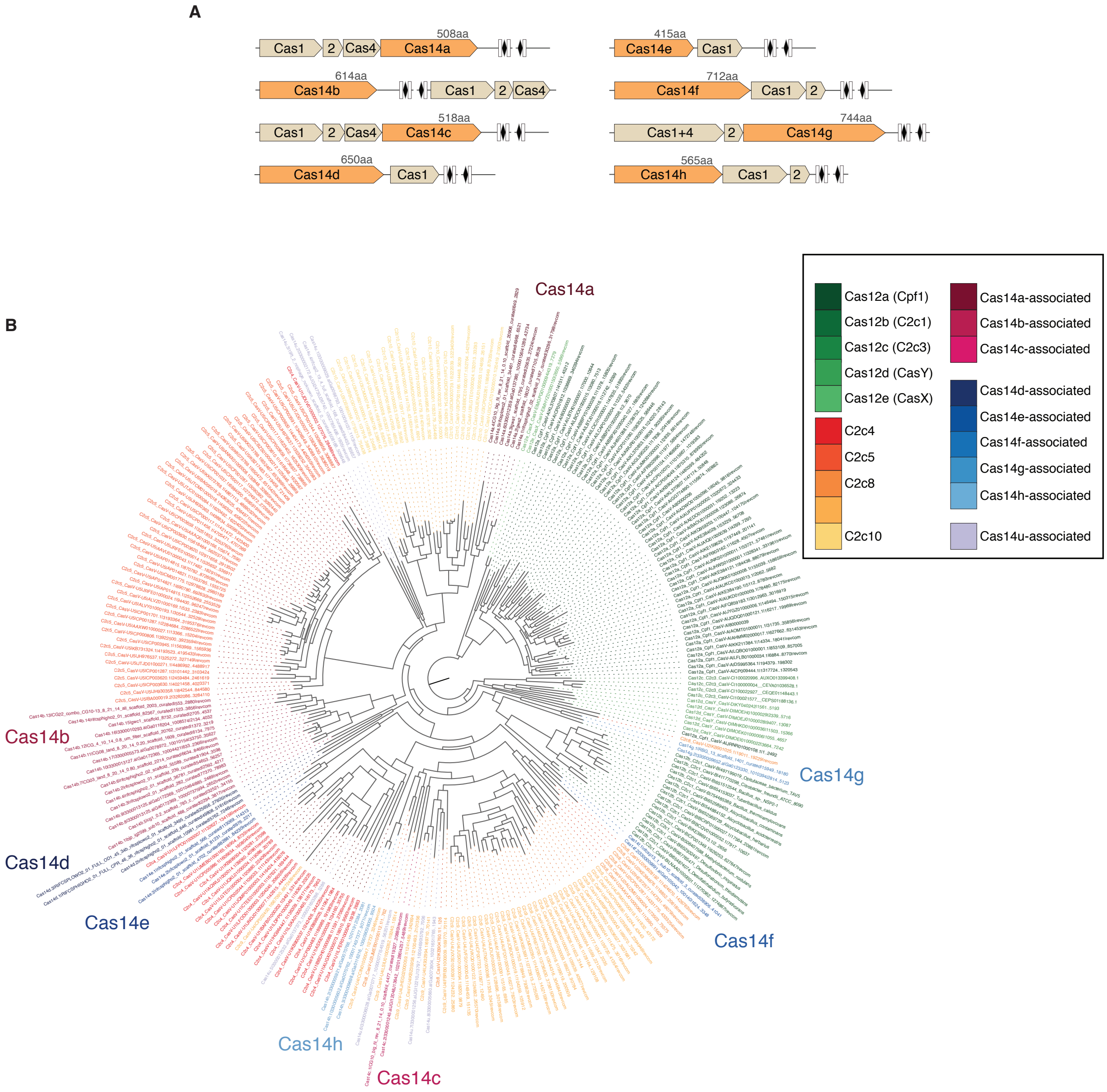


Fig. 6.15 Diversity of CRISPR-Cas14 systems

(A)Representative locus architecture for indicated Cas14 systems. Protein lengths are drawn to scale. The eight Cas14 subtypes. (B) Maximum likelihood tree for Type V effectors including all identified subtypes of Cas14. Detailed tree in Newick format provided in Data S3.

6.6 Acknowledgements

We thank Nan Ma and Kaihong Zhou for technical assistance. We thank Parker Harrington for graphic design assistance. D.B. was supported by a grant from the Paul Allen Frontiers Group, L.B.H., J.C.C. and J.S.C. by US National Science Foundation Graduate Research Fellowships. J.A.D. is an Investigator of the Howard Hughes Medical Institute. This work was supported in part by a Frontiers Science award from the Paul Allen Institute to J.A.D. and J.F.B., a grant from the National Science Foundation (MCB-1244557 to J.A.D.), the Lawrence Berkeley National Laboratory's Sustainable Systems Scientific Focus Area funded by the U.S. Department of Energy, (DE-AC02-05CH11231 to J.F.B.) and the Office of Science of the US Department of Energy under Contract No. DE-AC02–05CH11231.

6.7 Author Contributions

D. B. and D. P. E. conducted the computational analysis. L.B.H., J.S.C., I.P.W., E.M., and J.C.C. designed and executed biochemical investigation of Cas14. L.B.H. designed and conducted experiments investigating Cas14 activity and assembly in *E. coli*. L.B.H. and D.B. conceived of the study. N.C.K. J.F.B. and J.A.D supervised research and experimental design. J.A.D., L.B.H. and D.B. wrote and revised the manuscript. The manuscript was read, edited, and approved by all authors.

