

IDENTIFICATION OF CAS6 AS AN ENDORIBONUCLEASE INVOLVED IN CRISPR-
MEDIATED GENOME DEFENSE IN PROKARYOTES

by

JASON MICHAEL CARTE

(Under the Direction of Michael P. Terns)

ABSTRACT

The CRISPR-Cas system provides many prokaryotes with acquired resistance to genome invaders such as viruses and other mobile genetic elements. This system uses small RNAs derived from Clustered Regularly Interspaced Short Palindromic Repet (CRISPR) loci encoded in the host genome as well as Cas (CRISPR-associated) proteins. We have identified an endoribonuclease, Cas6, that plays a role in CRISPR (cr)RNA biogenesis. Cas6 interacts with sequence elements in the 5' region of precursor crRNAs and cleaves within the 3' region of the repeat. The cleavage was found to be divalent metal ion independent and occurs on the 5' side of the phosphodiester bond. Furthermore, the active site of Cas6 appears to consist of a triad of amino acids composed of conserved tyrosine, histidine, and lysine residues.

INDEX WORDS: CRISPR, Cas, Cas6, endoribonuclease, RNA, RNA processing, viral defense

IDENTIFICATION OF CAS6 AS AN ENDORIBONUCLEASE INVOLVED IN CRISPR-
MEDIATED GENOME DEFENSE IN PROKARYOTES

by

JASON MICHAEL CARTE

B.S., Georgia Southern University, 2006

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment
of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2009

© 2009

Jason Michael Carte

All Rights Reserved

IDENTIFICATION OF CAS6 AS AN ENDORIBONUCLEASE INVOLVED IN CRISPR-
MEDIATED GENOME DEFENSE IN PROKARYOTES

by

JASON MICHAEL CARTE

Major Professor:	Michael P. Terns
Committee:	Rebecca M. Terns Stephen L. Hajduk Sidney R. Kushner

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2009

ACKNOWLEDGEMENTS

I would first like to express my tremendous gratitude to my major professors Dr. Michael Terns and Dr. Rebecca Terns. I am very grateful for your guidance, support, encouragement and the exemplary example you both set for the members of your lab. It has been a true honor to have worked in the Terns' lab. I would like to thank my committee members, Dr. Stephen Hajduk and Dr. Sidney Kushner for their support and advice.

I would like to thank all the members of the Terns' lab, both past and present. In particular, I would like to thank Dr. Dan Baker, Dr. Osama Youssef, and Caryn Hale for their selfless mentorship and guidance. A special thanks to Caryn Hale, whose tireless efforts laid the groundwork for my research project. I would like to thank Neil Pfister, who was in my view, one of the most outstanding undergraduate researchers imaginable.

I would like to thank my family, especially my parents, Alan and Cindy Carte. Without your love and support I would not be where I am today. Finally, I would like to thank my beautiful wife Stephanie. I am eternally thankful to have you in my life. You mean everything to me and I love you.

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS	iv
CHAPTER	
1 INTRODUCTION AND LITERATURE REVIEW	1
2 CAS6 IS AN ENDORIBONUCLEASE THAT GENERATES GUIDE RNAS FOR INVADER DEFENSE IN PROKARYOTES	13
3 CHARACTERIZATION OF CAS6/CRISPR REPEAT INTERACTION AND CLEAVAGE MECHANISM.....	50
4 DISCUSSION	82

Chapter 1

INTRODUCTION AND LITERATURE REVIEW

Introduction

Maintenance of genomic integrity is critical to proper functioning and propagation of organisms and includes effective defense against mobile genetic elements such as transposons and viruses. In eukaryotes, the RNA interference (RNAi) system plays a role in this defense. Small RNAs (typically 21-30 nucleotides in length) guide protein effectors (Argonautes) to target RNAs through sequence-specific interactions. Short interfering (si)RNAs (21-24 nucleotides in length) are generated from double-stranded RNA precursors by dicer endonucleases and are then loaded onto Argonaute-containing complexes, which silence viral mRNAs, often through cleavage of the target [1,2]. Another class of small silencing RNAs, the Piwi-interacting (pi)RNAs, are typically 26-30 nucleotides in length and target and silence transposable elements in mammals and flies [2,3]. A growing body of evidence indicates that a similar RNA-guided genome defense mechanism exists in prokaryotes known as the CRISPR-Cas system.

The CRISPR-Cas system

The CRISPR-Cas system is a widespread and rapidly evolving heritable defense system found in approximately 40% of bacteria and 90% of archaea for which complete genome sequences are available [4]. The components of the CRISPR-Cas system are CRISPR (cr)RNAs

and Cas (CRISPR-associated) proteins [4-6]. The crRNAs are derived from CRISPR (Clustered Regularly Interspaced Short Palindromic Repet) loci within the host genome [5,7,8]. Cas protein genes, often physically linked to CRISPR loci, encode proteins with predicted nuclease, helicase, polymerase, and nucleic acid binding functions [4,9,10]. Several Cas proteins have been identified as nucleases ([11-13] and this thesis).

The CRISPR loci are arranged as arrays of short direct repeats 24-40 nucleotides in length separated by similarly sized unique spacer sequences ([4]and Figure 1.1). The number of repeat-spacer units in a given locus varies from a few to over two hundred [4,6]. Also, the number of CRISPR loci in a given genome varies between one and eighteen. The majority of repeats show weak dyad symmetry at the 5' and 3 ends, and when transcribed may form stable hairpin-loop structures [4,14]. Through bioinformatic analyses, it was found that many of the spacer sequences are identical to known viruses and plasmids [15-17]. It has been shown that CRISPR loci are transcribed from a conserved leader sequence and processed into small crRNAs corresponding roughly to one spacer and one repeat (crRNAs) ([5,7,8,13,18] and Figure 1.1). It has also been shown that these RNAs associate with multiple protein complexes (crRNPs) [8,13]. Based on these observations and the predicted functions of Cas proteins, it was hypothesized that the CRISPR-Cas system functions as an RNA-guided defense against foreign nucleic acids. It is hypothesized that crRNAs, once processed, associate with Cas proteins and guide these proteins to foreign nucleic acids through sequence specific recognition, leading to silencing [4,18].

Cas proteins are a very diverse class of proteins found in association with CRISPRs, often located in clusters directly adjacent to CRISPR loci [4,6,10]. The most widely distributed proteins are regarded as the “core” Cas proteins, Cas1-6 [4,6]. Cas1 is found in all CRISPR-

containing organisms (except *Pyrococcus abyssi*) and as a result has become a marker for CRISPR-Cas [4]. A typical CRISPR-containing organism possesses one or more of the “core” Cas proteins along with a number of other Cas proteins, often present as groups of one of 7 or 8 subtypes [4,6,10]. A very widely distributed subtype of Cas proteins are the so called RAMPs (Repet Associated Mysterious Proteins), which share a conserved glycine-rich C-terminus and are predicted nucleic acid binding proteins [4,9].

Evidence for CRISPR function in genome defense

The first experimental evidence that CRISPR-Cas provides acquired resistance to phage was provided by Barrangou and colleagues in 2007 [19]. In this work it was shown that in response to phage challenge, a phage-sensitive strain of *Streptococcus thermophilus* produced phage-resistant mutants that had integrated novel spacer units into the CRISPR locus that perfectly matched the infecting phage. When these spacers were deleted from the host genome, phage resistance was lost. This was the first demonstration that the presence of a phage-derived spacer in the host genome could result in phage resistance. In a subsequent study in *S. thermophilus*, it was shown that the most frequent outcome of phage challenge resulting in resistance was the integration of a single spacer directly downstream of the leader sequence [20]. It was also shown that the incorporation of multiple spacers led to enhanced resistance. The short sequence in the phage genome that is incorporated into the host genome has been termed a proto-spacer ([21] and Figure 1.1). While the process leading to novel spacer incorporation into CRISPR loci is not understood, recent evidence indicates that sequence elements directly adjacent to the proto-spacer may be important in this process.

A role for Cas proteins in CRISPR-mediated genome defense was first established by generating Cas gene knockouts in phage resistant mutants in *S. thermophilus* [19]. When Cas5 (denoted Csn1 by Haft et al [2005]), a predicted nuclease, was knocked out, phage resistance was abolished. When another protein, denoted Cas7 by the authors, was knocked out, the ability to acquire resistance to novel phage was lost, implicating this protein in the process of spacer incorporation. In a recent study done in *Escherichia coli*, Brouns and co-workers showed that a complex of five Cas proteins called Cascade, including the endoribonuclease Cse3 (CasE), are responsible for processing the CRISPR transcript into crRNAs [13]. Cascade, along with Cas3, was found to be required for resistance to bacteriophage Lambda infection in *E. coli*. Thus it appears that crRNAs and at least a subset of Cas proteins function to recognize and destroy invading phages. The precise roles of the Cas proteins involved remains to be fully elucidated.

The CRISPR-Cas system has been shown to inhibit plasmid conjugation by targeting DNA. *Staphylococcus epidermidis* ATCC strain 12228 contains a CRISPR locus harboring a spacer matching the *nickase* (*nes*) gene found on all staphylococcal plasmids [22]. This strain is resistant to plasmid conjugation. When the target sequence is mutated on the plasmid, the ability to block conjugation is lost. This result supports the hypothesis that base pairing must occur between the crRNA and the target in order for silencing to occur. As evidence that CRISPR targets DNA, a self-splicing intron was inserted within the target sequence on the plasmid and the ability to block conjugation was lost [22]. This result indicates that the crRNP is able to recognize and silence foreign DNA and not RNA in *S. epidermidis*.

If the CRISPR-Cas system targets foreign DNA rather than RNA, then how do crRNPs distinguish self from non-self, given that target sequences are located in the host genome? To begin to answer this question several recent studies have determined that sequences adjacent to

the proto-spacer may be important for recognition by crRNPs [20,23,24]. In *S. thermophilus*, for example, sequence analysis of the phage genome revealed a conserved NNAGAAW motif located downstream of the proto-spacer [20]. Evidence for the importance of this motif was then provided by the observation that when this sequence was mutated in the phage genome, CRISPR-mediated resistance was lost [20]. In archaea, conserved dinucleotides located upstream of the proto-spacer have been observed [23,24]. The importance of this dinucleotide has not been verified experimentally. In a more comprehensive analysis, Mojica and coworkers [23] analyzed phage and viral sequences relative to spacer sequences in over 150 organisms and found conserved di- and trinucleotide nucleotide motifs, and named these proto-spacer adjacent motifs (PAMs) (Figure 1.1). These PAMs are believed to serve as recognition motifs for Cas proteins involved in selecting new spacers for integration into the host genome and may be important for recognition by effector crRNPs, which would allow distinction between self and non-self.

Biochemical Characterization of Cas proteins

Uncovering the function of individual Cas proteins is vitally important to gaining an understanding of how the CRISPR-Cas system functions at the molecular level to defend the host against foreign nucleic acids. Functional characterization of several Cas proteins has recently been reported ([11-13,25] and this thesis). The first Cas protein to be characterized *in vitro* was the “core” Cas protein Cas2 (SSO1404) from *Sulfolobus sulfataricus* [11]. This protein was shown to possess endoribonuclease activity against single-stranded RNA, with a preference for uridine-rich regions [11]. Another “core” Cas protein, Cas1 from *Pseudomonas aeruginosa*

(PA14_33350) was later shown to be a double-stranded DNA specific deoxyribonuclease, cleaving long double-stranded DNA into ~80 nucleotide products [12].

Cas3, another of the conserved “core” Cas protein is present in several isoforms [4]. Cas3 may be present as a larger protein with two distinct, fused domains, one a predicted DExD box helicase, and the other a predicted HD nuclease [4]. Alternatively, the genes encoding the two Cas3 domains can be separate from one another in the host genome [4]. Recently, a Cas3 protein containing only the predicted HD nuclease domain (SSO2001 from *Sulfolobus sulfotaricus*) was shown to be an endonuclease that cleaves both double-stranded RNA and double-stranded DNA with a preference for G:C base pairs [26]. Cas1, Cas2, and Cas3 were all predicted to be nucleases, and appear to be divalent metal ion dependent [4,11,12,26]. The roles played by these enzymes in CRISPR defense have not been determined.

Brouns and coworkers [13] provided the first biochemical characterization of Cas proteins with corresponding *in vivo* data supporting a specific role for Cas proteins in the context of overall CRISPR function. *E. coli* Cse3 (called CasE by the authors) was shown to be a divalent metal ion independent endoribonuclease that, as part of a large protein complex named Cascade, cleaves CRISPR primary transcripts within the repeat region into crRNAs. A role for these proteins in phage resistance was then established when the components of Cascade, different combinations of Cas1-3, and crRNAs that had been engineered to target Lambda phage were expressed in a strain of *E. coli* that lacks the CRISPR-Cas system. It was shown that Cascade, along with Cas3 and crRNAs were sufficient to convey resistance to infection by Lambda phage. The function of Cas3 and the other members of Cascade have not been determined.

Overview

The goal of this research was to determine the Cas protein factor(s) involved in crRNA biogenesis. To that end, several recombinant Cas proteins from *Pyrococcus furiosus* were screened for their ability to cleave crRNA precursor RNAs. A single protein, Cas6, was found to cleave pre-crRNAs within the repeat region to generate a crRNA intermediate composed of a single spacer flanked by fragments of repeat (Ch. 2). The site of cleavage was predicted based on analysis of native crRNAs and intermediates from a *Pyrococcus furiosus* extract [8].

Cas6, a member of the RAMP superfamily and “core” Cas protein, is a member of a group of widely distributed orthologous proteins characterized by a conserved C-terminal GhGxxxGhG motif (where h is a hydrophobic residue and xxxx contains at least one lysine or arginine) and was predicted to be an RNA-binding protein [4,6]. Cas6 was not predicted to possess nuclease activity.

We found that Cas6 specifically recognizes sequence elements within the 5' region of the CRISPR repeat RNA and then cleaves downstream within the 3' region of the repeat (Ch. 2). The determinants for substrate recognition were determined by RNA mutagenesis and confirmed by RNA footprinting (Ch. 2 and 3). The site of cleavage was mapped and found to correspond to a predicted cleavage site based on native crRNAs from *Pyrococcus furiosus* (Ch. 2).

We determined that the cleavage mechanism is divalent metal ion independent and the crystal structure revealed a conserved trio of Tyr31, His46, and Lys52 that cluster near the conserved Gly-rich C-terminal loop (Ch. 2). Mutational analysis of these amino acids establishes that they are crucial for efficient cleavage (Ch. 3). Finally, evidence is presented that suggests that native Cas6 cleaves crRNA precursors and stably associates with its cleavage product in *P. furiosus*.

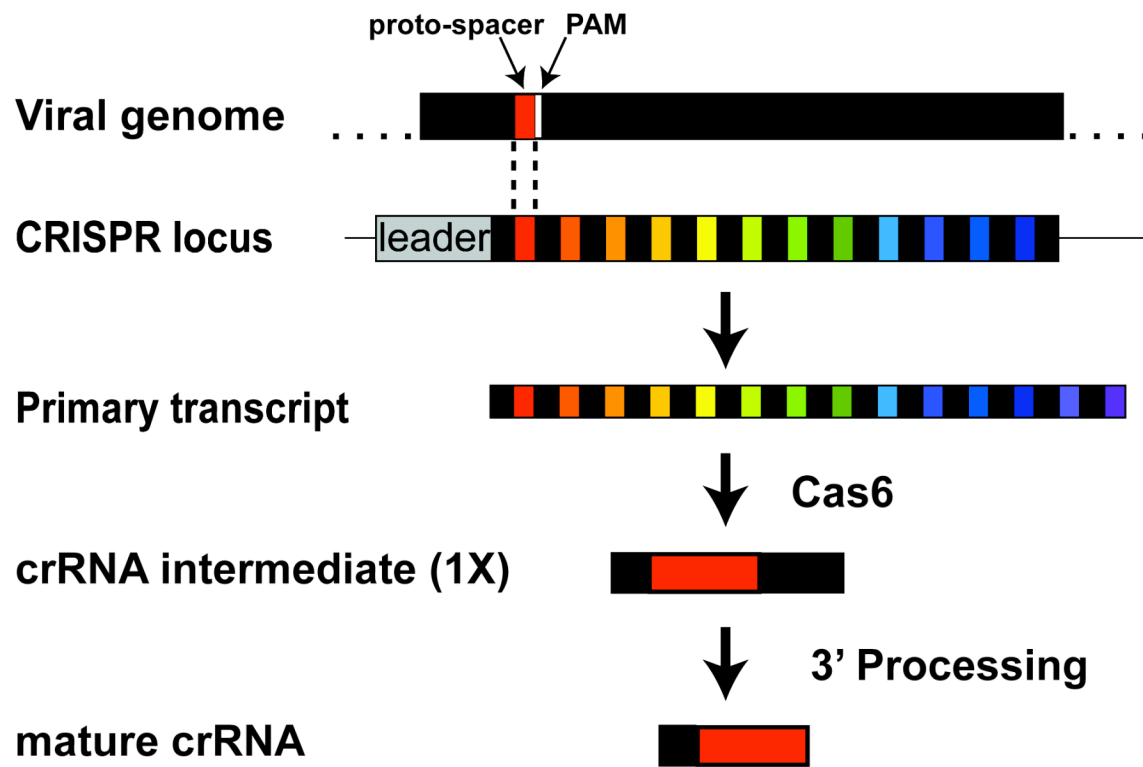
References

1. Jaskiewicz L, Filipowicz W: **Role of Dicer in posttranscriptional RNA silencing.** *Curr Top Microbiol Immunol* 2008, **320**:77-97.
2. Girard A, Hannon GJ: **Conserved themes in small-RNA-mediated transposon control.** *Trends Cell Biol* 2008, **18**:136-148.
3. Klattenhoff C, Theurkauf W: **Biogenesis and germline functions of piRNAs.** *Development* 2008, **135**:3-9.
4. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV: **A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action.** *Biol Direct* 2006, **1**:7.
5. Tang TH, Polacek N, Zywicki M, Huber H, Brugger K, Garrett R, Bachellerie JP, Huttenhofer A: **Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*.** *Mol Microbiol* 2005, **55**:469-481.
6. Haft DH, Selengut J, Mongodin EF, Nelson KE: **A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes.** *PLoS Comput Biol* 2005, **1**:e60.
7. Tang TH, Bachellerie JP, Rozhdestvensky T, Bortolin ML, Huber H, Drungowski M, Elge T, Brosius J, Huttenhofer A: **Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*.** *Proc Natl Acad Sci U S A* 2002, **99**:7536-7541.
8. Hale C, Kleppe K, Terns RM, Terns MP: **Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*.** *RNA* 2008, **14**:2572-2579.
9. Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV: **A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis.** *Nucleic Acids Res* 2002, **30**:482-496.
10. Jansen R, Embden JD, Gaastra W, Schouls LM: **Identification of genes that are associated with DNA repeats in prokaryotes.** *Mol Microbiol* 2002, **43**:1565-1575.
11. Beloglazova N, Brown G, Zimmerman MD, Proudfoot M, Makarova KS, Kudritska M, Kochinyan S, Wang S, Chruszcz M, Minor W, et al.: **A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats.** *J Biol Chem* 2008, **283**:20361-20371.