COMPETING INTERESTS

UC Regents have filed patents related to this work on which D.B., J.F.B., L.B.H., D.P.E., J.S.C. and J.A.D are inventors. L.B.H. and J.S.C. are co-founders of Mammoth Biosciences. I.P.W. is a consultant for Mammoth Biosciences. J.F.B. is a founder of Metagenomi. J.A.D. is a co-founder of Caribou Biosciences, Editas Medicine, Intellia Therapeutics, Scribe Therapeutics, and Mammoth Biosciences. J.A.D. is a scientific advisory board member of Caribous Biosciences, Intellia Therapeutics, eFFECTOR Therapeutics, Scribe Therapeutics, Synthego, Metagenomi, Mammoth Biosciences and Inari. J.A.D is a member of the board of directors at Driver and Johnson & Johnson and has sponsored research projects by Roche Biopharma and Biogen.

BIBLIOGRAPHY

Abudayyeh, O. O., Gootenberg, J. S., Konermann, S., Joung, J., Slaymaker, I. M., Cox, D. B. T., … Zhang, F. (2016a). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science*, aaf5573. https://doi.org/10.1126/science.aaf5573

Abudayyeh, O. O., Gootenberg, J. S., Konermann, S., Joung, J., Slaymaker, I. M., Cox, D. B. T., … Zhang, F. (2016b). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science*, *353*(6299), 1–9. https://doi.org/10.1126/science.aaf5573

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., … Zwart, P. H. (2010). PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D: Biological Crystallography*, *66*(2), 213–221. https://doi.org/10.1107/S0907444909052925

Amrani, N., Gao, X. D., Liu, P., Gupta, A., Edraki, A., Ibraheim, R., … Sontheimer, E. J. (2017). NmeCas9 is an intrinsically high-fidelity genome editing platform. *BioRxiv*. Retrieved from http://biorxiv.org/content/early/2017/08/04/172650.abstract

Anantharaman, K., Brown, C. T., Hug, L. A., Sharon, I., Castelle, C. J., Probst, A. J., … Banfield, J. F. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications*, *7*, 13219. https://doi.org/10.1038/ncomms13219

Bao, W., & Jurka, J. (2013). Homologues of bacterial TnpB _ IS605 are widespread in diverse eukaryotic transposable elements. *Mobile DNA*, *4*(1), 1. https://doi.org/10.1186/1759-8753-4-12

Barabas, O., Ronning, D. R., Guynet, C., Hickman, A. B., Ton-hoang, B., Chandler, M., & Dyda, F. (2008). Mechanism of IS 200 / IS 605 Family DNA Transposases : Activation and Transposon- Directed Target Site Selection, 208–220. https://doi.org/10.1016/j.cell.2007.12.029

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., … Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science (New York, N.Y.)*, *315*(5819), 1709–1712. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17379808

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., … Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, *315*(5819), 1709–1712. https://doi.org/10.1126/science.1138140

Biswas, A., Gagnon, J. N., Brouns, S. J. J., Fineran, P. C., & Brown, C. M. (2013). CRISPRTarget. *RNA Biology*, *10*(5), 817–827. https://doi.org/10.4161/rna.24046

Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C., & Brown, C. M. (2016). CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics*, *17*(1), 1–14. https://doi.org/10.1186/s12864-016-2627-0

Bolotin, A., Quinquis, B., Sorokin, A., & Dusko Ehrlich, S. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, *151*(8), 2551–2561. https://doi.org/10.1099/mic.0.28048-0

Bondy-Denomy, J., Garcia, B., Strum, S., Du, M., Rollins, M. F., Hidalgo-Reyes, Y., …

Davidson, A. R. (2015). Multiple mechanisms for CRISPR–Cas inhibition by anti-CRISPR proteins. *Nature*, *526*(7571), 136–139. https://doi.org/10.1038/nature15254

Bondy-Denomy, J., Pawluk, A., Maxwell, K. L., & Davidson, A. R. (2013). Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature*, *493*(7432), 429–432. https://doi.org/10.1038/nature11723

Briner, A. E., Donohoue, P. D., Gomaa, A. A., Selle, K., Slorach, E. M., Nye, C. H., … Barrangou, R. (2014). Guide RNA Functional Modules Direct Cas9 Activity and Orthogonality. *Molecular Cell*, *56*(2), 333–339. https://doi.org/10.1016/j.molcel.2014.09.019

Brouns, S. J. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J. H., Snijders, A. P. L., … van der Oost, J. (2008). Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science*, *321*(5891), 960 LP-964. Retrieved from http://science.sciencemag.org/content/321/5891/960.abstract

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., … Banfield, J. F. (2015a). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, *523*(7559), 208–211. https://doi.org/10.1038/nature14486

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., … Banfield, J. F. (2015b). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, *523*(7559), 208–211. https://doi.org/10.1038/nature14486

Burstein, D., Harrington, L. B., Strutt, S. C., & Probst, A. J. (2017). New CRISPR-Cas systems from uncultivated microbes. *Nature Publishing Group*, *542*(7640), 237–241. https://doi.org/10.1038/nature21059

Burstein, D., Harrington, L. B., Strutt, S. C., Probst, A. J., Anantharaman, K., Thomas, B. C., … Banfield, J. F. (2017). New CRISPR-Cas systems from uncultivated microbes. *Nature*, *542*(7640), 237–241. https://doi.org/10.1038/nature21059

Burstein, D., Sun, C. L., Brown, C. T., Sharon, I., Anantharaman, K., Probst, A. J., … Banfield, J. F. (2016). Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. *Nature Communications*, *7*, 10613. https://doi.org/10.1038/ncomms10613

Castelle, C. J., Wrighton, K. C., Thomas, B. C., Hug, L. A., Brown, C. T., Wilkins, M. J., … Banfield, J. F. (2015). Genomic expansion of domain archaea highlights roles for organisms from new phyla in anaerobic carbon cycling. *Current Biology*, *25*(6), 690–701. https://doi.org/10.1016/j.cub.2015.01.014

Cencic, R., Miura, H., Malina, A., Robert, F., Ethier, S., Schmeing, T. M., … Pelletier, J. (2014). Protospacer Adjacent Motif (PAM)-Distal Sequences Engage CRISPR Cas9 DNA Target Cleavage. *PloS One*, *9*(10), e109213. https://doi.org/10.1371/journal.pone.0109213