12. Wiedenheft B, Zhou K, Jinek M, Coyle SM, Ma W, Doudna JA: **Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense.** *Structure* 2009, **17**:904-912.
13. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J: **Small CRISPR RNAs guide antiviral defense in prokaryotes.** *Science* 2008, **321**:960-964.
14. Kunin V, Sorek R, Hugenholtz P: **Evolutionary conservation of sequence and secondary structures in CRISPR repeats.** *Genome Biol* 2007, **8**:R61.
15. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD: **Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin.** *Microbiology* 2005, **151**:2551-2561.
16. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E: **Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements.** *J Mol Evol* 2005, **60**:174-182.
17. Pourcel C, Salvignol G, Vergnaud G: **CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies.** *Microbiology* 2005, **151**:653-663.
18. Lillestol RK, Redder P, Garrett RA, Brugger K: **A putative viral defence mechanism in archaeal cells.** *Archaea* 2006, **2**:59-72.
19. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P: **CRISPR provides acquired resistance against viruses in prokaryotes.** *Science* 2007, **315**:1709-1712.
20. Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S: **Phage response to CRISPR-encoded resistance in Streptococcus thermophilus.** *J Bacteriol* 2008, **190**:1390-1400.
21. Horvath P, Romero DA, Coute-Monvoisin AC, Richards M, Deveau H, Moineau S, Boyaval P, Fremaux C, Barrangou R: **Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus.** *J Bacteriol* 2008, **190**:1401-1412.
22. Marraffini LA, Sontheimer EJ: **CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA.** *Science* 2008, **322**:1843-1845.
23. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Almendros C: **Short motif sequences determine the targets of the prokaryotic CRISPR defence system.** *Microbiology* 2009, **155**:733-740.

24. Lillestol RK, Shah SA, Brugger K, Redder P, Phan H, Christiansen J, Garrett RA: **CRISPR families of the crenarchaeal genus *Sulfolobus*: bidirectional transcription and dynamic properties.** *Mol Microbiol* 2009, **72**:259-272.
25. Han D, Lehmann K, Krauss G: **SSO1450--a CAS1 protein from *Sulfolobus solfataricus* P2 with high affinity for RNA and DNA.** *FEBS Lett* 2009, **583**:1928-1932.
26. Han D, Krauss G: **Characterization of the endonuclease SSO2001 from *Sulfolobus solfataricus* P2.** *FEBS Lett* 2009, **583**:771-776.

Figure 1.1. Proto-spacer incorporation and crRNA biogenesis. Short sequence elements on invading nucleic acids (proto-spacers) are incorporated in CRISPR loci. This process appears to involve sequence elements directly adjacent to the proto-spacer known as proto-spacer adjacent motifs (PAMs). The CRISPR locus is transcribed from a conserved leader sequence to produce a CRISPR primary transcript. In *P. furiosus* and likely many other organisms, Cas6 cleaves this primary transcript to yield the 1X crRNA intermediate. This RNA undergoes 3' processing to produce stable mature crRNAs.



Chapter 2

CAS6 IS AN ENDORIBONUCLEASE THAT GENERATES GUIDE RNAS FOR INVADER
DEFENSE IN PROKARYOTES¹

¹ Carte, J., R. Wang, H. Li, R.M. Terns, M.P. Terns. *Genes Dev.* 2008 Dec 15; 22(24):3489-96.
Reproduced with permission from the publishers.

Abstract

An RNA-based gene silencing pathway that protects bacteria and archaea from viruses and other genome invaders is hypothesized to arise from guide RNAs encoded by CRISPR loci and proteins encoded by the *cas* genes. CRISPR loci contain multiple short invader-derived sequences separated by short repeats. The presence of virus-specific sequences within CRISPR loci of prokaryotic genomes confers resistance against corresponding viruses. The CRISPR loci are transcribed as long RNAs that must be processed to smaller guide RNAs. Here we have identified *Pyrococcus furiosus* Cas6 as a novel endoribonuclease that cleaves CRISPR RNAs within the repeat sequences to release individual invader-targeting RNAs. Cas6 interacts with a specific sequence motif in the 5' region of the CRISPR repeat element and cleaves at a defined site within the 3' region of the repeat. The 1.8 angstrom crystal structure of the enzyme reveals two ferredoxin-like folds that are also found in other RNA binding proteins. The predicted active site of the enzyme is similar to that of tRNA splicing endonucleases, and concordantly, Cas6 activity is metal-independent. *cas6* is one of the most widely distributed CRISPR-associated genes. Our findings indicate that Cas6 functions in the generation of CRISPR-derived guide RNAs in numerous bacteria and archaea.

Introduction

All genomes are potential targets of invasion by molecular parasites such as viruses and transposable elements, and organisms have evolved RNA-directed defense mechanisms to cope with the constant threat of genome invaders [1,2]. The well-known sub-pathway of RNA silencing referred to as RNAi (RNA interference) functions in defense against viruses in eukaryotes [3]. The RNAi defense response is mediated by short (~22 nucleotide) RNAs termed

small interfering (si)RNAs. The siRNAs are generated from invading viral RNAs by double-stranded RNA-specific, RNase III-like endonucleases called Dicers [4]. The mature siRNAs are assembled with host effector proteins and target them to corresponding viral target RNAs to effect viral gene silencing via RNA destruction or other mechanisms [1,2].

Compelling evidence has recently emerged for the existence of an RNA-mediated genome defense pathway in archaea and numerous bacteria that has been hypothesized to parallel the eukaryotic RNAi pathway (reviewed in [5-8]). Known as the CRISPR-Cas system or prokaryotic RNA interference (pRNAi), the pathway is proposed to arise from two evolutionarily and often physically linked gene loci: the CRISPR (clustered regularly interspaced short palindromic repeats) locus, which encodes RNA components of the system, and the *cas* (CRISPR-associated) locus, which encodes proteins [6,9-11]. The individual Cas proteins do not share significant sequence similarity with protein components of the eukaryotic RNAi machinery, but have analogous predicted functions (e.g. RNA binding, nuclease, helicase, etc.) [6].

Unlike the siRNAs of the eukaryotic RNAi system, the effector RNAs of pRNAi are encoded in the host genome. CRISPR loci encode short (typically ~30-35 nucleotide) invader-derived sequences interspersed between short (typically ~30-35 nucleotide) direct repeat sequences [5-8,12-15]. Recent studies have provided clear experimental evidence that correlates the presence of virus-specific CRISPR sequences with viral immunity [16-18]. Furthermore, viral infection has been shown to result in the appearance of new corresponding CRISPR elements in surviving strains [16,17]. This rapidly adapting CRISPR-based immunity acts within natural microbial populations to promote host cell fitness and to influence microbial ecology [19,20].

The primary products of the CRISPR loci appear to be short RNAs that contain the invader-targeting sequences, and are termed guide RNAs or prokaryotic silencing (psi)RNAs based on their hypothesized role in the pathway [6,21]. RNA analysis indicates that CRISPR locus transcripts are cleaved within the repeat sequences to release ~60-70 nucleotide RNA intermediates that contain individual invader-targeting sequences and flanking repeat fragments ([7,18,21-23] and Figure 1A). In the archaeon *Pyrococcus furiosus*, these intermediate RNAs are further processed to abundant, stable ~35-45 nucleotide mature psiRNAs [21].

The primary goal of this study was to begin to understand the biogenesis of psiRNAs through identification and characterization of the enzyme that cleaves within the repeat sequences of CRISPR RNA transcripts to liberate the many individual psiRNA species that function in defense against molecular invaders. Our results indicate that Cas6, one of the six highly conserved or “core” Cas proteins (Haft et al., 2005), functions as a CRISPR repeat RNA-specific endoribonuclease in *P. furiosus* and likely numerous other archaea and bacteria.

Results

The psiRNAs, which are thought to be primary agents in prokaryotic genome defense, are derived from CRISPR RNA transcripts that consist of a series of individual invader-targeting sequences separated by a common repeat sequence (Figure 1A). To identify the enzyme required for dicing CRISPR RNA transcripts and releasing the individual embedded psiRNAs, we screened a number of recombinant *P. furiosus* Cas proteins for the ability to cleave CRISPR repeat sequences. We identified a single protein, Cas6 (PF1131), that cleaves specifically within the repeat sequence of radiolabeled substrate RNAs consisting of either a guide (invader-targeting or “spacer”) sequence flanked by two repeat sequences or the repeat sequence alone

(Figure 1B and C). Examination of the cleavage products generated from uniformly-labeled and 5' end-labeled RNA substrates indicates that cleavage occurs ~20-25 nucleotides from the 5' end of the repeat. Cleavage also occurs within each repeat of an extended substrate RNA containing two guide sequences and flanking repeats (Figure 2).

More than 40 CRISPR-associated genes have been identified, however only a subset of the *cas* genes is found in any given genome and no *cas* gene appears to be present in all organisms that possess the CRISPR-Cas system [6,11]. Cas6 is among the most widely distributed Cas proteins and is found in both bacteria and archaea [11]. A distinct protein with similar activity was very recently reported in *Escherichia coli* [18]. This protein, Cse3 (CRISPR-Cas system subtype *E. coli*, also referred to as CasE), is found in some bacteria that lack Cas6 [11]. Both Cas6 and Cse3 are members of the RAMP (repeat associated mysterious protein) superfamily, as are a large number of the Cas proteins [6,10]. RAMP proteins contain G-rich loops and are predicted to be RNA binding proteins [6,10]. Cas6 is distinguished from the many other RAMP family members by a conserved sequence motif within the predicted C-terminal G-rich loop (consensus GhGxxxxxGhG, where h is hydrophobic and xxxxx has at least one lysine or arginine) [10]. Nuclease activity was not predicted for Cas6 based on sequence analysis.

To determine the precise PfCas6 cleavage site within the CRISPR repeat sequence, 5' end-labeled repeat RNA was incubated with the purified enzyme and the 5' cleavage product was mapped relative to RNase T1 (cuts after guanosines) and alkaline hydrolysis (cuts after each nucleotide) cleavage products (Figure 3A). A 22-nucleotide 5' cleavage product was identified indicating that cleavage occurs between adenosine 22 and adenosine 23 of the 30-nucleotide repeat sequence (Figure 3A and B). The resulting 5' end generated by PfCas6 is the same as that observed in mature psiRNA species isolated from *P. furiosus* cells (C. Hale, R. Terns and M.

Terns, unpublished). Mutation of the two nucleotides spanning the cleavage site (AA to GG) drastically reduced the cleavage activity of PfCas6 (Figure 3C) without preventing binding of the enzyme to the RNA (assayed by RNA gel mobility shift; Figure 3D). The site of cleavage is at a junction within a potential stem-loop structure that may form by base-pairing between weakly palindromic sequences commonly found at the 5' and 3' termini of CRISPR repeat sequences ([5,24] and Figure 3B).

We next investigated the RNA sequence requirements of Cas6 binding and endonucleolytic cleavage. To identify the RNA binding determinants, we performed gel mobility shift assays with a series of RNAs (Figure 4A). The results indicate that sequences in the 5' region of the CRISPR repeat are important for PfCas6 binding. Under normal assay conditions, rapid cleavage prevents unambiguous observation of PfCas6 binding to the intact repeat (Figure 3C and D), though binding can be observed with the cleavage site mutant (Figure 3D) and at reduced temperatures where PfCas6 cleavage activity is inhibited (Figure S1). However, incubation of PfCas6 with the repeat RNA (Figure 3D) or with a guide sequence flanked by two repeat sequences (Figure 4A, panel a) under conditions compatible with cleavage reveals interaction of the protein with the 5' cleavage product generated during incubation. PfCas6 also interacts with the gel-purified 5' cleavage product, but not with the 3' cleavage product (Figure 4B). Furthermore, we found that PfCas6 binds each tested RNA that contains the repeat sequences found upstream of the cleavage site (i.e. the first 22 nucleotides of the repeat; Figure 4A, panels c, f, g), but not an RNA that contains only the downstream region (last 8 nucleotides) of the repeat (Figure 4A, panel b).

Further analysis indicates that the first 12 nucleotides of the 5' region of the CRISPR repeat play a critical role in Cas6 binding. PfCas6 binds to an RNA comprised of the first 12

nucleotides of the repeat with similar affinity as the 5' cleavage product (Figure 4A, panel h). Furthermore, protein binding is abolished by substitution or deletion of the first 8 nucleotides of the repeat (Figure 4A, panels d and e). In addition, substitution, insertion or deletion in the region of nucleotides 9-12 appears to have slightly reduced interaction (Figure 4A, panels i, j, k). No binding was observed with a DNA repeat sequence (Figure 4A, panel l). Taken together, the results indicate that PfCas6 requires sequence and/or structure information present within the first 12 nucleotides of the CRISPR repeat RNA for stable interaction (Figure 4C).

While nucleotides at the 5' end of the CRISPR repeat are sufficient for robust PfCas6 binding, cleavage appears to involve additional elements. As expected, mutations that disrupt protein binding also eliminate cleavage activity (Figure 5, panels d and e). However, other mutations dramatically reduced cleavage efficiency without disrupting PfCas6 binding. As indicated above, substitution of the 2 adenosines at the cleavage site disrupts cleavage but not binding (Figure 3C and 3D). In addition, substitution of the last 8 nucleotides of the repeat specifically disrupted cleavage (Figure 5, panel f). PfCas6 cleavage activity was also significantly reduced by small (4 nucleotide) insertions or deletions between the PfCas6 binding site and cleavage site (Figure 5, panels i and j). Substitution of 6 nucleotides between the binding and cleavage sites also disrupted cleavage (Figure 5, panel k). No cleavage activity was observed with a DNA repeat sequence (Figure 5, panel l). These results suggest that cleavage depends upon sequence elements along the length of the repeat and perhaps upon the distance between the binding and cleavage sites, and are consistent with a requirement for a specific RNA fold such as the predicted hairpin structure ([5,24] and Figure 3B)).

P. furiosus has 7 CRISPR loci with 5 slightly varied repeat sequences, and the elements that we have identified as most important for Cas6 recognition and cleavage map to the regions of

greatest sequence conservation. Variation is observed at only one position within each the first 12 and last 11 nucleotides of the *P. furiosus* repeat sequences, consistent with the importance of these two regions in Cas6 binding and cleavage. On the other hand, variation occurs at 3 positions between the binding and cleavage sites (positions 14, 16, and 19), suggesting that nucleotide identities are less important in this region.

To gain a more detailed understanding of PfCas6, we obtained a crystal structure of the protein at 1.8 Å resolution (Figure 5 and see Table S3 for structure determination details). PfCas6 contains a duplicated ferredoxin-like fold linked by an extended peptide (residues 118-123). The close arrangement of the β -sheets of the two ferredoxin-like folds creates a well-formed central cleft (Figure 6A). The ferredoxin fold is a common protein fold also found in the structures of other RNA binding proteins including the well-characterized RNA recognition motif (RRM), which primarily functions in single-stranded RNA binding [25]. However, PfCas6 appears to exploit a distinct mechanism of base-specific single-stranded RNA recognition. Most notably, PfCas6 lacks the prevalent aromatic and positive residues that characterize the β -sheets of RRMs [25]. The central regions of both the front and back surfaces of PfCas6 display positive potential that coincides with regions of conserved amino acids (Figure 6) suggesting that the composite surfaces formed by the tandem ferredoxin-like folds correspond to RNA binding sites.