Chase, J. W., & Williams, K. R. (1986). Single-stranded DNA binding proteins required for DNA replication. *Annual Review of Biochemistry*, *55*, 103–136. https://doi.org/10.1146/annurev.bi.55.070186.000535

Chen, I. M. A., Markowitz, V. M., Chu, K., Palaniappan, K., Szeto, E., Pillay, M., … Kyrpides, N. C. (2017). IMG/M: Integrated genome and metagenome comparative data analysis system. *Nucleic Acids Research*, *45*(D1), D507–D516.

https://doi.org/10.1093/nar/gkw929

Chen, J. S., & Doudna, J. A. (2017). The chemistry of Cas9 and its CRISPR colleagues. *Nature Reviews Chemistry*, *1*(10), 0078. https://doi.org/10.1038/s41570-017-0078

Chen, J. S., Ma, E., Harrington, L. B., Da Costa, M., Tian, X., Palefsky, J. M., & Doudna, J. A. (2018). CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science (New York, N.Y.)*, *360*(6387), 436–439. https://doi.org/10.1126/science.aar6245

Chylinski, K., Le Rhun, A., & Charpentier, E. (2013). The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA Biology*, *10*(5), 726–737. https://doi.org/10.4161/rna.24321

Chylinski, K., Makarova, K. S., Charpentier, E., & Koonin, E. V. (2014a). Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Research*, *42*(10), 6091–6105. https://doi.org/10.1093/nar/gku241

Chylinski, K., Makarova, K. S., Charpentier, E., & Koonin, E. V. (2014b). Classification and evolution of type II CRISPR-Cas systems. *Nucleic Acids Research*, *42*(10), 6091–6105. https://doi.org/10.1093/nar/gku241

Cong, L., Ran, F., Cox, D., Lin, S., & Barretto, R. (2013). Multiplex Genome Engineering Using CRISPR / Cas Systems. *Science*, (January). https://doi.org/10.1038/nbt1319

Cordova, L. T., Long, C. P., Venkataramanan, K. P., & Antoniewicz, M. R. (2015). Complete genome sequence, metabolic model construction and phenotypic characterization of Geobacillus LC300, an extremely thermophilic, fast growing, xylose-utilizing bacterium. *Metabolic Engineering*, *32*, 74–81. https://doi.org/10.1016/j.ymben.2015.09.009

Crooks, G., Hon, G., Chandonia, J., & Brenner, S. (2004). NCBI GenBank FTP Site\nWebLogo: a sequence logo generator. *Genome Res*, *14*, 1188–1190. https://doi.org/10.1101/gr.849004.1

Dagdas, Y. S., Chen, J. S., Sternberg, S. H., Doudna, J. A., & Yildiz, A. (2017). A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9. *BioRxiv*. https://doi.org/10.1101/122242

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., … Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, *471*(7340), 602–607. https://doi.org/10.1038/nature09886

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. a, … Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, *471*(7340), 602–607. https://doi.org/10.1038/nature09886

Deng, L., Garrett, R. A., Shah, S. A., Peng, X., & She, Q. (2013). A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus. *Molecular Microbiology*, *87*(5), 1088–1099. https://doi.org/10.1111/mmi.12152

Deveau, H., Barrangou, R., Garneau, J. E., Labonté, J., Fremaux, C., Boyaval, P., … Moineau, S. (2008). Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *Journal of Bacteriology*, *190*(4), 1390–1400. https://doi.org/10.1128/JB.01412-07

Dong, D., Guo, M., Wang, S., Zhu, Y., Wang, S., Xiong, Z., … Huang, Z. (2017). Structural basis of CRISPR–SpyCas9 inhibition by an anti-CRISPR protein. *Nature*, 1–15. https://doi.org/10.1038/nature22377

Donk, P. J. (1920). A Highly Resistant Thermophilic Organism. *Journal of Bacteriology*, *5*(4), 373–374.

Doudna, J. A., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science* , *346*(6213). https://doi.org/10.1126/science.1258096

East-Seletsky, A., O'Connell, M. R., Knight, S. C., Burstein, D., Cate, J. H. D., Tjian, R., & Doudna, J. A. (2016). Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature*, *538*(7624), 270–273. https://doi.org/10.1038/nature19802

Eiberg, H., Troelsen, J., Nielsen, M., Mikkelsen, A., Mengel-From, J., Kjaer, K. W., & Hansen, L. (2008). Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Human Genetics*, *123*(2), 177–187. https://doi.org/10.1007/s00439-007-0460-x

Emsley, P., & Cowtan, K. (2004). Coot: Model-building tools for molecular graphics. *Acta Crystallographica Section D: Biological Crystallography*, *60*(12 I), 2126–2132. https://doi.org/10.1107/S0907444904019158

Esvelt, K. M., Mali, P., Braff, J. L., Moosburner, M., Yaung, S. J., & Church, G. M. (2013a). Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nature Methods*, *10*(11), 1116–1121. https://doi.org/10.1038/nmeth.2681

Esvelt, K. M., Mali, P., Braff, J. L., Moosburner, M., Yaung, S. J., & Church, G. M. (2013b). Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat Meth*, *10*(11), 1116–1121. Retrieved from http://dx.doi.org/10.1038/nmeth.2681

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, *39*(suppl 2), W29–W37. https://doi.org/10.1093/nar/gkr367

Fonfara, I., Le Rhun, A., Chylinski, K., Makarova, K. S., Lécrivain, A.-L., Bzdrenga, J., … Charpentier, E. (2014). Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic Acids Research*, *42*(4), 2577–2590. https://doi.org/10.1093/nar/gkt1074

Franzosa, E. A., & Xia, Y. (2011). Structural principles within the human-virus protein-protein interaction network. *Proceedings of the National Academy of Sciences* , *108*(26), 10538–10543. https://doi.org/10.1073/pnas.1101440108

Friedland, A. E., Baral, R., Singhal, P., Loveluck, K., Shen, S., Sanchez, M., … Bumcrot, D. (2015). Characterization of Staphylococcus aureus Cas9 : a smaller Cas9 for all-in-one adeno-associated virus delivery and paired nickase applications. *Genome Biology*, 1–10. https://doi.org/10.1186/s13059-015-0817-8

Fujii, M., Takagi, M., Imanaka, T., & Aiba, S. (1983). Molecular Cloning of a Thermostable Neutral Protease Gene from Bacillus stearothermophilus in a Vector Plasmid and Its Expression in Bacillus stearothermophilus and Bacillus subtilis. *Microbiology*, *154*(2), 831–837.