The structure of PfCas6 allows us to predict the site of catalysis and catalytic mechanism of the enzyme. Several candidate catalytic residues are evident as strictly conserved residues in aligned Cas6 sequences (Figure S2). These include Tyr31, His46, and Lys52, which cluster within 6 Å of each other and are found in close proximity to the G-rich loop that contains the Cas6 signature motif (Figure 6B). We suggest that these three residues form a catalytic triad for RNA cleavage similar to that of the tRNA intron splicing endonuclease [26]. The G-rich loop is

located immediately above the putative catalytic triad and may facilitate the placement of CRISPR repeat RNA substrates. Consistent with the corresponding predicted general acid-base catalytic mechanism (proposed for the splicing endonuclease [26]), PfCas6 does not require divalent metals and like other metal-independent nucleases cleaves on the 5' side of the phosphodiester bond, likely generating 5' hydroxyl (OH) and 2', 3' cyclic phosphate RNA end groups (Figure 7). Finally, while binding of the enzyme occurs over a wide temperature range, PfCas6 cleavage activity is sharply temperature dependent with significantly more activity at 70°C than 37°C (Figure S1).

Discussion

The results presented here indicate that Cas6 plays a central role in the production of the psiRNAs in the emerging prokaryotic RNAi pathway. Cas6 is a novel riboendonuclease. Through direct binding and cleavage of CRISPR repeat sequences, Cas6 is capable of dicing long, single-stranded CRISPR primary transcripts into units that consist of an individual guide sequence flanked by a short (8 nucleotide) repeat sequence at the 5' end and by the remaining repeat sequence at the 3' end of the RNA (Figure 1A). Mature psiRNAs retain the short repeat-derived sequence established by Cas6 at their 5' ends in *P. furiosus* (C. Hale, R. Terns and M. Terns, unpublished), which we speculate functions as a psiRNA identity tag that allows recognition of the guide RNAs by components of the pRNAi machinery. A repeat sequence of the same length was observed on the 5' ends of RNAs associated with *E. coli* Cse3, indicating that this may indeed be a conserved feature [18]. The 3' ends of Cas6 cleavage products appear to be further processed since mature psiRNAs lack repeat sequences at their 3' termini in *P. furiosus* (C. Hale, R. Terns and M. Terns, unpublished). Because Cas6 remains bound to the

CRISPR repeat sequences at the 3' end of the cleavage product (Figure 3 and 4B), Cas6 could influence the subsequent 3' end processing of the RNA. Additional studies may reveal if Cas6 is also an important component of pRNAi effector complexes (serving to couple biogenesis and function), as is the case for eukaryotic Dicer enzymes [4].

Cas6 is evolutionarily, structurally and catalytically distinct from the Dicer proteins that function in the release of individual RNAs that mediate gene silencing in eukaryotes [4,27]. However, Cas6 is one of three different ferredoxin fold Cas proteins recently found to possess nuclease activity. Cas2, another protein found in many of the prokaryotes that possess the CRISPR-Cas system, cleaves U-rich single-stranded RNA [28]. The mechanism of action of Cas6 seems to be distinct from that of Cas2, which appears to be a metal-dependent, hydrolytic enzyme [28]. The role of Cas2 in the pRNAi pathway is currently unknown. The *E. coli* Cse3 protein functions like Cas6 as a CRISPR repeat cleaving enzyme [18]. Cse3 also cleaves RNA in a divalent metal independent manner [18]. The substrate RNA recognition requirements and the precise cleavage site have not yet been defined for Cse3. Interestingly, despite the lack of significant sequence homology, the Cas6 and Cse3 proteins appear to adopt similar structures to perform a common function in psiRNA biogenesis. Moreover, some bacteria with the CRISPR-Cas system do not appear to contain either a *cas6* or a *cse3* gene, suggesting that there is another Cas6 functional homolog among the Cas proteins, and illustrating the diversity of the CRISPR-Cas systems present in prokaryotes.

Materials and Methods

Purification of PF1131 protein for cleavage and RNA binding assays

N-terminal, 6x histidine-tagged PF1131 protein (PfCas6 from *P. furiosus* DSM 3638 strain) was expressed in *Escherichia coli* BL21 codon + (DE3, Invitrogen) cells harboring a pET24d plasmid containing the appropriate gene insert (gift of Michael Adams, University of Georgia). Protein expression was induced by growing the cells to an OD₆₀₀ of 0.6 and adding isopropylthio-b-D-galactoside (IPTG) to a final concentration of 1 mM. The cells were disrupted by sonication (Misonix Sonicator 3000) in buffer A (20 mM sodium phosphate (pH 7.0), 500 mM NaCl and 0.1 mM phenylmethylsulfonyl fluoride). The lysate was then cleared by centrifugation and the supernatant was incubated for 20 minutes at 70°C. This sample was centrifuged and the supernatant was applied to a Ni-NTA agarose column (Qiagen) that had been equilibrated with Buffer A. The protein was eluted from the column with Buffer A containing 350 mM imidazole. The purity of the protein was evaluated by SDS-PAGE and staining with coomassie blue. Buffer exchange into 40 mM HEPES-KOH (pH 7.0), 500 mM KCL was carried out using Microcon PL-10 filter columns (Millipore). The protein concentration was determined by the BCA assay (Pierce).

Generation of RNA substrates

Synthetic RNAs (listed in Supplemental Table 1) and the RNA size standards (Decade™ Markers) were purchased from Integrated DNA Technologies (IDT) and Ambion, respectively. These RNAs were 5' end labeled with T4 Polynucleotide kinase (Ambion) in a 20 µL reaction containing 20 pmol of RNA, 500 µCi of [³²P] ATP (3000 Ci/mmol, MP Biomedicals) and 20 U of T4 kinase. The RNAs were separated by electrophoresis on denaturing (7 M urea) 15%

polyacrylamide gels and the appropriate RNA species were excised from the gel with a sterile razor blade guided by a brief autoradiographic exposure. The RNAs were eluted from the gel slices by end-over-end rotation in 400 μ L of RNA elution buffer (500 mM NH_4OAc , 0.1% SDS, 0.5 mM EDTA) at 4° C for 12-14 hours. The RNA was then extracted with phenol/chloroform/isoamyl alcohol (PCI, 25:24:1, pH 5.2), and precipitated with 2.5 volumes of 100% ethanol in the presence of 0.3 M sodium acetate and 20 μ g of glycogen after incubation at -20° C for one hour.

All other RNAs were generated by *in vitro* transcription using T7 RNA polymerase (Ambion) and uniformly labeled with [α - ^{32}P] UTP (700 Ci/mmol, MP Biomedicals) as described [29]. The templates used were either annealed DNA oligonucleotides or PCR products (see Tables 1 and 2), both containing the T7 promoter sequence. A typical reaction contained 200 ng of PCR product or annealed deoxyoligonucleotides, 1 mM DTT, 10 U SUPERase-IN™ RNase inhibitor (Ambion), 500 μ M ATP, CTP, and GTP, 50 μ M UTP, 30 μ Ci [α - ^{32}P] UTP, 1X transcription buffer (Ambion), and 40 U T7 RNA polymerase in a total volume of 20 μ L.

RNA binding and cleavage reactions

Typically, identical reaction conditions were used to assay the ability of PfCas6 protein to bind to and to cleave substrate RNAs. These reactions were initiated by incubating 0.05 pmol of ^{32}P -radiolabeled RNAs (either uniformly or 5' end labeled) with up to 1 μ M (as indicated in figure legends) of PfCas6 protein in 20 mM HEPES-KOH (pH 7.0), 250 mM KCl, 0.75 mM DTT, 1.5 mM MgCl_2 , 5 μ g *E. coli* tRNA, and 10% glycerol in a 20 μ L reaction volume for 30 minutes at 70° C. Half of the reactions were directly run on native 8% polyacrylamide gels to assay RNA binding by gel mobility shift essentially as described [29]. RNA cleavage was assayed using the

remaining half of the reaction by deproteinizing (PCI extraction and ethanol precipitation) the RNAs and separating them by electrophoresis on denaturing (7 M urea), 12-15% polyacrylamide gels. Gels were dried and the radiolabeled RNAs visualized by phosphor imaging.

Cleavage site mapping

In order to map the site of RNA cleavage by Cas6, a standard cleavage reaction was set up using 5' end labeled repeat RNA as described above. Alkaline hydrolysis and RNase T1 (0.1 U) ladders were generated as described previously [30]. Following the reactions, the RNAs were extracted with PCI, ethanol precipitated, and separated by electrophoresis on large, denaturing (7 M urea), 15% polyacrylamide (19:1 acrylamide:bis) gels. The gels were dried and the RNAs visualized by phosphor imaging.

Purification of PfCas6 for structure determination

N-terminal polyhistidine-tagged wildtype and selenomethionine-labeled PF1131 protein was expressed in *Escherichia coli* and purified from cell extract by heat-denaturation and two chromatography steps. The cells were disrupted by sonication in a buffer containing 25 mM sodium phosphate (pH 7.5), 5% (v/v) glycerol, 1 M NaCl, 5 mM b-mercaptoethanol (bME) and 0.2 mM phenylmethylsulfonyl fluoride. The cell lysate was heated to 70°C for 15 minutes before being pelleted. The supernatant was then directly loaded at room temperature onto a Ni-NTA (Qiagen) column equilibrated with 25 mM sodium phosphate (pH 7.5), 5% (v/v) glycerol, 1 M NaCl, and 5 mM imidazole. The column was washed with the loading buffer containing 25 mM imidazole and then the bound protein was eluted using the loading buffer containing 350 mM imidazole. Fractions containing PF1131 were pooled and loaded onto a Superdex™ 200

(HiloadTM 26/60, Pharmacia) size-exclusion column equilibrated with 20 mM Tris-HCl (pH 7.4), 500 mM KCl, 5% glycerol, 0.5 mM ethylenediaminetetraacetic acid (EDTA) and 5 mM bME. The fractions corresponding to PF1131 were pooled and concentrated to 100 mg/ml for crystallization.

Crystallization of PF1131 and selenomethionine-labeled PF1131

Both the wildtype and selenomethionine-labeled PF1131 protein were crystallized using vapor diffusion in a hanging drop at 30°C. The droplets of PF1131 at 40 mg/ml were combined in equal volume with a well solution that contained 50 mM MES (pH 6.0), 30 mM MgCl₂, and 15% (v/v) isopropanol. The crystals formed in 1-5 days with a cubic shape and to a size of approximately 0.4 mm x 0.4 mm x 0.4 mm.

Data collection and structure determination

Crystals were soaked briefly in a cryo-protecting solution containing the mother liquor plus 20% (w/v) polyethylene glycol 4000 before being flash frozen in a nitrogen stream at 100 Kelvin. The crystals of the native and selenomethionine-labeled PF1131 diffracted to $d_{\min} = 1.8\text{--}2.2$ Å at the Southeast Regional Collaborative Access Team (SER-CAT) beamline 22ID. The space group of the crystals was determined to be P3₂21 and the cell dimensions are listed in Table 3. A single wavelength data set was collected at the anomalous peak of selenine from a selenomethionine-labeled crystal. The solvent content was calculated to be 54.9% if the crystal was assumed to contain one PF1131 in one asymmetric unit. The structure of PF1131 was solved by a SAD phasing method using the automated crystallographic structure solution program SOLVE [31]. The initial model traced by SOLVE was further improved by the

program COOT [32], followed by refinement using CNS [33] and REFMAC5 [34] to $R_{\text{work}}/R_{\text{free}}$ of 23.6/27.3. The quality of the structure model was checked by PROCHECK [35] and was found to be of satisfactory stereochemical properties.

Acknowledgments

We thank Caryn Hale (Terns lab) for contributions to the early stages of this project, Caryn Hale and Claiborne Glover (University of Georgia) for critical review of the manuscript, and Michael Adams (University of Georgia) for providing a PF1131 expression construct. This work was supported by NIH grant R01 GM54682 to M.T. and R.T. and NIH grant R01 GM66958 to H.L.. X-ray diffraction data were collected from the Southeast Regional Collaborative Access Team (SER-CAT) 22-ID beamline at the Advanced Photon Source, Argonne National Laboratory. Supporting institutions for APS beamlines may be found at <http://necat.chem.cornell.edu/> and www.ser-cat.org/members.html. Use of the Advanced Photon Source was supported by the U. S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. W-31-109-Eng-38.

References

1. Farazi TA, Juranek SA, Tuschl T: **The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members.** *Development* 2008, **135**:1201-1214.
2. Girard A, Hannon GJ: **Conserved themes in small-RNA-mediated transposon control.** *Trends Cell Biol* 2008, **18**:136-148.
3. Ding SW, Voinnet O: **Antiviral immunity directed by small RNAs.** *Cell* 2007, **130**:413-426.
4. Jaskiewicz L, Filipowicz W: **Role of Dicer in posttranscriptional RNA silencing.** *Curr Top Microbiol Immunol* 2008, **320**:77-97.
5. Godde JS, Bickerton A: **The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes.** *J Mol Evol* 2006, **62**:718-729.
6. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV: **A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action.** *Biol Direct* 2006, **1**:7.
7. Liljestol RK, Redder P, Garrett RA, Brugger K: **A putative viral defence mechanism in archaeal cells.** *Archaea* 2006, **2**:59-72.
8. Sorek R, Kounin V, Hugenholtz P: **CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea.** *Nat Rev Microbiol* 2008, **6**:181-186.
9. Jansen R, Embden JD, Gaastra W, Schouls LM: **Identification of genes that are associated with DNA repeats in prokaryotes.** *Mol Microbiol* 2002, **43**:1565-1575.
10. Makarova KS, Aravind L, Grishin NV, Rogozin IB, Koonin EV: **A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis.** *Nucleic Acids Res* 2002, **30**:482-496.
11. Haft DH, Selengut J, Mongodin EF, Nelson KE: **A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes.** *PLoS Comput Biol* 2005, **1**:e60.
12. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD: **Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin.** *Microbiology* 2005, **151**:2551-2561.

13. Horvath P, Romero DA, Coute-Monvoisin AC, Richards M, Deveau H, Moineau S, Boyaval P, Fremaux C, Barrangou R: **Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus***. *J Bacteriol* 2008, **190**:1401-1412.
14. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E: **Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements**. *J Mol Evol* 2005, **60**:174-182.
15. Pourcel C, Salvignol G, Vergnaud G: **CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies**. *Microbiology* 2005, **151**:653-663.
16. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P: **CRISPR provides acquired resistance against viruses in prokaryotes**. *Science* 2007, **315**:1709-1712.
17. Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S: **Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus***. *J Bacteriol* 2008, **190**:1390-1400.
18. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J: **Small CRISPR RNAs guide antiviral defense in prokaryotes**. *Science* 2008, **321**:960-964.
19. Andersson AF, Banfield JF: **Virus population dynamics and acquired virus resistance in natural microbial communities**. *Science* 2008, **320**:1047-1050.
20. Tyson GW, Banfield JF: **Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses**. *Environ Microbiol* 2008, **10**:200-207.
21. Hale C, Kleppe, K., Terns, R. M., Terns, M. P.: **Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus***. *RNA* 2008, **In Press**.
22. Tang TH, Bachellerie JP, Rozhdestvensky T, Bortolin ML, Huber H, Drungowski M, Elge T, Brosius J, Huttenhofer A: **Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus***. *Proc Natl Acad Sci U S A* 2002, **99**:7536-7541.
23. Tang TH, Polacek N, Zywicki M, Huber H, Brugger K, Garrett R, Bachellerie JP, Huttenhofer A: **Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus***. *Mol Microbiol* 2005, **55**:469-481.
24. Kunin V, Sorek R, Hugenholtz P: **Evolutionary conservation of sequence and secondary structures in CRISPR repeats**. *Genome Biol* 2007, **8**:R61.

25. Maris C, Dominguez C, Allain FH: **The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression.** *Febs J* 2005, **272**:2118-2131.
26. Calvin K, Li H: **RNA-splicing endonuclease structure and function.** *Cell Mol Life Sci* 2008, **65**:1176-1185.
27. Hammond SM: **Dicing and slicing: the core machinery of the RNA interference pathway.** *FEBS Lett* 2005, **579**:5822-5829.
28. Beloglazova N, Brown G, Zimmerman MD, Proudfoot M, Makarova KS, Kudritska M, Kochinyan S, Wang S, Chruszcz M, Minor W, et al.: **A Novel Family of Sequence-specific Endoribonucleases Associated with the Clustered Regularly Interspaced Short Palindromic Repeats.** *J Biol Chem* 2008, **283**:20361-20371.
29. Baker DL, Youssef OA, Chastkofsky MI, Dy DA, Terns RM, Terns MP: **RNA-guided RNA modification: functional organization of the archaeal H/ACA RNP.** *Genes Dev* 2005, **19**:1238-1248.
30. Youssef OA, Terns RM, Terns MP: **Dynamic interactions within sub-complexes of the H/ACA pseudouridylation guide RNP.** *Nucleic Acids Res* 2007, **35**:6196-6206.
31. Terwilliger TC, Berendzen J: **Automated MAD and MIR structure solution.** *Acta Crystallogr D Biol Crystallogr* 1999, **55**:849-861.
32. Emsley P, Cowtan K: **Coot: model-building tools for molecular graphics.** *Acta Crystallogr D Biol Crystallogr* 2004, **60**:2126-2132.
33. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, et al.: **Crystallography & NMR system: A new software suite for macromolecular structure determination.** *Acta Crystallogr D Biol Crystallogr* 1998, **54**:905-921.
34. Murshudov GN, Vagin AA, Dodson EJ: **Refinement of macromolecular structures by the maximum-likelihood method.** *Acta Crystallogr D Biol Crystallogr* 1997, **53**:240-255.
35. Laskowski RA, MacArthur MW, Moss DS, Thornton JM: **PROCHECK: a program to check the stereochemical quality of protein structures.** *Journal of Applied Crystallography* 1993, **26**:283-291.
36. Petrey D, Honig B: **GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences.** *Methods Enzymol* 2003, **374**:492-509.

Fig. 2.1. Cas6 is an endoribonuclease that cleaves CRISPR RNAs within repeat sequences.

(A) psiRNA biogenesis pathway model. The primary CRISPR transcript contains unique invader-targeting or guide sequences (colored blocks) flanked by direct repeat sequences (R). Cas6 catalyzes site-specific cleavage within each repeat, releasing individual invader-targeting units. The Cas6 cleavage products undergo further processing to generate smaller mature psiRNA species. **(B)** Purified recombinant PfCas6 expressed in *E. coli*. The sizes (kDa) of protein markers (M) are indicated. **(C)** Radiolabeled RNAs (repeat-guide-repeat (R-g-R) or repeat alone (R), as diagrammed) were either uniformly or 5' end labeled and incubated in the absence (-) or presence (+) of PfCas6 protein (500 nM). Products were resolved by denaturing gel electrophoresis and visualized using a phosphorimager. The main cleavage products are indicated by a star or asterisk on the gel and in the diagram.

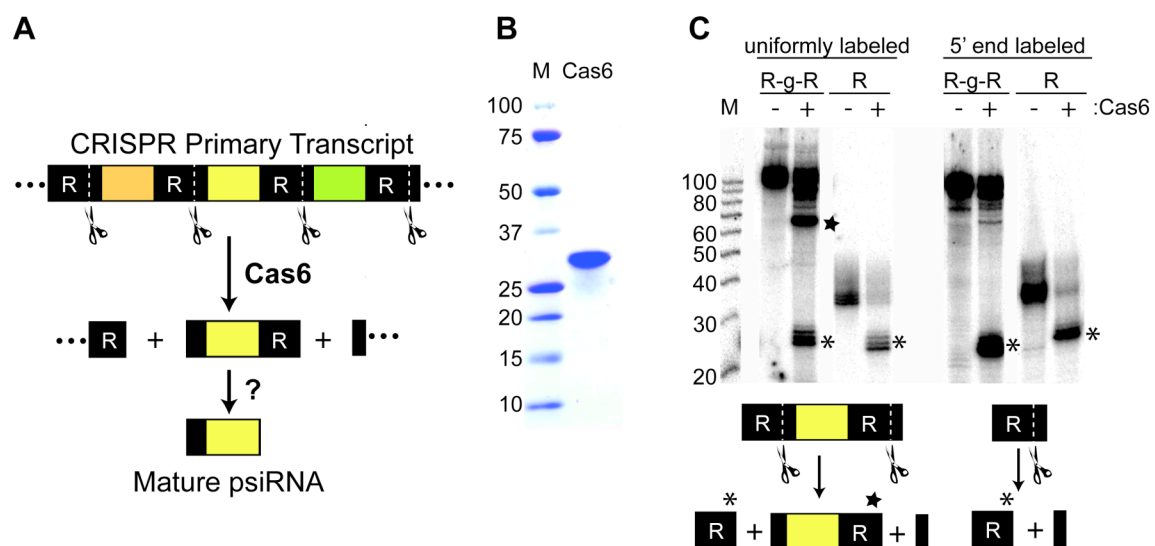


Fig. 2.2. PfCas6 cleavage of a CRISPR RNA containing two repeat-guide RNA units. A uniformly radiolabeled substrate RNA containing two guide (invader targeting) sequences (yellow and green), two repeats (R) and a short (natural) 5' leader (L) sequence was incubated with 1 μ M PfCas6 protein and samples were analyzed by denaturing gel electrophoresis at the indicated times. The expected sizes and compositions of the RNA products (based on site-specific cleavage within each repeat) are indicated, as are the sizes of the marker RNAs (M).

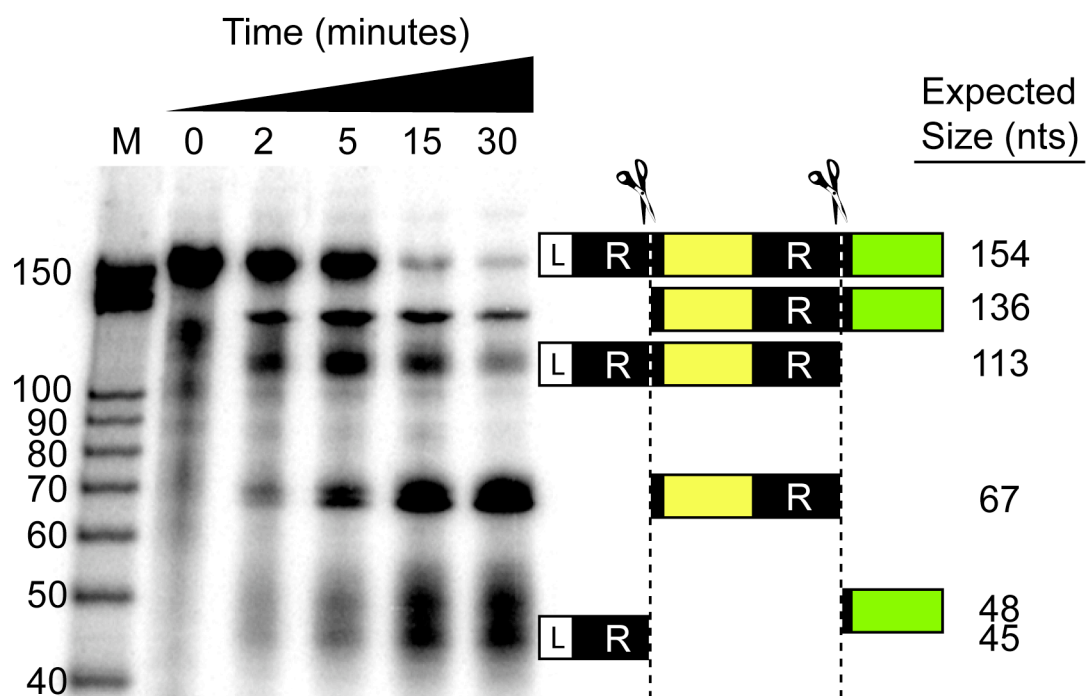


Fig. 2.3. Identification of the site of PfCas6 cleavage within the CRISPR repeat RNA. (A)

The site of PfCas6 cleavage within the CRISPR repeat RNA was mapped by incubating 5' end labeled repeat RNA with PfCas6 nuclease and comparing the size of the 5' RNA cleavage product (arrow) with RNase T1 (T1) and alkaline hydrolysis (OH) sequence ladders. **(B)** Potential secondary structure of *P. furiosus* repeat RNA with cleavage site indicated. **(C)** Analysis of cleavage of wildtype and cleavage site mutant (AA to GG) repeat RNAs with increasing concentrations (0, 1, 50, 200, and 500 nM) of PfCas6. **(D)** Native gel mobility shift analysis of wildtype and mutant repeat RNAs with increasing concentrations of PfCas6. The positions of the free (RNA) and protein-bound (RNP) RNAs are indicated. 5' and 3' cleavage products are indicated in both panels C and D. The sizes of RNA markers (M) are indicated in panels c, d and f.

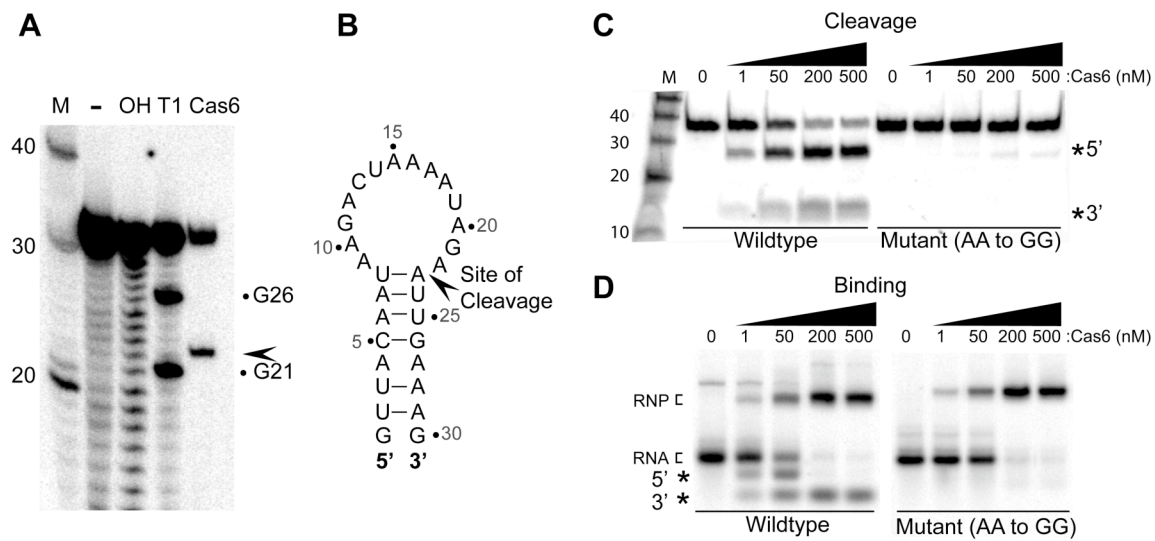
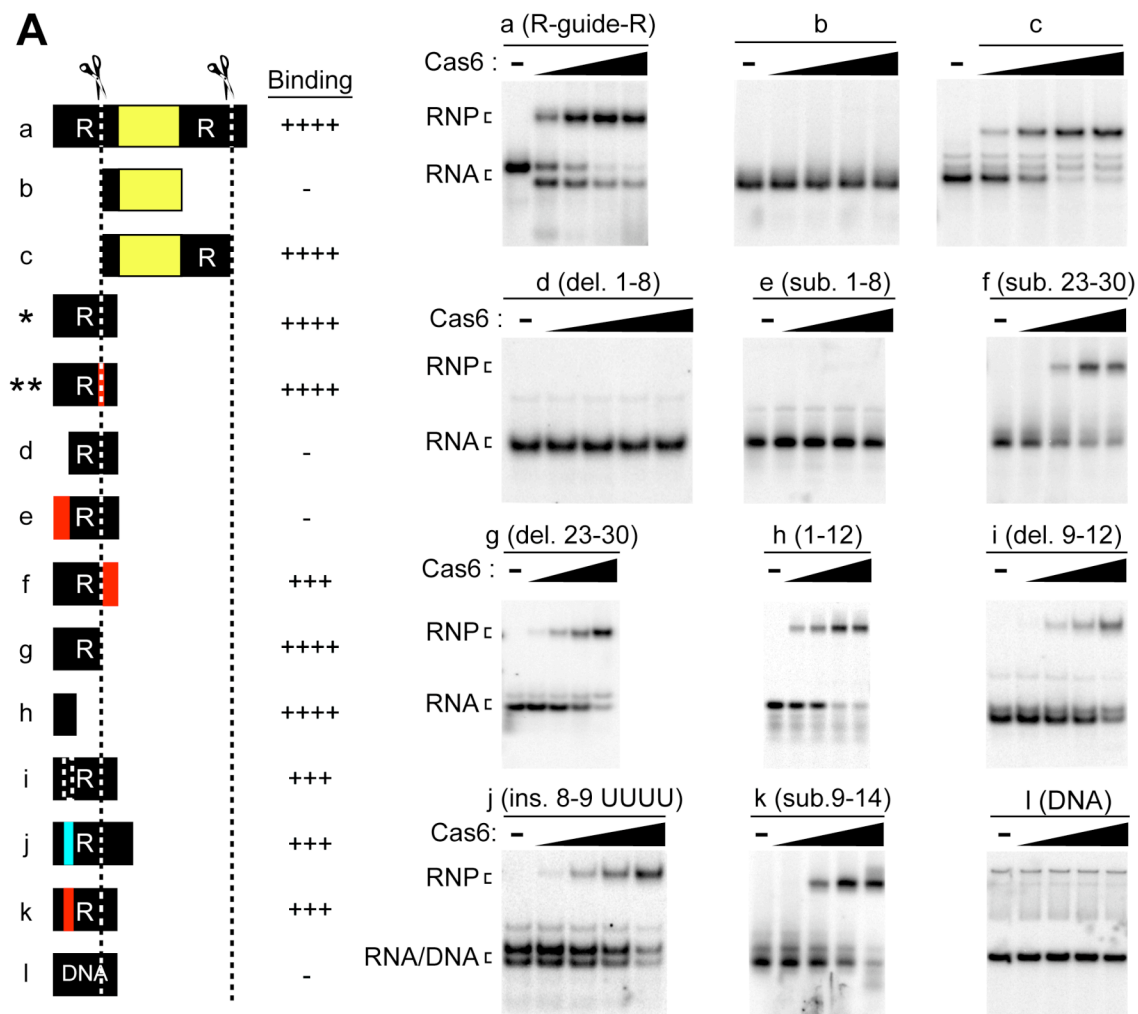
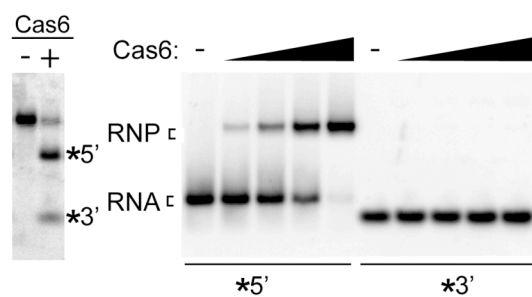


Fig. 2.4. CRISPR repeat sequence requirements for PfCas6 binding. (A) Detailed analysis of binding with a series of CRISPR-derived RNAs and mutants. Left panel illustrates the RNAs tested, with repeat (R) and invader targeting (yellow blocks) sequences, and PfCas6 cleavage site (dashed lines) indicated. Blue block denotes an insertion, dashed block denotes an internal deletion and red blocks denote substitutions (with complementary sequence). DNA indicates a DNA repeat sequence substrate. PfCas6 binding is summarized relative to binding to the 5' cleavage product (++++). Corresponding RNA diagrams and data panels are designated with lowercase letters. Right panels show gel mobility shift analysis of the indicated RNAs with increasing concentrations (0, 1, 50, 200, and 500 nM) of PfCas6. Substrates are uniformly radiolabeled except for those shown in panels a, b, c, and l, which are 5' end labeled. Data for the intact repeat (*) and cleavage site mutant (**) are shown in Figure 3D. (B) PfCas6 interacts with the gel-purified 5' cleavage product. Left panel shows the products of incubation of uniformly-radiolabeled repeat RNA with (+) or without (-) PfCas6 (1 μ M). The positions of the 5' and 3' cleavage products are indicated. Right panel shows native gel mobility shift analysis of the gel-purified 5' and 3' PfCas6 cleavage products (from left panel) with increasing concentrations (0, 1, 50, 200, and 500 nM) of PfCas6. The positions of free (RNA) and protein-bound RNA (RNP) are indicated. (C) Model summarizing the minimal PfCas6 binding site within the CRISPR repeat RNA relative to the cleavage site.



B



C

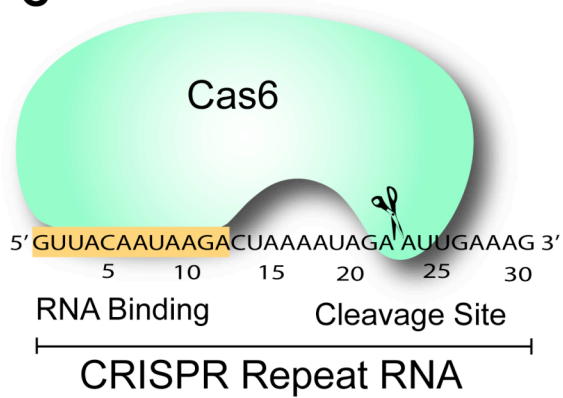


Fig. 2.5. CRISPR repeat sequence requirements for PfCas6 cleavage. Detailed analysis of cleavage with a series of CRISPR-derived RNAs and mutants. Left panel illustrates the RNAs tested as in Figure 4. PfCas6 cleavage is summarized relative to cleavage of the intact repeat RNA (++++). PfCas6 binding is summarized from Figure 4. Corresponding RNA diagrams and data panels are designated with lowercase letters. Right panels show cleavage assays using uniformly-radiolabeled repeat RNA with (+) or without (-) PfCas6 (500 nM). Data for the intact repeat (*) is shown on right and data for the cleavage site mutant (***) is shown in Figure 3C.