Garneau, J. E., Dupuis, M.-È., Villion, M., Romero, D. a, Barrangou, R., Boyaval, P., … Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, *468*(7320), 67–71. https://doi.org/10.1038/nature09523

Gasiunas, G., Barrangou, R., Horvath, P., & Siksnys, V. (2012). Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences* , *109*(39), E2579–E2586. https://doi.org/10.1073/pnas.1208507109

Godde, J. S., & Bickerton, A. (2006). The repetitive DNA elements called CRISPRs and their associated genes: Evidence of horizontal transfer among prokaryotes. *Journal of Molecular Evolution*, *62*(6), 718–729. https://doi.org/10.1007/s00239-005-0223-z

Grissa, I., Vergnaud, G., & Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research*, *35*(suppl 2), W52–W57. https://doi.org/10.1093/nar/gkm360

Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., … Terns, M. P. (2009). RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. *Cell*, *139*(5), 945–956. https://doi.org/10.1016/j.cell.2009.07.040

Harrington, L. B., Doxzen, K. W., Ma, E., Liu, J. J., Knott, G. J., Edraki, A., … Doudna, J. A. (2017). A Broad-Spectrum Inhibitor of CRISPR-Cas9. *Cell*, *170*(6), 1224–1233.e15. https://doi.org/10.1016/j.cell.2017.07.037

Harrington, L. B., Paez-espino, D., Chen, J. S., Staahl, B. T., Kyrpides, N. C., & Doudna, J. (2017). A thermostable Cas9 with increased lifetime in human plasma. *BioRxiv*.

Harrington, L. B., Paez-Espino, D., Staahl, B. T., Chen, J. S., Ma, E., Kyrpides, N. C., & Doudna, J. A. (2017). A thermostable Cas9 with increased lifetime in human plasma. *Nature Communications*, *8*(1), 1–7. https://doi.org/10.1038/s41467-017-01408-4

He, S., Lavatine, L., Dyda, F., Siguier, P., Caumont-Sarcos, A., Chandler, M., … Ton Hoang, B. (2015). The IS200/IS605 Family and “Peel and Paste” Single-strand Transposition Mechanism. *Microbiology Spectrum*, *3*(4), 1–21. https://doi.org/10.1128/microbiolspec.MDNA3-0039-2014

Hirano, H., Gootenberg, J. S., Horii, T., Abudayyeh, O. O., Kimura, M., Hsu, P. D., … Nureki, O. (2016). Structure and Engineering of Francisella novicida Cas9. *Cell*, *164*, 1–12. https://doi.org/10.1016/j.cell.2016.01.039

Holm, L., & Laakso, L. M. (2016). Dali server update. *Nucleic Acids Research*, *44*(W1), W351–W355. https://doi.org/10.1093/nar/gkw357

Hou, Z., Zhang, Y., Propson, N. E., Howden, S. E., Chu, L.-F., Sontheimer, E. J., & Thomson, J. A. (2013). Efficient genome engineering in human pluripotent stem cells using Cas9 from Neisseria meningitidis. *Proceedings of the National Academy of Sciences* , *110*(39), 15644–15649. https://doi.org/10.1073/pnas.1313587110

Houte, S. van, Ekroth, A. K. E., Broniewski, J. M., Chabas, H., Ashby, B., Gandon, S., … Westra, E. R. (2016). The diversity-generating benefits of a prokaryotic adaptive immune system. *Nature*, *532*(7599), 385–388. https://doi.org/10.1038/nature17436

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., … Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, *1*(5), 16048. https://doi.org/10.1038/nmicrobiol.2016.48

Ingram, L. O., Jarboe, L. R., Zhang, X., Wang, X., Moore, J. C., & Shanmugam, K. T. (2010). Metabolic engineering for production of biorenewable fuels and chemicals: Contributions of synthetic biology. *Journal of Biomedicine and Biotechnology*, *2010*. https://doi.org/10.1155/2010/761042

Jackson, S. A., McKenzie, R. E., Fagerlund, R. D., Kieper, S. N., Fineran, P. C., & Brouns, S. J. J. (2017). *CRISPR-Cas: Adapting to change*. *Science*. https://doi.org/10.1126/science.aal5056

Jiang, F., Taylor, D. W., Chen, J. S., Kornfeld, J. E., Zhou, K., Thompson, A. J., … Doudna, J. A. (2016). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science*, *8282*(January), 1–8. https://doi.org/10.1126/science.aad8282

Jiang, F., Zhou, K., Ma, L., Gressel, S., & Doudna, J. a. (2015). A Cas9-guide RNA complex preorganized for target DNA recognition. *Science*, *348*(6242), 1477–1481. https://doi.org/10.1126/science.aab1452

Jiang, W., & Marraffini, L. a. (2015). CRISPR-Cas: New Tools for Genetic Manipulations from Bacterial Immunity Systems. *Annual Review of Microbiology*, *69*(1), 150724172101001. https://doi.org/10.1146/annurev-micro-091014-104441

Jiang, W., Samai, P., & Marraffini, L. A. (2016). Degradation of Phage Transcripts by CRISPR-Associated RNases Enables Type III CRISPR-Cas Immunity. *Cell*, 1–12. https://doi.org/10.1016/j.cell.2015.12.053

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, *337*(6096), 816–821. https://doi.org/10.1126/science.1225829

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012a). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* , *337*(6096), 816–821. https://doi.org/10.1126/science.1225829

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012b). A Programmable Dual-RNA – Guided, *337*(August), 816–822.