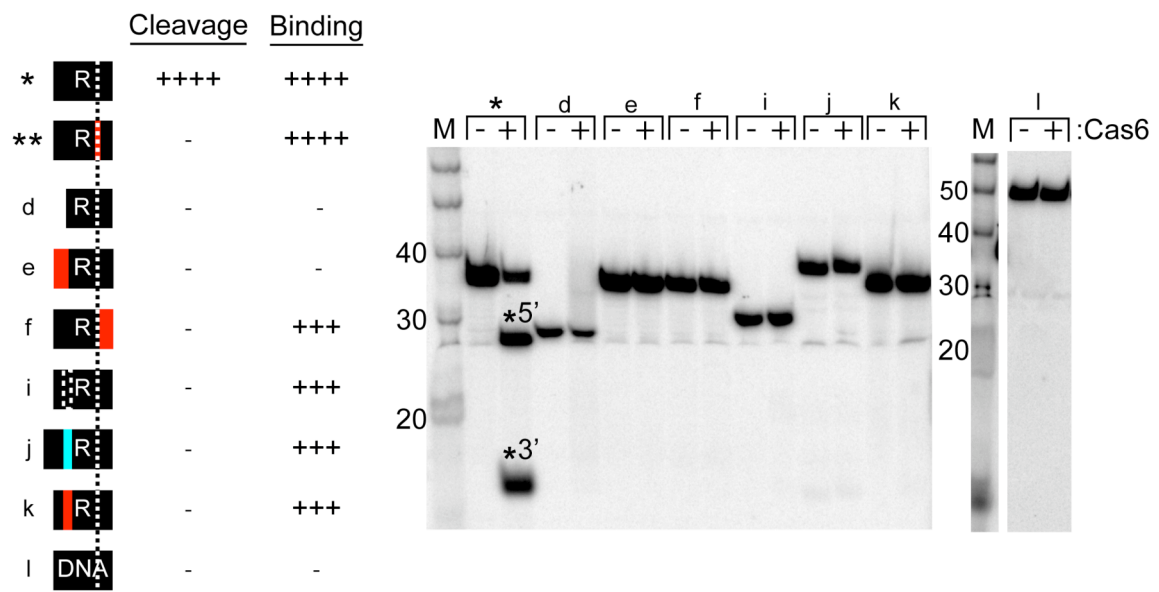


Fig. 2.6. Structural features of PfCas6. Front (**A**) and back (**B**) back views of the structure of PfCas6 represented in ribbon diagrams (left) and colored electrostatic surface potential (right). In the center, the fold topology is illustrated with arrows (β -strands) and circles (α -helices). In the ribbon diagrams, the G-rich loop characteristic of RAMP proteins is designated in red and the predicted catalytic triad residues are indicated in green. The electrostatic potential was computed using the GRASP2 program [36] and is colored red and blue, for negative and positive potentials, respectively.

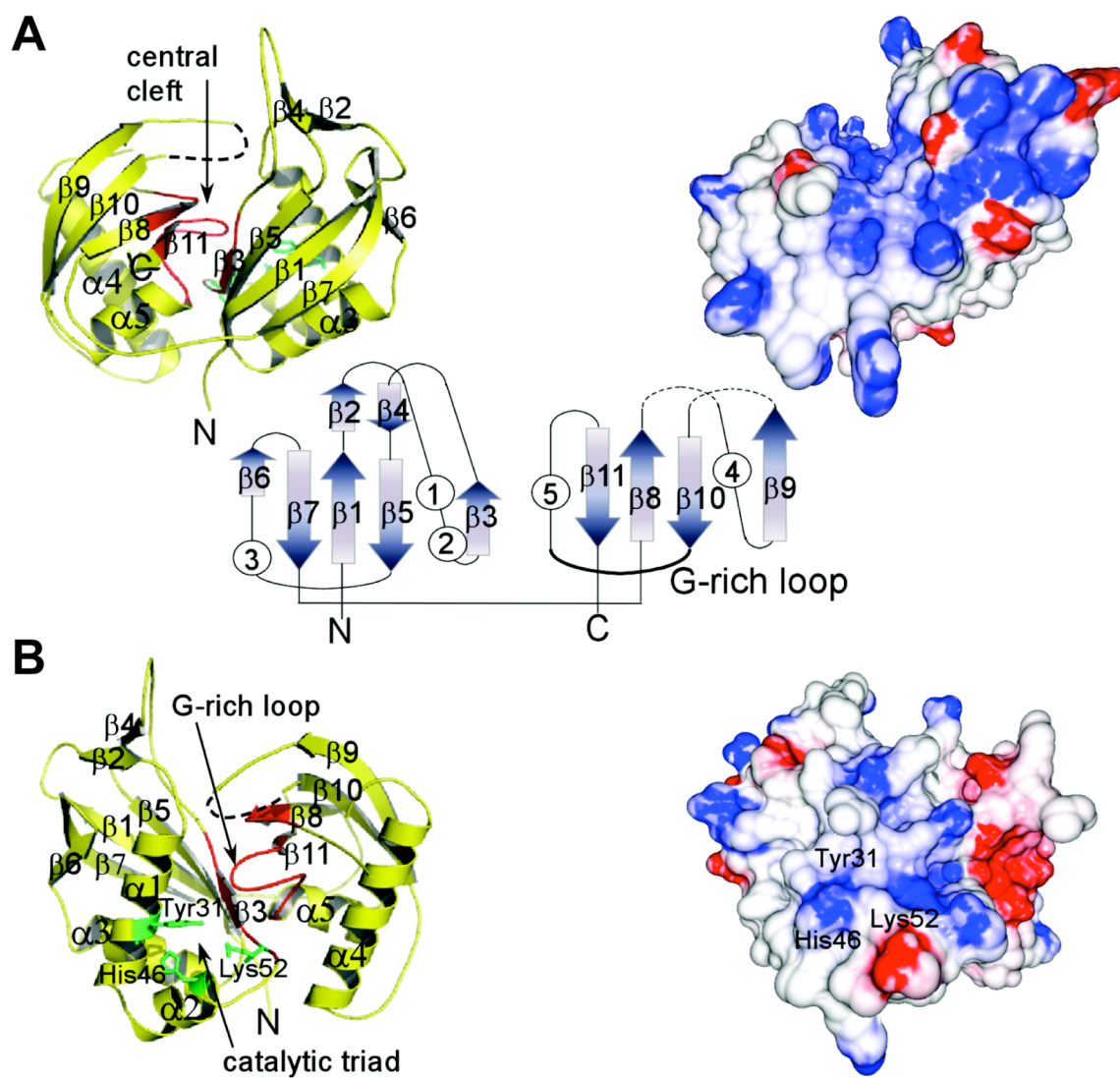


Fig. 2.7. Catalytic features of PfCas6 cleavage activity. (A) Cleavage activity is not dependent on divalent metal ions. Uniformly-radiolabeled repeat RNA was incubated with 1 μ M PfCas6 in the absence (-) or presence (+) of 1.5 mM MgCl₂ or 20 mM metal chelator EDTA as indicated. (B) Analysis of the termini of PfCas6 cleavage products. The products of cleavage reactions performed with unlabeled repeat RNA substrates (initially containing hydroxyl groups at both the 5' and 3' termini) were radiolabeled at either their 5' ends (using ³²P-ATP and polynucleotide kinase) or 3' ends (using ³²pCp and RNA ligase). The positions of the 5' and 3' cleavage products are indicated in (A) and (B). (C) The pattern of radiolabeling of the RNA cleavage products (B) indicates that PfCas6 cleaves on the 5' side of the phosphodiester bond, as is the case for other metal-independent ribonucleases. Cleavage likely generates 5' hydroxyl (OH) and 2', 3' cyclic phosphate (>P) RNA termini.

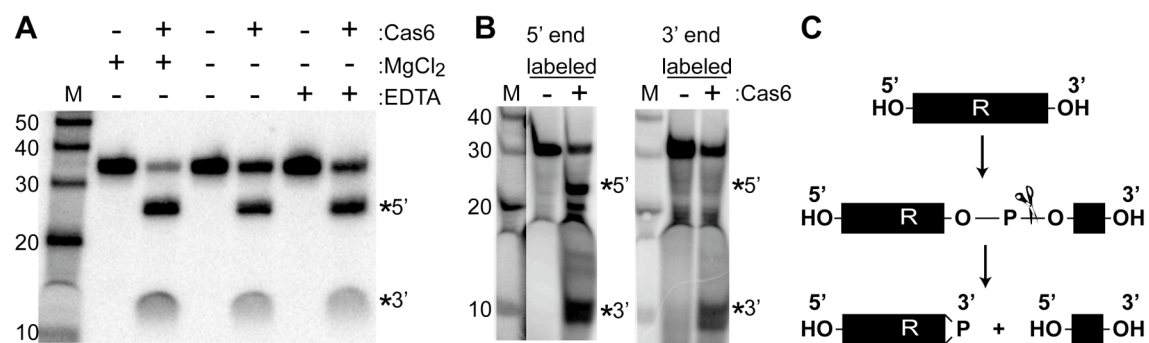


Fig. 2.S1. Influence of temperature on the ability of PfCas6 to bind and cleave CRISPR repeat RNA. Repeat RNA (uniformly radiolabeled) was incubated with (+) or without (-) 1 μ M PfCas6 protein at the indicated temperatures and the products were resolved by electrophoresis on denaturing (**A**) or native (**B**) polyacrylamide gels to assess RNA binding or cleavage, respectively. The positions of the 5' and 3' cleavage products are indicated. The positions of the free (RNA) and protein-bound (RNP) RNAs are indicated in panel B. Based on the data shown in panel A, the RNPs in panel B include primarily the 5' cleavage product at higher temperatures and the intact repeat at lower temperatures.

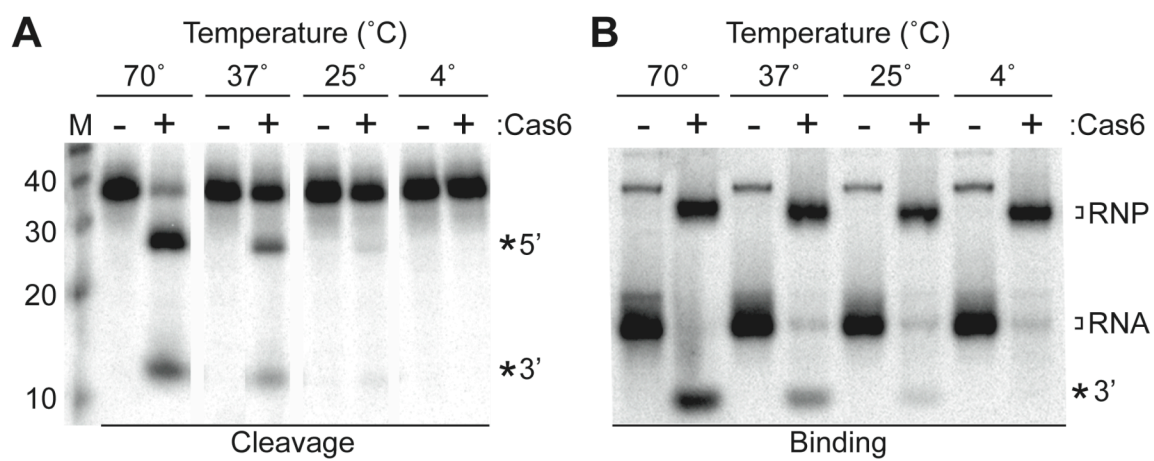
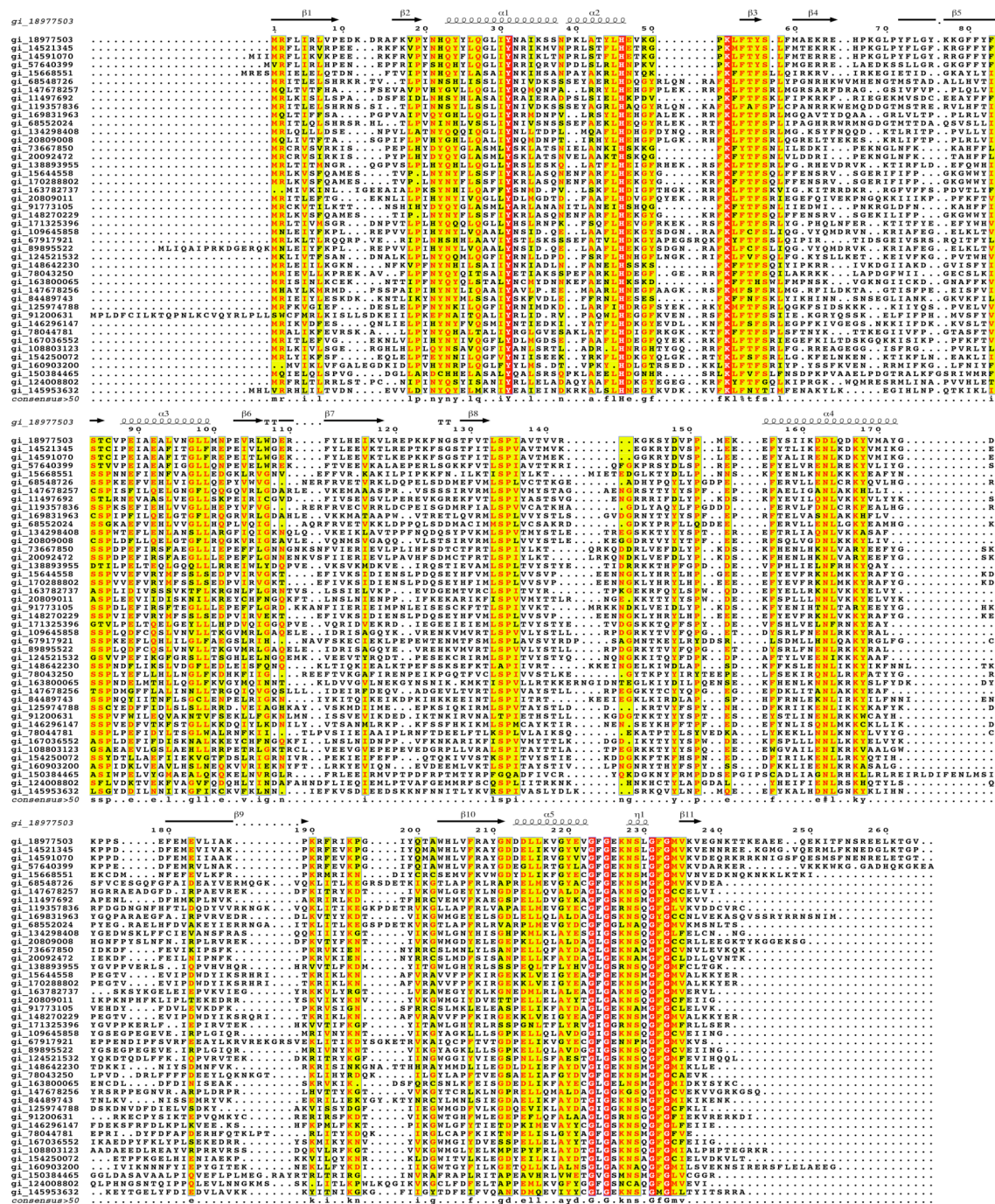


Fig. 2.S2. Amino acid sequence alignment of Cas6 proteins. PSI-BLAST of PF1131 amino acid sequence against the non-redundant protein database (nr) at NCBI yielded 151 protein sequences that have *E*-values of less than 10^{-4} . It was immediately clear that many organisms contain more than one Cas6-related sequence. These fell into two distinctive classes: one that includes the conserved triad residues (like PfCas6) and one that does not. We aligned 42 Cas6 homologs that appear to belong to the first class and have *E*-values of less than 10^{-23} . In this alignment, the only strictly conserved residues are the putative catalytic triad residues and the four glycine residues in the G-rich loop. Organisms and genes listed include: *Pyrococcus furiosus* DSM 3638 (gi_18977503), *Pyrococcus abyssi* GE5 (gi_14521345), *Pyrococcus horikoshii* OT3 (gi_14591070), *Thermococcus kodakarensis* KOD1 (gi_57640399), *Methanocaldococcus jannaschii* DSM 2661 (gi_15668551), *Pelodictyon phaeoclathratiforme* BU-1 (gi_68548726), *Archaeoglobus fulgidus* DSM 4304 (gi_11497692), *Chlorobium phaeobacteroides* DSM 266 (gi_119357836), *Candidatus Desulforudis audaxviator* MP104C (gi_169831963), *Prosthecochloris aestuarii* DSM 271 (gi_68552024), *Desulfotomaculum reducens* MI-1 (gi_134298408), *Thermoanaerobacter tengcongensis* MB4 (gi_20809008), *Methanosarcina barkeri* str. Fusaro (gi_73667850), *Methanosarcina acetivorans* C2A (gi_20092472), *Geobacillus thermodenitrificans* NG80-2 (gi_138893955), *Thermotoga maritima* MSB8 (gi_15644558), *Thermotoga* sp. RQ2 (gi_170288802), *Hydrogenivirga* sp. 128-5-R1-1 (gi_163782737), *Thermoanaerobacter tengcongensis* MB4 (gi_20809011), *Methanococcoides burtonii* DSM 6242 (gi_91773105), *Thermotoga petrophila* RKU-1 (gi_148270229), *Geobacillus* sp. WCH70 (gi_171325396), *Desulfitobacterium hafniense* DCB-2 (gi_109645858), *Chlorobium limicola* DSM 245 (gi_67917921), *Desulfitobacterium hafniense* Y51 (gi_124521532), *Methanobrevibacter smithii* ATCC 35061 (gi_148642230), *Carboxydotherrmus*

hydrogenoformans Z-2901 (gi_78043250), Methanococcus voltae A3 (gi_163800065), Pelotomaculum thermopropionicum SI (gi_147678256), Methanosphaera stadtmanae DSM 3091 (gi_84489743), Clostridium thermocellum ATCC 27405 (gi_125974788), Candidatus Kuenenia stuttgartiensis (gi_91200631), Caldicellulosiruptor saccharolyticus DSM 8903 (gi_146296147), Carboxydotherrnus hydrogenoformans Z-2901 (gi_78044781), Thermoanaerobacter pseudethanolicus ATCC 33223 (gi_167036552), Rubrobacter xylanophilus DSM 9941 (gi_108803123), Fervidobacterium nodosum Rt17-B1 (gi_154250072), Petrotoga mobilis SJ95 (gi_160903200), Victivallis vadensis ATCC BAA-548 (gi_150384465), Microscilla marina ATCC 23134 (gi_124008802), Clostridium difficile QCD-32g58 (gi_145953632).