Jinek, M., Jiang, F., Taylor, D. W., Sternberg, S. H., Kaya, E., Ma, E., … Doudna, J. a. (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science (New York, N.Y.)*, *343*(6176), 1247997. https://doi.org/10.1126/science.1247997

Kabsch, W. (2010). Xds. *Acta Crystallographica Section D: Biological Crystallography*, *66*(2), 125–132. https://doi.org/10.1107/S0907444909047337

Kantor, R. S., Wrighton, K. C., Handley, K. M., Sharon, I., Hug, L. A., Castelle, C. J., … Banfield, J. F. (2013). Small genomes and sparse metabolisms of sediment-associated bacteria from four candidate phyla. *MBio*, *4*(5), e00708-13. https://doi.org/10.1128/mBio.00708-13

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, *10*(6), 845–858. https://doi.org/10.1038/nprot.2015.053

Kelly, L. A., Mezulis, S., Yates, C., Wass, M., & Sternberg, M. (2015). The Phyre2 web portal for protein modelling, prediction, and analysis. *Nature Protocols*, *10*(6), 845–858. https://doi.org/10.1038/nprot.2015-053

Kim, E., Koo, T., Park, S. W., Kim, D., Kim, K., Cho, H., … Kim, J. (2017). In vivo genome editing with a small Cas9 orthologue derived from Campylobacter jejuni. *Nature Communications*, *8*, 1–12. https://doi.org/10.1038/ncomms14500

Kleinstiver, B. P., Prew, M. S., Tsai, S. Q., Nguyen, N. T., Topkar, V. V, Zheng, Z., & Joung, J. K. (2015). Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. *Nature Biotechnology*, (November), 1–7. https://doi.org/10.1038/nbt.3404

Kleinstiver, B. P., Prew, M. S., Tsai, S. Q., Topkar, V. V., Nguyen, N. T., Zheng, Z., … Joung, J. K. (2015). Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature*. https://doi.org/10.1038/nature14592

Knott, G. J., East-Seletsky, A., Cofsky, J. C., Holton, J. M., Charles, E., O'Connell, M. R., & Doudna, J. A. (2017). Guide-bound structures of an RNA-targeting A-cleaving CRISPR-Cas13a enzyme. *Nature Structural and Molecular Biology*, *24*(10), 825–833. https://doi.org/10.1038/nsmb.3466

Konarev, P. V., Petoukhov, M. V., Volkov, V. V., & Svergun, D. I. (2006). ATSAS 2.1, a program package for small-angle scattering data analysis. *Journal of Applied Crystallography*, *39*(2), 277–286. https://doi.org/10.1107/S0021889806004699

Konarev, P. V., Volkov, V. V., Sokolova, A. V., Koch, M. H. J., & Svergun, D. I. (2003). *PRIMUS* : a Windows PC-based system for small-angle scattering data analysis. *Journal of Applied Crystallography*, *36*(5), 1277–1282. https://doi.org/10.1107/S0021889803012779

Konermann, S., Brigham, M. D., Trevino, A. E., Joung, J., Abudayyeh, O. O., Barcena, C., … Zhang, F. (2014). Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature*, *517*(7536), 583–588. https://doi.org/10.1038/nature14136

Kontermann, R. E. (2011). Strategies for extended serum half-life of protein therapeutics. *Current Opinion in Biotechnology*, *22*(6), 868–876. https://doi.org/10.1016/j.copbio.2011.06.012

Koonin, E. V, Makarova, K. S., & Zhang, F. (2017). Diversity, classification and evolution of CRISPR-Cas systems. *Current Opinion in Microbiology*, *37*, 67–78. https://doi.org/10.1016/j.mib.2017.05.008

Lander, G. C., Stagg, S. M., Voss, N. R., Cheng, A., Fellmann, D., Yoshioka, C., … Carragher, B. (2009). Image Processing. *Access*, *166*(1), 95–102. https://doi.org/10.1016/j.jsb.2009.01.002.Appion

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

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 Research*, *28*(4), 491–493. https://doi.org/10.1038/s41422-018-0022-x

Li, Y., Pan, S., Zhang, Y., Ren, M., Feng, M., Peng, N., … She, Q. (2015). Harnessing Type i and Type III CRISPR-Cas systems for genome editing. *Nucleic Acids Research*, *44*(4). https://doi.org/10.1093/nar/gkv1044

Lin, S., Staahl, B., Alla, R. K., & Doudna, J. a. (2014). Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *ELife*, *3*(December), 1–13. https://doi.org/10.7554/eLife.04766

Liu, J.-J., Bratkowski, M. a, Liu, X., Niu, C.-Y., Ke, A., & Wang, H.-W. (2014). Visualization of distinct substrate-recruitment pathways in the yeast exosome by EM. *Nature Structural & Molecular Biology*, *21*(1), 95–102. https://doi.org/10.1038/nsmb.2736

Liu, L., Li, X., Ma, J., Li, Z., You, L., Wang, J., … Wang, Y. (2017). The Molecular Architecture for RNA-Guided RNA Cleavage by Cas13a. *Cell*, *170*(4), 714–726.e10. https://doi.org/10.1016/j.cell.2017.06.050

Luef, B., Frischkorn, K. R., Wrighton, K. C., Holman, H.-Y. N., Birarda, G., Thomas, B. C., … Banfield, J. F. (2015). Diverse uncultivated ultra-small bacterial cells in groundwater. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms7372

Ma, E., Harrington, L. B., O'Connell, M. R., Zhou, K., & Doudna, J. A. (2015). Single-Stranded DNA Cleavage by Divergent CRISPR-Cas9 Enzymes. *Molecular Cell*, *60*(3), 398–407. https://doi.org/10.1016/j.molcel.2015.10.030

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., … Koonin, E. V. (2015). An updated evolutionary classification of CRISPR-Cas systems. *Nature Reviews Microbiology*, *advance on*. https://doi.org/10.1038/nrmicro3569

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., … Koonin, E. V. (2015). An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol*, 1–15. https://doi.org/10.1038/nrmicro3569

Mali, P., Aach, J., Stranges, P. B., Esvelt, K. M., Moosburner, M., Kosuri, S., … Church, G. M. (2013). CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nature Biotechnology*, *31*(9), 833–838. https://doi.org/10.1038/nbt.2675

Markowitz, V. M., Chen, I. M. A., Chu, K., Szeto, E., Palaniappan, K., Pillay, M., … Kyrpides, N. C. (2014). IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Research*, *42*(D1), 568–573. https://doi.org/10.1093/nar/gkt919

Markowitz, V. M., Chen, I. M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., … Kyrpides, N. C. (2012). IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research*, *40*(D1), 115–122. https://doi.org/10.1093/nar/gkr1044

Marraffini, L. a, & Sontheimer, E. J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science*, *322*(5909), 1843–1845. https://doi.org/10.1126/science.1165771

Mekler, V., Minakhin, L., & Severinov, K. (2017). Mechanism of duplex DNA destabilization by RNA-guided Cas9 nuclease during target interrogation. *Proceedings of the National Academy of Sciences*, 201619926. https://doi.org/10.1073/pnas.1619926114

Mojica, F. J. M., Díez-Villaseñor, C., García-Martínez, J., & Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, *155*(3), 733–740. https://doi.org/10.1099/mic.0.023960-0

Mojica, F. J. M., Diez-Villasenor, C., Garcia-Martinez, J., Almendros, C., Díez-Villaseñor, C., García-Martínez, J., … Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, *155*(Pt 3), 733–740. https://doi.org/10.1099/mic.0.023960-0

Morgan, G. J., Hatfull, G. F., Casjens, S., & Hendrix, R. W. (2002). Bacteriophage Mu Genome Sequence : Analysis and Comparison with Mu-like Prophages in Haemophilus , Neisseria and Deinococcus. https://doi.org/10.1006/jmbi.2001.5437

Mougiakos, I., Bosma, E. F., Weenink, K., Vossen, E. M., Goijvaerts, K., Van Der Oost, J., & van Kranenburg, R. (2017). Efficient genome editing of a facultative thermophile using the mesophilic spCas9. *ACS Synthetic Biology*, acssynbio.6b00339. https://doi.org/10.1021/acssynbio.6b00339

Narasimhan, D., Nance, M. R., Gao, D., Ko, M. C., MacDonald, J., Tamburi, P., … Sunahara, R. K. (2010). Structural analysis of thermostabilizing mutations of cocaine esterase. *Protein Engineering, Design and Selection*, *23*(7), 537–547. https://doi.org/10.1093/protein/gzq025

Nelson, W. C., & Stegen, J. C. (2015). The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle. *Frontiers in Microbiology*, *6*. https://doi.org/10.3389/fmicb.2015.00713

Nishimasu, H., Ran, F. A., Hsu, P. D., Konermann, S., Shehata, S. I., Dohmae, N., … Nureki, O. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, *156*(5), 935–949. https://doi.org/10.1016/j.cell.2014.02.001

Notomi, T., Okayama, H., Masubuchi, H., Yonekawa, T., Watanabe, K., Amino, N., & Hase, T. (2000). Loop-mediated isothermal amplification of DNA. *Nucleic Acids Research*, *28*(12), E63. https://doi.org/10.1093/nar/28.12.e63

Nuñez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N., & Doudna, J. a. (2015). Foreign DNA capture during CRISPR–Cas adaptive immunity. *Nature*, 1–13. https://doi.org/10.1038/nature15760

Nuñez, J. K., Kranzusch, P. J., Noeske, J., Wright, A. V., Davies, C. W., & Doudna, J. A. (2014a). Cas1–Cas2 complex formation mediates spacer acquisition during CRISPR–Cas adaptive immunity. *Nature Structural & Molecular Biology*, *21*(6), 528–534. https://doi.org/10.1038/nsmb.2820

Nuñez, J. K., Kranzusch, P. J., Noeske, J., Wright, A. V, Davies, C. W., & Doudna, J. a. (2014b). Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nature Structural & Molecular Biology*, *21*(6), 528–534. https://doi.org/10.1038/nsmb.2820

Nuñez, J. K., Lee, A. S. Y., Engelman, A., & Doudna, J. a. (2015). Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. *Nature*, *519*(7542), 193–198. https://doi.org/10.1038/nature14237

O'Connell, M. R., Oakes, B. L., Sternberg, S. H., East-Seletsky, A., Kaplan, M., & Doudna, J. A. (2014). Programmable RNA recognition and cleavage by CRISPR/Cas9. *Nature*, *516*(7530), 263–266. https://doi.org/10.1038/nature13769

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., & Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *27*(1), 29–34. https://doi.org/10.1093/nar/27.1.29

Paez-Espino, D., Sharon, I., Morovic, W., Stahl, B., Thomas, B. C., Barrangou, R., & Banfielda, J. F. (2015). CRISPR immunity drives rapid phage genome evolution in streptococcus thermophilus. *MBio*, *6*(2), 1–9. https://doi.org/10.1128/mBio.00262-15

Pawluk, A., Amrani, N., Zhang, Y., Sontheimer, E. J., Maxwell, K. L., Davidson, A. R., … Edraki, A. (2016). Naturally Occurring Off-Switches for CRISPR-Cas9 Article Naturally Occurring Off-Switches for CRISPR-Cas9. *Cell*, *167*(7), 1–10. https://doi.org/10.1016/j.cell.2016.11.017