Chapter 3

CHARACTERIZATION OF CAS6/CRISPR REPEAT INTERACTION AND CLEAVAGE
MECHANISM¹

¹ Carte, J., N.T. Pfister, M.M. Compton, R.M. Terns, M.P. Terns. To be submitted for publication in *RNA Biology*.

Abstract

The CRISPR-Cas system provides many prokaryotes with rapidly evolving acquired resistance to viruses and other mobile genetic elements. The two components of this system are host encoded CRISPR (cr)RNAs and Cas proteins. Invader-derived sequence elements present on the crRNAs guide effector Cas proteins to an intruding nucleic acid, leading to silencing of the target. Critical in this process is the biogenesis of mature crRNAs from primary CRISPR RNA transcripts. Cas6, a previously identified endoribonuclease present in many organisms, performs the initial step in this biogenesis pathway. Previous mutational analysis indicated that Cas6 recognizes sequence elements in the 5' region of CRISPR repeat RNA. In the present study, we map at high resolution the nucleotides involved in a direct interaction with Cas6 and find that nucleotides 2-8 appear to contact Cas6 directly. Additionally, a critical role in catalysis for Tyr31, His46, and Lys52 is established, suggesting a general acid/base catalytic mechanism. Finally, we show that native Cas6 from *Pyrococcus furiosus* cleaves CRISPR repeat RNA and stably associates with its predicted cleavage products but only weakly with mature crRNAs, consistent with primary role in crRNA biogenesis.

Introduction

Recently, it has become clear that many prokaryotes possess a defense mechanism known as the CRISPR-Cas system that provides resistance to genome invaders. Found in approximately 40% of bacteria and 90% of archaea, this system is composed of small invader targeting RNAs derived from Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) loci and Cas (CRISPR-associated) proteins [1,2]. CRISPR loci are composed of short (25-40 nucleotide) direct repeats separated by similarly sized, invader derived sequences called spacers [3-7]. It has

been shown experimentally that organisms harboring spacers derived from a known phage are resistant to infection by that phage [8-10]. In addition, infection of phage-sensitive strains of *Streptococcus thermophilus* has been shown to result in incorporation of novel spacers and phage resistance [8,9,11]. Thus, the CRISPR-Cas system provides a rapidly evolving defense to combat attacks on the genome and has been shown to promote host cell survival in microbial communities [12,13].

CRISPR (cr)RNAs are transcribed from a conserved leader sequence and are processed into small RNAs that contain a single invader targeting spacer and a portion of the repeat [7,10,14-16]. Previously we showed that in *Pyrococcus furiosus* the initial processing step is performed by one of the highly conserved “core” Cas proteins, Cas6 [17]. Cas6 cleaves endonucleolytically and site-specifically within the repeat region to produce a crRNA intermediate which, in *P. furiosus*, undergoes further processing into stable 39 and 45 nucleotide mature crRNAs [16,17].

Cas6 is one of the most widely distributed Cas proteins, found in both bacteria and archaea [6,18]. Cas6 proteins contain a characteristic C-terminal GhGxxxxGhG motif (h indicates a hydrophobic amino acid and xxxx is composed of at least one lysine or arginine) [18]. Previous mutational analysis indicated that Cas6 selectively interacts with sequence elements in the 5' region of CRISPR repeat RNA [17]. Cas6 cleaves repeat RNA at a site eight nucleotides upstream of the 3' end of the repeat [17]. Cleavage was determined to be divalent metal ion independent, generating cleavage products with a 5' hydroxyl and a likely 2'-3' cyclic phosphate, characteristic of such enzymes. The crystal structure of Cas6 from *P. furiosus* revealed that a conserved trio of Tyr, His, and Lys that cluster directly adjacent to the conserved

Gly-rich loop [17]. We proposed that these three amino acids compose a catalytic triad required for cleavage of the CRISPR repeat RNA [17].

In the present study, Cas6 substrate recognition was probed at single nucleotide resolution using RNA footprinting. The results of this analysis confirm that sequence elements in the 5' region of the repeat are the primary determinants for recognition by Cas6 and that nucleotides 2-8 likely have direct contact with Cas6. Also, through mutational analysis, a critical role of the predicted catalytic triad was established and we propose an acid/base catalytic mechanism involving these three amino acids. Finally, native Cas6 was isolated from *P. furiosus* extract, was shown to cleave CRISPR repeat RNA, and was found to co-purify with several crRNA species.

Results

Mapping the Cas6/CRISPR repeat RNA interaction

We previously reported that the 5' region of CRISPR repeat RNA was important for recognition by Cas6. Substitution or deletion of these nucleotides prevented detectable binding [17]. Additionally, an RNA consisting of nucleotides 1-12 of the repeat displayed a binding affinity comparable to that of the full length repeat RNA [17]. Sequence elements in the middle and 3' regions of the repeat did not appear to be important for binding given that Cas binding was insensitive to deletions or substitutions in these regions of the repeat RNA [17]. In order to gain a more detailed understanding of Cas6 recognition of CRISPR repeat RNA, RNA footprinting was performed with radioactively labeled (either 3' or 5' end-labeled) CRISPR repeat RNA and recombinant PfCas6 protein. Lead (II) acetate (cleaves single-stranded and

tertiary interactions) and RNase A (cleaves after unpaired Cs and Us) were chosen as probing reagents.

A strong protection was observed in the 5' region of the repeat with both lead (II) acetate and RNase A using 3' end-labeled repeat RNA (Fig 3.1A and 3.1C). Specifically, nucleotides 2-8 were protected from lead induced cleavage in a Cas6 concentration dependent manner. A similar protection profile was observed with RNase A, with cleavage products at nucleotides 3, 5, and 8 becoming less susceptible to degradation in the presence of Cas6. No protection was observed with either lead (II) acetate or RNase A within the 3' region of the repeat using 5' end labeled repeat (Fig 3.1B and 3.1C). Similar results were obtained when RNase T1 was used as a cleavage reagent for RNA footprinting (data not shown).

These findings are in good agreement with previous results in which RNA mutagenesis revealed that sequence elements in the 5' region of the repeat are the primary determinants for recognition by Cas6 [17]. Despite weak potential for the 5' and 3' regions of repeat RNA to base-pair, consistent with predictions made by *in silico* analysis [19], repeats from *P. furiosus* appear to be mostly unstructured in solution (RNA alone in Fig 3.1A and 3.1B and data not shown).

Mutational analysis of a putative catalytic amino acid triad

Cleavage of CRISPR RNA repeats by Cas6 has been predicted to involve a conserved Tyr, His, Lys triad [17]. This prediction was based both on the high degree of conservation of these amino acids and the observation that in the crystal structure these amino acids clustered in close proximity to one another in a similar configuration to that observed in the archaeal tRNA splicing endonuclease [18,20]. In order to determine whether these amino acids are required for

catalysis, Cas6 proteins containing single amino acid substitutions (Y31A, Y31F, H46A, H46Q, K52A, and K52E) were expressed and purified in *Escherichia coli* (Fig 3.2B) and assessed for their ability to cleave radiolabeled CRISPR repeat RNA.

Mutation of any of the three triad amino acids led to a significant decrease or complete loss of cleavage activity relative to wild type Cas6 (Fig. 3.2A). Cleavage of the repeat was abolished in the Y31A, H46A, and H46Q mutants indicating that Tyr31 and His46 likely play a critical role in catalysis. In the K52A and K52E mutants, cleavage was reduced >40 and >150 fold at the highest concentration tested (500 nM), respectively, suggesting a key role for this residue in catalysis. Significant cleavage activity was retained in the Y31F mutant, with only ~2-fold reduction in cleavage compared to the wild type. Similar results were obtained when residues of the catalytic triad of the tRNA splicing endonuclease from *Archaeoglobus fulgidus* and *Thermoplasma acidophilum* were mutated [21,22].

Next, we tested whether the loss or reduction of cleavage observed in each of the Cas6 mutants was due to an inability of the mutants to bind to the substrate RNA. To this end, native gel mobility shift assays were performed with each Cas6 mutant and radiolabeled CRISPR repeat RNA (Fig. 3.3). The ability of Cas6 to bind CRISPR repeat RNA was largely unaffected by mutations in the proposed catalytic triad. Each of the six Cas6 mutants was able to form a stable complex with CRISPR repeat RNA with a similar binding affinity as wild type Cas6 (Fig. 3.3). Thus Tyr31, His46, and Lys52, are required for efficient cleavage of the repeat and likely play direct roles in catalysis.

Native Cas6 cleaves CRISPR repeat RNA and associates with crRNA intermediates.

In order to determine whether native Cas6 behaves similarly to the recombinant protein, polyclonal antibodies were raised against recombinant Cas6 and used to immunoprecipitate the native protein from a *P. furiosus* extract. The immunoprecipitation samples, along with whole cell extract (WCE) were then tested for Cas6 cleavage activity by incubation with uniformly labeled repeat RNA (Fig. 3.4A). Remarkably, in the WCE, the CRISPR repeat RNA was cleaved into same products generated by recombinant Cas6 with no other cleavage products evident. Additionally, Cas6 cleavage activity was present in the immune, but not preimmune pellet following immunoprecipitation, indicating that the cleavage activity observed in WCE was carried out by native Cas6. The cleavage activity observed in *P. furiosus* extract was found to be divalent metal ion independent, as was shown to be the case for recombinant Cas6.

It has been shown that a similar CRISPR repeat RNA-cleaving endoribonuclease, Cse3 from *E. coli*, is found as part of a large ribonucleoprotein complex (RNP) that contains both Cas proteins and crRNAs [10]. In order to determine whether Cas6 was also part of an RNP, RNAs were extracted from immunoprecipitation samples and probed by northern blot analysis for a crRNA spacer. Cas6 co-purifies with several crRNA species including the 2X and 1X intermediates [16], which correspond to cleavage products generated by Cas6 cleavage of the CRISPR primary transcript (Fig. 3.4B). Cas6 also weakly co-purified with mature crRNAs (Fig. 3.4B).

Discussion

The biogenesis of mature crRNAs is critical to CRISPR-Cas mediated resistance to genome invaders. The initial processing step, endonucleolytic cleavage of the primary transcript

within the repeat region is performed by Cas6 in *P. furiosus* [17]. This cleavage results in a crRNA intermediate that retains eight nucleotides of the repeat at the 5' end and ~22 nucleotides of the next repeat at the 3' end. In *P. furiosus*, it appears that this crRNA intermediate is then processed at the 3' end to yield two mature crRNA species that retain the eight nucleotide, repeat derived "tag" that we propose serves as a recognition sequence for other Cas proteins [16].

In an *E. coli* K12 strain, a ribonuclease that performs a similar function as Cas6, Cse3, was shown to be required for crRNA biogenesis [10]. Cse3 is a divalent metal ion independent endoribonuclease that cleaves CRISPR repeat RNAs within the 3' region of the repeat. Although the sequences of the repeat RNAs differ, the position of cleavage on *E. coli* derived CRISPR repeat RNAs and Cas6 cleavage of *P. furiosus* repeat RNAs, occurs eight nucleotides upstream of the 3' end of the repeat. This cleavage generates an eight nucleotide tag which is retained on the mature crRNAs in both *E. coli* and *P. furiosus* [10,16]. The presence of these eight nucleotides may be a universal feature of crRNAs, serving as a recognition sequence for effector Cas proteins that rely on the spacer sequence to guide the complex to invading mobile genetic elements.

Cse3 from *E. coli* was found to be a component of a large RNP containing a number of other Cas proteins as well as mature crRNAs [10]. In the present study we have shown that Cas6 appears to associate with several crRNA species, including its predicted cleavage products, the 1X crRNA intermediate (Fig. 3.4B). Because it remains bound to its cleavage product, Cas6 may influence the 3' end processing of the 1X crRNA intermediate. Cas6 is not likely to be a structural component of the invader-targeting effector complex because the protein only weakly associates with mature crRNAs (Fig. 3.4B).

The structure of Cse3 from *Thermus thermophilus* revealed a very similar overall architecture as that shown in the structure of Cas6 from *P. furiosus* [23]. That is, Cse3 is composed of duplicated ferredoxin folds separated by a central cleft which contains a conserved Gly-rich loop [23]. Located adjacent to this loop is an invariant His residue that was shown to be required for Cse3 cleavage activity [10,23]. In our study, Cas6 mediated cleavage of CRISPR repeat RNA was shown to require the highly conserved His46 residue, which is also located adjacent to the conserved Gly-rich loop characteristic of Cas6 proteins. Cse3, however, lacks conserved Tyr and Lys residues that were shown in the current study to be important for Cas6 cleavage activity. Therefore it seems that despite the aforementioned similarities between the two, Cas6 and Cse3 likely employ distinct catalytic mechanisms.

Proposed Cas6 cleavage mechanism.

We propose a general acid/base catalytic mechanism for Cas6 based on similar active site architecture and reaction characteristics to the archaeal tRNA splicing endonuclease. In this proposed mechanism, a proton is abstracted from the 2' hydroxyl of the ribose ring by the hydroxyl group of Tyr31 (Fig. 3.5). The ability of Y31F mutant to support significant cleavage activity is consistent with previous studies involving archaeal tRNA splicing endonuclease [21,22]. It has been proposed that the stereochemistry of the catalytic Tyr, rather than its hydroxyl group, may account for its important role in cleavage. In addition to its role as a general base, Tyr31 may also be required for proper substrate positioning in the active site. Removal of a proton from the 2' hydroxyl of the ribose ring leads to nucleophilic attack by the 2' oxygen of the ribose on the phosphate backbone, resulting in a pentavalent transition state whose negative charge is stabilized by the positive charge of the amine group of Lys52. Cleavage of

the scissile phosphate bond is facilitated by proton donation from the imidazole ring of His46. This would result in cleavage products with a 5' hydroxyl and a 2'-3' cyclic phosphate, consistent with previous findings [17,20].

Cas6 employs a distinct method of substrate recognition and cleavage. Initially, Cas6 binds to sequence elements in the 5' region of CRISPR repeat RNA and then cleavage occurs site specifically at a location outside the binding site. One model for how this might occur is that Cas6 binding to the repeat RNA induces a conformational change in the RNA, possibly involving base pairing between the 5' and 3' palindromic regions of the repeat, resulting in the proper positioning of the scissile phosphate bond in the active site. Alternatively, following substrate recognition by Cas6, the RNA may wrap itself around Cas6 through a series of weak contacts to position the cleavage site in the active site. In this were the case, the weak interactions that occur outside the primary binding site could not be detected by the techniques used in this study. Further studies are required to determine the molecular mechanisms and dynamics that allow substrate recognition and catalysis to occur in distinct regions of both the protein and substrate.

Materials and methods

Expression and purification of PfCas6 and mutants

Primers to generate site-specific mutants were designed and ordered from Eurofins MWG Operon (listed in table S1). Mutant *cas6* genes were generated from a pET24d plasmid containing the PF1131 (Cas6 from *P. furiosus*) insert using QuikChangeTM site-directed mutagenesis (Stratagene) and the DNA sequences were confirmed by sequencing. N-terminal,

6x histidine-tagged proteins were expressed in *E. coli* BL21 codon + (DE3, Invitrogen) and purified to homogeneity as described previously [17].