Pawluk, A., Bondy-denomy, J., Cheung, V. H. W., Maxwell, K. L., & Davidson, R. (2014). A New Group of Phage Anti-CRISPR Genes Inhibits the Type I-E CRISPR-Cas System of Pseudomonas aeruginosa, *5*(2), 1–7. https://doi.org/10.1128/mBio.00896-14.Editor

Pawluk, A., Staals, R. H. J., Taylor, C., Watson, B. N. J., Saha, S., Fineran, P. C., …

Davidson, A. R. (2016). Inactivation of CRISPR-Cas systems by anti-CRISPR proteins in diverse bacterial species. *Nature Microbiology*, *1*(June), 16085. https://doi.org/10.1038/nmicrobiol.2016.85

Porteus, M. (2016). Genome Editing: A New Approach to Human Therapeutics. *Annual Review of Pharmacology and Toxicology*, *56*(1), 163–190. https://doi.org/10.1146/annurev-pharmtox-010814-124454

Probst, A. J., Castelle, C. J., Singh, A., Brown, C. T., Anantharaman, K., Sharon, I., … Banfield, J. F. (2017). Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO2concentrations. *Environmental Microbiology*, *19*(2), 459–474. https://doi.org/10.1111/1462-2920.13362

Rambo, R. P., & Tainer, J. a. (2013). Accurate assessment of mass, models and resolution by small-angle scattering. *Nature*, *496*(7446), 477–481. https://doi.org/10.1038/nature12070

Ran, F. A., Cong, L., Yan, W. X., Scott, D. a., Gootenberg, J. S., Kriz, A. J., … Zhang, F. (2015a). In vivo genome editing using Staphylococcus aureus Cas9. *Nature*. https://doi.org/10.1038/nature14299

Ran, F. A., Cong, L., Yan, W. X., Scott, D. A., Gootenberg, J. S., Kriz, A. J., … Zhang, F. (2015b). In vivo genome editing using Staphylococcus aureus Cas9. *Nature*, *520*(7546), 186–191. Retrieved from http://dx.doi.org/10.1038/nature14299

Rauch, B. J., Silvis, M. R., Hultquist, J. F., Waters, C. S., McGregor, M. J., Krogan, N. J., & Bondy-Denomy, J. (2017). Inhibition of CRISPR-Cas9 with Bacteriophage Proteins. *Cell*, *168*(1), 150–158.e10. https://doi.org/10.1016/j.cell.2016.12.009

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, *9*(2), 173–175. https://doi.org/10.1038/nmeth.1818

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N. N., Anderson, I. J., Cheng, J. F., … Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, *499*(7459), 431–437. https://doi.org/10.1038/nature12352

Samai, P., Pyenson, N., Jiang, W., Goldberg, G. W., Hatoum-Aslan, A., & Marraffini, L. A. (2015a). Co-transcriptional DNA and RNA cleavage during type III CRISPR-cas immunity. *Cell*, *161*(5), 1164–1174. https://doi.org/10.1016/j.cell.2015.04.027

Samai, P., Pyenson, N., Jiang, W., Goldberg, G. W., Hatoum-Aslan, A., & Marraffini, L. A. (2015b). Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. *Cell*, *161*(5), 1164–1174. https://doi.org/10.1016/j.cell.2015.04.027

Sampson, T. R., Saroj, S. D., Llewellyn, A. C., Tzeng, Y.-L., & Weiss, D. S. (2013). A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. *Nature*, *497*(7448), 254–257. https://doi.org/10.1038/nature12048

Savell, K. E., & Day, J. J. (2017). Applications of CRISPR/CAS9 in the mammalian central nervous system. *Yale Journal of Biology and Medicine*. https://doi.org/10.1126/science.1225829

Sawle, L., & Ghosh, K. (2011). How do thermophilic proteins and proteomes withstand high temperature? *Biophysical Journal*, *101*(1), 217–227. https://doi.org/10.1016/j.bpj.2011.05.059

Schellenberger, V., Wang, C., Geething, N. C., Spink, B. J., Campbell, A., To, W., … Stemmer, W. P. C. (2009). A recombinant polypeptide extends the in vivo half-life of peptides and proteins in a tunable manner. *Nature Biotechnology*, *27*(12), 1186–1190. https://doi.org/10.1038/nbt.1588

Semenyuk, A. V., & Svergun, D. I. (1991). GNOM. A program package for small-angle scattering data processing. *Journal of Applied Crystallography*, *24*(pt 5), 537–540. https://doi.org/10.1107/S002188989100081X

Shin, J., Jiang, F., Liu, J.-J., Bray, N. L., Rauch, B. J., Baik, S. H., … Doudna, J. A. (2017). Disabling Cas9 by an anti-CRISPR DNA mimic. *BioRxiv*. Retrieved from http://biorxiv.org/content/early/2017/04/23/129627.abstract

Shmakov, S., Abudayyeh, O. O., Makarova, K. S., Wolf, Y. I., Gootenberg, J. S., Semenova, E., … Koonin, E. V. (2015). Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. *Molecular Cell*, *60*(3), 385–397. https://doi.org/10.1016/j.molcel.2015.10.008

Shmakov, S., Smargon, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., … Koonin, E. V. (2017a). Diversity and evolution of class 2 CRISPR–Cas systems. *Nature Reviews Microbiology*, *15*(3), 169–182. https://doi.org/10.1038/nrmicro.2016.184

Shmakov, S., Smargon, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., … Koonin, E. V. (2017b). Diversity and evolution of class 2 CRISPR–Cas systems. *Nature Reviews Microbiology*, *15*(3), 169–182. https://doi.org/10.1038/nrmicro.2016.184

Sontheimer, E., & Barrangou, R. (2015). The Bacterial Origins of the CRISPR Genome Editing Revolution. *Human Gene Therapy*. https://doi.org/10.1089/hum.2015.091

Staahl, B. T., Benekareddy, M., Coulon-Bainier, C., Banfal, A. A., Floor, S. N., Sabo, J. K., … Doudna, J. A. (2017). Efficient genome editing in the mouse brain by local delivery of engineered Cas9 ribonucleoprotein complexes. *Nature Biotechnology*, (August 2016). https://doi.org/10.1038/nbt.3806

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Sternberg, S. H., LaFrance, B., Kaplan, M., & Doudna, J. A. (2015). Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature*, *527*(7576), 1–14. https://doi.org/10.1038/nature15544

Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C., & Doudna, J. a. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature*, *507*(7490), 62–67. https://doi.org/10.1038/nature13011

Szczelkun, M. D., Tikhomirova, M. S., Sinkunas, T., Gasiunas, G., Karvelis, T., Pschera, P., … Seidel, R. (2014). Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(27), 9798–9803. https://doi.org/10.1073/pnas.1402597111

Terns, R. M., & Terns, M. P. (2015). CRISPR-based technologies: prokaryotic defense weapons repurposed. *Trends in Genetics*, *30*(3), 111–118. https://doi.org/10.1016/j.tig.2014.01.003

Terwilliger, T. C. (1999). Reciprocal-space solvent flattening. *Acta Crystallographica Section D: Biological Crystallography*, *55*(11), 1863–1871. https://doi.org/10.1107/S0907444999010033

Thyme, S. B., Akhmetova, L., Montague, T. G., Valen, E., & Schier, A. F. (2016). Internal guide RNA interactions interfere with Cas9-mediated cleavage. *Nature*

Communications, *7*, 11750. https://doi.org/10.1038/ncomms11750

van der Oost, J., Westra, E. R., Jackson, R. N., & Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nature Reviews. Microbiology*, *12*(7), 479–492. https://doi.org/10.1038/nrmicro3279

Wang, H., La Russa, M., & Qi, L. S. (2016). CRISPR/Cas9 in Genome Editing and Beyond. *Annual Review of Biochemistry*, *85*(1), 227–264. https://doi.org/10.1146/annurev-biochem-060815-014607

Wang, R., Preamplume, G., Terns, M. P., Terns, R. M., & Li, H. (2011). Interaction of the Cas6 riboendonuclease with CRISPR RNAs: Recognition and cleavage. *Structure*, *19*(2), 257–264. https://doi.org/10.1016/j.str.2010.11.014

Weinberger, A. D., Wolf, Y. I., Lobkovsky, A. E., Gilmore, M. S., & Koonin, E. V. (2012). Viral diversity threshold for adaptive immunity in prokaryotes. *MBio*, *3*(6), 1–10. https://doi.org/10.1128/mBio.00456-12

Wright, A. V, Sternberg, S. H., Taylor, D. W., Staahl, B. T., Bardales, J. A., Kornfeld, J. E., & Doudna, J. A. (2015). Rational design of a split-Cas9 enzyme complex. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(10), 2984–2989. https://doi.org/10.1073/pnas.1501698112

Xiang, G., Zhang, X., An, C., Cheng, C., & Wang, H. (2016). Temperature effect on CRISPR-Cas9 mediated genome editing. *Journal of Genetics and Genomics*, pp. 3–9. https://doi.org/10.1016/j.jgg.2017.03.004

Yamada, M., Watanabe, Y., Gootenberg, J. S., Hirano, H., Ran, F. A., Nakane, T., … Nureki, O. (2017). Crystal Structure of the Minimal Cas9 from Campylobacter jejuni Reveals the Molecular Diversity in the CRISPR-Cas9 Systems. *Molecular Cell*, *65*(6), 1109–1121.e3. https://doi.org/10.1016/j.molcel.2017.02.007

Yosef, I., Goren, M. G., & Qimron, U. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Research*, *40*(12), 5569–5576. https://doi.org/10.1093/nar/gks216

Yoshida, M., Mochizuki, T., Urayama, S. I., Yoshida-Takashima, Y., Nishi, S., Hirai, M., … Takai, K. (2018). Quantitative viral community DNA analysis reveals the dominance of single-stranded DNA viruses in offshore upper bathyal sediment from Tohoku, Japan. *Frontiers in Microbiology*, *9*(FEB), 1–10. https://doi.org/10.3389/fmicb.2018.00075

Zeldes, B. M., Keller, M. W., Loder, A. J., Straub, C. T., Adams, M. W. W., & Kelly, R. M. (2015). Extremely thermophilic microorganisms as metabolic engineering platforms for production of fuels and industrial chemicals. *Frontiers in Microbiology*, *6*(NOV), 1–17. https://doi.org/10.3389/fmicb.2015.01209

Zetsche, B., Gootenberg, J. S., Abudayyeh, O. O., Slaymaker, I. M., Makarova, K. S., Essletzbichler, P., … Zhang, F. (n.d.). Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell*. https://doi.org/10.1016/j.cell.2015.09.038

Zetsche, B., Gootenberg, J. S., Abudayyeh, O. O., Slaymaker, I. M., Makarova, K. S., Essletzbichler, P., … Zhang, F. (2015). Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell*, *163*(3), 759–771. https://doi.org/10.1016/j.cell.2015.09.038

Zhang, Y., Heidrich, N., Ampattu, B. J., Gunderson, C. W., Seifert, H. S., Schoen, C., … Sontheimer, E. J. (2013). Processing-Independent CRISPR RNAs Limit Natural Transformation in Neisseria meningitidis. *Molecular Cell*, *50*(4), 488–503.

https://doi.org/10.1016/j.molcel.2013.05.001

Zhang, Y., Rajan, R., Seifert, H. S., Mondragón, A., & Sontheimer, E. J. (2015). DNase H Activity of Neisseria meningitidis Cas9. *Molecular Cell*, *60*(2), 242–255. https://doi.org/10.1016/j.molcel.2015.09.020