Generation of radiolabeled RNAs

The synthetic RNA (see table 3.S1 for sequence) and RNA size standards (DecadeTM markers) used in this study were purchased from Integrated DNA technologies (IDT) and Applied Biosystems, respectively. The northern probe used in this study (see Table 3.S1 for sequence) was purchased from Eurofins MWG Operon. RNAs were 5' end-labeled with T4 polynucleotide kinase (Applied Biosystems) and [γ ³²P] ATP (7000 Ci/mmol; MP Biomedicals) as described previously [17]. End-labeling at the 3' end was performed with T4 RNA ligase (Promega) and [α ³²P] pCp (2500 Ci/mmol; MP Biomedicals). A typical reaction contained 10 pmol of RNA, 20 U T4 RNA ligase, 10 U SUPERase-INTM RNase inhibitor (Applied Biosystems), 1X T4 RNA ligase buffer (Promega), 20% polyethylene glycol 3350, and ~12 pmol [α ³²P] pCp. The uniformly labeled CRISPR repeat RNA substrate was generated by *in vitro* transcription by T7 polymerase using annealed DNA oligos containing the T7 promoter sequence as a template (see table 3.S1 for sequences) in the presence of [α ³²P] UTP (800 Ci/mmol; MP Biomedicals) as described previously [17]. All radiolabeled RNAs were extracted with phenol/chloroform/isoamyl alcohol (PCI), precipitated with ethanol, and gel purified as described previously [17].

RNA footprinting

Lead (II) induced and RNase A cleavage were carried out essentially as described previously [24]. Briefly, 0.1 pmol of ³²P end-labeled RNA (either 5' or 3') were incubated in the

absence (free RNA) or presence of increasing concentrations of Cas6 at 65-70°C for 30 minutes in buffer A (20 mM HEPES-KOH [pH 7.0], 500 mM KCl). Lead (II) induced cleavage was initiated by the addition of 15 mM Pb(II) acetate prepared fresh in sterile water. Reactions were carried out at room temperature for 10 minutes and were stopped by the addition of EDTA to a final concentration of 20 mM followed by PCI extraction and ethanol precipitation. RNase A cleavage was initiated by the addition of .01 ng of RNase A (Applied Biosystems) and incubated at 37°C for 15 minutes. Reactions were stopped by PCI extraction followed by ethanol precipitation. Alkaline hydrolysis ladders (cleavage after each nucleotide) were generated as described previously [24]. In each case, precipitated RNAs were resuspended in RNA loading dye (10 M urea, 2 mM EDTA, .5% SDS, and .02% [w/v] each bromophenol blue and xylene cyanol) and separated on 38x30 cm 15% polyacrylamide (acrylamide:bis ratio 19:1), 7 M urea containing gels. The gels were dried and RNAs visualized by phosphor imaging.

RNA-binding and cleavage reactions

RNA binding and cleavage reactions were carried out as described previously [17]. Briefly, .05 pmol of uniformly ^{32}P -labeled RNA was incubated in the absence (free RNA) or presence of increasing concentrations of Cas6 (as indicated in figure legends) in buffer A for 30 minutes at 65-70°C. Half of each reaction was run directly on 8% native polyacrylamide gels to assess RNA binding by gel mobility shift analysis. RNA cleavage was assessed by separation of the RNAs on denaturing, 7 M urea containing, 15% polyacrylamide gels following PCI extraction and ethanol precipitation. For analysis of native Cas6 cleavage activity, either ~40 μg of whole cell extract (WCE) or supernatant from an immunoprecipitation reaction (see below) was incubated with .05 pmol of uniformly ^{32}P labeled RNA for 30 minutes at 70°C.

Alternatively, 10 μ L of resin from an immunoprecipitation reaction (see below) was added. In this case, samples were mixed every five minutes by pipetting up and down during the 30-minute incubation at 70°C. The gels were dried and RNAs visualized by phosphor imaging.

Quantitation of cleavage was performed using ImageQuantTM TL software (GE Life Sciences).

Preparation of *P. furiosus* cell extract

Four grams of *P. furiosus* cells were lysed in 10 mL 50 mM Tris (pH 8.0) in the presence of 100 U RQ1 DNase (Promega) and .1 mM phenylmethanesulfonylfluoride (PMSF). The extract was then subjected to ultracentrifugation at 100,000xg for 90 minutes. The resulting S100 was then stored at -80°C until use.

Preparation of polyclonal antibodies against PfCas6 in *Gallus gallus*

Specific antibodies against PfCas6 were raised in egg laying hens (*Gallus gallus*). For immunization, three injections of 200 μ g of 6x histidine-tagged PfCas6 in buffer B (20 mM sodium phosphate [pH 7.0], 500 mM NaCl) were done. Each injection was separated by two weeks. For the initial injection, 500 μ L antigen (.4 mg/mL) was emulsified in 200 μ L of Freund's complete adjuvant prior to injection in the breast muscle. For the two booster injections, 500 μ L of antigen (.4 mg/mL) was emulsified in 200 μ L of Freund's incomplete adjuvant prior to injection. One week following the final injection, immune eggs were collected daily for three months.

IgY was purified from the egg yolks by polyethylene glycol (PEG) precipitation as described previously [25]. Briefly, egg yolks were separated from the whites and washed with dH₂O. In a typical purification, three egg yolks were punctured, combined, and then

resuspended in approximately 250 mL of lysis buffer (10 mM Tris [pH 7.5], 100 mM NaCl). Polyethylene glycol (PEG) 8000 (Fisher Scientific) was then added to 3.5 % w/v. The sample was then mixed by shaking and centrifuged at 10,000g for ten minutes. The supernatant was then filtered through 100% cotton cheesecloth and then PEG 8000 was added 9% w/v. The sample was mixed by shaking and then centrifuged at 10,000g for ten minutes. The supernatant was removed and discarded. The pellet was resuspended in approximately 35 mL of Lysis buffer by incubation at 4° C overnight. The PEG precipitation was then repeated and the pellet from the 9% w/v PEG step was resuspended in approximately 7 mL Lysis buffer and stored at either 4° C or -80° C until use. The protein concentration was determined by the BCA assay (Pierce).

Immunoprecipitation of Cas6 from *P. Furiosus* extract

Immunoprecipitations (IP) were performed using anti-Cas6 IgY antibodies conjugated to CarboLink™ coupling gel (Pierce). Coupling was performed according to the manufacturer's protocol and was verified by A²⁶⁰ absorbance readings.

A *P. furiosus* S100 cell extract was pre-cleared in a reaction containing ~8 mg of total protein, ~550 µg of non-immune IgY coupled to CarboLink™ resin, 1X Complete™ Mini protease inhibitor (Roche), 50 U SUPERas-IN™ RNase inhibitor, and brought up to a total volume of 1 mL with IPP-300 (10 mM Tris [pH 8.0], 300 mM NaCl, .05% Igepal). The pre-clearing reaction was incubated at room temperature for two hours with end-over-end rotation. The sample was then centrifuged at 3000xg for two minutes and the supernatant was split between preimmune and immune IP reactions. A typical IP reaction contained 500 µL of pre-cleared S100 cell extract (~4 mg total protein), 1X Complete™ Mini protease inhibitor, 50 U

SUPERas-INTM RNase inhibitor, ~270-550 µg antibody (either preimmune or immune) coupled CarboLinkTM resin, and brought up to a total volume of 1 mL with IPP-300. The reactions were incubated at room temperature for 2 hours with end-over-end rotation and then washed four times with IPP-300. The pellets were resuspended in an equal volume of buffer A and stored at 4° C for later analysis.

Northern analysis

For northern blot analysis, RNAs were extracted from immunoprecipitation samples (both immune and preimmune) and WCE using TRIzolTM LS Reagent (Invitrogen) according to manufacturer's recommendations. Northern blots were performed essentially as described previously [16]. Briefly, RNAs were separated on a 15% polyacrylamide, 7 M urea containing gel (CriterionTM, Bio-Rad), then transferred onto Zeta-ProbeTM nylon membranes (Bio-Rad) using a Trans-Blot SD Semi-Dry CellTM (Bio-Rad). The membranes were then baked at 80°C for one hour before prehybridization in a ProBlot hybridization oven (LabNet) at 42°C for one hour. Prehybridization and hybridization were performed in Oligo-UltraHybTM solution (Applied Biosystems). Hybridization was initiated by adding 5' end labeled probe to the prehybridization buffer, and hybridization was carried out at 42°C overnight. Following hybridization, the membrane was washed twice with 2X SSC (30 mM sodium citrate [pH 7.0], 300 mM NaCl) with .5% SDS for 30 minutes at 42°C. RNAs were visualized by phosphor imaging.

References

1. Jansen R, Embden JD, Gaastra W, Schouls LM: **Identification of genes that are associated with DNA repeats in prokaryotes.** *Mol Microbiol* 2002, **43**:1565-1575.
2. Mojica FJ, Ferrer C, Juez G, Rodriguez-Valera F: **Long stretches of short tandem repeats are present in the largest replicons of the Archaea *Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning.** *Mol Microbiol* 1995, **17**:85-93.
3. Mojica FJ, Diez-Villasenor C, Garcia-Martinez J, Soria E: **Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements.** *J Mol Evol* 2005, **60**:174-182.
4. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD: **Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin.** *Microbiology* 2005, **151**:2551-2561.
5. Pourcel C, Salvignol G, Vergnaud G: **CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies.** *Microbiology* 2005, **151**:653-663.
6. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV: **A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action.** *Biol Direct* 2006, **1**:7.
7. Lillestol RK, Redder P, Garrett RA, Brugger K: **A putative viral defence mechanism in archaeal cells.** *Archaea* 2006, **2**:59-72.
8. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P: **CRISPR provides acquired resistance against viruses in prokaryotes.** *Science* 2007, **315**:1709-1712.
9. Deveau H, Barrangou R, Garneau JE, Labonte J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S: **Phage response to CRISPR-encoded resistance in *Streptococcus thermophilus*.** *J Bacteriol* 2008, **190**:1390-1400.
10. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slikhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J: **Small CRISPR RNAs guide antiviral defense in prokaryotes.** *Science* 2008, **321**:960-964.
11. Mills S, Griffin C, Coffey A, Meijer WC, Hafkamp B, Ross RP: **CRISPR analysis of bacteriophage-insensitive mutants (BIMs) of industrial *Streptococcus thermophilus*-implications for starter design.** *J Appl Microbiol* 2009.

12. Andersson AF, Banfield JF: **Virus population dynamics and acquired virus resistance in natural microbial communities.** *Science* 2008, **320**:1047-1050.
13. Tyson GW, Banfield JF: **Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses.** *Environ Microbiol* 2008, **10**:200-207.
14. Tang TH, Bachelier JP, Rozhdestvensky T, Bortolin ML, Huber H, Drungowski M, Elge T, Brosius J, Huttenhofer A: **Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*.** *Proc Natl Acad Sci U S A* 2002, **99**:7536-7541.
15. Tang TH, Polacek N, Zywicki M, Huber H, Brugger K, Garrett R, Bachelier JP, Huttenhofer A: **Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*.** *Mol Microbiol* 2005, **55**:469-481.
16. Hale C, Kleppe K, Terns RM, Terns MP: **Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*.** *RNA* 2008, **14**:2572-2579.
17. Carte J, Wang R, Li H, Terns RM, Terns MP: **Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes.** *Genes Dev* 2008, **22**:3489-3496.
18. Haft DH, Selengut J, Mongodin EF, Nelson KE: **A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes.** *PLoS Comput Biol* 2005, **1**:e60.
19. Kunin V, Sorek R, Hugenholtz P: **Evolutionary conservation of sequence and secondary structures in CRISPR repeats.** *Genome Biol* 2007, **8**:R61.
20. Calvin K, Li H: **RNA-splicing endonuclease structure and function.** *Cell Mol Life Sci* 2008, **65**:1176-1185.
21. Calvin K, Xue S, Ellis C, Mitchell MH, Li H: **Probing the catalytic triad of an archaeal RNA splicing endonuclease.** *Biochemistry* 2008, **47**:13659-13665.
22. Kim YK, Mizutani K, Rhee KH, Nam KH, Lee WH, Lee EH, Kim EE, Park SY, Hwang KY: **Structural and mutational analysis of tRNA intron-splicing endonuclease from *Thermoplasma acidophilum* DSM 1728: catalytic mechanism of tRNA intron-splicing endonucleases.** *J Bacteriol* 2007, **189**:8339-8346.
23. Ebihara A, Yao M, Masui R, Tanaka I, Yokoyama S, Kuramitsu S: **Crystal structure of hypothetical protein TTHB192 from *Thermus thermophilus* HB8 reveals a new protein family with an RNA recognition motif-like domain.** *Protein Sci* 2006, **15**:1494-1499.

24. Youssef OA, Terns RM, Terns MP: **Dynamic interactions within sub-complexes of the H/ACA pseudouridylation guide RNP**. *Nucleic Acids Res* 2007, **35**:6196-6206.
25. Polson A, von Wechmar MB, van Regenmortel MH: **Isolation of viral IgY antibodies from yolks of immunized hens**. *Immunol Commun* 1980, **9**:475-493.

Figure 3.1. Lead-induced and RNase A cleavage footprinting with CRISPR repeat RNA and PfCas6. (A) 3' end-labeled CRISPR repeat RNA was incubated in the absence (RNA) or presence of increasing concentrations of PfCas6 (indicated in μM) and then subjected to RNase A cleavage (left panel) or lead-induced cleavage (right panel). RNAs were separated by 15% denaturing (7 M urea) polyacrylamide gels. Size markers include 5' end-labeled RNA markers (M) and alkaline hydrolysis ladders (OH). Blue bars indicate strong protections. (B) 5' end-labeled CRISPR repeat RNA was used for lead-induced and RNase A cleavage as was done in (A). (C) A summary of cleavage protection is shown. The Cas6 cleavage site is shown by an asterisk (*).

Fig 3.2. Cleavage activity of Cas6 mutants. (A) Uniformly ^{32}P -labeled CRISPR repeat RNA was incubated in the absence (-) or presence of increasing concentrations of wild type or mutant Cas6 (.001, .05, and .5 μM) followed by separation on a 15% denaturing (7 M urea) polyacrylamide gel to assess cleavage along with 5' end-labeled RNA markers (M). The 5' and 3' cleavage products are indicated. (B) Purified wild type (wt) and mutant Cas6 proteins (as indicated above) were separated by SDS-PAGE. Molecular weight markers are indicated in kDa.

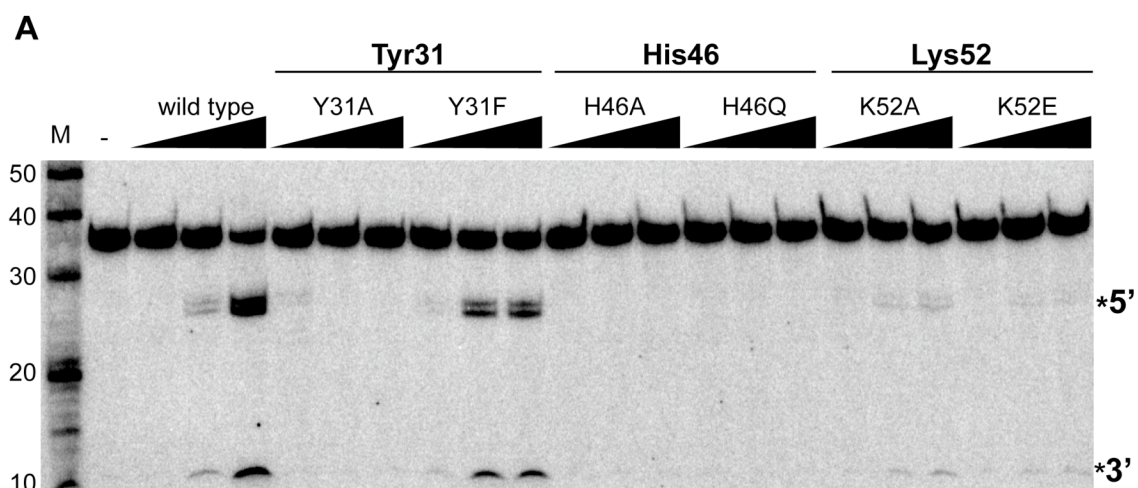


Fig. 3.3. Substrate recognition by Cas6 mutants. Uniformly ^{32}P -labeled CRISPR repeat RNA was incubated in the absence (-) or presence of increasing concentrations of wild type or mutant Cas6 (.001, .05, .2, and .5 μM) and then assessed for their ability to form a stable complex with the substrate RNA employing native gel mobility shift analysis. The positions of the free (RNA) and bound (RNP) substrate RNA are indicated. The Cas6 cleavage product is indicated with an asterisk (*).

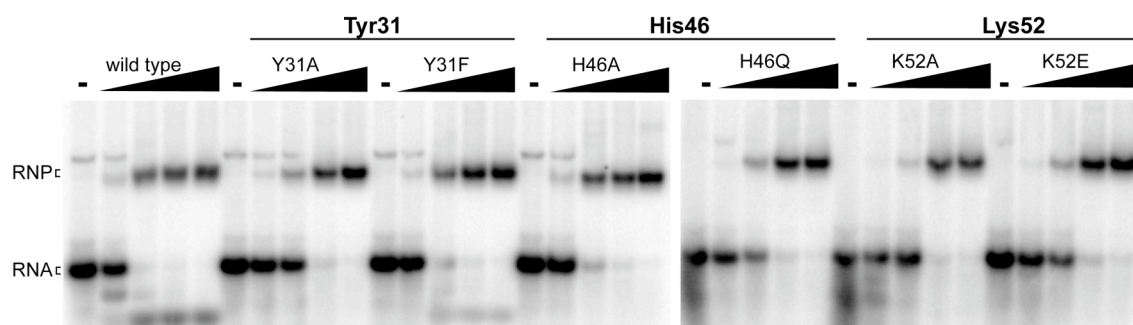
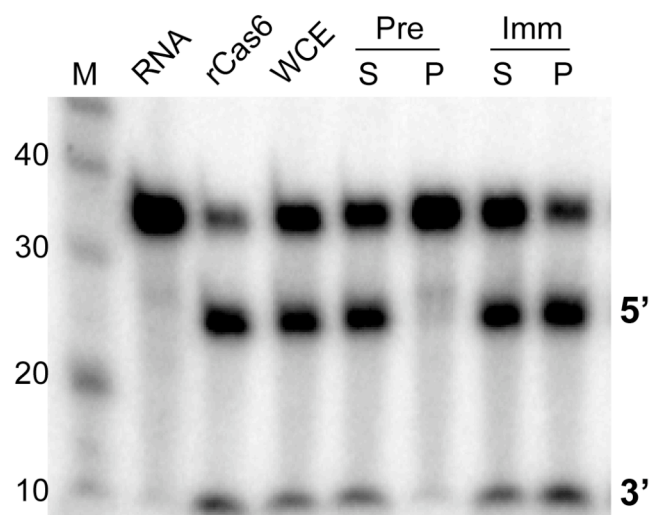


Fig. 3.4. Native Cas6 cleaves CRISPR repeat RNA and associates with crRNAs. (A) Uniformly ^{32}P -labeled CRISPR repeat RNA was incubated in the absence (RNA) or presence of recombinant Cas6 (rCas6), whole cell extract (WCE), or samples from immunoprecipitation reactions using anti-Cas6 antibodies (Pre, preimmune; Imm, immune; S, supernatant; P, pellet). The RNAs were separated on a 15% denaturing, 7 M urea, containing polyacrylamide gel along with 5' end-labeled RNA markers (M). (B) Northern blot analysis of Cas6 immunoprecipitation. RNAs extracted from WCE, preimmune (Pre) and immune (Imm) supernatants (Sup, left panel), and pellets (Pel, right panel) from an immunoprecipitation using anti-Cas6 antibodies were separated on 15% denaturing, 7 M urea containing, polyacrylamide gel along with 5' end-labeled RNA markers (M). A 5' end-labeled DNA oligonucleotide that was antisense to crRNA spacer 6.01 from *P. furiosus* was used as a probe. The positions of the 2X intermediate, 1X intermediate, and mature crRNAs are indicated.

A



B

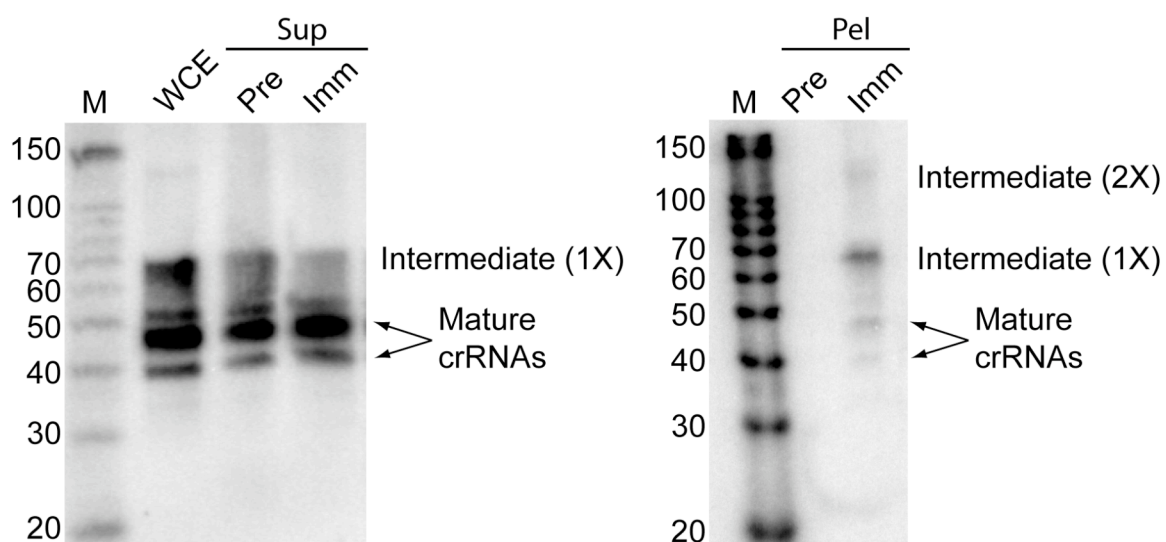


Fig. 3.5. The proposed catalytic mechanism of Cas6. Tyr31 acts as a general base and His46 as a general acid, while Lys52 stabilizes a predicted pentavalent intermediate. The cleavage products generated contain a 5' OH and likely 2'-3' cyclic phosphate.

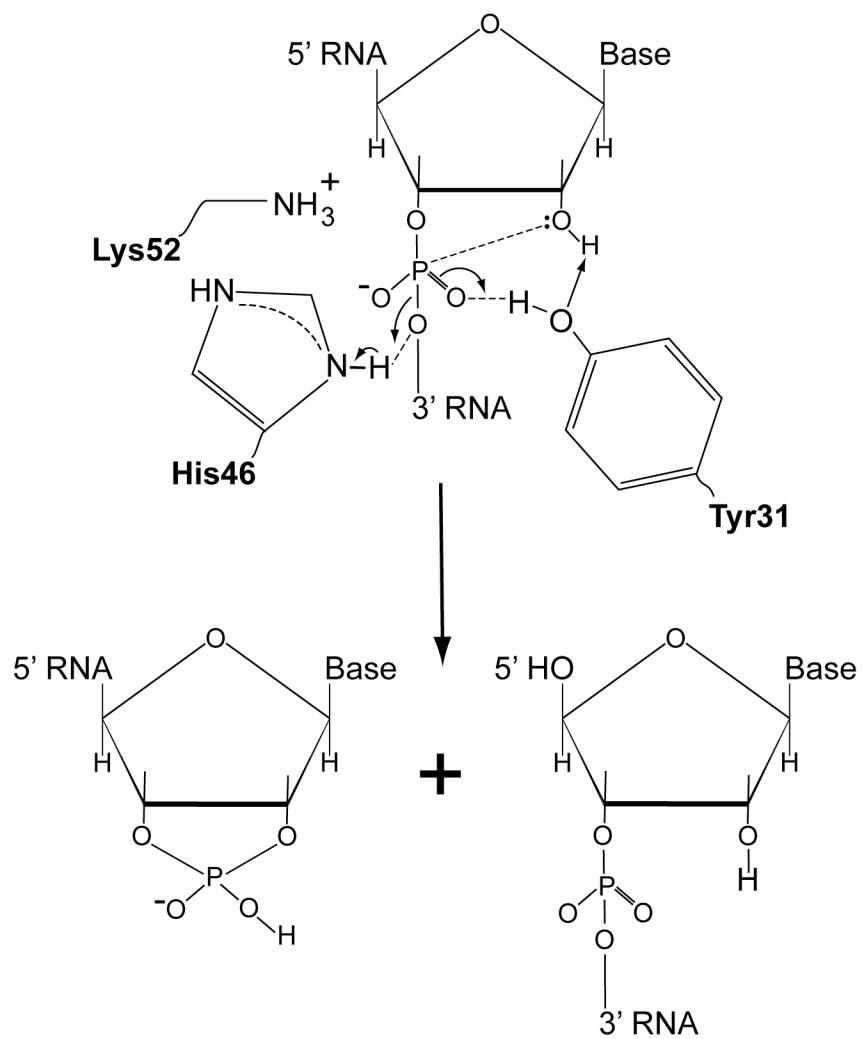


Table 3.S1. Oligonucleotides used in this study.

#	DNA oligonucleotides (5'-3')
1	GGAAGATTTAATAGCGTTGGCTATTAAACCCTGGAGGTAGTATTGATGATTG
2	CAATCATCAATACTACCTCCAGGGTTTAATAGCCAACGCTATTAAATCTTCC
3	GGAAGATTTAATAGCGTTGAATATTAAACCCTGGAGGTAGTATTGATGATTG
4	CAATCATCAATACTACCTCCAGGGTTTAATATTCAACGCTATTAAATCTTCC
5	GTTTAGGGCCCTTAACTTCAGCGAGATATGTTGCAAGCTTCG
6	CGAAGCTTGCAACATATCTCGCTGAAGTTAAGGGCCCTAAAC
7	CGAAGCTTGCAACATATCTCGCTGAAGTTAAGGGCCCTAAAC
8	CGAAGCTTGCAACATATCTCCAGGAAGTTAAGGGCCCTAAAC
9	CGGCCATAAAAAGTGAATACGTAAAGAGTGCAGGGCCCTTAACTTCATGGAG
10	CTCCATGAAGTTAAGGGCCCTGCACTCTTTACGTATTCACTTTTTATGGCCG
11	CGGCCATAAAAAGTGAATACGTAAAGAGTTCAGGGCCCTTAACTTCATGGAGATATG
12	CATATCTCCATGAAGTTAAGGGCCCTGAACTCTTTACGTATTCACTTTTTATGGCCG
13	TAATACGACTCACTATAGGGTTACAATAAGACCAAAATAGAATTGAAAG
14	CTTTCAATTCTATTTTGGTCTTATTGTAACCCTATAGTGAGTCGTATTA
15	GTATGCGGTCCTTGCGGCTGAGAGC
#	RNA oligonucleotides (5'-3')
16	GUUACAAUAAGACCAAAAUAGAAUUGAAAG

Table 3.S2. Combinations of deoxynucleotides used in this study. The oligos were either used to generate site-directed mutant PfCas6 constructs (PCR) or annealed directly and used as templates for in vitro transcription (IVT).

Product	5' and 3' oligos	PCR or IVT
Y31A	1 + 2	PCR
Y31F	3 + 4	PCR
H46A	5 + 6	PCR
H46Q	7 + 8	PCR
K52A	9 + 10	PCR
K52E	11 + 12	PCR
Repeat RNA	13 + 14	IVT

Chapter 4

DISCUSSION

Identification of Cas6 as an endoribonuclease involved in crRNA biogenesis

This work describes the identification of Cas6 as an endoribonuclease that cleaves crRNA precursors within the repeat region. Cas6 was shown to specifically recognize sequence elements in the 5' region of the repeat, while cleavage occurs downstream within in the 3' region of the repeat (Ch. 2). Cleavage was shown to produce a crRNA intermediate composed of a single CRISPR-derived spacer flanked by fragments of the repeat (Ch. 2). The eight nucleotide fragment of the repeat located at the 5' end is retained in mature crRNAs in *Pyrococcus furiosus*, and is believed to be functionally important and may serve as a recognition sequence for other Cas proteins [1]. The 3' repeat fragment is removed by further processing to produce mature crRNAs [1]. The factor(s) involved in the 3' processing remain to be identified.

Characterization of Cas6 from *P. furiosus*

The binding site of Cas6 on CRISPR repeat RNA was determined by RNA mutagenesis and supported by RNA footprinting (Ch. 2 and Ch. 3). Through mutational analysis of a single CRISPR-derived repeat, it was determined that the first eight nucleotides of the repeat are required for the formation of a stable complex with Cas6 (Ch. 2). When these sequence elements were altered by either deletion or substitution, detectable binding was lost (Ch. 2). Subsequent RNA footprinting with lead-induced and RNase A cleavage indicated that nucleotides 2-8

directly contact Cas6 (Ch. 3). Sequences in the middle and 3' region of the repeat do not appear to be important for recognition; however cleavage occurs within the 3' region (Ch. 2 and 3).

This indicates that Cas6 has a distinct mechanism of action for a ribonuclease: sequence-specific substrate recognition and cleavage of the substrate that occurs distally to the binding site. One possible model for how this could occur is that Cas6 binding to the 5' region of the CRISPR repeat RNA could induce a conformational change in the RNA, possibly resulting in base-pairing between the palindromic regions of the repeat RNA (located in the 5' and 3' regions), which could lead to the proper positioning of the scissile phosphate bond in the active site for cleavage. Alternatively, following the establishment of a high affinity interaction between the protein and the 5' region of the repeat RNA, the unbound portion of the RNA may wrap completely around the protein through a series of low affinity interactions to position the scissile bond in the active site. If this is true, it is possible that these low affinity interactions could not be detected by the techniques used in this work.

Cas6 cleavage of CRISPR repeat RNA has been shown to be divalent metal ion independent (Ch. 2). As is the case for other metal-independent nucleases including RNase A [2], RNase T2 [3], and the archaeal splicing endonuclease [4], cleavage occurs on the 5' side of the phosphodiester bond, generating products with a 5' hydroxyl and likely 2'-3' cyclic phosphate (Ch. 2).

The crystal structure of PfCas6 was determined at 1.8 Å (Ch. 2). The structure revealed duplicated ferredoxin folds separated by a well-formed central cleft. Ferredoxin folds, including the well studied RNA recognition motif (RRM), are common structural features of single stranded RNA binding proteins [5]. The conserved GhGxxxxGhG motif (where h is hydrophobic and xxxx contains at least one lysine or arginine) characteristic of Cas6s, forms a

loop located in the central cleft [6]. Located adjacent to this Gly-rich loop are a trio of highly conserved Tyr, His, and Lys residues that have been verified by mutational analysis of cleavage activity to be involved in catalysis (Ch. 2 and 3). This finding, along with the observation that cleavage is divalent metal ion independent led us to propose a general acid/base catalytic cleavage mechanism based on similarity to the well characterized archaeal tRNA splicing endonuclease [4]. In this proposed cleavage mechanism, Tyr31 acts as a general base and His46 as a general acid, while Lys52 is predicted to stabilize a putative pentavalent transition state.

Comparative analysis of Cas6 and Cse3

Cse3 (also known as CasE) from *Escherichia coli* was shown to possess ribonuclease activity very similar to Cas6. Cse3 cleaves crRNA precursors within the 3' region of CRISPR repeat RNA in a metal-independent fashion [7]. Cse3 cleavage appears to occur eight nucleotides upstream of the 3' end of the CRISPR repeat RNAs from *E. coli*, generating what is believed to be the mature crRNA 5' end [7]. Cas6 also cleaves at a site eight nucleotides upstream of the 3' end of repeats from *P. furiosus* creating the mature 5' ends of crRNAs from this organism. The repeat sequences from these two organisms differ substantially, however the length of repeat retained at the 5' end on mature crRNAs in both organisms is eight nucleotides. These eight, repeat derived nucleotides, may be a conserved feature of all mature crRNAs that we hypothesize serves as a recognition sequence for other Cas proteins. Some organisms lack both Cas6 and Cse3. It seems likely that there are one or more other yet to be characterized Cas proteins that perform this same function in these organisms, which highlights the high level of diversity in the CRISPR-Cas system.

The Cse3 protein from *E. coli* has been shown to associate with a complex of five other Cas proteins as well as mature crRNAs, and the complex has been named Cascade [7]. In the present work, we have shown that Cas6 appears to associate several crRNA species, including the 1X crRNA intermediate (its cleavage product) and weakly with mature crRNAs. Since mature crRNAs in *P. furiosus* lack the Cas6 binding site and these molecules were co-purified from native cell extract, this protein is also likely a component of a larger protein-RNA complex. Given that Cas6 remains bound to the crRNA intermediate following cleavage of the repeat, Cas6 may play a role in the subsequent 3' processing of this RNA that leads to the accumulation of mature crRNAs in *P. furiosus*.

This work begins to address the molecular processes required for the generation of mature crRNA that are used in CRISPR-mediated genome defense. Future studies will address the downstream processing of crRNA intermediates into mature crRNAs and how these crRNAs, along with Cas proteins, silence genome invaders.

At this point, there are many more questions than answers regarding the molecular basis of CRISPR-Cas function. The mechanism by which novel spacers are incorporated into CRISPR loci remains to be elucidated, but is believed to involve the “core” (most highly conserved) Cas proteins [8]. The nature of the effector complex that targets and silences invading nucleic acids has not been characterized. In *P. furiosus*, the crRNA intermediate generated by Cas6 cleavage undergoes further processing at the 3' end and likely involves yet to be identified Cas nuclease(s) [1]. The answers to these exiting questions will be likely be determined in the very near future as the interest in this field has intensified greatly in recent years.

References

1. Hale C, Kleppe K, Terns RM, Terns MP: **Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus***. *RNA* 2008, **14**:2572-2579.
2. Raines RT: **Ribonuclease A**. *Chem Rev* 1998, **98**:1045-1066.
3. Deshpande RA, Shankar V: **Ribonucleases from T2 family**. *Crit Rev Microbiol* 2002, **28**:79-122.
4. Calvin K, Li H: **RNA-splicing endonuclease structure and function**. *Cell Mol Life Sci* 2008, **65**:1176-1185.
5. Maris C, Dominguez C, Allain FH: **The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression**. *FEBS J* 2005, **272**:2118-2131.
6. Haft DH, Selengut J, Mongodin EF, Nelson KE: **A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes**. *PLoS Comput Biol* 2005, **1**:e60.
7. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV, van der Oost J: **Small CRISPR RNAs guide antiviral defense in prokaryotes**. *Science* 2008, **321**:960-964.
8. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P: **CRISPR provides acquired resistance against viruses in prokaryotes**. *Science* 2007, **315**:1709-1712.